AN INVESTIGATION OF EXON REPETITION

By
Roberto Rigatti BSc (Hons) (Padova, Italy)

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

February 2004
I would like to dedicate this work to my parents
ACKNOWLEDGEMENTS

I would like to thank the following people:

My supervisor Prof. Ian Eperon for giving me the opportunity to do this PhD in his lab and for his constant scientific help and encouragement.

My PhD committee members, Dr David Heery and Prof Nilesh Samani.

Dr Juan-hua Jia who worked on the exon repetition project and also all the people of labs 214 and 216, in particular Helen Reed for reading chapters 3 and 4.

All the friends in the Department of Biochemistry for their help and friendship.

Finally, my dearest friend Marilena, and many other friends, in particular Jenny, Barbara, Tiago, Gabi, Cristina, JP, Mario, Eliana, Dave, Dolores, Tim, Amel, Elena, Laura, and Christian.
# TABLE OF CONTENTS

ABSTRACT ....................................................................................................................................1  
CHAPTER 1. INTRODUCTION .............................................................................................2  
  1.1. Structure of eukaryotic genes ................................................................. 2  
  1.2. Steps in gene expression in protein-encoding genes ............................... 3  
  1.3. Roles of introns ......................................................................................... 4  
  1.4. Transcript heterogeneity .......................................................................... 5  
  1.5. Splicing and the spliceosome ................................................................. 5  
  1.6. SR proteins ............................................................................................... 9  
  1.7. Alternative splicing ............................................................................... 11  
  1.8. The carboxy-terminal domain of RNA polymerase II large subunit ...... 15  
  1.9. Coupling between transcription and splicing ........................................... 16  
  1.10. Trans splicing ......................................................................................... 18  
  1.11. Genes that undergo exon repetition ..................................................... 23  
  1.12. Exon repetition in the SA gene .............................................................. 24  
  1.13. Exon repetition in the COT gene ............................................................ 25  
OBJECTIVES ............................................................................................................................ 26  
CHAPTER 2. MATERIALS AND METHODS ..................................................................27  
  2.1. Culturing of cell lines (Hela and 293 cells) .............................................. 27  
  2.2. DNA miniprep .......................................................................................... 31  
  2.3. Growth media for bacteria .................................................................... 28  
  2.4. Preparation of cloning vectors ............................................................... 28  
  2.5. Agarose gel electrophoresis .................................................................... 29  
  2.6. Elution of DNA fragments from agarose gel ........................................... 29  
  2.7. Ligation reactions ................................................................................... 30  
  2.8. Precipitation of ligation .......................................................................... 30  
  2.9. Preparation of electro-competent cells .................................................. 30  
  2.10. Electroporation ...................................................................................... 31  
  2.11. DNA miniprep ........................................................................................ 31  
  2.12. DNA maxiprep ....................................................................................... 32  
  2.13. Quantification of nucleic acids ............................................................. 33  
  2.14. RNA extraction from tissues ............................................................... 33  
  2.15. Transfection of cell lines ..................................................................... 34  
  2.16. RNA isolation from cell lines ............................................................... 35  
  2.17. Reverse transcription .......................................................................... 36  
  2.18. Polymerase Chain reaction (PCR) ........................................................ 36  
  2.19. DNA sequencing .................................................................................. 37  
  2.20. Firefly luciferase assay ......................................................................... 37  
  2.21. Extraction of genomic DNA from tissues ............................................ 38  
  2.22. MATERIALS AND METHODS FOR CHAPTER 3 ......................... 39  
    2.22.1. PCR conditions .............................................................................. 39  
    2.22.2. Southern blot .................................................................................. 40  
    2.22.3. Real-time PCR ............................................................................... 42  
    2.22.4. Cycling curve ................................................................................... 44  
  2.23. MATERIALS AND METHODS FOR CHAPTER 4 .......................... 44  
    2.23.1. Construct preparation ................................................................... 44  
    2.23.2. Transfection of mammalian cells .................................................. 47  
  2.24. MATERIALS AND METHODS FOR CHAPTER 5 .......................... 47
CHAPTER 3. EXON REPETITION IN SHR AND WKY. STRAIN-SPECIFIC OR ALLELE-SPECIFIC? ..............................................................................................................

3.1. INTRODUCTION ........................................................................................................... 56
3.1.1. Exon repetition of the SA gene............................................................................... 56
3.1.2. Different ways of obtaining exon repetition.......................................................... 56
3.1.3. Exon repetition in SA is strain-specific and tissue-specific ................................ 58
3.2. RESULTS..................................................................................................................... 60
3.2.1. Exon repetition in the COT gene is strain-specific ............................................... 60
3.2.2. Exon repetition in SA and COT segregate independently in an F2 population ....... 63
3.2.3. Exon repetition in SA is allele-specific .................................................................. 64
3.2.4. Exon repetition in COT is allele-specific ............................................................... 68
3.2.5. The pattern of exon repetition is allele-specific in COT ....................................... 71
3.2.6. Genomic organisation of the COT gene ............................................................... 74
3.2.7. The promoter of the COT gene ............................................................................... 77
3.3. DISCUSSION............................................................................................................... 79

CHAPTER 4. FUNCTIONAL SIGNIFICANCE OF EXON REPETITION ................................ 83
4.1. INTRODUCTION ........................................................................................................... 83
4.1.1. Exon repetition can create new proteins ................................................................. 83
4.1.2. Non sense mediated mRNA decay ........................................................................ 84
4.1.3. Exon repetition can potentially decrease the production of the normal protein product from a gene ................................................................. 86
4.1.4. Factors that influence the efficiency of translation ............................................... 87
4.1.5. Selection of a downstream AUG as translational initiation codon ..................... 88
4.1.6. Effects of exon repetition on translation ............................................................... 89
4.2. RESULTS..................................................................................................................... 90
4.3. DISCUSSION............................................................................................................... 93

CHAPTER 5. MECHANISM OF EXON REPETITION ......................................................... 97
5.1. INTRODUCTION ........................................................................................................... 97
5.1.1. Characteristics of exon repetition .......................................................................... 97
5.1.2. Possible mechanisms for generating exon repetition ............................................ 98
5.2. RESULTS..................................................................................................................... 101
5.2.1. Cloning of the SHR SA allele ............................................................................... 101
5.2.2. Comparison between the DNA sequence of the SHR and WKY SA alleles .. 102
5.2.3. Mechanism 1. Cryptic promoter and pause sites. Is there a cryptic promoter in intron 1 of the WKY SA gene? ................................................................. 103
5.2.4. Mechanism 2. Complementary sequences .......................................................... 107
5.2.5. Mechanism 3. Exonic splicing enhancers ............................................................... 108
5.2.6. Mechanism 4. Weak exon and 5' splice site of the exon 5' of the repeated exons ..................................................................................................................... 109
CMV  cytomegalovirus
COT  carnitine octanoyltransferase
CP  cryptic promoter
CTD  carboxy-terminal domain
dATP  2’ deoxyadenosine 5’ triphosphate
dCTP  2’ deoxycytidine 5’ triphosphate
DEPC  Diethyl pyrocarbonate
dGTP  2’ deoxyguanosine 5’ triphosphate
DNA  deoxyribonucleic acid
dNTP  2’ deoxynucleotides 5’ triphosphates
DTT  Dithiothreitol
dTTP  2’ deoxythymidine 5’ triphosphate
ECFP  enhanced cyano fluorescent protein
EDTA  Ethylene diamino tetracacetate
eIF-2  eukaryotic initiation factor 2
ESE  exonic splicing enhancer
FCS  fetal calf serum
GFP  green fluorescent protein
GTP  guanosine 5’ triphosphate
HEPES  N(2-hydroxyethyl) pireazine-N’-2-ethane sulphonic acid
IPTG  Isopropyl-β-D-thiogalactoside
kDa  kilodalton
mRNA  messenger RNA
NMD  nonsense-mediated mRNA decay
NP  natural promoter
OD    optical density
ORF   open reading frame
PAGE  polyacrylamide gel electrophoresis
PCR   polymerase chain reaction
Pre-mRNA  precursor mRNA
PTB   polypyrimidine-tract-binding protein
PTC   premature termination codon
RNA   ribonucleic acid
rpm   revolutions per minute
RT-PCR reverse transcription-polymerase chain reaction
SD    Sprague-Dawley
SDS-PAGE sodium dodecyl sulphate polyacrylamide-gel electrophoresis
SHR   Spontaneously hypertensive rat
SNP   single nucleotide polymorphism
snRNA small nuclear RNA
snRNP small nuclear ribonucleoprotein particle
TBS   Tris buffered solution
TEMED N,N,N',N'-Tetramethylethylenediamine
Tris  Tris(hydroxymethyl) aminomethane
U2AF  U2 auxiliary factor
uORF  upstream open reading frame
UTR   untranslated region
WKY   Wystar-Kyoto
AN INVESTIGATION OF EXON REPETITION

ABSTRACT

Exon repetition consists in the presence of tandemly repeated exons in mRNA. In all the cases reported so far, the trivial explanation that the phenomenon arises from a duplication of specific exons in the gene has been ruled out by southern blot analysis. It is therefore assumed that exon repetition is the result of a post-transcriptional process involving two pre-mRNA molecules transcribed from the same gene. Exon repetition of SA is of particular interest, since it is both tissue and strain-specific. This at first suggested that repetition in SA might be caused by a tissue and strain-specific *trans*-acting factor. This work has shown that the property of exon repetition is allele-specific in the two best characterised examples, represented by the SA and COT (carnitine octanoyltransferase) genes in rat. Allele-specificity of exon repetition is therefore the underlying cause of strain-specificity of exon repetition in SA and also in COT.

The biological significance of exon repetition remains unknown. One possible function is the generation of new proteins. Quite often though, repetition of the exon that contains the AUG translational start codon creates a short open reading frame upstream of the one encoding the full length protein. The effect of this has been investigated in this work and the results have shown that translation of the downstream open reading frame is significantly reduced.

The mechanism that generates exon repetition is not well understood. It was initially proposed in COT that a putative exonic splicing enhancer (ESE) in exon 2 was required for exon repetition. This work demonstrates that the ESE on its own cannot determine exon repetition *in-vivo*.
CHAPTER 1
INTRODUCTION

1.1. Structure of eukaryotic genes

Mammalian genes are composed of a promoter region, coding sequences named exons and non-coding sequences called introns. The promoter is composed of basal promoter elements, promoter proximal elements and distal enhancer (or silencer) elements. A well characterised example of core promoter element is the TATA element which is located 25 bp upstream of the transcription start site. The consensus sequence for the TATA element is TATAa/tAa/t. At the start site of transcription there is a pyrimidine initiator element. Promoter proximal elements are situated up to 200 bp from the transcription start site. Activators of transcription bind to these elements and regulate transcription of a particular gene. It has been shown that promoters can affect splicing (Cramer P 1997). Enhancers and silencers can be located thousands of nucleotides away from the gene they control. They can also be positioned upstream, downstream or within the gene and function in either orientation. These DNA elements are bound by transcription factors, which can also bind to transcription factors assembled on promoters (Li R 1991).

Exons represent the coding sequences of a gene. Generally exons are relatively short. For example in human, the average exon is 150 nucleotides long (Deutsch M 1999). Introns are non-coding sequences and are usually much larger than exons. Again, in human, the average size of an intron is 3,500 nucleotides (Deutsch M 1999). A diagrammatic representation of a eukaryotic gene is depicted in figure 1.1.

Introns are quite rare in single-cell eukaryotes such as yeasts. For example, in the yeast Saccharomyces cerevisiae there are approximately 300 introns in its entire genome. In contrast, in mammals, the coding sequences of almost every gene are interrupted by introns. Interestingly, analysis of mRNA in Saccharomyces cerevisiae revealed that more than 25% of mRNA derives from intron-containing genes (Ares MJR 1999).

The boundaries between exons and introns are characterised by the presence of very short and highly conserved sequences. The majority of introns begin with the dinucleotide sequence GT and end with the sequence AG. In mammals, the consensus sequence for the 5’ splice site is CAG|GTRAGT (where R is a purine residue), whereas the consensus sequence for the 3’ splice site is NCAG|G. The vertical lines indicate the exon-intron boundaries. Variations from these consensus sequences are well known and quite
Figure 1.1. Diagrammatic representation of a typical mammalian gene. It is composed of exons, which are shown as white and numbered boxes and introns, represented by the black lines between the exons. The numbers inside the boxes indicate the exon number. The arrow represents the promoter region of the gene. The upstream box represents an enhancer sequence. Enhancers can be found also downstream of the promoter and can work in both orientations.
common. For example, from a detailed analysis of more than 22,000 splice site pairs in mammals, it emerged that 99.24% were of the GT-AG type, 0.69% were GC-AG, 0.05% were AT-AC and 0.02% were of other types (Burset M 2000).

The other important sequences of an intron are the branch site and a polypirimidin tract. The branch site has the consensus (T/C)NCT(A/G)AC. The bold and underlined A nucleotide represents the adenosine residue of the branch site and its 2' OH group is involved in the first catalytic step of the splicing reaction. The polypirimidin tract is positioned between the branch site and the 3' splice site. The distance between the branch site and the 3’ splice site is 11-40 nucleotides (nt). A detailed structure of an intron with the flanking exons is shown in figure 1.2.

### 1.2. Steps in gene expression in protein-encoding genes

Gene expression is the process by which the genetic information stored in the DNA, is utilised to synthesise proteins. This is carried out through a number of events that are linked to one another. The first step is transcription in which the enzyme RNA polymerase II synthesises a pre-mRNA molecule using DNA as template. Before the RNA molecule can be translated into a protein product, it must undergo several modifications that will transform it into an mRNA (messenger RNA) molecule. These modifications are:

- **Capping.** This is the first modification to take place. The initial step is the removal of a phosphate group from the first nucleotide of the nascent pre-mRNA molecule, leaving two phosphate groups. Then a covalent bond is formed between the second phosphate group and a GTP molecule. The bond is atypical since it is a 5'-5' triphosphate linkage. The last step is the addition of a methyl group at the N7 position of the added GMP. The cap structure is thought to stabilise the RNA molecule (Beelman CA 1995).

- **Splicing.** The coding sequences of genes in higher eukaryotes are interrupted by the presence of introns. The intronic sequences must be removed so that the coding portions can be joined together to form an mRNA. The process of removing introns from pre-mRNA molecules is termed splicing. Splicing is carried out by a macromolecular complex known as the spliceosome.

- **3’ end formation.** Virtually all mRNA molecules contain at their 3’ end a poly(A) tail composed of about 200 adenosine residues. Before a poly(A) tail is added, the nascent pre-mRNA molecule is cleaved.
Figure 1.2. Schematic representation of two exons separated by the intervening sequence (intron). The most important features are the 5’ and 3’ splice sites which are located at the boundaries between the exons and the intron, the branch site (BS) and the polypirimidine tract (PPT).
1.3. Roles of introns

Introns used to be considered as selfish DNA sequences without any particular function (Cavalier-Smith 1985). However it was later discovered that introns can have several functions. First of all, they can contain regulatory elements, some of which are involved in gene expression. For example, in the human apolipoprotein B gene, the second intron is required for expression of this gene in liver (Brooks AR 1994). Another example is the cis-acting element present in the second intron of the murine interleukin IL-4 gene. This element regulates transcription and also influences chromatin structure by causing methylation of the gene (Hural JA 2000). The regulatory elements present in introns not only act upon gene expression but some of them can also have an effect on splicing. For example, alternative splicing of the hnRNP A1 gene is regulated by intronic elements (Simard MJ 2000). It must be borne in mind that alternative splicing can occur only in intron-containing genes. Hence, just with their presence, introns allow the expansion of the proteome complexity by allowing alternative splicing.

Another function of introns is that they can contain genes, such as those that are transcribed into non-coding RNAs. For example, the mouse U14 small nucleolar RNA, also called snoRNA, is encoded by intronic sequences of the hsc70 heat shock gene (Liu J 1990). Moreover, around 200 snoRNA encoding genes were found in vertebrates and the majority of them are encoded by intronic regions.

Another important function of introns is to increase the rate of meiotic recombination between the coding sequences of a gene.

Introns have also a general positive effect on the rate of transcription. For example, in intron-containing transgenes, the expression level is increased 10-100 times compared to the same transgene without introns (Brinster RL 1988). Several different mechanisms are responsible for this effect. First of all, introns can contain enhancers and repressors of transcription. Furthermore, introns can regulate transcription by affecting the position of nucleosomes, since this affects the accessibility of the RNA polymerase to the promoter. Recently, it has been shown that the presence of an intron, or more precisely a promoter proximal splice site, can modulate transcription while the RNA is being transcribed (Furger A 2002). As introns are transcribed, spliceosomes start to assemble at the splice sites. A component of the spliceosome known as U1 snRNP (U1 small nuclear ribonucleoprotein particle) is known to recruit a protein called Tat-specific factor 1 (TAT-SF1). This factor interacts with a second protein, the positive transcription elongation factor b (pTEFb) which in turn phosphorylates the CTD of RNA pol II. Upon phosphorylation, the
elongation rate of RNA pol II is enhanced (Fong YW 2001) and this suggests that introns can increase the transcription rate regardless of their sequence.

1.4. Transcript heterogeneity

The idea that a single gene encodes one polypeptide chain, the so-called one gene-one polypeptide hypothesis (Beadle and Tatum) has become obsolete. It is now known that in higher eukaryotes, a single gene can produce many different mRNA molecules and some of them encode functionally different proteins. Several different mechanisms contribute to transcript heterogeneity. For instance, multiple transcription start sites can be found in a single gene. This leads to the production of mRNA isoforms that differ at their 5’ end (Jansen E 1991). The same concept is true for 3’ end formation. In fact, many genes have alternative polyadenylation sites and can therefore produce transcripts with a different 3’ end (Tosi M 1981). The major contributor to transcript heterogeneity is alternative splicing. Alternative splicing consists in the alternative use of 5’ and 3’ splice sites in a pre-mRNA molecule. Through alternative splicing, several different pre-mRNA isoforms can be produced from a single gene and many of these isoforms encode a different protein product. It has been postulated that between 35-59 % of human genes are alternatively spliced (Modrek B 2002).

Recently, another phenomenon that increases transcript heterogeneity has been discovered in a limited number of genes. In this case, some mRNA isoforms contain tandem repeats of specific exons and this phenomenon was named exon repetition.

1.5. Splicing and the spliceosome

The splicing reaction discussed in this section is catalysed by the spliceosome and is responsible for the removal of introns from pre-mRNA molecules transcribed from protein-encoding genes. But some introns are known to catalyse their own removal without the requirement any external catalytic machinery (spliceosome). This type of introns called self-splicing introns, are classified into two groups. In group I, a free guanylyl nucleotide carries out the first nucleophylic attack. In group II, the splicing reaction resembles the one that characterises protein encoding genes, since it has the same types of intermediates.

Pre-mRNA molecules undergo several modifications that transform them into mRNAs. One of these modifications is the removal of introns which is known as splicing and is performed by a highly complex macromolecular machine known as the spliceosome.

The spliceosome is composed of five snRNPs (small nuclear ribonucleoprotein particles) named U1, U2, U4-U6 and U5 and also by a large number of non-snRNP
proteins. Each snRNP consists of a snRNA molecule and a set of seven Sm or Sm-like proteins (B'/B, D3, D2, D1, E, F and G) as well as many particle-specific proteins. All snRNAs (with the exception of U6, which is transcribed by RNA polymerase III) are transcribed by RNA polymerase II. After transcription, the snRNA molecules undergo modifications such as capping and addition of nucleotides at their 3' end. After being exported to the cytoplasm, Sm proteins are added on to the snRNA molecules. Quite recently, a detailed study of the components of human spliceosomes was carried out (Zhou Z 2002). The authors identified 145 different proteins, of which 88 turned out to be factors known to be involved in splicing. The high number of components identified makes the spliceosome the most complex cellular machine known to date. The spliceosome seems to undergo a change in protein composition during spliceosome activation (Makarov EM 2002). For example, a number of proteins known as the Prp 19 complex, is associated with the U5 snRNP. After the splicing reaction, the U5 snRNP seems to be released as a 35S complex. It is postulated that the Prp19 complex is later released from the dissociated U5 snRNP which then joins the spliceosome as a 20S particle.

It is generally believed that spliceosomes are assembled in a stepwise manner onto the pre-mRNA molecule. According to this model, the first step in splicing is the recognition of the 5' splice site (the junction between the upstream exon and the intron) by the U1 snRNA and binding of U2AF (U2 auxiliary factor) to the polypyrimidine tract and the 3'splice site (the boundary between the intron and the downstream exon). This leads to the formation of the so-called E (early) complex. At this stage, U2 snRNP is also associated with the pre-mRNA, but very loosely (Das R 2000). The U2 snRNP then associates stably with the branch point sequence. This is the first step in splicing that is known to require ATP and leads to the formation of the A complex. Upon binding of the tri-snRNP U4/U6-U5, the B complex is formed. At this stage, a number of RNA-RNA and RNA-protein rearrangements take place together with the recruitment of additional proteins and the B complex switches to the C complex which represents the catalytically competent spliceosome. Many RNA-RNA interactions that take place during spliceosome assembly are mutually exclusive. For example in the tri-snRNP, U6 is base paired to U4 via stems I and II. As the spliceosome becomes activated for catalysis, both stems between U4 and U6 are unwound enabling U6 to base pair with U2 and with the 5'splice site while U4 is no longer needed and leaves the spliceosome assembly. The U1 snRNP is also released from the spliceosome because it is replaced by U6 in its interaction with the 5' splice site.

In spite of its general acceptance, this model for spliceosome assembly cannot account for certain experimental findings. For example, some snRNPs are found associated
in the spliceosome at earlier stages. In particular it was shown that the tri-snRNP
U4/U6.U5 interacts with the 5′ splice site prior to the binding of the U2 snRNP to the
branch point sequence (Maroney PA 2000). The authors proposed that U1snRNP and the
tri-snRNP U4/U6.U5 cooperate to recognise and define the 5′ splice site before the
downstream 3′ splice site is defined. The generally accepted idea that the spliceosome is
assembled on a pre-mRNA molecule in a stepwise manner has recently been challenged
(Stevens SW 2002). In previous works, snRNPs had been purified using salt concentrations
incompatible with \textit{in vitro} splicing, which takes place at a salt concentration of 50-120
mM. Furthermore, the use of heparin in many previous experiments might have disrupted
many weak interactions. When Stevens and collaborators used a salt concentration of 250
mM, they could only recover the tri-snRNP in yeast. By lowering the salt concentration
down to 150 mM, they were able to isolate a tetra-snRNP (U2•U4/U6•U5). Finally, at a salt
concentration of 50 mM, which is compatible with splicing conditions \textit{in vitro}, they
successfully isolated a 45S penta-snRNP particle. The particle was shown to contain the 5
snRNP particles in stoichiometric proportions and also more than 60 factors involved in
splicing (approximately 85% of all the known splicing factors in yeast). Interestingly, the
particle was shown to catalyse splicing \textit{in vitro}. These results suggested that the
spliceosome might be completely assembled before it interacts with the pre-mRNA, rather
than assembling by addition of one snRNP at a time on a pre-mRNA molecule. Further
supporting evidence came from an elegant experiment. Stevens and coworkers used two
different yeast strains. In one strain, the protein Prp4 (which is part of the tri-snRNP) was
epitope-tagged, whereas the other strain contained mutants U1 and U2 snRNAs (named
U1δ and U2δ, respectively) characterised by large deletions that do not affect the function
of the two snRNA molecules. Two extracts, one from each strain, were mixed together and
incubated with biotynilated pre-mRNA. Spliceosomes were then purified in two steps. The
first one consisted in the use of an antibody that recognises the epitope-tag of Prp4. In this
way, only the spliceosomes that contain the epitope-tagged Prp4 were selected. In the
second step, functionally engaged spliceosomes (bound to pre-mRNA) were isolated with
streptavidin. All the spliceosomes isolated in this way were found to contain only U1 and
U2 but not U1δ or U2δ, thus demonstrating that there had not been any exchange of
snRNPs between the two types of spliceosomes. This experiment strengthened the idea that
spliceosomes are fully assembled before they engage with a pre-mRNA molecule.
Interestingly, multi-snRNP complexes had been observed in the past, in human extracts
under modified salt conditions (Konarska MM 1988), suggesting that spliceosomes might
be pre-assembled before they interact with a pre-mRNA molecule in mammals as well.
The spliceosome that has been described so far is the so-called U2-dependent spliceosome, but this is not the only type of spliceosome. In fact, a second type of spliceosome exists and is called U12-dependent spliceosome. The five snRNPs that compose the U12-dependent spliceosome are the U11, U12, U4atac/U6atac and U5 snRNP (Tam WY 1996a; Tam WY 1996b). U5 is the only snRNP in common between the two different spliceosomes. While U1 and U2 snRNPs are isolated as single particles from nuclear extracts, U11 and U12 snRNPs can form a highly stable U11/U12 complex (Wassarman KM 1992). The analysis of proteins in isolated U11/U12 complexes revealed the presence of unique proteins in this particle together with other factors that are in common with those found in U1 and U2 snRNPs (Will CL 1999). In particular, twenty proteins were found. Eight of them, corresponded to the snRNP Sm proteins B', B, D3, D2, D1, E, F and G of the U2-type spliceosome. Four proteins were found to correspond to the components of the essential splicing factor SF3b. A distinct component, consisting of a 35 kD polypeptide, was found that seemed to be related to the U1 70 kD protein.

As a general rule, introns with the consensus sequence GT-AG are excised by the U2-dependent spliceosome whereas introns with the consensus sequence AC-AT are excised by the U12-dependent spliceosome. Hybrid introns (GT-AC or AC-AG) have not been found yet. It should be pointed out that there are exceptions to this general rule of intron excision. In particular, some GT-AG introns are known to be excised by the U12-dependent spliceosome and at the same time, there are AT-AC introns which are removed by the U2-dependent spliceosome. It has been proposed that introns should be classified based upon the type of spliceosome that carries out their excision rather than the consensus dinucleotide sequences at their 5' and 3' ends (Dietrich RC 1997).

Splicing is accomplished by two transesterification reactions. In the first reaction, the 2' hydroxyl group of the adenosine residue that is part of the branch point carries out a nucleophilic attack at the phosphate group of the 5'splice site. This generates a free 5' exon and a lariat intermediate (intron and 3'exon) as shown in figure 1.3. In the second reaction, the 3' hydroxyl of the 5'exon is responsible for the nucleophilic attack at the 3' splice site. With the second reaction, the two exons are joined together and the intron is excised as a lariat which is subsequently degraded. In the two catalytic steps of the splicing reaction, bonds are neither broken nor formed. Instead, both steps consist of an SN2 reaction that involves a pentacoordinate transition state. It is not known which components of the spliceosome catalyse the splicing reaction; it could be the snRNA components or some of the many proteins that are found in the spliceosome. Both U1 and U4 leave the spliceosome assembly before any catalytic reaction takes place and therefore they can be
Figure 1.3. Schematic representation of the two trans-esterification reactions in cis-splicing. Both the 5' and 3' splice sites are part of the same mRNA precursor. In the first reaction, the 2' OH of the adenosine residue of the branch site attacks the 5' splice site and the intron takes the form of a lariat structure. In the second reaction, the OH of the 5' exon attacks the 3' splice site and the two exons are joined together whereas the intron is excised as a lariat structure which is subsequently degraded. The two nucleophilic reactions are represented by the dotted arrows.
eliminated as possible candidates possessing catalytic activity. Preliminary data suggest that it is the snRNAs (in particular U6) that form the active site. Divalent cations are known to be required for ribozyme activity (Narlikar GJ 1997). Phosphate groups in which phosphorus has been substituted with sulphur are less efficient at coordinating Mg$^{2+}$ but they are more efficient at coordinating Cd$^{2+}$ or Mn$^{2+}$. Interestingly, one of such substitutions in the U6 snRNA impaired splicing in the presence of Mg$^{2+}$. Moreover, splicing activity was recovered by addition of Cd$^{2+}$ or Mn$^{2+}$, which are better coordinated by sulphur (Yean SL 2000). This indicates that the U6 snRNA is involved in the catalytic activity of the spliceosome. It is not known whether the catalytic activity resides exclusively in U6 snRNA or if proteins are required as well.

A splicing reaction can be classified based on how many pre-mRNA molecules take part in it. In the vast majority of cases, splicing is an intramolecular process in which both the 5' and 3' splice sites are part of the same pre-mRNA molecule. This type of splicing is known as cis-splicing. Some cases are known in which two pre-mRNA molecules are involved in the splicing reaction. In this case, one molecule provides the 5' splice site whereas the other molecule contributes the 3' splice site as shown in figure 1.4. This reaction is called trans-splicing and will be described in detail in another section.

1.6. SR proteins
SR proteins are involved in both constitutive and alternative splicing.

Structure of SR proteins: all the members of the SR family of proteins have at least one RNA binding domain (RBD) at their N-terminus, while the C-terminus consists of an arginine-serine rich domain, the so called RS domain. Some SR proteins such as SF2/ASF, SRp40, SRp55 and SRp75 have a second RBD, while others like SC35 and SRp20 possess a single RBD domain. The RBD is responsible for substrate specificity. Different SR proteins have different RBDs and hence bind different RNA sequences.

The RS domain undergoes reversible phosphorylation which in turns affects protein-RNA and protein-protein interactions, and also affects the localisation of the SR proteins within the cell, and their recruitment to active transcription sites.

Phosphorylation: as mentioned before, SR proteins undergo reversible phosphorylation. Two kinases that phosphorylate SR proteins have been well characterised and are known as SRPK1 (Gui JF 1994a; GuiJF 1994b) and Clk/Sty (Ben-David Y 1991; Howell BW 1991). Clk/Sty has a broader range of target proteins compared to SRPK1. In fact not only can Clk/Sty phosphorylate SR proteins, but it can also phosphorylate other types of proteins. In contrast, the kinase activity of SRPK1 appears to be more specific for SR proteins.
Figure 1.4. Schematic representation of the two trans-esterification reactions in trans-splicing. The two exons are part of different pre-mRNA molecules. The two reactions are essentially the same as in cis-splicing. In the first reaction, the 2' OH of the adenosine residue of the branch site attacks the 5' splice site and the intron takes the form of a Y-shaped structure. In the second reaction, the OH of the 5' exon attacks the 3' splice site and the two exons are joined together whereas the intron is excised as a Y-shaped structure which is subsequently degraded. The two nucleophilic reactions are represented by the dotted arrows.
Phosphorylation is a very important process because it controls the localisation of SR proteins as well as their activities in splicing.

It was shown that SR proteins are localised in nuclear compartments called speckles. Based on their different appearance, speckles can be classified into two types (Fakan 1994):

1) Interchromatin granule clusters (IGCs) which have the appearance of granules and are found in the interchromatin space.

2) Perichromatin fibrils (PFs) of about 5nm in diameter.

Generally genes that are being transcribed are found within PFs but not in IGCs (Jackson DA 1998). It has been shown that phosphorylation of specific serine residues within the RS domain of SR proteins is essential for the migration of SR proteins from nuclear speckles to sites of active transcription (Misteli T 1998).

**Role of SR proteins:** SR proteins bind sequences in RNA known as exonic splicing enhancers (ESEs) and stimulate the use of weak splice sites and poorly defined exons. In some cases, different SR proteins can have antagonistic effects on alternative splicing (Gallego ME 1997; Jumaa H 1997).

The SR protein SF2/ASF can act antagonistically with hnRNPA1 to determine splice site selection (Eperon IC 2000). Interestingly, it has been shown that there is significant variation in the ratio between SF2/ASF and hnRNPA1 in different rat tissues (Hanamura A 1998). The activity of SR and hnRNP proteins is regulated by phosphorylation. Therefore, by affecting the activity of kinases that phosphorylate SR and hnRNP proteins, alternative splicing could be regulated by extracellular signals. RNA-interference (RNAi) experiments carried out in *C. elegans* have demonstrated that the ortholog gene of the mammalian SF2/ASF is essential. In contrast, knock out experiments of other members of the SR protein family by RNAi did not cause any obvious phenotypic abnormalities, demonstrating redundancy in these genes (Kawano T 2000; Longman D 2000). Interestingly, the simultaneous knock out of more than one SR protein resulted in some cases in lethality or abnormalities.

The effect of SR proteins on alternative splicing and their ability to bind ESEs has been recently investigated as a potential therapeutic strategy to cure genetic diseases that are due to a fault in splicing (Skordis LA 2003) and will be discussed in the section that deals with alternative splicing.

**Involvement of SR proteins in trans-splicing:** the effect of SR proteins was studied with a trans-splicing substrate derived from a nematode. These experiments showed that SR proteins are required for the addition of the U2 snRNP onto the branch site of a naturally trans-spliced pre-mRNA (Romfo CM 2001). In order to stimulate U2 snRNP binding, SR
proteins must be phosphorylated (Furuyama S 2002). SR proteins were also shown to be required for the subsequent steps of trans-splicing such as the addition of the spliced leader (SL) and the U4/U6.U5 snRNP and again their phosphorylation state was important. For example, dephosphorylation of the SR protein 9G8 abrogated its ability to promote the trans-splicing reaction (Furuyama S 2002).

1.7. Alternative splicing

The sequencing of the entire genome from several higher eukaryotes indicates that the number of genes in relatively complex organisms such as Drosophila, is lower compared to the number of genes in much simpler organisms such as the nematodes. Another discrepancy is that the number of genes in a particular organism is believed to be insufficient to account for the complexity of its proteome. In order to obtain such coding potential, it is estimated that a single gene should encode for more than one protein. This is believed to be achieved mainly through alternative splicing. As mentioned before, alternative splicing is the alternative use of 5' and 3' splice sites in pre-mRNA molecules. One of the most striking example that underscores the role of alternative splicing as the main source of protein diversity is the Drosophila gene homolog of human Down Syndrome cell adhesion molecule (dscam). In this gene, exons 4, 6, 9 and 7 are encoded by sets of mutually exclusive exons. In particular, there are 12 alternative copies of exon 4, 48 copies of exon 6, 33 copies of exon 9 and 2 copies of exon 12. If all combinations were to be used, this gene would encode 38,016 different mRNA isoforms, which is far more than the number of genes in Drosophila.

Alternative splicing can not only increase the protein repertoire of an organism but it can also act as an on-off switch of gene expression through the introduction of premature stop codons in mRNA (Bopp D 1991).

Alternative splicing can operate in several different ways such as exon inclusion or exclusion, mutually exclusive exons, the use of alternative 5' or 3' splice sites and intron retention. Several examples of alternative splicing are reported in figure 1.5. In each case, different mRNA molecules are produced that can potentially encode functionally or structurally different proteins. Alternatively, a specific mRNA isoform can be targeted for NMD through the insertion of premature termination codons (PTCs).

Alternative splicing is said to be constitutive when the ratio between different mRNA isoforms remains constant, but in most cases, the production of different mRNA isoforms by alternative splicing is regulated in response to developmental and physiological stimuli.
Figure 1.5. Alternative splicing can be accomplished in several different ways. (A) Exon inclusion/exclusion. (B) Mutually exclusive exons. (C) Alternative 5’ splice sites. (D) Alternative 3’ splice site. (E) Intron retention.
Alternative splicing is the result of a sophisticated interplay between trans-acting factors and cis-acting sequences. The most important cis-acting sequences are intronic splicing enhancers (ISEs), exonic splicing enhancers (ESEs), intronic splicing silencers (ISSs) and exonic splicing silencers (ESSs). This classification is based on the location of the sequence (exonic or intronic) and on the effect that it has on splicing (stimulation or repression).

Exonic splicing enhancers are believed to be bound by SR proteins through the RRM (RNA recognition motif). As mentioned before, SR proteins stimulate exon inclusion. Currently, there are two different mechanisms that could explain the effect of SR proteins on splicing which are summarised as follows:

1) RS-domain-dependent mechanism. The RRM domain of an SR protein binds to an ESE. The bound protein contacts U2AF$^{35}$ and or U1-70K via its RS domain. The U1-70K protein is part of the U1 snRNP. This model is illustrated in figure 1.6 A.

2) RS-domain-independent mechanism. In this case the SR protein exerts its effects by antagonising an inhibitory protein that binds an ESS which juxtaposes to the ESE. This mechanism is shown in figure 1.6 B.

These two models are not mutually exclusive and in some cases it is believed that both mechanisms are implemented simultaneously (Zhu J 2000).

It has long been known that point mutations that reside in exons can cause genetic diseases. Sometimes this is the result of a premature termination codon created by the mutation, or alternatively a mutation can cause the substitution of an amino acid residue that is essential for the function of the protein. Interestingly, some sense mutations are known to cause genetic disorders. These types of mutations are believed to disrupt cis-acting elements that are important in splicing, such as exonic splicing enhancers. Exonic splicing enhancers are often found in exons that are flanked by weak splice sites (Fairbrother WG 2002). In these circumstances, retention of the alternative exon in the mRNA requires the presence of a functional exonic splicing enhancer. One example of a genetic disease that can be caused by mutation of an ESE is spinal muscular atrophy (SMA). SMA is an autosomal recessive disorder which is caused by mutation in the SMN gene. There are two SMN genes named SMN1 and SMN2. Mutations in the SMN1 gene cause SMA whereas the SMN2 gene acts as a modifying gene. Both SMN genes are located on chromosome 5 and are composed of 8 exons. SMN1 produces full length mRNA whereas in SMN2, exon 7 is mainly skipped from the mRNA. This is due to a mutation in exon 7 that disrupts an ESE that is bound by the SR protein SF2/ASF. A strategy with
Figure 1.6 Different mechanisms by which SR proteins regulate splicing through binding of exonic splicing enhancer sequences (adapted from Cartegni et al, 2002). (A) RS-domain-dependent mechanism. In this case the SR protein binds to the ESE through its RRM motifs and then interacts with U2AF35(U2 auxiliary factor). U2AF65 binds to the polypurimidine tract, which in this specific case is weak because it is interrupted by purines (R). U2AF35 binds the 3’ splice site. There are three sets of interactions represented by the U2AF-3’ splice site, U1snRNP-5’ splice site and finally the SR protein-ESE. These interactions are strengthened by protein-protein interactions which are represented by the blue arrows and mediated by the SR protein. In some cases, the splicing co-activator Sm160 forms a bridge between the various interactions. (B) RS-domain-independent mechanism. The SR protein acts by inhibiting the binding of a negative regulator of splicing. This is due to the close proximity of the sites to which the SR protein and the negative regulator bind (ESE and ESS, respectively). The RS domain of the SR protein is not shown even though it is present, and can still mediate protein-protein interactions. The inhibitory interactions are shown in red whereas the question mark represents a putative stimulatory interaction. It is thought that splicing of some introns involves a combination of the two mechanisms.
Inhibitor

ESE

YRRYYR-AG

ESE

U1 snRNP

U2 snRNP

Srm 160

70K

EXON

GU

A

B

U2AF65

RRM

Inhibitor

YRRYYR-AG

ESE

ESS

U2AF65

35
potential therapeutic applications has recently been developed to correct the splicing defect that causes SMA (Skordis LA 2003). It consists in the use of modified oligonucleotides composed of two parts. The 5’ end of the modified oligonucleotide is made of a noncomplementary tail containing ESE sequences that can be bound by SF2/ASF whereas the 3’ end of the oligo is complementary to exon 7 of the SMN2 gene. The region of complementarity was designed in a way that would not block any splice sites. It was shown that the oligonucleotide containing the enhancer sequence stimulates incorporation of exon 7 in vitro to a level comparable with that of SMN1. Moreover, treatment of fibroblasts derived from SMA patients with the same oligonucleotide increases incorporation of exon 7 to the levels found in normal individuals. These findings underline the importance of exonic splicing enhancers in ensuring exon incorporation when it is needed.

Another important class of cis-acting sequences involved in alternative splicing is represented by the exonic splicing silencers (ESS). Their mechanism of action is not well understood. It is believed that they operate by binding negative regulators of splicing, which usually belong to the heterogeneous nuclear ribonucleoprotein (hnRNP) family. This is a class of RNA-binding proteins that bind to nascent RNA molecules. hnRNPs are composed of an RNA-binding domain and an auxiliary domain, which often mediates protein-protein interactions. The best characterised members of this family of proteins are the hnRNP I, which is also known as polypyrimidine-tract-binding protein or PTB) and the hnRNP A/B and hnRNP H (Chen CD 1999).

Currently, two different mechanisms of action have been envisaged for ESSs. The first mechanism is by direct competition. It involves the presence of an ESE and an ESS in close proximity. Binding of a negative regulator to the ESS prevents binding of a positive regulator to the ESE and this leads to exon exclusion. The second mechanism is by nucleation and cooperative binding. This can take place when several ESEs and ESSs are present in an exon. Initially, a negative regulator of splicing binds to a high affinity ESS. This is followed by the cooperative binding of more negative regulator molecules, which in turn prevents the binding of positive regulators to ESEs and as a result, the exon is not included in the mRNA molecule. The two mechanisms of action of the ESSs are illustrated in figure 1.7.

It has recently been shown that 5’ splice site selection can be determined by competition between proteins that possess antagonistic effects (Eperon IC 2000). More specifically, hnRNP A1 causes a reduction in the binding of U1 snRNP to all 5’ splice sites whereas the splicing factor SF2/ASF increases the binding of U1 snRNP to all 5’ splice sites. When the concentration of hnRNP A1 is high, binding of U1snRNP to the 5’ splice sites is
Figure 1.7 Two different models for splicing silencing (exon inclusion). (A) Silencing by direct competition. The SR protein, which stimulates splicing, binds to an exonic splicing enhancer (ESE). On the other hand, a negative regulator of splicing such as an heterogeneous nuclear ribonucleoprotein (hnRNP) binds to an exonic splicing silencer which overlaps with the ESE. The overlapping of ESE and ESS makes the binding of the SR protein and hnRNP mutually exclusive. Which factor will bind to the exon depends on the relative affinity of the proteins for their cognate sequences, and their concentration. If the SR protein is bound the exon will be retained in the mRNA. In contrast, if the hnRNP protein is bound, the exon will be skipped from the mRNA. (B) Silencing by nucleation and cooperative binding. The alternative exon contains multiple ESE and ESS sequences. When the inhibitory factor is absent the SR proteins binds the ESE and stimulates inclusion of the alternative exon. In contrast, when the inhibitory factor is present, it first binds to a high affinity site and then proceeds to nucleate cooperative binding of other inhibitory factor molecules. In these conditions, the SR protein cannot bind to the ESE and the exon is excluded from mRNA.
A  Direct competition

Exon inclusion

Exon skipping

B  Nucleation and cooperative binding

Exon inclusion

Exon skipping
determined by their affinity for U1, and this determines which 5' splice site will be selected. In contrast, with high concentrations of SF2/ASF, all 5' splice sites are bound by U1 and the downstream 5' splice site is favoured.

Alternative splicing is not only influenced by the presence of cis-acting sequences and the concentration of their cognate binding factors. In particular, new evidence has emerged indicating that the rate of transcriptional elongation can influence splice site selection. This was shown by using a mutant pol II enzyme that is characterised by a reduced elongation rate. When a minigene containing the alternatively spliced exon EDI (extra domain I) from the human fibronectin gene was transiently transfected into mammalian cells, transcription driven by the slow mutant pol II enzyme increased inclusion of the alternatively spliced exon by approximately 4-fold (de la Mata M 2003). Similarly, in constructs that contained several 5' splice sites that could splice to a common downstream 3' splice site, the slow polymerase increased the use of the upstream 5' splice site approximately 3-fold (de la Mata M 2003). Inclusion of the EDI exon has been shown to depend on an exonic splicing enhancer that binds the splicing factor SF2/ASF. Mutation of the ESE abolished inclusion of the EDI exon and even the slow pol II enzyme could not restore inclusion and this demonstrated that the change in the rate of transcriptional elongation could not replace the effect of SF2. Interestingly, the slow pol II was shown not to affect inclusion of the fibronectin exon EDII, suggesting that transcription rate affects alternative splicing of some exons but not others. It is believed that a slow transcriptional elongation favours usage of the upstream splice sites because they will have more time to undergo splicing before a downstream splice sites is synthesised. Consistently, alternative splicing is also influenced by the promoter (Cramer P 1997; Cramer P 1999; Kadener S 2001; Kadener S 2002). Promoters that stimulate processive elongation tend to increase the skipping of weak alternative exons because the stronger downstream splice sites will be transcribed before the weak ones have had time to undergo splicing (Kadener S 2001; Kadener S 2002). Moreover, promoters are believed to affect alternative splicing through the transcription factors that are bound to them. This was inferred by studying the effects on alternative splicing of the alternative exon EDI from the fibronectin (FN) promoter (region comprised between -220 and +44) and a mutant of the FN promoter in which some binding sites for transcription factors have been mutated so that they are no longer bound. It emerged that exon inclusion from the mutant promoter was 2.7-fold higher than that from the wild-type promoter (Cramer P 1997). It has therefore been proposed that alternative splicing is regulated by the binding sites for transcription factors that are present in a promoter.
1.8. The carboxy-terminal domain of RNA polymerase II large subunit

The enzyme RNA polymerase II transcribes protein-encoding genes and also many genes that encode small nuclear RNAs. In eukaryotes, the enzyme is composed of 12 subunits (Sentenac 1985). The large subunit (Rpbl) has been well characterised. The carboxy-terminal domain (CTD) of the large subunit is an evolutionarily conserved feature. It consists of a heptapeptide repeat with the consensus sequence YSPTSPS. In mouse, the heptapeptide is present in 52 copies whereas the yeast S. cerevisiae has only 26 or 27 copies. The heptad repeats are followed by 10 amino acid residues (ISPDDSDEEN), which are highly conserved among mammals. Interestingly, the CTD is positioned near the RNA exit channel (Cramer P 2001). The serine residues at positions number 2 and 5 are important because they can be phosphorylated and dephosphorylated during transcription (Komarnitsky P 2000; Puttarajan M 1998; Schroeder SC 2000; Zhang J 1991). The CTD of pol II is phosphorylated by several kinases. Among these are the cyclin H-Cdk7 and cyclin T-Cdk9. Cyclin H-Cdk7 is part of the TFIIH complex and is involved in Ser-5 phosphorylation. In contrast, cyclin T-Cdk9 is a component of P-TEFb, which phosphorylates Ser-2 and also stimulates pol II elongation. It is believed that shortly after transcription initiation, trans-acting factors can induce RNA pol II to abort transcription. Instead, P-TEFb acts upon RNA pol II and stimulates transcription elongation. The kinase activity of P-TEFb is required for the stimulation of elongation of RNA pol II. For example, the molecule 5,6-dichloro-1-β-D-ribofuranosyl-benzimidazole (DRB) inhibits transcription elongation by affecting the kinase and elongation activities of P-TEFb (Marshall NF 1995; Marshall NF 1996). Two other factors have been found that are sensitive to DRB (Renner DB 2001). One is called DRB sensitivity-inducing factor (DSIF) and the other is known as negative elongation factor (NELF). These two factors act together to increase the time that RNA polymerase II spends at pause sites in an in vitro system (Renner DB 2001) and their action can be abolished by P-TEFb.

In Saccharomyces cerevisiae, it was shown by chromatin immunoprecipitation that pol II molecules are equally distributed along a gene (Komarnitsky P 2000; Schroeder SC 2000). However, polymerase molecules engaged at a different stage of transcription possess a different phosphorylation pattern (Komarnitsky P 2000). In particular in the promoter-proximal region, RNA pol II molecules are mainly phosphorylated at Ser-5. In contrast, during elongation, Ser-5 becomes dephosphorylated and there is a concomitant phosphorylation of Ser-2 (Komarnitsky P 2000). Interestingly it was shown that Ser-5 phosphorylation is required for the recruitment of the capping enzymes (Cho EJ 1997;
McCracken S 1997). Other pre-mRNA processing events, such as splicing and polyadenylation are known to be coupled to transcription by mean of the CTD of pol II. Interestingly, it is the hypophosphorylated form of pol II that interacts with splicing factors such as SR proteins (Hirose Y 1998; Hirose Y 1999; Misteli T 1999; Fong N 2001). The distribution of RNA polymerase II and its different phosphorylation states were also studied in mammalian genes (Cheng C 2003). The authors investigated the distribution of pol II in two genes: dihydrofolate reductase and γ-actin. They found that in both genes, pol II was more concentrated in the promoter-proximal region. Moreover, pol II molecules with a high Ser-5 phosphorylation were found mainly in the promoter-proximal region. In contrast, pol II molecules in which Ser-2 is phosphorylated appeared to be uniformly distributed along the genes. In conclusion, the phosphorylation state affects the properties of RNA polymerase II. In particular, based on the phosphorylation extent of the serine residues in the heptapeptide repeats, two distinct forms of RNA pol II are found. RNA pol IIa has a hypophosphorylated CTD and is the form that is found in the transcription initiation complexes which assemble in the promoter region of a gene. Shortly after transcription initiation, the CTD becomes hyperphosphorylated and this converts RNA pol IIa into RNA pol IIo. It is believed that the phosphorylation of the CTD transforms the RNA pol II enzyme from a form that is prone to pausing or even transcription termination, to a form that is capable of processive elongation. Once RNA pol II transcribed a gene, it is believed that dephosphorylation converts it back into RNA pol IIa.

It has been shown that the link between transcription and pre-mRNA processing is mediated by the CTD of Rpb1. In particular, the N-terminal heptads 1-15 or 1-25 support capping of nascent pre-mRNA whereas heptads 27-52 including the last 10 amino acid residues support splicing and 3’ end formation (Fong N 2001). A more detailed analysis of the effect of the CTD of RNA pol II on splicing and polyadenylation was carried out by deleting the last 10 amino acid residues of Rbp1. This deletion mutant was unable to support efficient transcription, splicing and 3’ end formation (Fong N 2003).

1.9. Coupling between transcription and splicing

The first step in gene expression is transcription, which is carried out by the enzyme RNA polymerase II. Transcription yields a pre-mRNA molecule which undergoes several modifications (capping, splicing and polyadenylation) that transform it into a messenger RNA molecule (mRNA) that can be translated into protein. A considerable amount of experimental evidence suggests that pre-mRNA processing is coupled to transcription. In this section, only the link between splicing and transcription will be discussed. It has been
shown that introns can be spliced while the pre-mRNA molecule is being transcribed. In particular, electron micrographs of Drosophila embryo genes showed that introns are removed by splicing while the pre-mRNA molecules are still attached to the DNA template and some introns near the 5' end of the gene are removed before introns at the 3' end of the gene have been synthesised (Beyer AL 1988).

Another indication of the coupling between transcription and pre-mRNA processing is the effect of secondary structures on splice site selection in vivo but not in vitro (Eperon LP 1988). This was discovered with constructs that contained two competing 5' splice sites. The downstream splice site was named the reference 5' splice site and corresponded to the authentic 5' splice site of the second intron of the rabbit β-globin gene whereas the upstream 5' splice site was termed test 5' splice site (see figure 1.8.). Transient transfection experiments in HeLa cells revealed that the reference and test splice sites were used with efficiencies of 36% and 64% respectively. In a second construct, a sequence that can base pair to the test 5' splice site was placed just upstream of it and this caused the exclusive usage of the reference 5' splice site. Interestingly, separation of the complementary sequence and the test splice site increased the use of this splice site. Moreover, there seemed to be a threshold effect (around 60 nucleotides) of the distance between the two sequences. Remarkably, the separation of the test 5' splice site and its complementary sequence had no effect on splice site selection in an in vitro experiment. The interpretation was that proteins and possibly snRNPs associate with the RNA molecule while it is being transcribed and once factors are bound to the RNA molecule, base pairing can no longer take place. When the sequence separating the two complementary sequences is too short, the factors might not have sufficient time to bind to the RNA molecule and it was postulated that in these conditions the secondary structure can be formed.

More direct evidence supporting the cotranscriptional nature of splicing has come from experiments aimed at determining the excision of introns from nascent pre-mRNA molecules. In particular, the presence of introns in RNA molecules produced from a particular gene was compared between RNA molecules still attached to the DNA template and RNA molecules that had been released from the transcribing enzyme RNA pol II (Bauren G 1994). For this purpose, the BR (balbiani ring) genes in the salivary glands of Chironomus tentans were chosen. With this system, it is possible to separate the nucleoplasm which contains released RNAs, from the chromatin with its attached nascent RNAs. Another advantage of this system is the length of exon 4 (around 35 Kb) in the BR gene, as shown in figure 1.9. The kinetic of the removal of introns 3 and 4 was studied. It was estimated that in about 70-75% of nascent transcripts, intron 3 was co-transcriptionally
Figure 1.8. Splice site choice in vivo is influenced by the distance between the two alternative splice sites. A portion of the rabbit β-globin gene is shown. Exons are represented as boxes whereas the intron is shown as a black line. (A) Construct consisting of a portion of the rabbit β-globin gene with two alternative splice sites. The authentic (or reference) 5' splice site is indicated by the black arrow whereas the other 5' splice site, named the test splice site, is indicated by the dotted arrow and has been introduced artificially. It consists of 9 nucleotides from the authentic 5' splice site. (B) New construct in which the same sequence containing the test 5' splice site was inserted in the construct shown in A, upstream of the test splice site but in inverted orientation. The two sequences can base pair in the RNA molecule transcribed from this construct. (C) The size of the putative loop structure between the two inserts can be increased by inserting a DNA segment between the test 5' splice site and its complementary sequence. Sequences comprised between 16-118 nucleotides in length were inserted (represented by the grey box) and tested. Adapted from Eperon et al, 1988).
removed. Accordingly, only 2-4% of released BR1 RNAs were found to be unspliced. These data represent direct proof of the cotranscriptional removal of introns.

There is also indirect evidence supporting the coupling between transcription and splicing. For example, it was shown that antibodies against snRNPs localise to the transcribing chromatin loops, suggesting that snRNPs are recruited onto nascent pre-mRNA molecules (Gall 1991). Splicing factors appear to be recruited to sites of active transcription as well. These proteins are normally concentrated in several nuclear regions called speckles (Spector 1993). In order to study the movement of splicing factors in living cells, the splicing factor SF2/ASF was fused to the green fluorescent protein (GFP) and the construct was transfected into mammalian cells. The fusion protein localised at first in nuclear speckles. Interestingly, the fusion protein migrated to sites of active transcription, suggesting that splicing factors interact with nascent pre-mRNA molecules (Misteli T 1997). Speckles are highly dynamic structures, and their shape changes with time. Quite remarkably, in cells treated with the transcription inhibitor α-amanitin, the movements in speckles were abrogated. Other indirect evidence of the link between transcription and splicing comes from the observation that the rate of transcription of pre-mRNA has an effect on splicing and this will be discussed in detail in the section dealing with alternative splicing.

The coupling between transcription and splicing is believed to be mediated by the carboxy-terminal domain (CTD) of the large subunit of RNA polymerase II. Among the evidence suggesting this is the observation that splicing is inhibited by antibodies against the CTD of RNA pol II (Chabot B 1995; Yurvey A 1996), and also by peptides corresponding to the CTD repeats (Yurvey A 1996; Du L 1997). Moreover, phosphorylated RNA polymerase II stimulates pre-mRNA splicing directly (Hirose Y 1999). Interestingly, RNA polymerase mutants carrying a deletion of the CTD are unable to support splicing as was explained in the section dealing with the CTD of RNA polymerase II.

1.10. Trans splicing

Trans splicing has been observed for the first time in trypanosomes (Murphy WJ 1986; Sutton RE 1986). In these organisms, all mRNA molecules have at their 5' end a common sequence called SL (spliced leader). The SL RNA is transcribed from a gene present in about 200 copies (Nelson RG 1983). Transcription of the SL genes produces a small RNA molecule of 135 nucleotides. The first 35 nucleotides represent the SL sequence, which will then be spliced to the mRNAs whereas the remaining 100 nucleotides constitute an outron (i.e. an intron with only one splice site, in this case a 5' splice site).
Figure 1.9. Diagrammatic representation of the balbiani ring 1 (BR1) gene. It is composed of 5 exons which are represented by the white boxes. The exon numbers are shown under each box. Exon 4 is unusually long (approximately 35 Kb). With this gene, it is possible to physically separate the RNA molecules that have become detached from the gene from those that are still being transcribed (nascent pre-mRNA molecules). This in turn, allowed to establish that splicing is a co-transcriptional process as explained in the text.
Two possible mechanisms for the addition of SL to mRNA could be conceived. The first one is primed transcription, in which the SL RNA is used to prime transcription of the pre-mRNAs. This would lead to the production of pre-mRNAs containing both the SL and the coding exons. These pre-mRNAs could then be processed linearly by cis-splicing to create chimeric mRNA molecules. The other mechanism is by trans-splicing. In this case the two pre-mRNAs are transcribed separately and subsequently joined together. The two mechanisms differ in the type of intermediate molecules that they produce. With the first mechanism, a lariat intron would be generated. In contrast, if the mechanism is trans-splicing, a Y shaped intermediate would be formed. Experimental evidence confirmed the presence of Y shaped intermediates. For example, incubation of poly(A)\textsuperscript{+} RNAs from Trypanosoma brucei with debranching enzyme released the 100-mer outron that was part of the SL RNA (Murphy WJ 1986). These data demonstrated that addition of the SL RNA to the mRNAs occurs by trans-splicing.

Trans-splicing has been well characterised also in \textit{C.elegans}. This organism has two SL (splice leaders) molecules called SL1 and SL2. SL1 trans-splices to RNA molecules that possess an outron, in this case an AU-rich sequence that contains a 3' splice site but lacks an upstream 5' splice site. It was shown that a normal gene (one that does not contain an outron) undergoes trans-splicing when an outron is added at the 5'end (Conrad R 1991). On the other hand, the removal of an outron from a transcript that normally undergoes trans-splicing abolishes this reaction (Conrad R 1993). In \textit{C elegans}, approximately 25 % of genes are part of polycistronic transcription units. SL2 is added to the downstream gene products of such transcripts (Spieth J 1993).

Trans-splicing has also been observed \textit{in vitro} as well. When two pre-mRNA molecules, one carrying a 5'splice site and the second one carrying the branch site and a 3' splice site are mixed together in the presence of a nuclear extract, the exons can be joined by trans-splicing. It has been shown that the presence of complementary sequences in two pre-mRNA molecules can significantly increase the efficiency at which they undergo trans-splicing (Konarska MM 1985). These experiments were made with two sets of RNA molecules. In the first set, one molecule was composed of the L1 exon of adenovirus and the downstream intron. The second molecule consisted of a part of intron and the L2 exon. The second set of molecules was similar to the first one except for the presence of a complementary 38 nt sequence at either the 3' or 5' end. The two RNA molecules could associate through base pairing of the 38 nt complementary sequences. The two annealed RNAs underwent splicing at an efficiency of about 15% of that obtained from a pre-mRNA in which the two exons were part of the same molecule (cis-splicing). Without pre-
annealing, the efficiency of trans-splicing was about 8 times lower. Furthermore, it was shown that after two hours of incubation in reaction conditions, the two RNAs were associated like with the pre-annealing conditions. It was also demonstrated that the two pre-mRNA molecules were not covalently associated prior to splicing. The efficiency of trans-splicing between the two RNAs without complementary sequences was only 0.2% compared to cis-splicing. In another paper (Solnick 1985), the efficiency of trans-splicing versus cis-splicing was investigated in vitro, in molecules capable of base-pairing to one another. For this purpose, two chimeric plasmids were constructed. The first one contained the adenovirus first exon (AdE1) followed by 156 nucleotides of flanking intron joined to 53 nt of intron 1 of the β-globin gene and 53 nt of β-globin exon 2. The other plasmid consisted of 85 nt of the β-globin exon 1 and 56 nt of the flanking intron, joined downstream to 80 nt of the adenovirus first intron and the entire second exon of the adenovirus (AdE2). Each plasmid contained a sequence of 105 nucleotides at the junction between adenovirus and beta globin in one orientation, or in the opposite orientation, so that they could base pair to one another, as shown in figure 1.10. Transcripts were produced from these plasmids, annealed and incubated in a splicing reaction. A significant amount of mRNA composed of two adenovirus exons (AdEx1 and AdEx2) was produced by trans-splicing. Interestingly, when the two RNA molecules were mixed with a 14-fold excess of an RNA molecule corresponding to the 105 nucleotide sequence, trans-splicing was nearly abolished. Furthermore, switching of the same sequence in one construct prevented trans-splicing. All these results taken together demonstrated that trans-splicing between the two pre-mRNA molecules depended on the presence of complementary sequences.

The property of complementary sequences to favour trans-splicing has been exploited to devise a tool with potential applications in gene therapy (Puttaraju M 1999). This system was utilised to target the pre-mRNA of the β-subunit of the human chorionic gonadotropin gene 6 (βhCG6), which is expressed in many types of tumours but not in normal cells. Puttaraju and co-workers constructed a prototypical trans-splicing molecule (PTM) consisting of a binding domain (BD) composed of 18 nucleotides which are complementary to the branch site of intron 1 of the target gene and therefore can base pair to it. Downstream of the binding domain there is a spacer region of 30 nt, a branch point sequence, a polypyrimidine tract and a 3' splice site followed by an exon that encodes the A subunit of the diphtheria toxin. The system is depicted in figure 1.11. A pre-mRNA molecule corresponding to the PTM was capable of trans-splicing to an in vitro transcribed βhCG6 pre-mRNA molecule. In contrast, trans-splicing from PTM that did not contain the
Figure 1.10. Competition between cis-splicing and trans-splicing in two mRNA precursors that can base pair to each other through complementary sequences (represented by the two arrows). Two hybrid pre-mRNA molecules are used as substrates for the splicing reaction. Both molecules contain one exon from the adenovirus (black box) and one exon from the human β-globin gene (white box). Trans-splicing between the two molecules generates mRNAs with both exons of the adenovirus or the β-globin. In contrast, cis-splicing creates mRNAs composed of one exon from adenovirus and one from β-globin. Adapted from Solnick, 1985.
Figure 1.11. Diagrammatic representation of SMaRT (spliceosome-mediated RNA trans-splicing). Modified from Puttaraju et al, 1999. Two pre-mRNA molecules are shown. E1 is exon 1 of the β-subunit of the human chorionic gonadotropin gene 6 and E2 is exon 2 of the same gene. BD is the binding domain of the therapeutic RNA molecule. BP is its branch point, PTT is the polypyrimidine tract and finally DT-A is the A subunit of the diphtheria toxin. E1 can be joined to E2 by cis-splicing (shown by the blue arrow). Alternatively, E1 can be trans-spliced to the DT-A exon (shown by the red arrow).
BD (and therefore could not base-pair to the target molecule) was reduced approximately eightfold. It was also shown that increasing concentrations of the PTM molecule could reduce the efficiency of cis-splicing. These data were confirmed in transient transfection experiments and in this case trans-splicing occurred with an efficiency that was about 10% compared to that of cis-splicing. The authors tried to restore protein function by inserting intron 1 of βhCG6 into the gene that encodes β-galactosidase. They then inserted non-sense mutations in the 3’ portion of the β-galactosidase gene. Cotransfection of a PTM that could base-pair to intron 1 of βhCG6 and that had the 3’ end of the β-galactosidase as exon produced a functional enzyme, demonstrating that spliceosome-mediated RNA trans-splicing (SMaRT) can restore protein function in a non-functional gene. Finally, trans-spliced mRNA was also obtained in an in vivo model.

Naturally occurring examples of trans-splicing have also been observed. For example, one mRNA isoform from the Drosophila gene mod(mdg4) was found to be encoded on both DNA strands and the most obvious explanation was that the isoform was produced by trans-splicing (Labrador M 2001). Later on, more mRNA isoforms were discovered that were encoded on both strands of the same gene (Dorn R 2001). All the isoforms begin with the first four exons that are located at the 5’ end of the gene. The other exons are organized in five groups and are encoded on both DNA strands. The proof that the isoforms were originated by trans-splicing was obtained from transgenic flies in which the variable exons had been integrated on a different chromosome from the one where exons 1-4 are found. The expected mRNA isoforms were obtained and since the exons had been transcribed from different chromosomes, the mRNA molecules had to be produced by trans-splicing. The same results were obtained by positioning exons that were encoded on the same strand as exons 1-4 on a different chromosome, demonstrating that these isoforms as well were the result of trans-splicing. Apparently, transcription of many of the exons was driven by their own promoter elements (Dorn R 2001).

It was discovered that mammalian cells are capable of producing functional mRNA molecules by trans-splicing (Eul J 1995). A fragment of the early SV40 region comprised between BstXI and BamHI (see figure 1.12) was used to transform mammalian cells. Cell lines produced two types of mRNA which were named T1 and T2. The T1 and T2 mRNAs are translated into T1 and T2 antigens, respectively. T2 mRNA is derived from the removal of the small intron located between the 5’ splice site of the small t antigen and the 3’ splice site. In contrast, T1 mRNA contains the portion of the exon comprised between the 3’ splice site and the cryptic 5’ splice site, joined to the beginning of the exon. Genomic DNA analysis showed that a single SV40 fragment had been inserted into the cell lines. Since the
Figure 1.12. Diagrammatic representation of the large T antigen coding region of SV40 (modified from Eul et al, 1995). Two exons separated by one intron are shown. Inside the second exon there is a cryptic 5' splice site. The 5' splice site of the small t antigen, which is located inside the intron, is also shown. Downstream of it, is the 3' splice site associated with the second exon. The position of the recognition sequences of BstXI and BamHI are also shown.
cryptic 5' splice site is downstream of the 3' splice site, the T1 mRNA molecule had to be produced by trans-splicing between two SV40 pre-mRNA molecules. Interestingly, the presence of the T1 antigen correlated with a transformed phenotype, demonstrating a biological role for the protein derived by translation of the mRNA produced by trans-splicing. Surprisingly, some cell lines could produce T2 mRNA but not T1 mRNA. As shown in figure 1.12, there is a 5' splice site upstream of the 3' splice site. Therefore, if these two splice sites take part in a cis-splicing reaction, the 3' splice site will no longer be available for trans-splicing. From analysis of the RNA, it emerged that in the cell lines that were not producing the T1 mRNA, cis-splicing occurred with high efficiency. Therefore, cis-splicing and trans-splicing seemed to be competing against each other in this system.

Examples of mRNA molecules composed of exons derived from different genes were also discovered in mammals. One such example is the CYP2C mRNAs (Zaphiropoulos 1999). The cytochrome P450 2C (CYP2C) proteins are involved in the metabolism of foreign chemicals such as drugs. There are four known human CYP2C genes named 2C8, 2C9, 2C18 and 2C19 and they are clustered on chromosome 10. All four genes are composed of 9 exons and have a high degree of identity. The order of the genes on the chromosome is 2C18-2C19-2C9-2C8. 2C18 is the closest to the centromere whereas 2C8 is the closest to the telomere. RNA molecules containing exons from genes 2C18 and 2C19 were identified in liver and epidermis (Zaphiropoulos 1999). In order to see if RNA molecules containing exons from the most distal genes (2C18 and 2C8) were present, RT-PCR analysis was carried out on RNA extracted from liver and testis using primers on exon 1 and exon 9 of genes 2C8 and 2C18 respectively. Hybrid mRNA molecules, which are shown in figure 1.13, were identified in liver (Finta C 2000). One PCR product contained exons from 3 genes (2C18, 2C8 and 2C19). The same experiments carried out on RNA extracted from testis failed to detect hybrid molecules. Therefore, there is tissue-specificity in the generation of CYP2C8 hybrid mRNA molecules. The gene retinol binding protein 4 (RBP4) is adjacent to the CYC2C genes but RT-PCR reactions with primers specifically designed to detect hybrid RNA molecules between the CYP2C genes and the RBP4 gene failed to amplify any PCR products. This demonstrated that the production of hybrid RNA molecules was restricted to the CYP2C genes. A detailed analysis of the RNA molecules from genes 2C8 and 2C18 revealed the presence of scrambled exons and also some examples of exon repetition. In particular, a tandem repeat of exon 4 or exon 5 was found in CYP2C8 RNA. Since CYP2C8 is a single copy gene, the most likely explanation for the mRNA isoforms with repeated exons is that they are derived by trans-splicing. These
Figure 1.13. CYP2C hybrid mRNAs (modified from Finta et al, 1999). cDNA synthesis was performed with an oligo(dT) primer. Exons derived from gene 2C8 are shown as grey boxes. Exons from gene 2C18 are represented as white boxes. The black box is an exon derived from gene 2C9. (A) PCR products obtained with 2C8 exon 1 forward and 2C18 exon 9 reverse primers. (B) PCR products from amplification carried out with 2C18 exon 1 forward and 2C8 exon 9 reverse primers.
molecules are particularly complex since they contain both scrambled exons and repeated exons. Therefore, it is likely that they are derived from two trans-splicing reactions.

Another example of mRNA molecules containing exons from different genes is found in the cytochrome CYP3A genes. There are four known human CYP3A genes which are known as CYP3A4, CYP3A5, CYP3A7 and CYP3A43. All four genes are located on chromosome 7 and are shown in figure 1.14. CYP3A4, CYP3A5 and CYP3A7 are in the same orientation whereas CYP3A43 is in opposite orientation. Each gene is composed of 13 exons. Chimeric mRNA molecules were discovered in which exon 1 from CYP3A43 was joined to exons 2-13, or 4-13 or 7-13 of CYP3A4 and also to exons derived from CYP3A5. Since genes CYP3A4 and CYP3A5 are in opposite orientation with respect to gene CYP3A43, the possibility that a bicistronic pre-mRNA molecule is produced by bypassing transcription termination can be ruled out. The abundance of the CYP3A43/CYP3A4 chimeric mRNAs was estimated to be approximately 0.15% of the amount of CYP3A43 mRNA. Other examples of mRNA molecules derived from two different genes were found in rat (Joseph DR 1991), in plants (Kawasaki T 1999) and again in human (Chatterjee TK 2000; Fujieda S 1996; Tasic B 2002).

Another example of trans-splicing was found in a Drosophila gene named lola. In this case, mRNAs are produced by trans-splicing with high efficiency (Horiuchi T 2003). This process can also happen between transcripts transcribed from different alleles (interallelic trans-splicing), which is facilitated by chromosomal pairing.

Finally, there are also examples of mRNA molecules in which some of the exons are tandemly repeated in the mRNA, even though they are not duplicated in genomic DNA and this phenomenon is known as exon repetition.

1.11. Genes that undergo exon repetition

So far, exon repetition has been found only in a few mammalian genes (human and rat). In human, exon repetition has been observed in three genes. One of them, the Sp1 gene, which encodes for a transcription factor, produces an mRNA isoform that contains a tandem repeat of exons 2 and 3 (Takahara T 2000). Interestingly, exon 1 seems to be absent in the mRNA isoform containing the exon 2-3 repetition. The second gene is the cytochrome CYP2C8. In this gene, there is a repetition of exon 4 or exon 5 in mRNA (Finta C 2000). Exon repetition in the CYP2C8 gene is quite complex because the mRNA isoforms that contain repeated exons also contain scrambled exons (the exons in mRNA are found in a different order from that of the gene). The third gene is the human estrogen receptor alpha (hER alpha). In this case, exon 1A (which is the second exon in the gene) is
Figure 1.14. Schematic representation of the four human CYP3A genes (modified from Finta et al., 2002). All four genes are positioned on chromosome 7. Each gene is composed of 13 exons. Only exons 1, 2 and 13 are shown for each gene. The grey arrow indicates the orientation of the CYP3A43 gene whereas the orientation of genes CYP3A4, CYP3A7 and CYP3A5 is shown by the white arrows. Genes CYP3A4, CYP3A5 and CYP3A7 are in opposite orientation with respect to the CYP3A43 gene.
tandemly repeated in mRNA (Flouriot G 2002). Exon repetition has also been found in four genes in rat. In one gene named carnitine octanoyl transferase (COT), exon 2 alone or exons 2 and 3 are repeated (Caudevilla C 1998). In a second gene named SA, repetition involves either exon 2 alone or exons 2, 3 and 4 simultaneously (Frantz SA 1999). In the third gene, which encodes for a voltage gated sodium channel, exons 12, 13 and 14 are tandemly repeated in mRNA (Akopian AN 1999). Finally, the Sp1 gene is characterised by exon repetition in rat as well as in human. In rat, mRNA isoforms contain a tandem repetition of either exons 2-3 or exon 3 alone (Takahara T 2002).

In principle, exon repetition in mRNA could arise by a duplication of specific exons in the genomic DNA (Letunic I 2002). Alternatively, the whole gene could be duplicated and by by-passing transcription termination, a long pre-mRNA molecule containing a tandem duplication of all the exons would be produced. By eliminating most of the exons from the pre-mRNA, mRNA isoforms that contain a tandem repeat of specific exons could be produced as shown in figure 3.3. In the case of SA, a detailed analysis of the genomic structure of the gene did not detect exon duplication. In similar analysis of other genes for which exon repetition has been observed, exon/gene duplications have not been detected. Using this technique, it would not be possible to detect large duplicated regions. In all the cases in which the exons are not duplicated, the most likely explanation for exon repetition is that it arises during RNA processing (i.e. the repeated exons are derived from two pre-mRNA molecules that have been transcribed, or are being transcribed from the same gene).

1.12. Exon repetition in the SA gene

The SA gene is expressed at high levels in the proximal tubules of the kidney and in liver hepatocytes of the spontaneously hypertensive rat (SHR). In contrast, the level of expression observed in another strain of rat named Wistar-Kyoto (WKY) is much lower. The differences between the two strains extend beyond the level of expression. In particular, analysis of SA mRNA by RT-PCR revealed the presence of a single transcript in the kidney and liver of SHR and liver of the WKY rat. In contrast, parallel experiments from the proximal tubules of the WKY identified the presence of three SA mRNA isoforms of different size (Frantz SA 1999). One of them corresponded to the normal SA mRNA whereas the other two contained tandem repeats of specific exons (either exon 2 alone or exons 2-3-4). To rule out a PCR artefact, the presence of the mRNA isoforms with repeated exons was confirmed by RNase protection assay and RNase-H digestion. The presence of repeated exons in mRNA was termed exon repetition. Extensive analysis of genomic DNA demonstrated that exon repetition was not due to the duplication of specific exons or even
the entire SA gene. All this experimental evidence suggested that the most likely explanation for exon repetition was that it arised from an unusual RNA processing event.

1.13. Exon repetition in the COT gene

Exon repetition in the COT gene was discovered in the Sprague-Dawley rat (Caudevilla C 1998). In this case, either exon 2 alone or exons 2-3 were found to be repeated in mRNA. The duplication of exons 2 and 3 in the COT gene was ruled out by southern blot experiments. Therefore, it appeared that as in the case of SA, exon repetition in COT was also produced by a non-conventional post-transcriptional RNA processing event.

Western blot analysis with a polyclonal antibody specific for the COT protein detected two protein products of different molecular masses. One of them had a size consistent with the putative protein derived from translation of the COT mRNA isoform containing the tandem repeat of exons 2 and 3. This suggested that exon repetition might have a biological function. It was also shown that a putative exonic splicing enhancer in exon 2 of COT was sufficient to elicit the production of the mRNA isoform with a tandem repeat of exon 2 from a transiently transfected COT minigene (Caudevilla C 2001).
OBJECTIVES

Exon repetition is an extraordinary phenomenon by which apparently normal genes can insert a tandem repeat of specific exons in some, but not all the mRNA molecules that they produce. At present, very little is known about exon repetition. For example, it was discovered that exon repetition in SA is strain-specific, since it is observed only in one strain of rat (WKY) but not in another strain (SHR) (Frantz SA 1999). Considering the extent to which trans-acting factors affect pre-mRNA processing, the most likely explanation for the strain-specificity of exon repetition in SA is the presence of a strain-specific trans-acting factor that promotes exon repetition. But since this aspect has not been investigated, it remains purely speculative. Another intriguing question is why exon repetition of SA is tissue-specific (Frantz SA 1999). Again, the most likely explanation seems to be a trans-acting factor that causes exon repetition in SA and that is expressed in the proximal tubules but not in the liver.

A fundamental question of exon repetition that has not been sufficiently addressed is its biological significance. So far, it has only been proposed that exon repetition can potentially create new proteins (Caudevilla C 1998; Akopian AN 1999; Takahara T 2000). Other interesting aspects, such as a possible negative effect on the synthesis of the main protein product from mRNA isoforms that contain a short open reading frame (ORF) upstream of the ORF that encodes the full length protein has never been addressed.

The final question that needs to be answered is the mechanism underlying exon repetition. For example, in the COT gene, it was shown that a putative exonic splicing enhancer is required for exon repetition, but it is not known whether this ESE sequence is sufficient to cause exon repetition, or instead it is just one of the many DNA elements required. Moreover, it is not known if the presence of an ESE is necessary to produce exon repetition in every gene that exhibits this phenomenon, or if this is only limited to the COT gene.

The aim of this PhD thesis is to shed further light on some of the key questions relative to exon repetition, and this was done by studying the two best characterised examples of exon repetition, which are represented by two rat genes, SA and COT.
CHAPTER 2
MATERIALS AND METHODS

2.1. Culturing of cell lines (HeLa and 293 cells)

Cells were cultured in DMEM (Dulbecco’s Modified Eagle Medium, GIBCO, 4,500 mg/L glucose, with pyrodoxine, without sodium pyruvate) supplemented with 10% of FCS (foetal calf serum) and 1% of penicillin and streptomycin.

Conditions of growth were a humidified incubator at 37°C with 5% CO₂.

To split cells, the old medium was removed and cells were washed once with 1X PBS (phosphate buffered saline). A small amount of trypsin was added and cells were placed back in the incubator until they became loose. Cells were then completely detached by gently tapping the flask. Complete medium was added, and an appropriate number of cells (usually 1/8 of a confluent flask) were seeded in a new flask.

2.2. Bacterial strains

<table>
<thead>
<tr>
<th>Strain</th>
<th>Genotype</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>XL-1 Blue</td>
<td>SupE44 hsdR17 recA1 endA1 gyrA46 thi relA1 lac' F'[proAB' lacI8 lacZΔM15 Tn10(tet')]</td>
<td>Recombination-deficient strain that is used for DNA cloning. The F' in this strain allows blue/white screening on X-gal</td>
</tr>
<tr>
<td>LE392</td>
<td>SupE44 supF58 hsdR514 galK2 galT22 metB1 trpR55 lacY1</td>
<td>Suppressing strain commonly used to propagate λ vectors and their recombinants. LE392 is a derivative of ED8654</td>
</tr>
</tbody>
</table>
2.3. Growth media for bacteria

**LB (Luria-Bertani Medium)**
1% (w/v) tryptone (Oxoid)
0.5% (w/v) yeast extract (Oxoid)
1% NaCl

**LB agar**
The same as LB with the addition of 1.5% of Agar Bacteriological (Oxoid)

**T-broth**
1% Tryptone
0.5% NaCl

2.4. Preparation of cloning vectors

**Restriction digestion:** approximately 1 μg of cloning vector was cleaved with 3-5 units of the appropriate restriction enzyme in a final volume of 20 μl.

**Dephosphorylation:** to the restricted cloning vector, the following components were added:
- 10 μl of CIAP 10X Reaction Buffer (500 mM Tris-HCl, pH 9.3, 10 mM MgCl2, 1mM ZnCl2, 10 mM spermidine)
- 2 μl of CIAP (Calf Intestinal Alkaline Phosphatase, Promega)
- 68 μl of H2O, bringing the final volume up to 100 μl

The reaction was incubated at 37 °C for 2-3 hours, after which the volume was brought to 200 μl by adding 100 μl of H2O. Then 100 μl of phenol and 100 μl of IAC (chloroform and isoamyl alcohol, 24:1) were added. The tube was shaken vigorously and the two phases were separated by centrifugation at 13,000 rpm for 3 minutes. The upper phase was transferred to a new tube and extracted with 200 μl of IAC. The upper phase was transferred to a new tube and the cloning vector was precipitated by addition of 20 μl of 3 M Na acetate pH 5.2 and 200 μl of absolute ethanol. The tubes was inverted and centrifuged at 13,000 rpm for 15 minutes. The DNA pellet was washed once with 75% ethanol, air-dried briefly and suspended in 10 μl of TE.1.
2.5. Agarose gel electrophoresis

DNA fragments or PCR products were separated on agarose gel (0.7-1.5 % w/v analytical grade agarose, Helena Biosciences).

1) Agarose was melted in TAE 1X buffer (TAE 50X buffer: 242 g Tris base, 57.1 ml glacial acetic acid, 100 ml 0.5M EDTA pH 8.0, H₂O up to 1L).

2) Ethidium bromide was added to a final concentration of 0.2 μg/ml.

3) The melted agarose was poured and set at room temperature.

4) 10 μl of sample were mixed with 2 μl of 6X loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol FF, 30% glycerol in H₂O) and loaded on agarose gel.

5) Electrophoresis was performed at 70-80 volts.

6) DNA was visualised on a UV transilluminator (254 nm).

2.6. Elution of DNA fragments from agarose gel

Ultra Agarose Spin Kit (ABgene)

1) The DNA band was cut out from the agarose gel with a scalpel blade, weighed and transferred to a spin filter basket which was placed in a collection tube.

2) Three volumes of solution GB were added.

3) The collection tube was incubated at 65 °C for 2 minutes, inverted and incubated for a further minute at 65 °C or until the block of agarose was completely melted.

4) The collection tube was inverted to mix the solution.

5) Centrifugation was carried out at 13,000 rpm for 10 seconds.

6) The collection tube was vortexed to mix the solution which was then loaded back onto the spin filter basket.

7) The collection tube was centrifuged at 13,000 rpm for 10 seconds.

8) The flow-through was discarded and 300 μl of solution GW (gel wash) were added.

9) Centrifugation was carried out at 13,000 rpm for 10 seconds.

10) The flow-through was then discarded and the collection tube was centrifuged at 13,000 rpm for 30 seconds to eliminate all residual traces of ethanol.

11) The spin filter basket was transferred to a new collection tube.

12) 50 μl of solution GE (Elution buffer, 10 mM Tris) were added at the centre of the spin filter basket and DNA was collected by centrifugation at 13,000 rpm for 30 seconds.
SPIN-X Column (Costar)
1) The DNA band was cut out of the agarose gel with a scalpel blade and placed in a SPIN-X column. DNA was collected by centrifugation at 13,000 rpm for 20 minutes.

2.7. Ligation reactions
The insert was ligated to the restricted and dephosphorylated cloning vector as follows:
- 30-50 ng of restricted and dephosphorylated cloning vector
- a molar excess of insert
- 1 μl of T4 DNA ligase Buffer 10 X (1X T4 DNA ligase buffer: 50 mM Tris-HCl, 10 mM MgCl₂, 10 mM DTT, 1 mM ATP, 25 μg/ml BSA, pH 7.5 @ 25 °C)
- 1 μl of T4 DNA ligase (New England Biolabs)
- H₂O to a final volume of 10 μl
Ligation reactions were incubated for 1 hour at room temperature or over night at 16 °C.

2.8. Precipitation of ligation
1) Ten μl of ligation reaction were placed in a 1.5 ml test tube and DNA was precipitated by addition of 1 μl of 3 M Na acetate pH 5.2, 20 μl of absolute ethanol and 2 μl of pellet paint.
2) Tubes were centrifuged at 13,000 rpm for 20 minutes.
3) The pellet was washed twice with 75 % ethanol and air-dried.
4) The pellet was suspended in 2 μl of H₂O.

2.9. Preparation of electro-competent cells
1) A single colony of E. coli (strain XL1-blue) was grown over night in 10 ml of YENB medium (0.75% Yeast extract, 0.8% Bacto Nutrient) at 37°C under vigorous shaking (200 rpm).
2) The day after, 1 litre of YENB medium was inoculated with 5-10 ml of over night culture and grown at 37°C under vigorous shaking.
3) When the bacterial culture reached an OD₆₀₀ of 0.5-0.9, the flask was chilled on ice for about 10 minutes.
4) Cells were spun down at 4,000 rpm (Sorvall RC-5B centrifuge, GS-3 rotor) for 10 minutes at 4°C.
5) The medium was discarded and cells were suspended in 100 ml of cold sterile water and centrifuged as before for 2 times.

6) The supernatant was discarded and cells were suspended in 20 ml of cold 10% glycerol in H₂O.

7) A further centrifugation step was repeated as before and the supernatant was discarded.

8) Cells were finally suspended in 2-3 ml of cold 10% glycerol and aliquoted in a 96 well plate (50µl/well) which was then sealed and placed on dry ice until cells froze.

9) The plate was stored at -80°C.

2.10. Electroporation

1) 1 µl of precipitated ligation reaction was placed in a 1.5 ml test tube and cooled on ice.

2) 15 µl of electrocompetent cells were thawed on ice and added to the ligation and test tubes were incubated on ice for 3 minutes.

3) Cells were then transferred to a 0.1 mm pre-chilled electroporation cuvette (BioRad) and incubated for 3 minutes on ice.

4) Electroporation was performed with a micropulser (Biorad) using the settings Ec1.

5) 1 ml of LB was immediately added.

6) LB and cells were transferred to a 20 ml tube.

7) Cells were incubated at 37°C under gentle shaking (120 rpm) for 60 minutes.

8) An aliquot of transformed cells (usually 50-500 µl) was plated on an LB agar plate containing the appropriate selective agent. For white and blue screening, 25 µl IPTG (stock solution 0.4 M) and 60 µl of X-Gal (stock solution 20 mg/ml) were added to the plate, spread, and air-dried before plating the cells.

9) Plates were incubated over night at 37°C.

2.11. DNA miniprep

1) Single colonies were picked from an LB agar plate and inoculated in 2.5 ml of LB containing the appropriate selective agent, and then grown over night at 37°C under vigorous shaking (200 rpm).

2) The day after, 1.5 ml of bacterial culture were transferred to a 1.5 ml test tube and centrifuged at 13,000 rpm for 2 minutes.
3) The supernatant was discarded and 100 μl of ice cold solution I containing 1 μl of RNase A (stock concentration 10 mg/ml) were added.
4) Cells were suspended by brief vortexing.
5) 200 μl of room temperature solution II were added.
6) Tubes were gently inverted 5-6 times and incubated on ice for 3 minutes.
7) 150 μl of ice cold solution III were added.
8) Tubes were gently inverted 5-6 times and incubated on ice for 10 minutes.
9) Tubes were centrifuged at 13,000 rpm for 15 minutes.
10) The supernatant was transferred to a new tube and 900 μl of absolute ethanol were added.
11) Tubes were inverted 2-3 times and centrifuged at 13,000 rpm for 15 minutes.
12) The pellet was washed once with 1 ml of 75% ethanol, air-dried and suspended in 30 μl of TE.1.

Solutions
Solution I: 50mM, 25mM Tris pH 8.0, 10mM EDTA pH 8.0
Solution II: 0.2M NaOH, 1% SDS
Solution III: 29.4 g K acetate, 11.5 ml glacial acetic acid, H₂O up to 100 ml

2.12. DNA maxiprep
1) A single bacterial colony was inoculated in 50 ml of LB (containing the appropriate selective agent) and grown over night at 37°C under vigorous shaking (200-240 rpm).
2) The bacterial culture was transferred into a 50 ml tube. Bacteria were pelleted by centrifugation at 4,000 rpm for 15 minutes at 4°C.
3) The supernatant was discarded.
4) Cells were suspended in 4 ml of ice cold solution I.
5) Cells were lysed by adding 8 ml of solution II at room temperature. The tubes were then inverted for 10 times and incubated on ice for 5 minutes.
6) 6 ml of ice cold solution III were added and the tube was inverted 10 times and incubated on ice for 10 minutes.
7) A centrifugation step was carried out at 4,000 rpm for 15 minutes at 4°C.
8) The supernatant was filtered through a cheese-cloth (to eliminate all the solid particles) into a new tube.
9) 17 ml of isopropanol were added and the tube was placed at −80°C for 20 minutes.
10) The tube was centrifuged at 4,000 rpm for 20 minutes at 4°C.
11) The supernatant was decanted and after draining the tube, the pellet was suspended in 2 ml of TE.1 (10 mM Tris pH 8.0, 0.1 mM EDTA).
12) 2.5 ml of 4.4 M LiCl solution were added to precipitate ssDNA and RNA.
13) The tube was chilled on ice for 1 hour.
14) The tube was centrifuged at 4,000 rpm for 15 minutes at 4°C and the supernatant was transferred to a new tube.
15) 10 ml of absolute ethanol were added and the tube was inverted a few times and centrifuged at 4,000 rpm for 15 minutes at 4°C.
16) The pellet was washed with 10 ml of 75% ethanol, air-dried, suspended in 400 µl of TE.1 and transferred to a 1.5 ml tube.
17) 10 µl of RNase A were added and the tube was incubated at 37°C for 15 minutes.
18) 20 µl of 10% SDS were added and incubation was carried out at 70°C for 10 minutes.
19) The sample was extracted twice with an equal volume of phenol saturated with Tris pH 8.0, and once with a mixture of phenol/chloroform/isoamyl alcohol (25: 24:1).
20) DNA was precipitated by adding 40 µl of 3 M Na acetate pH 5.2 and 800 µl of absolute ethanol.
21) The tube was centrifuged at 13,000 rpm for 15 minutes.
22) The pellet was washed once with 75% ethanol and air-dried.
23) The DNA pellet was finally suspended in an appropriate volume of TE.1 (usually 100-300 µl).

2.13. Quantification of nucleic acids

The quantification of nucleic acids (DNA and RNA) was determined by measuring the absorbance at 260 nm (A_{260}) with a spectrophotometer. The concentration was calculated with the following formula:

\[ \text{Concentration (µg/ml)} = B \times A_{260} \times \text{dilution} \]

B is a constant equal to 50 for DNA and 40 for RNA.

The quality of the nucleic acid was determined by calculating the ratio between the absorbance at 260 nm (A_{260}) and that at 280 nm (A_{280}).

2.14. RNA extraction from tissues

1) Approximately 1 g of liver (or kidney) was homogenised in 2 ml of Tri-reagent (Sigma) using a tissue homogeniser.
2) 1 ml of homogenised tissue was placed into a 1.5 ml tube and then centrifuged at 13,000 rpm for 10 minutes at 4°C to remove the insoluble material.
3) The liquid phase was transferred to a new tube and left at room temperature for 5 minutes.
4) 200 µl of chloroform were added and tubes were shaken vigorously for 15 seconds.
5) Tubes were incubated at room temperature for 10 minutes.
6) Samples were centrifuged at 12,000 rpm for 15 minutes at 4°C.
7) The upper phase was transferred to a new tube and 500 µl of isopropanol were added.
8) Tubes were mixed and left at room temperature for 5-10 minutes.
9) Samples were centrifuged at 12,000 rpm for 10 minutes at room temperature.
10) The supernatant was discarded and the RNA pellet was washed with 1 ml of 75% ethanol in DEPC water.
11) Tubes were vortexed and centrifuged at 7,500 rpm for 5 minutes at 4°C.
12) The RNA pellet was air-dried briefly and suspended in a suitable volume of DEPC water.
13) The RNA concentration was calculated by measuring the absorbance at 260 nm.

2.15. Transfection of cell lines

Calcium phosphate method
1) On day 1, cells were seeded in a 6 or 12 well plate (250,000-500,000 cells/well in a 6 well plate or 100,000-200,000 cells/well in a 12 well plate).
2) On day 2, the culture medium was replaced 3 hours prior to transfection.
3) 250 µl of 2X HBS solution [10X stock solution: 8.18% NaCl (w/v), 5.94% Hepes (w/v), 0.2% NaHPO4 (w/v) pH 7.12] were placed in a 1.5 ml test tube (referred to as tube B).
4) An appropriate amount of plasmid DNA (usually 1-3 µg) was mixed with 31 µl of 2M CaCl2 and H2O in a final volume of 250 µl in a 1.5 ml test tube (referred to as tube A).
5) The content of tube A was added drop by drop to tube B and the mixture was incubated on ice for 5-10 minutes.
6) The precipitate was gently added to the cells and incubated at 37°C for 3.5-4 hours.
7) Cells were finally washed twice with 1X PBS and complete fresh medium was added to the cells.
Gene Jammer method

1) On day 1, cells were seeded in a 6 or 12 well plate (250,000-500,000 cells/well in a 6 well plate or 100,000-200,000 cells/well in a 12 well plate).
2) On day two, 100 μl of sterile room temperature, serum-free, antibiotic-free DMEM was placed into a polystyrene tube.
3) The Gene Jammer transfection reagent (1 or 3 μl for each transfection in a 12 or 6 well plate respectively) was added directly into the medium, mixed gently without pipetting and incubated at room temperature for 5-10 minutes.
4) The plasmid DNA was added, mixed gently as before and incubated at room temperature for 5-10 minutes.
5) The standard medium was removed from the plate, and replaced with fresh complete medium (900 μl/well).
6) The transfection mixture was added drop by drop to the cells, and the plate was rocked gently back and forth to distribute the transfection mixture evenly.
7) The plate was incubated at 37°C.
8) After 3 hours, 900 μl of complete medium/well were added.

2.16. RNA isolation from cell lines

All the operations were done at room temperature

1) The medium was removed from a 6 well plate.
2) 500 μl of Tri-reagent solution (Sigma) were added onto the cells and incubation at room temperature was carried out for 10 minutes.
3) The supernatant was transferred to a 1.5 ml test tube and 100 μl of chloroform were added and the test tubes were inverted gently for 10 times.
4) Samples were centrifuged at 13,000 rpm for 10 minutes.
5) The upper phase was carefully transferred to a new test tube and 250 μl of isopropanol were added.
6) The tubes were gently inverted 10 times and incubated for 10 minutes at room temperature.
7) The RNA was precipitated by centrifugation at 13,000 rpm for 20 minutes.
8) The pellet was washed once with 75% ethanol in DEPC treated water.
9) The RNA pellet was air-dried briefly.
10) RNA was suspended in 40 μl of DEPC-treated water.
11) The concentration of the RNA was calculated by measuring the absorbance at 260 nm with a spectrophotometer.

12) RNA was stored at -80°C.

### 2.17. Reverse transcription

1) Two μg of total RNA were mixed with 0.2 μl random hexamers (conc. 2.5 μg/μl) and DEPC-treated water was added to a final volume of 17 μl.

2) The mixture was heated at 70°C for 5 minutes to melt secondary structures in the RNA and immediately cooled on ice to prevent the formation of secondary structures.

3) Samples were spun down briefly.

4) The following reagents were added:
   - 5 μl of M-MLV 5X reaction buffer
   - 1.25 μl of 10 mM dNTPs
   - 25 units of rRNasin Ribonuclease Inhibitor (Promega)
   - 1 μl of Moloney Murine Leukemia Virus Reverse transcriptase (M-MLV reverse transcriptase, Promega)
   - DEPC-treated water to a final volume of 25 μl

5) Samples were mixed gently and incubated at 37°C for 60 minutes.

### 2.18. Polymerase Chain reaction (PCR)

**Taq polymerase (Promega)**

The PCR mixture was as follows:

<table>
<thead>
<tr>
<th>Component</th>
<th>Concentration/Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA template</td>
<td>10-50 ng of DNA or 1 μl of cDNA</td>
</tr>
<tr>
<td>10X Reaction Buffer with MgCl₂</td>
<td>1 μl</td>
</tr>
<tr>
<td>Primer 1</td>
<td>0.6 μM</td>
</tr>
<tr>
<td>Primer 2</td>
<td>0.6 μM</td>
</tr>
<tr>
<td>dNTP mix</td>
<td>0.25 mM</td>
</tr>
<tr>
<td>Taq Polymerase</td>
<td>1 unit</td>
</tr>
<tr>
<td>H₂O</td>
<td>To a final volume of 10 μl</td>
</tr>
</tbody>
</table>
10X Reaction Buffer with MgCl₂: 500 mM KCl, 100 mM Tris-HCl (pH 9.0 at 25 °C), 1% Triton® X-100 and 15mM MgCl₂.

The PCR mixture was covered with a drop of mineral oil (USB) and amplification was performed with a BIOMETRA TRIO-Thermoblock. Conditions varied with different primers and different PCR reactions and are specified in each case.

Expand Long Template PCR System (Roche).

Two PCR mixes were prepared as follows:

<table>
<thead>
<tr>
<th>Master mix No. 1</th>
<th>Master mix No. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>350 μM dNTPs</td>
<td>5 μl of 10X PCR buffer with MgCl₂</td>
</tr>
<tr>
<td>0.3 μM Primer 1</td>
<td>0.75 μl of enzyme mix</td>
</tr>
<tr>
<td>0.3 μM primer 2</td>
<td>H₂O up to 25 μl</td>
</tr>
<tr>
<td>Template DNA (same amount as for Taq)</td>
<td></td>
</tr>
<tr>
<td>H₂O up to 25 μl</td>
<td></td>
</tr>
</tbody>
</table>

Master mix No. 1 and No. 2 were mixed and a drop of mineral oil (USB) was layered on top. Amplification was performed with a BIOMETRA TRIO-Thermoblock. Conditions varied with different primers and different PCR reactions and are specified in each case.

2.19. DNA sequencing

DNA sequencing reactions were carried out using an ABI Prism® BigDye® Terminator V 3.1 Cycle Sequencing Kit and analysed with an ABI Prism 377 DNA sequencer. 0.2-0.5 μg of plasmid DNA or 50-100 ng of PCR product and 1 pmole of primer were used as template for the sequencing reaction.

2.20. Firefly luciferase assay

The firefly luciferase assays were performed with the dual-luciferase™ Reporter Assay System (Promega).

1) On day 1, cells were seeded in a 6 or 12 well plate.
2) On day two, cells were transfected with 250-500 ng of a plasmid DNA molecule encoding the firefly luciferase gene (containing the promoter or 5' UTR under investigation) and 2-20 ng of plasmid phRL-CMV Vector, which encodes the renilla luciferase gene driven by the CMV promoter.
3) On day three, cells were washed twice with 1X PBS.
4) 200 μl of freshly prepared 1X PLB (passive lysis buffer: freshly prepared from a 5X stock solution by diluting it with H2O just before use) were added to the cells and the plate was gently shaken at room temperature for 15 minutes.

5) The liquid was transferred to a 1.5 ml test tube and cell debris was spun down at 13,000 rpm for 2 minutes.

6) 5 μl of protein extract were mixed with 25 μl of Luciferase Assay Buffer II (LucII), and the intensity of the signal from firefly luciferase was determined with a luminometer (EG and G Berthold LB 9507).

7) 25 μl of Stop&Glow® Reagent were added in order to measure the signal from renilla luciferase.

2.21. Extraction of genomic DNA from tissues

1) Approximately 1 gram of liver or kidney, or a tail biopsy of approximately 0.5 cm were sliced with a sterile scalpel blade and placed in a 1.5 ml tube containing 700 μl of TNE buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl and 100mM EDTA).

2) 35 μl of 10 % SDS and 35 μl of Proteinase K (10mg/ml) were added

3) Tubes were rolled over night at 37°C.

4) The day after, 700 μl of Phenol (saturated with Tris pH 8.0) were added and tubes were shaken vigorously for 5 minutes and centrifuged at 13,000 rpm for 10 minutes at room temperature.

5) The upper phase was transferred to a new tube and 350 μl of phenol (saturated with Tris pH8.0) and 350 μl of chloroform were added.

6) Tubes were shaken for 5 minutes and centrifuged at 13,000 rpm for 10 minutes at room temperature.

7) The upper phase was transferred to a new tube and 700 μl of chloroform were added.

8) Tubes were shaken vigorously for 5 minutes and centrifuged at 13,000 rpm for 10 minutes at room temperature.

9) The upper phase was transferred to a 2.0 ml tube and DNA was precipitated by addition of 2 volumes of absolute ethanol.

10) Tubes were inverted gently a few times, and the cotton-like DNA was collected with a syringe needle and dipped into a 1.5 ml tube containing 75% ethanol.

11) The DNA pellet was air-dried briefly on the tip of the syringe needle and was then placed in 100 μl of TE.1.
12) The DNA pellet was left to dissolve overnight at 4°C.
13) The DNA concentration was estimated by measuring the absorbance of the samples at 260 nm with a spectrophotometer.

2.22. MATERIALS AND METHODS
FOR CHAPTER 3

2.22.1. PCR conditions
All PCR reactions for the figures of chapter 3 were performed with the enzyme Taq DNA polymerase with the exception of figure 3.27 A in which the Expand Long Template PCR System (Roche) was used. Below are the conditions for each amplification experiment.

Figure 3.7 A: 94°C 2 minutes, 94°C 10 seconds, 55°C 30 seconds, 72°C 90 seconds.
Amplification was carried out for 25 cycles. A final incubation at 72°C was carried out for 10 minutes.

Figure 3.7 B: 94°C 2 minutes, 94°C 45 seconds, 55°C 45 seconds, 72°C 45 seconds.
Amplification was carried out for 28 cycles. A final incubation at 72°C was carried out for 10 minutes.

Figure 3.12: same as in figure 3.7 A.

Figure 3.14 A and 3.14 B: same as in figure 3.7 A.

Figure 3.14 C: same as in figure 3.7 B. Instead of 25 cycles, amplification was performed with 30 cycles.

Figure 3.15: same as in figure 3.7 A.

Figure 3.16 A and 3.16 C: same as in figure 3.7 A, but with 30 cycles instead of 25.

Figure 3.16 B: 94°C 2 minutes, 94°C 10 seconds, 60°C 30 seconds, 72°C 90 seconds.
Amplification was carried out for 30 cycles. A final incubation at 72°C was carried out for 10 minutes.

Figure 3.19 A: 94°C 2 minutes, 94°C 10 seconds, 55°C 30 seconds, 72°C 60 seconds.
Amplification was carried out for 35 cycles. A final incubation at 72°C was carried out for 10 minutes.

Figure 3.19 B: two-step PCR. The first amplification for the lanes indicated as “T” was:
94°C 2 minutes, 94°C 10 seconds, 55°C 30 seconds, 72°C 180 seconds. Amplification was carried out for 38 cycles. A final incubation at 72°C was carried out for 10 minutes. For the lanes indicated as “ER” the first amplification was the following: 94°C 2 minutes, 94°C 10
seconds, 65°C 30 seconds, 72°C 180 seconds, for the first 5 cycles. The following 5 cycles were carried out with an annealing temperature of 63°C and then another 5 cycles with an annealing temperature of 60°C. Then, the following 25 cycles were carried out with an annealing temperature of 60°C. An aliquot of the first PCR reaction was used as template for the second PCR reaction that was carried out with the following conditions: 94°C 2 minutes, 94°C 10 seconds, 55°C 30 seconds, 72°C 60 seconds. Amplification was carried out for 30 cycles. A final incubation at 72°C was carried out for 10 minutes.

**Figure 3.20:** same conditions as in figure 3.7 A, but with 35 cycles instead of 25.

**Figure 3.23 A:** same as in figure 3.7 A.

**Figure 3.23 B:** same as in figure 3.7 B.

**Figure 3.23 C:** same as in figure 3.7 A, but with 28 cycles instead of 25.

**Figure 3.24 A, B and C:** same as in figure 3.23 A, B and C.

**Figure 3.26 A:** two-step PCR. For lanes “T” and “2-3”, the first amplification was performed as follows: 94°C 2 minutes, 94°C 10 seconds, 55°C 30 seconds 72°C 180 seconds. Amplification was done for 38 cycles. A final step at 72°C for 10 minutes was carried out. For lanes “2-2” the first amplification was carried out as in figure 3.18 B. An aliquot of the first PCR was used as template for a second amplification in which the following conditions were used: 94°C 2 minutes, 94°C 10 seconds, 55°C 30 seconds 72°C 60 seconds. Amplification was carried out for 35 cycles. A final incubation at 72°C was carried out for 10 minutes.

**Figure 3.28 A:** the conditions for the first five cycles were 94°C 2 minutes, 94°C 10 seconds, 63°C 30 seconds, 68°C 240 seconds. The conditions for the next 33 cycles were 94°C 2 minutes, 94°C 10 seconds, 60°C 30 seconds, 68°C 240 seconds. A final incubation at 68°C was carried out for 10 minutes.

### 2.22.2. Southern blot

**Digestion of genomic DNA**

1) 10 μg of genomic DNA were mixed with 10 μl of buffer 10X, 1 μl BSA, 2 μl restriction enzyme (EcoRI or HindIII) and H₂O was added to a final volume of 100 μl.

2) Samples were incubated at 37°C for several hours. Another 2 μl of restriction enzyme were added and incubation was extended over night.

3) The digestion was monitored by loading 1/10 of the digestion on a 0.7 % agarose gel.
4) Once ascertained that the DNA had been properly digested, the remaining DNA was precipitated. 1/10 of the original volume of 3 M sodium acetate pH 5.3 and 2 volumes of absolute ethanol were added and tubes were centrifuged at 13,000 rpm for 15 minutes.

5) The supernatant was discarded.

6) The DNA pellet was washed once with 75% ethanol, air-dried briefly and suspended in 10 µl of H₂O.

**Separation of DNA fragments**

1) The digested genomic DNA was loaded on a 0.7% agarose gel and electrophoresis was carried out overnight at 25 volts.

2) The gel was then incubated twice in denaturing solution (0.4M NaOH and 0.6M NaCl) for 15 minutes each time under constant shaking and at room temperature.

3) The gel was neutralised with two incubations of 20 minutes each in neutralising solution (0.5M Tris pH 7.5, 1.5M NaCl) under constant shaking at room temperature.

4) The DNA was then transferred overnight to a Hybond N+ membrane (Amersham Pharmacia Biotech) by capillary transfer with 10X SSC.

5) The next day the filter was air-dried and DNA was fixed to the filter by UV cross-linking for 30 seconds on each side of the filter.

6) Pre-hybridisation was carried out overnight at 42°C in 10 ml of pre-hybridisation solution (3 ml SSC 20X, 1 ml SDS 10%, 400 µl Denhardt’s 25X, 100 µl of herring sperm DNA and H₂O to a final volume of 10 ml). Herring sperm DNA was boiled for 10 minutes and immediately chilled on ice before being added to the pre-hybridisation solution.

**Synthesis of the probe**

The probes were synthesised with the Prime-a-Gene kit (Promega).

1) 30-50 ng of eluted PCR product was mixed with H₂O to a final volume of 9 µl. DNA was then denatured by incubation at 100 °C for 10 minutes. To prevent formation of double stranded DNA, after boiling the sample was immediately cooled on ice.
2) The following components were added:
   • 8 μl of 8X Buffer
   • 8 μl 2.5mM dATP dTTP dGTP
   • 1.0 μl BSA (10 mg/ml)
   • 5 μl 32Pα-dCTP
   • 1.0 μl klenow DNA polymerase
   • H₂O to a final volume of 40 μl
3) The reaction was incubated at 37°C for 2 hours.
4) Unincorporated nucleotides were eliminated by loading the probe onto a G50 column (Probe Quant™ G-50 Micro columns, Amersham Pharmacia) The column was vortexed to resuspend the resin. The end was snapped off and the column was placed in a 1.5 ml tube. Centrifugation was carried out at 2,000 rpm for 1 minute. The probe was loaded onto the centre of the resin and the column was centrifuged at 2,000 rpm for 2 minutes.
5) The probe was boiled for 10 minutes, cooled on ice and then added to the hybridisation solution (0.9 ml H₂O, 5.0 ml formamide, 1 ml SDS 10%, 3 ml SSC 20X, 100 μl of herring sperm DNA boiled and cooled on ice).
6) Hybridisation was carried out over night at 42°C.
7) The day after the filter was washed twice for 15 minutes each time in SSC 6X and 0.1% SDS at room temperature.
8) A final wash in 0.1X SSC and 0.1% SDS was performed at 65°C for 10 minutes.
9) The filter was finally exposed for an appropriate period of time (usually over night) to a phosphoimager screen and the signal was then detected with a Cyclone storage phosphor screen Packard or a Phosphorimager (Molecular Dynamics).

2.22.3. Real-time PCR
Real-time PCR was performed with a kit (LightCycler-FastStart DNA Master SYBR Green I, Roche).
   1) Vial LightCycler-fastStart Enzyme (vial 1a) and vial LightCycler-FastStart Reaction Mix SYBR green I (vial 1b) were briefly centrifuged.
   2) The total volume of vial 1b (60 μl) was transferred into vial 1a.
3) The two solutions were mixed by gently pipetting up and down. This solution was labelled LightCycler-FastStart DNA Master SYBR Green I.

4) A master mix was then prepared as follows:

<table>
<thead>
<tr>
<th>Component/Stock conc.</th>
<th>Volume (µl)</th>
<th>Final conc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer 1 (10 µM)</td>
<td>6</td>
<td>0.5 µM</td>
</tr>
<tr>
<td>Primer 2 (10 µM)</td>
<td>6</td>
<td>0.5 µM</td>
</tr>
<tr>
<td>LightCycler-FastStart DNA Master SYBR Green I (10 X)</td>
<td>12</td>
<td>1 X</td>
</tr>
<tr>
<td>MgCl₂ (25 mM)</td>
<td>14.4</td>
<td>3 mM</td>
</tr>
<tr>
<td>H₂O</td>
<td>75.6</td>
<td></td>
</tr>
</tbody>
</table>

5) One µl of cDNA was mixed with 19 µl of Master Mix.

6) Amplification was carried out with the following protocol:
   - 95 °C for 10 minutes
   - 67-55 °C for 5 seconds (step size of 0.8 °C / cycle)
   - 72 °C for 10 seconds
   Amplification was carried out for 45 cycles.

Real-time PCR reactions were performed with a LightCycler instrument (Roche Diagnostics, Mannheim, Germany). Results were analysed using the LDCA software supplied with the machine.
2.2.2.4. Cycling curve

To establish a cycling curve PCR, reactions were carried out as normal with the addition of 0.25 µl of [α³²P] dCTP (Amersham). Amplifications were performed in separate tubes for 18, 20, 22 or 24 cycles. PCR reactions were then mixed with the same volume of formamide/EDTA/XC/BPB gel loading buffer (composition: 10 ml formamide, 10 mg xylene cyanol FF, 10 mg bromphenol blue, 200 µl 0.5 M EDTA pH 8.0), heated at 94°C for 5 minutes and immediately cooled on ice. DNA fragments were then separated on a 6% acrylamide gel prepared as follows:

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urea</td>
<td>16.8 g</td>
</tr>
<tr>
<td>Acrylamide (Ultra Pure Accugel™ sequencing grade 40% (w/v) 19:1 Acrylamide:Bis-Acrylamide)</td>
<td>6 ml</td>
</tr>
<tr>
<td>TBE 10X</td>
<td>4 ml</td>
</tr>
<tr>
<td>Formamide (Sigma)</td>
<td>12 ml</td>
</tr>
<tr>
<td>H₂O</td>
<td>up to 40 ml</td>
</tr>
</tbody>
</table>

The mixture was stirred with a magnetic stirrer until the urea was completely dissolved. 400 µl of 10 % APS (Ammonium persulfate, Sigma) and 40 µl of TEMED (N, N', N'-Tetramethylethylendiamine, Sigma) were added and mixed thoroughly. A mixture was poured between two glass plates until acrylamide polymerised. A pre-run of 20 minutes was performed after which the samples were loaded and electrophoresis was carried out at 1000 V. The gel was dried with a gel dryer and finally exposed to a phosphorimager screen.

2.2.3. MATERIALS AND METHODS

FOR CHAPTER 4

2.2.3.1. Construct preparation

pGL3-Normal COT 5' UTR

The 5' UTR from the normal COT mRNA isoform, was amplified with primers COT E1F HindIII and COT E3R, using the Expand long template PCR system. Amplification conditions were 94°C 120 seconds, 94°C 10 seconds, 55°C 30 seconds, 68°C 60 seconds.
Amplification was carried out for 25 cycles. A final incubation at 68 °C for 7 minutes was carried out. The template was cDNA from SHR kidney.

The PCR product was eluted from agarose gel and extensively digested with the restriction endonucleases HindIII and Ncol in buffer B (Promega). The recognition sequence for HindIII was artificially introduced in primer COT E1 HindIII whereas the recognition sequence for Ncol was naturally present in correspondence of the ATG translational start codon of COT. The restriction enzymes were heat-inactivated at 65°C for 15 minutes. The insert was then ligated over night at 16°C into pGL3-Control Vector (Promega) which had been previously digested with the same restriction endonucleases.

pGL3-Rep Exon 2 COT 5’ UTR

The insert corresponds to the 5’ UTR of the COT mRNA isoform with a tandem repeat of exon 2. COT mRNA was amplified by RT-PCR from WKY liver RNA with primers COT E1F and COT E3R. The PCR product corresponding to the isoform containing a tandem repeat of exon 2 was eluted from agarose gel with the ABgene kit and was then digested with the restriction endonuclease Ncol in buffer D (Promega). The released fragment comprised between the two ATG translational start codons was eluted from agarose gel and ligated into the previous construct (pGL3-Normal COT 5’ UTR) that had been previously digested with Ncol.

pGL3-Mut ATG COT 5’ UTR

This construct was obtained by mutating the first ATG codon of construct pGL3-Rep Exon 2 COT 5’ UTR to a TTG codon. This point mutation was created by overlapping PCR. The two PCR reactions were as follows:

PCR 1: primers pGL3 FORW and COT ATG MUT R
PCR 2: primer LUC REV and primer COT ATG MUT F.

PCR 1 and PCR 2 were then mixed together and used as template for the final PCR which was performed with primers pGL3 FORW and LUC REV. The PCR product was eluted from agarose gel and digested with the restriction endonucleases HindIII and Ncol. After heat inactivation of the restriction endonucleases, the insert was ligated into construct pGL3-Control Vector that had been previously digested with the same restriction endonucleases.

pGL3-Normal SA 5’ UTR

The insert consists of the 5’ UTR of the normal SA mRNA isoform. SA mRNA was amplified by RT-PCR from SHR kidney RNA with primers SA E1F HindIII and SA E5R. The PCR product was then eluted from agarose gel, digested with the restriction
endonucleases HindIII and NcoI and cloned into pGL3-Control Vector that had been previously digested with the same restriction endonucleases.

**pGL3-Rep Exon 2 SA 5’ UTR**

The insert corresponds to the 5’ UTR of the SA mRNA isoform with a tandem repeat of exon 2. SA mRNA was amplified from WKY kidney RNA by RT-PCR with primers SA E1F HindIII and SA E5R. The PCR product containing the tandem repeat of exon 2 was eluted from agarose gel, digested with HindIII and NcoI and cloned into pGL3-Control Vector that had been previously digested with the same restriction endonucleases.

**pGL3-Rep Exons 2-3-4 SA 5’ UTR**

The insert corresponds to the 5’ UTR of the SA mRNA isoform that contains a tandem repeat of exons 2, 3 and 4. SA mRNA was amplified from WKY kidney RNA by RT-PCR with primers SA E1F and SA E5R. The PCR product containing the tandem repeat of exons 2, 3 and 4 was then eluted from agarose gel and digested with NcoI. The recognition sequence for NcoI is located at the ATG translational start codon located in exon 3 of SA. The fragment comprised between the two ATG translational start sites was eluted from agarose gel and cloned into the construct pGL3-Normal SA 5’ UTR that had been previously digested with the restriction endonuclease NcoI.

**pGL3-Mut ATG SA 5’ UTR**

This construct was obtained by mutating the upstream ATG translational start site codon of construct pGL3-Rep Exons 2-3-4 SA 5’ UTR, to a TTG codon. The insert was obtained by overlapping PCR. The upstream PCR product was amplified with primers pGL3 FORW and SA ATG MUT R whereas the downstream PCR product was amplified with primers SA ATG MUT F and LUC REV. The template for these PCR reactions was clone pGL3-Rep Exons 2-3-4 SA 5’ UTR. PCR conditions were as follows: 94°C 120 seconds, 94°C 10 seconds, 55°C 30 seconds, 68°C 60 seconds.

Amplification was carried out for 18 cycles. A final incubation at 68 °C for 10 minutes was carried out. The two DNA fragments were eluted from agarose gel, mixed together and utilised as template for the next PCR reaction with primers pGL3 FORW and LUC REV. The PCR cycles were as follows: 94°C 120 seconds, 94°C 10 seconds, 55°C 30 seconds, 68°C 90 seconds.

Amplification was carried out for 20 cycles. A final incubation at 68 °C for 10 minutes was carried out. The PCR product was eluted from agarose gel with the ABgene kit. The eluted DNA fragment was then digested with the restriction endonucleases HindIII and NcoI and cloned into pGL3-Control Vector that had been previously digested with the same restriction endonucleases.
2.23.2. Transfection of mammalian cells

1) Hela cells were seeded in a 12-well plate at a density of 200,000 cells/well.
2) After 24 hours, cells were transfected with 0.25 µg or 0.5 µg of the pGL3 constructs and 10 ng of phRL-CMV Vector that contains the renilla luciferase gene. The experiment was then performed as described previously. For each construct, the experiment was performed in triplicate.
The ratio between firefly and renilla luciferase was calculated and then multiplied by a factor of 1,000.

2.24. MATERIALS AND METHODS
FOR CHAPTER 5

2.24.1. Preparation of EMBL-3 phage DNA

Isolation of single phage plaques

1) Serial dilutions of phage particles were mixed with LE392 cells (that had been previously grown overnight at 37°C under vigorous shaking in 5 ml of LB containing 10 mM MgSO_{4}), and incubated at 37 °C for 20 minutes.
2) 3 ml of top agar pre-warmed at 48°C were added and plated onto an LB agar plate containing 10mM MgSO_{4} that had been pre-warmed at 37°C.
3) The plate was incubated over night at 37°C.
4) The day after, a single plaque was isolated, grown in liquid culture and plated again as before.
5) Phage particles were collected by adding 3 ml of phage buffer (20mM Tris-HCl pH 7.4, 100mM NaCl, 10 mM MgSO_{4}) and shaking for 3 hours.
6) The phage buffer was collected and extracted once with an equal volume of chloroform isoamyl alcohol (24:1) and titration of phage particles was carried out by plating a serial dilution.

λ phage DNA prep

1) 300 µl of LE392 cells, prepared as before, were infected at a m.o.i of 0.01 with EMBL-3 phage particles and incubated at 37°C for 15 minutes without shaking.
2) 15 ml of T-broth containing 10mM MgSO_{4} were added, and incubation at 37°C was carried on under constant shaking, until most of the cells were lysed.
3) 15 µl of chloroform were added and the medium was vortexed thoroughly.
4) Cells were spun down at 8,000 rpm for 20 minutes at 4°C in a Sorvall RC-5B centrifuge with a Sorvall SS 34 rotor.

5) The liquid phase (10ml) was transferred to a new tube and 100 µl of RNase A (10 mg/ml) and 10 µl DNase RQ1 (Promega, conc. 1unit/µl) were added.

6) Incubation was carried out at 37°C for 90 minutes.

7) 2.15 ml of 5 M NaCl and 0.64 g PEG 6000 were added.

8) Incubation was then carried out for 30 minutes on ice and then over night at +4°C.

9) The following day, phage particles were spun down at 10,000 rpm for 20 minutes.

10) The supernatant was aspirated and the phage pellet was suspended in 250 µl of 20 mM MOPS buffer containing 5mM MgCl₂ and placed in a 1.5 ml test tube.

11) 5µl of 5 M NaCl and 5 µl of 10% (w/v) SDS were added.

12) The mixture was heated at 65°C for 10 minutes.

13) 250µl of phenol were added and tubes were vortexed briefly.

14) Tubes were then placed at 65°C for 15 minutes.

15) The two phases were separated by centrifugation at 13,000 rpm for 3 minutes.

16) The upper phase was transferred to a new tube and extracted twice with chloroform.

17) The upper phase was transferred to a new tube and 90 µl of 8.3M ammonium acetate and 600 µl of absolute ethanol were added.

18) The DNA was precipitated by centrifugation at 13,000 rpm for 15 minutes.

19) The pellet was washed once with 75% ethanol, air-dried and suspended in a suitable volume of TE.1.

2.24.2. Cloning of phage inserts into pBluescript SK (+) (Stratagene)

1) λ DNA was digested with the restriction enzyme XhoI in a 20 µl final volume reaction.

2) The restriction enzyme was heat inactivated by incubation at 70°C for 15 minutes.

3) The DNA was precipitated by adding 1/10 of the original volume of sodium acetate 3 M pH 5.2 and 2 volumes of absolute ethanol and centrifugation at 13,000 rpm for 15 minutes. The pellet was washed once with 75% ethanol, air-dried briefly and suspended in 10 µl of TE.1.

4) 2.5 µl of digested λ DNA were ligated with approximately 50 ng of pBluescript SK (+) that had been previously digested with XhoI and dephosphorylated. 1 µl of T4
DNA ligase (New England Biolabs) and 1 μl of T4 DNA ligase buffer (1X T4 DNA ligase buffer: 50 mM Tris-HCl, 10 mM MgCl₂, 10 mM DTT, 1 mM ATP, 25 μg/ml BSA, pH 7.5 @ 25 °C) were added. Finally, H₂O was added to a final volume of 10 μl.

5) The ligation reactions were incubated at 16°C over night.
6) DNA was then precipitated by adding 1 μl of Na acetate 3 M pH 5.2, 20 μl absolute ethanol and 2 μl of pellet paint (Novagen). The tubes were centrifuged at 13,000 rpm for 15 minutes.
7) The pellet was washed twice with 75% ethanol, air-dried and dissolved in 2 μl of H₂O.
8) 1 μl was used for electroporation.

2.24.3. Cloning of PCR products into pGEM®-T Easy Vector

1) PCR products were first separated by electrophoresis on agarose gel and eluted.
2) 7.8 μl of eluted PCR product were mixed with 0.2μl dATP 10 mM, 1μl Taq polymerase and 1μl of its 10X buffer.
3) The reaction mixture was incubated at 70°C for 30 minutes.
4) 3.8 μl of Taq polymerase-treated PCR product were mixed with 0.2 μl pGEM®-T Easy Vector, 5 μl of 2X Buffer for pGEM-T Easy (Promega) and 1 μl of T4 DNA ligase.
5) The ligation reactions were incubated at room temperature for 1 hour or over night at + 4°C.
6) Ligations reactions were precipitated and transformed by electroporation.

2.24.4. Cloning of the constructs for the promoter studies

WKY SA NP pECFP-1

The SA promoter region (approximately 1.2 Kbp) was amplified from WKY genomic DNA with primers SA Prom Sac II F and SA Prom Sac II R. The PCR product was purified from agarose gel and cleaved with the restriction endonuclease Sac II. After heat-inactivating the enzyme Sac II, the DNA fragment was cloned into the plasmid pECFP-1 (Clontech) promoter-less vector containing an open reading frame encoding the enhanced cyano fluorescent protein) that had been previously cleaved with the restriction endonuclease Sac II and dephosphorylated.
**WKY SA intron 1 pECFP-1**
The insert, corresponding to intron 1 of the WKY SA gene was amplified from WKY genomic DNA by PCR with primers WKY SA Int 1 Sac II F and WKY SA Int 1 Sac II R. After elution from agarose gel, the DNA amplicon was cleaved with the restriction endonuclease Sac II and cloned into plasmid pECFP-1 that had been previously digested with the same restriction endonuclease. In order to identify a recombinant clone containing the insert in the right orientation, restriction enzyme analysis of plasmid DNA derived from recombinant clones was carried out.

**WKY SA NP pGL3**
The natural promoter of the WKY SA gene was excised from construct WKY SA NP pEFCP-1 with the restriction endonucleases Sac I and Sma I and sub-cloned into the promoter-less vector pGL3-Basic vector that had been cleaved with the same restriction endonucleases. This vector contains an open reading frame that encodes for firefly luciferase.

**WKY SA intron 1 pGL3**
The insert corresponding to intron 1 of the WKY SA gene was amplified from WKY genomic DNA by PCR with primers WKY SA Int 1 KpnI and WKY SA Int 1 XhoI. After elution from agarose gel, the DNA amplicon was cleaved with the restriction endonucleases KpnI and XhoI and cloned into plasmid pGL3-Basic vector that had been previously digested with the same restriction endonucleases.

**WKY CP pGL3**
This vector contains the putative promoter sequence from intron 1 of the SA gene. This sequence was amplified by PCR from genomic DNA with primers WKY SA CP F and WKY SA CP R. After elution from agarose gel, the DNA fragment was cleaved with the restriction endonucleases SacI and SmaI and cloned into pGL3-Basic Vector that had been previously cleaved with the same restriction endonucleases.

**SHR 1-3 pECFP-1**
The insert, consisting of a portion of the SHR SA gene comprised between exon 1 and exon 3, was excised from the construct SHR 1-3 pECFP-N1 (provided by Jian-hua Jia) with the restriction endonuclease KpnI. It was then eluted and cloned into pECFP-1 that had been previously digested with KpnI.

**WKY 1-3 pECFP-1**
The insert, consisting of a portion of the WKY SA gene comprised between exon 1 and exon 3, was excised from the construct WKY 1-3 pECFP-N1 (provided by Jian-hua Jia)
with the restriction endonuclease KpnI. It was then eluted and cloned into pECFP-1 that had been previously digested with KpnI.

**WKY 2-3 pECFP-1**
The insert was amplified by PCR with primers SA E2F KpnI and SA E3R KpnI. The PCR product was then digested with the restriction endonuclease KpnI and cloned into pECFP-1 that had been previously digested with KpnI.

### 2.24.5. Cloning of the SA minigenes mutated at the 5’ end

Construct WKY 1-3 ECFP-N1 was digested with Nhel (it cleaves near the transcription start site) and BglII (cuts inside SA intron 1).

**WA1-3 ECFP-N1**
This construct was obtained from WKY 1-3 ECFP-N1 by removing exon 1 (except for the last 6 nucleotides). To construct it, a PCR fragment was amplified with primers SA Nhel F and SA BglII IIR from clone WKY 1-3 ECFP-N1. The amplicon was then digested with the restriction endonucleases Nhe I and Bgl II and then cloned into WKY 1-3 ECFP-N1 that had been previously digested with the same enzymes. This construct was used to prepare construct WA1-3 glob1 ECFP-N1.

**WKY 1-3 Glob 1 ECFP-N1**
Exon 1 from the human β-globin gene was amplified by PCR with primers β Glob E1F Nhel and β Glob E1R Nhel. It was then restricted with Nhel and cloned into plasmid W 1-3 ECFP-N1 previously digested with Nhel

**WKY 1-3 Nhel ECFP-N1**
The insert was amplified with primers SA BglII IIR and SA E1F Nhel, digested with Nhel and BglIII and cloned into WKY 1-3 ECFP previously digested with Nhel and BglIII.

**WKY 1-3 5’ SS MUTANT ECFP-N1**
The insert was prepared by overlapping PCR. The first two PCR reactions were the following:

- **PCR A**: primer pECFP F and primer SA E1R 5’ss mutant.
- **PCR B**: primer SA E1F 5’ss mutant and primer SA BglII IIR

2nd round of PCR: PCR A and PCR B were mixed and used as template. Amplification was carried out with primers pECFP F and SA BglII IIR. The PCR product was restricted with the enzyme BglIII and cloned into WKY 1-3 ECFP-N1 restricted with the same enzyme.
The insert was amplified by PCR with primers β Glob E1F Nhel and β Glob E1R Nhel. The PCR product was then digested with the restriction endonuclease Nhel and cloned into WΔ 1-3 ECFP-N1.

2.24.6. Detection of a cryptic promoter by western blot

1) On day 1, 200,000 cells (293 cells) were seeded in a 6 well plate.
2) On day 2, cells were transfected with 1 μg (or equivalent) of plasmid DNA using the Gene Jammer Kit.
3) After 48 hours, cells were washed twice in 1X PBS. The second wash was used to harvest the cells by pipetting.
4) Cells were centrifuged at 1,200 rpm for 30 seconds.
5) The supernatant was discarded and cells were suspended in 100 μl of 0.5X RIPA buffer (1X RIPA contains 50mM Tris pH 7.4, 150 mM NaCl, 1%
NP40, 1% sodium deoxycholate, 0.1% SDS) containing 20 μg/ml of aprotinin (stock solution 10mg/ml) and incubated on ice for 20 minutes.
6) Samples were centrifuged at 13,000 rpm at + 4°C for 10 minutes.
7) The supernatant containing the proteins was transferred to a new tube and stored at −20°C.

2.24.7. SDS-PAGE

SDS-PAGE was carried out according to the protocol described by Laemmli (Laemmli 1970). The composition of the resolving gel (RG) and stacking gel (SG) are shown below:

<table>
<thead>
<tr>
<th>Component</th>
<th>RG 10% (ml)</th>
<th>SG 5% (ml)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₂O</td>
<td>4.0</td>
<td>2.1</td>
</tr>
<tr>
<td>30% Acrylamide</td>
<td>3.3</td>
<td>0.5</td>
</tr>
<tr>
<td>1.5 M Tris (pH 8.8)</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>1 M Tris (pH 6.8)</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>10% SDS</td>
<td>0.1</td>
<td>0.03</td>
</tr>
<tr>
<td>10% APS</td>
<td>0.1</td>
<td>0.03</td>
</tr>
<tr>
<td>TEMED</td>
<td>0.006</td>
<td>0.003</td>
</tr>
</tbody>
</table>
Samples were mixed with 2X SDS gel-loading buffer (100 mM Tris-HCl pH 6.8, 200 mM DTT added just before use, 4% SDS, 0.2% bromophenol blue, 20 % glycerol) and denatured at 95 °C for 5 minutes. Electrophoresis was carried out using the Mini Protean II (Bio-Rad) system at 150 volts until the bromophenol blue dye reached the bottom of the gel. SDS-PAGE was performed in the following buffer: 25 mM Tris-HCl pH 8.6, 192 mM glycine, 0.15% SDS.

2.24.8. Western blot

1) Transfer of SDS-PAGE separated proteins to the membrane (Hybond, ECL) was performed by semi-dry electrophoresis with the Bio-Rad Fast Blot B33 apparatus. The SDS-PAGE and membrane were sandwiched in 3MM paper pre-wetted in Transfer Buffer (1X Transfer buffer: 48 mM Tris base, 39 mM glycine, 0.037% SDS, 20 % IMS). Proteins were transferred at 15V for 30 minutes.

2) The membrane was blocked over night in TBS (20 mM Tris-HCl pH 7.5, 150 mM NaCl) containing 0.1% Tween 20 and 5% dried skimmed milk.

3) The day after, the membrane was incubated for 60 minutes with an anti-GFP monoclonal antibody (CHEMICON # MAB 2510 MSXGFP) diluted 1:2,000 in TBS containing 0.1% Tween 20 and 1% dried skimmed milk.

4) Three washes of 15 minutes each were carried out with TBS containing 0.1% Tween 20 and 1% dried skimmed milk.

5) The membrane was incubated for 60 minutes with the secondary antibody α mouse IgGAM (Sigma) diluted 1:2,000 in TBS containing 0.1% Tween 20 and 1% dried skimmed milk.

6) Three washes of15 minutes each were carried out with TBS containing 0.1% Tween 20 and 1% dried skimmed milk.

7) The membrane was covered with ECL solution (Amersham Pharmacia Biotech).

8) The membrane was finally exposed to an X-ray film for a variable length of time after which the X-ray film was developed.
2.25. Synthetic oligonucleotides

2.25.1. Oligonucleotides used in chapter 3

<table>
<thead>
<tr>
<th>PRIMER NAME</th>
<th>PRIMER SEQUENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alb E12F</td>
<td>AAGGCTCGCTTCATAGTGCC7AG</td>
</tr>
<tr>
<td>Alb E12R</td>
<td>CGGTTCCTACTGTCATATACTATG</td>
</tr>
<tr>
<td>COT E1F</td>
<td>GAGTCAGAGGACCAAGGCTGCA</td>
</tr>
<tr>
<td>COT E2F</td>
<td>GGACTCTCTTCCGCCCTTGC</td>
</tr>
<tr>
<td>COT E2R</td>
<td>GGAATGTCGTTCTTCAATTGAC</td>
</tr>
<tr>
<td>COT E3F</td>
<td>TTTGCAATGAGACGAAATAC</td>
</tr>
<tr>
<td>COT E3R</td>
<td>TGATGCAATGCTTGCCAAC</td>
</tr>
<tr>
<td>COT E4R</td>
<td>TTGTGATGGAATACGCACATCCAAG</td>
</tr>
<tr>
<td>COT E7R</td>
<td>GGGCAACAGGTTATCAGGCA</td>
</tr>
<tr>
<td>COT 2-2 junction</td>
<td>CTTGAGTCAGTTTTTCTTACT</td>
</tr>
<tr>
<td>COT 3-2 junction</td>
<td>AAAGAAACTGCTTTTCTTACT</td>
</tr>
<tr>
<td>COT E16F</td>
<td>TTCCCATGTTACATAAATGATGAC</td>
</tr>
<tr>
<td>COT E2F</td>
<td>GGACTCTCTTCCGCCCTTGC</td>
</tr>
<tr>
<td>COT E3R</td>
<td>GGAATGTCGTTCTTCAATTGAC</td>
</tr>
<tr>
<td>COT E4R</td>
<td>TTGTGATGGAATACGCACATCCAAG</td>
</tr>
<tr>
<td>COT E7R</td>
<td>GGCCCAACAGGTTATCAGGCA</td>
</tr>
<tr>
<td>COT Prom F</td>
<td>AATACGCTTTAGGAGATGATGAC</td>
</tr>
<tr>
<td>COT Prom Aval F</td>
<td>AGTGAATCCTAAAATATTGCGACAAGCAAA</td>
</tr>
<tr>
<td>COT Prom R</td>
<td>CTCAGGCCATACCCGGTTAG</td>
</tr>
<tr>
<td>COT I1 F</td>
<td>CATGGGAGTCCTAAAATTCCA</td>
</tr>
<tr>
<td>COT I2 R</td>
<td>CTACACTAGAAAGATTATGAAGATAC</td>
</tr>
<tr>
<td>SA E1F</td>
<td>GAATTCCTAGTGTCGGTGTAGGCATCG</td>
</tr>
<tr>
<td>SA E4R</td>
<td>CACCACTCTGGGATCTTGCGAGAACTCCA</td>
</tr>
<tr>
<td>SA LINE F</td>
<td>ACAAAAGAATTCACAGCTGGAGATG</td>
</tr>
<tr>
<td>SA 2-2 junction</td>
<td>CCAAATCCATGTGAAAAACA</td>
</tr>
<tr>
<td>SA 4-2 junction</td>
<td>AAATGATAGCTGCTGCGAGTAAAAAC</td>
</tr>
<tr>
<td>SA 4-2junction reverse</td>
<td>TCCTGTTTTTACCTGTTGC</td>
</tr>
</tbody>
</table>

The bold letters represent recognition sequences for the restriction endonuclease HindIII. The bold and underlined nucleotides represent the point mutation in the ATG codon (A → T).

2.25.2. Oligonucleotides used in chapter 4

<table>
<thead>
<tr>
<th>PRIMER NAME</th>
<th>PRIMER SEQUENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>COT E1F HindIII</td>
<td>CCCAAGCTTTGAGTGCGAGAGCCCGAGC</td>
</tr>
<tr>
<td>COT E2F HindIII</td>
<td>AGGCAGCTCTGTCAATTAGTCAAGTGGATG</td>
</tr>
<tr>
<td>COT ATG MUT F</td>
<td>CTGTGACTATACCTGGGAAATCAATGGCTAAG</td>
</tr>
<tr>
<td>COT ATG MUT R</td>
<td>TCAAGGTTATAGTCAAGTCAAGGATG</td>
</tr>
<tr>
<td>LUC REV</td>
<td>GTCCCATCTTCAGCGGATAG</td>
</tr>
<tr>
<td>SA E1F</td>
<td>GGCTTTTCTTCCTCCATTAAGCGGTCTA</td>
</tr>
<tr>
<td>SA E1F HindIII</td>
<td>CCCAAGCTTTGAGTGCGAGAGCCCGAGC</td>
</tr>
<tr>
<td>SA E5R</td>
<td>CCAAGCTTTCTTGGAGTGTTG</td>
</tr>
<tr>
<td>SA ATG Mutant F</td>
<td>CTATCTTTGTCACCCTGGCAATGGTACCTTGCTGCTAG</td>
</tr>
<tr>
<td>SA ATG Mutant R</td>
<td>CAAGGTTAGCAAGAGATAGCTAG</td>
</tr>
<tr>
<td>pGL3 FORW</td>
<td>TGCAATCTGCCATCTCAATTAG</td>
</tr>
</tbody>
</table>

The bold letters represent recognition sequences for the restriction endonuclease HindIII. The bold and underlined nucleotides represent the point mutation in the ATG codon (A → T).
2.25.3. Oligonucleotides used in chapter 5

<table>
<thead>
<tr>
<th>PRIMER NAME</th>
<th>PRIMER SEQUENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCR 1 F</td>
<td>CAAAGAAAGCTCGTGTTCTGAC</td>
</tr>
<tr>
<td>PCR 1 R</td>
<td>GAAGCTCAAGTCAGTGATCTTG</td>
</tr>
<tr>
<td>PCR 2 F</td>
<td>TCCACAGTGTGAGCAGCTGAGG</td>
</tr>
<tr>
<td>PCR 2 R</td>
<td>CATTAAAGCAGGAGGAGGCTTCCCATTG</td>
</tr>
<tr>
<td>PCR 3 F</td>
<td>CAGGAGTAGCTGACACAAAT</td>
</tr>
<tr>
<td>PCR 3 R</td>
<td>CGATCGGAGAAACTGGAACAGCAATATC</td>
</tr>
<tr>
<td>PCR 4 F</td>
<td>TGTGTTCAAGTCTCTCAGTGAC</td>
</tr>
<tr>
<td>PCR 4 R</td>
<td>CTGACAACAGCTGACTCTGAG</td>
</tr>
<tr>
<td>PCR 5 F</td>
<td>TTAGAAGTCCTCAGTAGACATC</td>
</tr>
<tr>
<td>PCR 5 R</td>
<td>GGAGTTTTTGCTGATGAT</td>
</tr>
<tr>
<td>SA Prom Sac II F</td>
<td>TCCCGCGGTACCCACGACTGTAATAACAAAG</td>
</tr>
<tr>
<td>SA Promoter Sac II R</td>
<td>TCCCGCGGAGAATTTCAGCTGAGAGAG</td>
</tr>
<tr>
<td>WKY SA int 1 Sac II F</td>
<td>TCCCGCGGAGGCTTTATATAACGTTAGC</td>
</tr>
<tr>
<td>WKY SA int 1 Sac II R</td>
<td>TCCCGCGGAGGGTTTGAACTGCTTTAGA</td>
</tr>
<tr>
<td>WKY SA Int 1 KpnI</td>
<td>TCCGTTACAGGCTTTTTTATAACGTTAGC</td>
</tr>
<tr>
<td>WKY SA Int 1 Xhol</td>
<td>TCCCTCGAGGGAGGTGTAACGTTAGC</td>
</tr>
<tr>
<td>WKY SA CP F</td>
<td>CAGACTCTTTGGGAAATGGTGT</td>
</tr>
<tr>
<td>WKY SA CP R</td>
<td>TCCCGCGGAGCTCTCTCCAAGGACTAAAC</td>
</tr>
<tr>
<td>pECFP-1F</td>
<td>TAGTTATTAC TAGCGTACC</td>
</tr>
<tr>
<td>SA E2F KpnI</td>
<td>CGGGTGACCGTGAACACAGAGAGCTCAAGTG</td>
</tr>
<tr>
<td>SA E3R KpnI</td>
<td>CGGGTGACCGTGAACACAGAGAGCTCAAGTG</td>
</tr>
<tr>
<td>SA E1F Nhe I</td>
<td>CTACTAGCTAGCCAAATTGGCTTTTCTCCCAT</td>
</tr>
<tr>
<td>SA E1R 5' ss mutant</td>
<td>TAGGATCTACCTGAGATAGCTGCTTTCCACAGCT</td>
</tr>
<tr>
<td>SA E1F 5' ss mutant</td>
<td>TTTCTTTTTCTCTGTGATAGACCTCCCTG</td>
</tr>
<tr>
<td>SA Nheli F</td>
<td>CTACTAGCTAGCTTCTAGTGATCTCTCATTTTC</td>
</tr>
<tr>
<td>β Glob E1F Nhe I</td>
<td>CTACTAGCTAGCAGATTTGCTCTGACACAATGCTG</td>
</tr>
<tr>
<td>β Glob E1R Nhe I</td>
<td>CTACTAGCTAGCAGCTCATCTACACTACAC</td>
</tr>
<tr>
<td>β Glob E1F</td>
<td>AGTCTGCGTTAAGCTGCTG</td>
</tr>
<tr>
<td>SA Bgl II 1R</td>
<td>GTGTCTCTAGGAGATGCAG</td>
</tr>
<tr>
<td>SA 2-2 junction</td>
<td>CCAAATCCATGTGAAAAC</td>
</tr>
<tr>
<td>SA E3R</td>
<td>GTGCCATGTGCTCAAACACTTTT</td>
</tr>
<tr>
<td>SA E2F</td>
<td>TCAACAGAGACCTAAACCCAAATC</td>
</tr>
</tbody>
</table>

The bold letters represent recognition sequences for restriction endonucleases.
CHAPTER 3

EXON REPETITION IN SHR AND WKY.
STRAIN-SPECIFIC OR ALLELE-SPECIFIC?

3.1. INTRODUCTION

3.1.1. Exon repetition of the SA gene

SA, a putative candidate gene linked to hypertension, has been well characterised in two rat strains, the hypertensive SHR (spontaneously hypertensive rat) and the normotensive WKY (Wystar-Kyoto). SA was found to be expressed in the proximal tubule cells of the kidney and in liver hepatocytes. Northern blot analysis revealed that the expression level of SA was markedly increased in both tissues of the SHR compared to the same tissues from WKY. Another significant difference between the two strains was the presence of extra transcripts only in the WKY kidney. In order to characterise these extra transcripts, SA mRNA was amplified by RT-PCR between exon 1 and exon 15. This set of primers amplified a single PCR product of approximately 2 Kb from both tissues of the SHR and also from WKY liver. This PCR fragment was sequenced and was found to contain a single copy of each exon. In contrast, the same set of primers amplified three PCR products of different size from WKY kidney. One PCR product had the same size as the one in SHR and WKY liver as shown in figure 3.1 (Frantz SA 1999) whereas the other two amplicons were 2.1 and 2.7 Kb in size. These two PCR products were sequenced. The 2.1 Kb fragment was found to contain a tandem repeat of exon 2, whereas the 2.7 Kb amplicon contained a tandem repetition of exons 2, 3 and 4. This phenomenon was named exon repetition.

3.1.2. Different ways of obtaining exon repetition

Several explanations for the presence of exon repetition in SA mRNA are possible. The first one is that exons 2-3-4 are duplicated in genomic DNA of the WKY proximal tubule cells as shown in figure 3.2. This gene would be transcribed into a pre-mRNA molecule containing a tandem repeat of exons 2-3-4 and relative intronic sequences. During pre-mRNA processing, the three repeated exons could all be skipped (pathway I in panel B...
Figure 3.1. Exon repetition of the rat SA gene is strain and tissue-specific (from Frantz et al, 1999). (A) Northern blot analyses of SA mRNA in kidney and liver of the two rat strains SHR and WKY. (B) Analysis of SA mRNA by RT-PCR. Amplification was carried out between exon 1 and exon 15. Two tissues, liver (liv) and kidney (kid) from SHR and WKY were analysed. (C) Exon composition of the amplicons of panel B. The 2.0 Kb cDNA has a normal arrangement of exons whereas the 2.1 Kb and 2.7 Kb SA mRNA isoforms contain a tandem repetition of exon 2 or exons 2-3-4 respectively. The SA mRNA isoforms containing repeated exons are present only in kidney of WKY rats, whereas the normal SA mRNA isoform is found in both tissues of SHR and in WKY liver.
Figure 3.2. Exon repetition in SA could arise from the duplication of exons 2-3-4 in genomic DNA of the proximal tubules of WKY. (A) Schematic representation of the WKY SA gene. Exons 2-3-4 are duplicated (shown in blue and red). For simplicity, only ten exons are shown instead of the fifteen exons present in the rat SA gene. (B) Diagrammatic representation of the pre-mRNA derived by transcription of the gene in A. The exons that are duplicated in the gene are repeated in the pre-mRNA. A complex pattern of alternative cis-splicing produces several mRNA molecules, some of which have repeated exons. When cis-splicing occurs through pathway I, a normal mRNA molecule is generated. In contrast, pathway II causes the skipping of exons 3 and 4 and yields an mRNA isoform with a tandem repeat of exon 2. Finally, in pathway III, all the exons of the pre-mRNA are retained in the mRNA molecule, which will contain a tandem repeat of exon 2 2-3-4. (C) Schematic representation of the three mRNA isoforms. The symbols I, II and III indicate from which splicing event (see-B) a particular mRNA isoform was derived.
of figure 3.2), leading to the production of normal mRNA. Normal SA mRNA could also be produced whenever one repeated exon of the first repeat is spliced to the next exon of the second repeat. Skipping of exon 3 and exon 4 only (pathway II in panel B of figure 3.2) would lead to the production of an mRNA isoform containing a tandem repeat of exon 2. Finally, if none of the exons were skipped (pathway III in panel B of figure 3.2), exons 2, 3 and 4 would be repeated in the mRNA. In this way, from a single gene, several mRNA isoforms (some of which would contain repeated exons) would be produced by alternative cis-splicing. The possible duplication of exons 2, 3 and 4 in the SA gene of WKY was ruled out by a southern blot experiment which clearly showed that those exons are present in single copy in the WKY SA gene. Moreover, a southern blot performed with genomic DNA extracted from proximal tubule cells (the cells in which exon repetition of the SA gene is observed) gave the same result, thus ruling out the possibility of a tissue-specific duplication of certain exons.

Another possible explanation for the presence of repeated exons in WKY SA mRNA was the duplication of the entire SA gene in the proximal tubules of WKY as shown in figure 3.3. After transcribing the first copy of the gene, the enzyme RNA polymerase II could fail transcription termination and carry on transcription into the second copy of the gene, thus producing a single pre-mRNA molecule consisting of two tandem repeats of the SA gene. Through a complex pattern of alternative cis-splicing (see panel B of figure 3.3), the two mRNA isoforms containing repeated exons could be generated. In particular, skipping of the exons comprised between exon 4 of the upstream copy and exon 5 of the downstream copy (pathway I in panel B of figure 3.3) would lead to the production of an mRNA isoform with a normal exon organisation. Alternatively, splicing of exon 2 of the upstream copy to the downstream exon 2 (pathway II) would produce the mRNA isoform with the tandem repeat of exon 2. Finally, splicing of the upstream exon 4 to exon 2 of the downstream copy would generate the SA mRNA isoform containing the tandem repeat of exons 2-3-4 (pathway III in panel B of figure 3.3). Interestingly, hybrid mRNA molecules derived from tandemly arranged genes have been detected (Pradet-Balade B 2002). One example is a hybrid mRNA molecule composed of exons from two different tumour necrosis factor genes named TWEAK and APRIL. The TWEAK gene is only less than 1 Kb upstream from APRIL and both genes have the same orientation. The hybrid mRNA molecule is composed of exons 1-6 of TWEAK joined to exons 2-6 of APRIL. The hybrid molecule was named TWE-PRIL and was shown to have a biological function.

It was possible that that exon repetition in SA might arise through gene duplication and the by-pass of transcription termination. This was investigated by performing a dot blot
Figure 3.3. Exon repetition in SA could arise by duplication of the entire SA gene in the proximal tubules of WKY. (A) Schematic representation of the tandem duplication of the entire SA gene. For simplicity, only ten of the fifteen exons are shown. (B) Pre-mRNA molecule derived by transcription of the duplicated gene shown in A. RNA polymerase II failed transcription termination after transcribing the first copy of the gene, thus producing a pre-mRNA molecule containing a tandem duplication of the gene. Through a complex pattern of alternative splicing, different mRNA isoforms are produced from the same pre-mRNA molecule. Splicing from exon 2 of the upstream copy, to exon 2 of the downstream copy (pathway II) generates an mRNA isoform with a tandem repeat of exon 2. In contrast, when exon 4 of the upstream copy is spliced to exon 2 of the second copy, repetition of exons 2-3-4 occurs (pathway III). There are several ways by which a normal SA mRNA molecule could be produced. The first one is transcription termination after the first copy of the gene has been transcribed, combined to the retention of all the exons of the pre-mRNA. Secondly, from the pre-mRNA molecule illustrated in panel B, normal mRNA can be produced whenever an exon in the upstream copy of the gene is spliced to the next exon in the downstream copy of the gene (one example is shown by pathway I in which exon 4 of the upstream copy is spliced to exon 5 of the downstream copy). (C) Schematic representation of the mRNA isoforms produced by the alternative cis-splicing reactions shown in panel B.
experiment. Serial dilutions of genomic DNA from SHR and WKY were spotted on a filter which was then hybridised to a probe corresponding to exon 2 of SA and lately to a second probe corresponding to the GAPDH gene, which was known to be a single copy gene. By calculating the ratio between the intensities of the hybridisation signals from SA and GAPDH, it was demonstrated that the SA gene was not duplicated in WKY. Furthermore, the same result was obtained with genomic DNA from liver and kidney of WKY and this ruled out the possibility of a tissue-specific duplication of the entire SA gene in WKY kidney.

All the evidence mentioned above suggests that exon repetition in SA arises through a post-transcriptional (or possibly co-transcriptional) RNA processing event. One possible explanation is that the SA mRNA isoforms containing repeated exons are derived from two different SA pre-mRNA molecules. For example, the two pre-mRNA molecules could be spliced together through a process called trans-splicing. This is simply a splicing reaction in which two different pre-mRNA molecules are involved as shown in figure 3.4. If the 5' splice site associated with exon 2 in one molecule is spliced in trans to the 3' splice site adjacent to exon 2 of another molecule, an mRNA with a tandem repetition of exon 2 is generated (pathway I of panel A in figure 3.4). Instead, if the 5' splice site associated with exon 4 in one molecule, is trans-spliced to the 3'splice site of exon 2 of a second pre-mRNA, an mRNA isoform containing a tandem repeat of exons 2-3-4 is produced (pathway I in panel A of figure 3.4).

3.1.3. Exon repetition in SA is strain-specific and tissue-specific

Exon repetition in SA is very peculiar, since it is both strain and tissue specific. Three possible explanations could account for these peculiarities. The first one is that exon repetition is caused by a trans-acting factor (possibly a splicing factor that promotes trans-splicing) that is expressed only in WKY kidney. The second explanation is that there are differences in the genomic sequence between the WKY and SHR SA alleles (both strains are homozygous at the SA locus), and these differences are in turn responsible for the different behaviour of the two strains. This explanation on its own would fail to explain the tissue-specificity of exon repetition in the SA gene. The third explanation is that strain-specificity is due to allelic differences, whereas the tissue-specificity is caused by a tissue-specific trans-acting factor. In order to shed light on the causes underlying strain specificity, an F2 population was obtained by crossing F1 rats that had been derived from a cross between SHR and WKY as shown in figure 3.5. The rats of the F2 population were divided into two groups. The first group included all of the animals with two copies of the
Figure 3.4. The presence of repeated exons in SA mRNA could arise by trans-splicing. (A) Schematic representation of two pre-mRNA molecules (one shown with white exons, the other one with black exons). When the donor splice site associated with exon 2 of the molecule with white exons is trans-spliced to exon 2 of the molecule with black exons, an mRNA isoform containing a tandem repeat of exon 2 is produced (see pattern I). In contrast, when the donor splice site adjacent to exon 4 of the first molecule is trans-spliced to the acceptor site adjacent to exon 2 of the second molecule, an mRNA isoform with a tandem repeat of exons 2-3-4 is produced as illustrated by pattern II. (B) Diagrammatic representation of the two mRNA isoforms containing repeated exons that have been derived through trans-splicing between the two pre-mRNA molecules shown in A.
Figure 3.5. Obtainment of an F2 population from a cross between WKY and SHR. First of all, the parental strains WKY and SHR were crossed to generate F1 rats which are heterozygote at the SA locus, having inherited one SA allele from each parent. The WKY and SHR SA alleles are represented by the grey and black bars, respectively. The F1 rats were then crossed with each other to produce an F2 population which was made of rats homozygote for the WKY SA allele (shown on the left with the two grey bars), heterozygotes (shown in the middle with one grey and one black bar) and finally SHR SA homozygotes (shown on the right and containing two black bars). The expected ratio of WKY homozygote : heterozygote : SHR homozygote is 1:2:1.
Figure 3.6. Two possible explanations for the segregation of the property of exon repetition of SA with the chromosomal region around the WKY SA allele. **(A)** Exon repetition of SA is under the control of a gene (encoding an "exon repetition trans-acting factor") that lies near the WKY SA allele and thus co-segregates with it. It is represented by the black portion of the chromosome in WKY. In contrast, on the SHR chromosome, this gene is either missing or non-functional. Both the WKY and SHR SA alleles are depicted as a grey bar to show that they are both capable of producing SA mRNA isoforms with repeated exons, providing they are expressed in an environment containing the trans-acting factor that causes exon repetition. **(B)** Differences between the WKY and SHR SA alleles are responsible for the different behaviour of the two alleles. The WKY SA allele is depicted in grey, whereas the SHR SA allele is shown in black. The different colour represents a difference in behaviour of the two alleles, with the WKY allele being able to produce the normal SA mRNA and also SA mRNA isoforms that contain repeated exons. In contrast, the SHR SA allele can only produce the normal mRNA isoform.
WKY SA allele (WKY SA homozygotes). The second group included all of the rats that had two copies of the SHR SA allele (SHR SA homozygotes). The size difference of intron 1 of SA in WKY and SHR was exploited to genotype the animals by southern blot. Exon repetition of SA was evaluated in the two groups. All the rats belonging to the first group (WKY SA homozygotes) had exon repetition. In contrast, in the second group (SHR SA homozygotes), all the animals lacked exon repetition (Simon Frantz and Nilesh Samani, personal communication). These findings strongly suggested cis-acting elements as the determinants of strain-specificity, most likely located inside the SA gene. Another possible explanation involving a gene that encodes a trans-acting factor (possibly involved in trans-splicing) and that is in close proximity to the WKY SA allele, and hence co-segregates with could not be completely ruled out. The two possible explanations for the strain-specificity of exon repetition in the SA gene are illustrated in figure 3.6. The second explanation is somehow more exhaustive, because it could easily account for the tissue-specificity of exon repetition in SA by simply assuming that the trans-acting factor required for exon repetition is expressed in the kidney but not in the liver. In contrast, differences between the WKY and SHR SA alleles could easily explain the different behaviour of the two strains, but would fail to explain tissue specificity. As mentioned before, a third mechanism could explain all the facts of exon repetition in SA. In this case, the strain specificity could be due to differences between the WKY and SHR SA alleles, while the tissue-specificity could result from a trans-acting factor that is expressed in kidney but not in liver (tissue-specific epigenetic modifications might well cause the same gene to behave differently in different tissues). Experiments carried out with minigenes from both WKY and SHR showed that both minigenes recapitulated exon repetition upon transfection in mammalian cells (Jianhua Jia personal communication). These data suggested that strain-specificity was not due to allelic differences between the two strains and supported the hypothesis that the SHR strain lacks a trans-acting factor required for exon repetition of SA. If exon repetition is indeed due to the presence of a gene encoding a trans-acting factor that co-segregates with the WKY SA allele, it is possible that the same trans-acting factor might cause exon repetition in other genes that are prone to produce mRNA isoforms containing repeated exons. Another gene that had been shown to exhibit exon repetition with high efficiency in rat was the COT (carnitine octanoyltransferase) gene (Caudevilla C 1998). Exon repetition in COT was investigated only in the Sprague-Dawley rat. Hence no data regarding strain-specificity were available. Three COT mRNA isoforms were detected. In one isoform, all the exons were present in single copy whereas the other two isoforms contained a tandem repetition of exon 2 alone or exons 2 and 3.
Figure 3.7. Exon repetition of COT is strain-specific. The presence or absence of exon repetition in COT was evaluated by RT-PCR followed by separation on agarose gel. Eight WKY and seven SHR animals were used in the experiment. (A) Analysis by amplification with primers COT E1F and COT E7R. This set of primers amplifies both the normal COT mRNA and mRNA containing repeated exons. From all WKY rats, two PCR products were amplified. Their DNA sequence revealed that the shorter amplicon had a normal arrangement of exons, while the longer PCR product contained a tandem repetition of exon 2. The exon organization is depicted by the black boxes with white numbers. In contrast, from the seven SHR rats that were tested, the same RT-PCR reaction amplified a single PCR product that was found to correspond to the normal COT mRNA. (B) Analysis of COT mRNA by amplification with primers COT E2F and COT E2R, which are specifically designed to amplify the tandem repetition of exon 2. This set of primers confirmed the presence of exon repetition in the WKY animals and its absence in the SHR rats. (C) Schematic representation of the COT cDNA (normal cDNA and the one with a tandem repeat of exon 2) with the primers used in the PCR reactions.
Figure 3.8. DNA sequence chromatographs of the junction between the two tandemly repeated copies of exon 2 in COT mRNA. DNA fragments were obtained by RT-PCR from WKY kidney RNA. (A) Sequence chromatograph of the PCR product amplified with primers COT E1F and COT E7R. The junction between the two copies of exon 2 is marked by the vertical arrow. (B) Sequence chromatograph of the PCR product amplified with primers COT E2F and COT E2R which are specifically designed to detect the tandem repetition of exon 2. The junction between the two copies of exon 2 is marked by a vertical arrow.
3.2. RESULTS

3.2.1. Exon repetition in the COT gene is strain-specific

To test the hypothesis that the strain-specificity of exon repetition is caused by a trans-acting factor encoded by a gene that co-segregates with the WKY SA allele, exon repetition in a second gene called COT, was evaluated in WKY and SHR. The COT gene was chosen because it had been shown to produce mRNA isoforms containing repeated exons very efficiently in the Sprague-Dawley rat. Exon repetition of COT was examined in eight WKY and seven SHR animals by RT-PCR from kidney RNA. COT cDNA was amplified between exon 1 and exon 7 and the result is shown in panel A of figure 3.7. Two PCR products were amplified from the eight WKY animals. In contrast, from all seven SHR rats, a single PCR product was amplified. The DNA fragments were eluted from the agarose gel and sequenced. The short PCR product in WKY had a normal exon organisation, from exon 1 to exon 7. In contrast, the DNA sequence of the longer amplicon, which is present only in WKY, contained a tandem repeat of exon 2. The DNA sequence chromatographs of the junction between the two copies of exon 2 are shown in figure 3.8 A. The failure of primers COT E1F and COT E7R to detect exon repetition of COT in SHR could have been due to a very low level of exon repetition in SHR compared to WKY. To investigate this, a second PCR reaction was performed with primers COT E2F and COT E2R. These primers are specifically designed to amplify the tandem repetition of exon 2. From all WKY rats, a PCR product of the expected size was amplified (see figure 3.7 B) and its DNA sequence showed that it corresponded to an exact tandem repetition of exon 2 as shown in figure 3.8 B. The same primers failed to amplify a PCR product from SHR cDNA (see figure 3.7 B). This clearly demonstrated the absence of exon repetition of COT in the SHR strain. These data were confirmed by real-time PCR, which was carried out in the presence of SYBR Green. The fluorescence of this molecule is considerably increased when double stranded DNA is present. Figure 3.9 shows the result with primers COT E3F and COT E4R, which amplify from total COT mRNA (with or without the repetition of exon 2). For each sample, the cycle number at which the fluorescence reaches a particular threshold, which is known as crossing point, was calculated. As shown in table 3.1, the crossing points obtained from WKY and SHR are very similar. Since SYBR green binds to any double stranded DNA molecule, a melting curve was determined to ensure that the signal was not due to any aspecific amplification. Different DNA molecules have different length and base composition and therefore will be denatured at slightly different
Figure 3.9. Amplification of total COT mRNA by real-time PCR with primers COT E3F and COT E4R. (A) Amplification profile of the real-time PCR reaction. One microlitre of cDNA from kidney of WKY or SHR was used as a template. The amplification profile relative to the blank reaction is also shown. The horizontal red bar represents the threshold of the fluorescence signal. The crossing point is represented by the cycle number at which the intensity of the fluorescence signal coincides with the threshold and is calculated by the software. (B) Melting profile of the three PCR reactions. The samples obtained from WKY and SHR had the same melting profile indicating that the same amplicon had been amplified in these two samples. In contrast, the PCR reaction derived from the negative control had a lower melting temperature, indicating that the signal at the end of the amplification (see panel A) was likely to be due to the presence of primer dimers.
Figure 3.10. Real-time PCR with primers COT E2F and COT E2R. This set of primers was designed to amplify the tandem repeat of COT exon 2. (A) Amplification profiles of the real-time PCR reaction. One microlitre of cDNA from WKY or SHR kidney was used as template. The amplification profile relative to the negative control is also shown. The horizontal red bar represents the threshold of the fluorescence signal. The crossing point is represented by the cycle number at which the intensity of the fluorescence signal coincides with the threshold and is calculated by the software. (B) Melting profile of the three PCR reactions. Every sample had a different melting temperature, suggesting that a different amplicon had been amplified from each sample. All samples were loaded on an agarose gel. The size of the amplicon obtained from WKY corresponded to the size expected with this set of primers. In contrast the size of the amplicon from SHR differed from the expected size demonstrating that a non specific amplification had occurred (data not shown). Finally, the size of the amplicon obtained from the blank reaction was consistent with primer
temperatures. The presence of a single peak in figure 3.9 B indicates the absence of primer dimers or any other aspecific amplicons.

Non-specific amplifications in the sample would result in multiple peaks from the melting curve experiment.

<table>
<thead>
<tr>
<th>Crossing point</th>
<th>WKY kidney</th>
<th>SHR kidney</th>
<th>Blank</th>
</tr>
</thead>
<tbody>
<tr>
<td>23.32</td>
<td>23.20</td>
<td>&gt;41</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1 Analysis of total COT mRNA by real-time PCR. The amplification was carried out with primers COT E3F and COT E4R. As template, 1 µl of cDNA from WKY or SHR kidney was used. The crossing points for WKY, SHR and the blank reaction are shown.

Primers COT E2F and COT E2R, which specifically amplify the COT cDNA isoform that contains a tandem repeat of exon 2, were also tested by real-time PCR and the result is shown in figure 3.10 and table 3.2. This time, a significant difference was observed between the two strains. In WKY, the crossing point with these primers was similar to that obtained with the previous set of primers (compare the crossing points of tables 3.1 and 3.2), suggesting the presence of a significant amount of COT mRNA containing the tandem repeat of exon 2. In contrast, in SHR, not only was the crossing point much higher, but the amplicon generated was found to be a non-specific amplification as demonstrated by the aberrant melting profile (see figure 3.10 B, SHR). This was confirmed by separation of the PCR reaction on agarose gel, which showed that the size of the amplicon obtained from SHR was incorrect. Taken together, all these results indicate that exon repetition in COT mRNA is strain specific; it is present in WKY but not in SHR.

<table>
<thead>
<tr>
<th>Crossing point</th>
<th>WKY kidney</th>
<th>SHR Kidney</th>
<th>Blank</th>
</tr>
</thead>
<tbody>
<tr>
<td>22.6</td>
<td>28.97</td>
<td>34.74</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2 Analysis of the COT mRNA isoform with the tandem repeat of exon 2 by real-time PCR. The amplification was carried out with primers COT E2F and COT E2R, which are specifically designed to detect repetition of exon 2. As template, 1 µl of cDNA from WKY or SHR kidney was used. The crossing points for WKY, SHR and the blank reaction are shown.

The results in figure 3.7 showed that the COT mRNA isoform with the tandem repeat of exon 2 is more abundant than the normal isoform. These data were purely
Figure 3.11 Quantification of repetition of exon 2 in COT mRNA of WKY rat by PCR. Amplification was carried out with different cycle numbers. 1 µl of cDNA obtained by RT from total kidney RNA was used as template. PCR amplification was performed with primers COT E1F and COT E3R. PCR conditions were as before except that 0.25 µl of (α-^32P) dCTP (10 µCi/ml 3000mCi/mm) were added to each reaction. Prior to loading, PCR reactions were mixed with formamide loading dye, heated at 94°C for 5 minutes and cooled on ice. PCR products were then separated on a 6% polyacrylamide gel, and the result is shown in panel A. The radioactive signal was quantified with a phosphorimager, and log_{10} signal (arbitrary units) was plotted against the number of cycles as shown in panel B. The signal at cycle 1 was calculated from each slope, corrected for the numbers of dCMP moieties incorporated on both strands, and used to calculate the proportion of the exon repetition isoform (Eperon IC 1994).
Cycles 18 20 22 24

A

B

Number of cycles

Log_{10} radioactivity

Normal mRNA values and best fit line. Calculated amplification efficiency = 1.83

Exon repetition isoform values and best fit line. Calculated amplification efficiency = 1.76
indicative since the visualisation of DNA fragments with ethidium bromide is not quantitative. Furthermore, since the two PCR products differ in size, it was possible that they were amplified with different efficiencies. To investigate this possibility, a cycling curve with primers COT E1F and COT E3R was performed. Kidney cDNA from a WKY rat (the second from the left in figure 3.7) was utilised as template. PCR reactions were performed in the presence of $\alpha^{32}\text{P-}dCTP$. Amplification was carried out in separate tubes for 18, 20, 22 and 24 cycles. PCR products were separated on a denaturing acrylamide gel. The signal was detected with a Phosphorimager and the result is shown in figure 3.11 A. The intensity of each band was evaluated with Optiquant. The $\log_{10}$ of each value was then plotted against the number of cycles using the cricket graph software and the result is shown in figure 3.11 B. The existence of a linear relation between the number of cycles and the logarithm of the signal intensity demonstrated that the amplification reaction was still in the exponential phase. The best line fitting each set of values was then obtained and the equation that described each line was determined with cricket graph. The amount of each isoform at cycle 1 was extrapolated using the equation of the two lines as described in materials and methods. Finally, a correction was made that took into account the different length of the two PCR products (one DNA molecule with a tandem repeat of exon 2 contained more radioactivity than a DNA molecule with a normal arrangement of exons). According to these calculations, the normal isoform represented 46.3 % of the total COT mRNA whereas the isoform with the tandem repeat of exon 2 accounted for 53.7 % of total COT mRNA. All these data taken together demonstrated that the COT mRNA isoform with the tandem repeat of exon 2 was not a minor isoform due to a noise background but in contrast was even more abundant than the normal COT isoform.

The presence of exon repetition of the two genes tested (SA and COT) in one strain (WKY) but not in the other (SHR), supported the hypothesis that exon repetition could be caused by a trans-acting factor expressed only in the kidney of WKY. To further investigate this, exon repetition of COT was studied in liver by RT-PCR. The amplification was carried out between exon 1 and exon 7 and the result is shown in figure 3.12. Exon repetition of COT was detected in both tissues of WKY but not in the corresponding tissues of SHR. The fact that exon repetition of COT was found in both kidney and liver, while exon repetition of SA was present only in kidney, suggested that exon repetition of both genes could not be determined by a single trans-acting factor. If trans-acting factors are the cause of strain-specificity then at least two different ones must be involved. For example one factor could be gene-specific, either for COT or SA, the other one could be required by both genes. This model is illustrated in figure 3.13. The hypothetical trans-
Figure 3.12. Analysis of exon repetition of COT in liver and kidney of WKY and SHR by RT-PCR. Amplification was carried out between exon 1 and exon 7 with primers COT E1F and COT E7R. Exon repetition was detected in both tissues of WKY but not in the corresponding tissues of SHR. The exon organisation of the PCR products is depicted by the black boxes and white numbers.
Figure 3.13. Model for exon repetition in which two different trans-acting factors are required for each gene. According to this model, mRNA isoforms containing repeated exons are produced only if two trans-acting factors are present. One trans-acting factor (■) is required by both SA and COT and is expressed ubiquitously (liver and kidney). The second factor is gene-specific. The trans-acting factor specific for SA (○) is only expressed in the kidney (this explains the absence of exon repetition of SA in liver). In contrast, the trans-acting factor specific for COT (○) is expressed in kidney and liver and hence exon repetition of COT takes place in both tissues.
acting factor involved in exon repetition of both genes is expressed ubiquitously (in liver and kidney of WKY). The COT-specific trans-acting factor is also expressed in both tissues. In contrast, the trans-acting factor specific for SA is expressed only in kidney. This model would explain all the observations relative to strain and tissue-specificity of exon repetition in SA and COT.

3.2.2. Exon repetition in SA and COT segregate independently in an F2 population

One explanation for the presence of exon repetition of both genes in WKY but not in SHR is that a trans-acting factor required for exon repetition was present in WKY and absent in SHR. Of course it was still possible that the presence of exon repetition of two genes in one strain but not in the other could be purely a coincidence. This question could be addressed using two different strategies:

1) Investigate exon repetition in other genes in WKY and SHR.
2) Determine if in an F2 population obtained by crossing F1 rats (WKY X SHR) exon repetition of SA co-segregates with exon repetition of COT.

The second approach is certainly more informative than the first one (again having three genes with exon repetition in WKY but not in SHR might still be due to chance). For this reason, the second approach was adopted.

The rats of an F2 population had already been genotyped at the SA locus and divided into two groups by Simon Frantz. The first group included all the rats homozygote for the WKY SA allele whereas the second group was composed of rats homozygote for the SHR SA allele. First of all, exon repetition of SA was evaluated by RT-PCR from kidney RNA, with primers SA E1F and SA E2R. As shown in figure 3.14 A, from the seven WKY SA homozygotes (animals 1-7), three PCR products of the expected size were amplified, demonstrating the presence of SA mRNA isoforms with repeated exons in this group. In contrast, the same primers amplified only a single PCR product corresponding to the normal SA isoform from all nine SHR SA homozygotes (animals 8-16 in figure 3.14 A).

To verify whether exon repetition of SA and of COT co-segregated, exon repetition of COT was evaluated in the same rats. COT cDNA was amplified between exon 1 and exon 7 with primers COT E1F and COT E7R and the results are shown in figure 3.14 B. Out of seven WKY SA homozygotes, six had exon repetition in COT (animals 1-3 and 5-7) whereas only one lacked exon repetition (animal 4). In the second group (SHR SA homozygotes), the same set of primers detected COT exon repetition in 6 animals (animals
Figure 3.14. Exon repetition in SA and COT segregate independently in an F2 population. Animals from an F2 population derived by crossing F1 rats (WKY X SHR) were divided into two groups: WKY SA homozygotes (animals 1-7) and SHR SA homozygotes (animals 8-16). (A) Analysis of exon repetition in SA mRNA by RT-PCR with primers SA E1F and SA E2R. All seven rats that are homozygote for the WKY SA allele exhibit the three band pattern typical of exon repetition in SA (animals 1-7). In contrast all 9 rats homozygote for the SHR SA allele lack exon repetition in SA (animals 8-16). The exon organisation of the SA PCR products is shown by the white boxes near each PCR product. (B) Analysis of exon repetition in COT mRNA by RT-PCR with primers COT E1F and COT E7R. Out of the seven rats homozygote for the WKY SA allele, one did not show exon repetition of COT (animal number 4 in panel B). Of the nine rats homozygote for the SHR SA allele, six showed exon repetition of COT (rats number 8, 9, 10, 11, 13 and 16), but the remaining three did not (animals number 12, 14 and 15). The exon organisation of the PCR products is shown by the black boxes near each amplicon. (C) Further analysis of exon repetition in COT mRNA by PCR with primers COT E2F and COT E2R. The exon organisation of the PCR products is shown by the black boxes near each amplicon.
8-11, 13 and 16) but not in the other three rats (animals 12, 14 and 15). The presence (or absence) of exon repetition in COT, was confirmed with primers COT E2F and COT E2R, and the result is shown in figure 3.14 C. The results are summarised in tables 3.3 and 3.4.

<table>
<thead>
<tr>
<th>Repetition in SA gene</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

| Repetition in COT gene | yes | yes | yes | no | yes | yes | yes |

Table 3.3 Exon repetition of COT and SA in F2 rats homozygous for the WKY SA allele. Seven rats from the F2 population (indicated by numbers 1-7) were analysed. The presence of exon repetition is indicated by "yes", its absence by "no". The results in disagreement with the hypothesis that exon repetition of SA co-segregates with exon repetition of COT are shown in bold (animal # 4).

<table>
<thead>
<tr>
<th>Repetition in SA gene</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

| Repetition in COT gene | yes | yes | yes | yes | no  | yes | no  | no  | yes |

Table 3.4 Exon repetition of COT and SA in F2 rats homozygous for the SHR SA allele. Nine rats from the F2 population (indicated by numbers, 8-16) were analysed. The presence of exon repetition is indicated by "yes", its absence by "no". The results in disagreement with the hypothesis that exon repetition of SA co-segregates with exon repetition of COT are shown in bold (animals #8-11, 13 and 16).

These results clearly demonstrated that exon repetition in SA and COT segregate independently. For example in rat No. 4, exon repetition was present in SA mRNA but not in COT mRNA. Furthermore, animals number 8-11 exhibited exon repetition in COT but not in SA. The independent segregation of exon repetition in SA and COT was further supported by the absence of exon repetition of SA in Sprague-Dawley (see figure 3.15 A) in which exon repetition of COT is observed. An analysis of the SA genotype in Sprague-Dawley revealed that it contained the SHR SA allele as shown in figure 3.15 B. These results clearly show that exon repetition in SA and COT do not co-segregate. Hence, exon repetition in SA and COT cannot be determined by the same trans-acting factor. Interestingly, out of the sixteen F2 animals tested, twelve had exon repetition of COT. The remaining four did not, which was indicative of a dominant monogenic trait.

3.2.3. Exon repetition in SA is allele-specific

As mentioned before, exon repetition in SA co-segregates with the chromosomal region around the WKY SA allele. This can be explained in two ways:
Figure 3.15. Characterization of the SA gene and its expression in the Sprague-Dawley rat. (A) SA mRNA was analyzed in two Sprague-Dawley rats by RT-PCR with primers SA E1F and SA E2R. WKY (W) and SHR (S) were included in the experiment as positive and negative control respectively. From the two Sprague-Dawley rats, indicated as SD1 and SD2, only the PCR product corresponding to the normal SA mRNA was amplified. The exon organisation of the PCR products is shown by the white boxes. (B) Determination of the SA alleles in Sprague-Dawley. SA intron 1 was amplified by PCR from the genomic DNA extracted from one Sprague-Dawley animal with primers SA E1F and SA E2R. SHR (S) and WKY (W) were also included in the experiment and the result is shown on the left. The PCR product amplified from Sprague-Dawley is identical to that amplified from SHR, suggesting the presence of the LINE element in intron 1 of the SA gene in Sprague-Dawley. In contrast, the PCR product obtained from WKY was smaller in size, consistent with the absence of the LINE element in intron 1 of the SA gene in this strain. The data were confirmed with primers SA LINE F and SA E2R that were specifically designed to amplify a PCR product from the SA allele containing the LINE element in intron 1 of SA. These primers amplified the expected PCR product from SHR and Sprague-Dawley, but not from WKY as expected, and the result is shown in the right panel.
Figure 3.16 Exon repetition in SA is allele-specific in heterozygotes. Exon repetition in SA was evaluated by RT-PCR in the parental rat strain WKY (W) and SHR(S), and in four F2 animals, two of which were homozygote for the WKY SA allele (animals 1 and 2, the same as in figure 3.13); the other two were heterozygotes, having one copy of the SA gene from each parent (animals 17 and 18). (A) RT-PCR with primers SA E1F and SA E2R. The efficiency of exon repetition in the two heterozygotes is lower compared to that in the WKY SA homozygotes and the parental WKY. (B) RT-PCR with primer SA 2-2 junction (which is specifically designed to amplify from the SA cDNA isoform that contains a tandem repetition of exon 2) and primer SA E4R. The corresponding PCR product was amplified from all rats except the parental SHR, as expected. (C) RT-PCR with primers SA 4-2 junction and SA E4R, to specifically amplify SA cDNA molecules in which exons 2, 3 and 4 are repeated. From SHR, these primers failed to amplify a PCR product as expected. (D) An alignment of the first four exons of the WKY and SHR SA cDNA showing the presence of a SNP (single nucleotide polymorphism) in exon 3 (WKY→G, SHR→A). (E) DNA sequence chromatographs of WKY and SHR PCR products, and a mixture of the two in different proportions. (F) Sequence analysis of RT-PCR products from the two heterozygotes. The amplicons corresponding to the normal SA mRNA (top), exon 2 repetition (middle) and repetition of exons 2, 3 and 4 (bottom) were produced in separate reactions, purified from agarose gel and analyzed by sequencing. In the two heterozygotes, the expression of normal mRNA from the SHR SA allele (green signal) is much stronger compared to the level from the WKY SA allele (black signal). The parental WKY and SHR are also shown as control. Exon repetition products contained only the WKY allele. (G) Sequence analysis of the RT-PCR products of the two heterozygote rats corresponding to the first repeat of exons 2-3-4. Amplification was carried out with primers SA E1F and SA 4-2 junction reverse. This set of primers amplifies the first set of repeated exons from the mRNA isoform that contains the tandem repeat of exons 2-3-4. Only the G nucleotide (WKY SA allele) is detected.
A

B

C

D

E

F

G
1) Strain-specificity is due to differences between the WKY and SHR SA alleles; only the WKY SA allele produces mRNA isoforms with repeated exons.

2) Strain-specificity is due to the presence of a gene that co-segregates with the WKY SA allele. The \textit{trans}-acting factor encoded by this hypothetical gene acts specifically upon SA pre-mRNA and induces it to produce isoforms with repeated exons, but it has no effect on COT pre-mRNA.

In order to distinguish between these two possibilities, exon repetition in SA was investigated in two rats from the F2 population (animals 17 and 18) that were heterozygote at the SA locus, having one SHR SA allele and one WKY SA allele. In these animals, transcription and pre-mRNA processing of SA take place in the same environment, which is the nucleus of the proximal tubule cells of the same animal. Hence, if strain-specificity is due to the presence of a gene co-segregating with the WKY SA allele that encodes a diffusible \textit{trans}-acting factor responsible for exon repetition in SA pre-mRNA, in an animal that is heterozygote, both alleles would be expected to produce mRNA isoforms containing repeated exons. On the other hand, if strain-specificity of exon repetition in SA is due to differences between the WKY and SHR SA alleles, then it is expected that the SA mRNA isoforms with repeated exons would be derived exclusively from the WKY SA allele in heterozygote animals.

First of all, the efficiency of exon repetition was compared between two WKY SA homozygotes and two SA heterozygotes. SA mRNA was amplified by RT-PCR between exon 1 and exon 2. The parental strains WKY and SHR were also included as positive and negative control respectively, and the result is shown in figure 3.16 A. The efficiency of exon repetition was much higher in the two WKY SA homozygotes, compared to the heterozygotes. The lower level of exon repetition in the two heterozygotes could be explained in two ways:

1) Only the WKY SA allele produces SA mRNA isoforms containing repeated exons and since the heterozygotes possess only one WKY SA allele, the efficiency of exon repetition would be lower in heterozygotes compared to WKY SA homozygotes.

2) Exon repetition is determined by a gene that co-segregates with the WKY SA allele, and the concentration of the \textit{trans}-acting factor encoded by this gene is lower in the heterozygote because there is only one copy of the gene instead of the two copies of WKY in WKY SA homozygotes. If we assume that the level of exon repetition in SA correlates with the concentration of this \textit{trans}-acting factor, then the difference in the efficiency of exon repetition cannot discriminate between hypothesis 1 and 2.
Figure 3.17. Primers used in the RT-PCR reactions for the amplification of the different SA mRNA isoforms. Primer SA 2-2 junction allows the selective amplification of the SA mRNA isoform with the tandem repeat of exon 2. Primer SA 4-2 junction allows the selective amplification of the downstream repeat of exons 2-3-4 in the SA mRNA isoform with the tandem repeat of exons 2-3-4. Finally, primer SA 4-2 junction reverse specifically amplifies the upstream repeat of exons 2-3-4.
In principle, if there was a mutation that differentiates the SA mRNA of SHR from that of WKY, it would be possible to determine the proportion of each mRNA isoform that derives from the WKY and SHR SA alleles in heterozygotes. In order to determine if such mutation existed, SA mRNA was amplified by RT-PCR from SHR and WKY, and sequenced. By comparing the two sequences, a SNP (single nucleotide polymorphism) was identified at the end of exon 3 (WKY \( \rightarrow \) G; SHR \( \rightarrow \) A) and is shown in figure 3.16 D. In principle, by sequencing the RT-PCR products, it could be possible to determine the proportion of SA mRNA that was derived from each allele in heterozygotes. First of all, the suitability of this approach was tested. SA mRNA was amplified between exon 1 and exon 4 from WKY and SHR with primers SA E1F and SA E4R. The concentration of the two PCR products was estimated by loading several dilutions on agarose gel. The two PCR products were then mixed in different proportions and sequenced. The two original PCR products (WKY and SHR) were also sequenced separately and the result is shown in figure 3.16 E. As the proportion of PCR product from SHR was increased, there was a corresponding increase of the signal from the A nucleotide (SHR SA allele) and at the same time the signal from the G nucleotide (WKY SA allele) decreased. Hence, it was concluded that sequencing RT-PCR products could give a good estimate of the proportion of SA mRNA derived from each of the two alleles in heterozygotes.

First of all, SA mRNA from two heterozygotes and the parental lines WKY and SHR was amplified between exon 1 and exon 4. All the primers used in the PCR reactions are shown in figure 3.17. After separation of the DNA fragments by gel electrophoresis, the amplicon corresponding to the normal SA mRNA was eluted from agarose gel and sequenced. The DNA sequence chromatographs are shown in figure 3.16 F and indicated as "Normal mRNA". The DNA sequence clearly showed that normal SA mRNA in heterozygotes had been derived mainly from the SHR SA allele. In contrast, the contribution to normal SA mRNA from the WKY SA allele was rather small. As mentioned before, the expression level of SA is much higher in SHR than WKY. The result of figure 3.16 F relative to the normal mRNA demonstrates that the higher expression level of SA mRNA in SHR compared to WKY is caused by cis-acting elements, rather than strain-specific differences in trans-acting factors that affect the expression level of SA mRNA. If this were the case, then the expression level of the WKY SA allele would be similar to that of the SHR SA allele in heterozygotes. Allele-specific variations in the level of gene expression have been observed in several human genes (Yan H 2002) as well as in mouse genes (Cowles CR 2002) and in rat (Lin Z 2002).
Next, the SA mRNA isoforms containing repeated exons were analysed in a similar way. First of all, the SA mRNA isoform with the tandem repeat of exon 2 was amplified by RT-PCR using primers SA 2-2 junction and SA E4R and the result of this amplification is shown in figure 3.16 B. Primer SA 2-2 junction was specifically designed to amplify a PCR product only from SA mRNA that contains a tandem repeat of exon 2. In fact, this primer failed to amplify a PCR product from SHR (see figure 3.16B, lane S), indicating that the amplification was indeed specific. The PCR product amplified from WKY and the two heterozygotes was eluted from the agarose gel and sequenced and the DNA sequence chromatographs are shown in figure 3.16 F and indicated as “Repetition exon 2”. From WKY, only the signal corresponding to the WKY allele (G nucleotide) was detected, as expected. Interestingly, the same result was obtained from the two heterozygotes. This demonstrated that in these animals, the SA mRNA isoform with a tandem repeat of exon 2 was derived solely from the WKY SA allele. The simultaneous repetition of exons 2, 3 and 4 was investigated by RT-PCR with primers SA 4-2 junction and SA E4R. Primer SA 4-2 junction specifically amplifies a PCR product only from the SA mRNA isoform with a tandem repeat of exons 2-3-4. Amplification with this primer did not yield any PCR product from SHR (see figure 3.16 C, lane S), demonstrating that the amplification was specific. The PCR product from WKY and the two heterozygotes was eluted from the agarose gel and sequenced and the results are shown in figure 3.16 F indicated as “Repetition exons 2-3-4”. Again, from WKY, only the signal from the G nucleotide was detected, as expected. Interestingly, the sequence chromatograph relative to the PCR product amplified from the two heterozygotes, was identical to that of WKY, demonstrating that in heterozygotes, the mRNA isoform with the tandem repeat of exon 2-3-4 was derived exclusively from the WKY SA allele. The SA 4-2 junction primer specifically amplifies the second set of repeated exons (exons 2, 3 and 4). Hence, strictly speaking, the results so far have shown that the second set of repeated exons was derived from the WKY SA allele. In order to investigate whether the same was also true for the first set of repeated exons, an RT-PCR reaction from the same heterozygote rats was performed with primer SA 4-2 junction reverse and SA E1F. This set of primers amplifies the first set of repeated exons. The sequence chromatographs relative to these RT-PCR products are shown in figure 3.16 G. Again, only the nucleotide corresponding to the WKY SA allele was detected, thus demonstrating that also the first set of repeated exons in the SA mRNA isoform with the tandem repeat of exons 2-3-4 was also derived from the WKY SA allele. All these data show that in heterozygotes, even though both alleles are expressed, the SA mRNA isoforms with repeated exons were derived solely from the
WKY SA allele. If the SHR SA allele was generating SA mRNA molecules containing repeated exons, the sequence would have shown not only a signal from the G nucleotide (WKY allele), but also one from the A nucleotide (SHR allele). This is the first example of allele-specific differences in gene expression that involve fundamentally different reactions.

3.2.4. Exon repetition in COT is allele-specific

As seen before, exon repetition of COT segregated in the F2 population obtained by crossing F1 rats (WKY X SHR). Out of sixteen animals tested, twelve had exon repetition, while the remaining four did not. This corresponded to a 3:1 ratio, which is typical of a Mendelian segregation obtained when two heterozygotes are crossed and the trait is dominant and monogenic. Hence, exon repetition in COT seems to be controlled by a single gene. One possibility was that WKY and SHR possessed two different COT alleles and only the WKY COT allele was capable of producing the COT mRNA isoform with the tandem repeat of exon 2, whereas the SHR COT allele could only produce normal COT mRNA. Alternatively, the strain-specificity of COT exon repetition, could be caused by another gene, possibly encoding a trans-acting factor involved in exon repetition, that is present in WKY but absent in SHR. In the Sprague-Dawley rat, the presence of an exonic splicing enhancer (ESE) in exon 2 of COT was shown to be required for exon repetition in vitro and with COT minigenes transfected into Hela cells. One obvious explanation for the lack of exon repetition of COT in SHR would be the mutation of the ESE of exon 2 in the SHR COT allele. In order to address this question, COT mRNA from WKY and SHR was amplified by RT-PCR between exon 1 and exon 7 and sequenced. An alignment of the two DNA sequences revealed that they were identical, and both contain the putative exonic splicing enhancer that was originally found in Sprague-Dawley. Thus, the absence of exon repetition of COT in SHR was not due to the mutation of the exonic splicing enhancer in exon 2 of the COT gene in the SHR strain.

To determine whether WKY and SHR had different COT alleles, the entire COT cDNA from WKY and SHR was amplified by RT-PCR and sequenced. While the coding region was identical between the two strains, several SNPs (single nucleotide polymorphisms) and a short deletion of 2 nucleotides were found in the 3'UTR. Therefore, WKY and SHR did indeed possess two different COT alleles. The alignment of the WKY and SHR COT 3' UTR is shown in figure 3.18.
Figure 3.18. Alignment of the DNA sequence of the COT 3' UTR from WKY and SHR. The differences are shown bold. The TAG codon shown with a grey background is the translational stop codon. The six nucleotide sequence (CTTAAG) underlined is the recognition sequence for the restriction endonuclease AflIII, which is present in WKY but not in SHR (mutation C→T).
Figure 3.19. Exon repetition in COT is allele-specific. (A) Segregation of the WKY and SHR COT alleles in the F2 population of figure 3.13. The animals are numbered as in Fig. 3.13. The results are shown also for the parental lines (S, W) and an F1 heterozygote (H). A SNP (single nucleotide polymorphism) in the COT 3'UTR between the WKY and SHR COT alleles (see figure 3.17) was utilized to analyse the distribution of the COT alleles in the F2 population. The point mutation creates a recognition sequence for the restriction endonuclease Afl II in the WKY COT allele. Thus, the PCR product derived from the WKY COT allele is cleaved into two fragments (indicated by the two red arrows), whereas the corresponding fragment from the SHR COT allele is not cleaved (indicated by the black arrow). The mRNA was amplified by RT-PCR with primers COT E16F and COT E17R, which amplify from total COT mRNA (normal mRNA and the mRNA isoform containing the tandem repetition of exon 2). An aliquot of the PCR reaction was extensively digested with the restriction endonuclease Afl II and the DNA fragments were separated on agarose gel and stained with ethidium bromide. (B) Analysis of COT mRNA in two COT heterozygote rats (animals 7 and 13 of panel A). Total COT mRNA (T) was amplified between exons 1 and 17; it was then re-amplified by PCR between exons 16 and 17, and analyzed by digestion with Afl II and agarose gel electrophoresis. This procedure is similar to that in (A), with prior amplification between exon 1 and 17. Exon repetition COT mRNA (ER) was amplified between the exon 2-2 junction and exon 17 with primers COT 2-2 junction and COT E17R and then re-amplified and analyzed as for T. Only the WKY COT allele was detected. (C) Analysis of products of amplification between exons 16 and 17 by sequencing. As in (B), the PCR products in T were derived from re-amplification from PCR products derived initially by RT-PCR between exons 1-17, whereas those labelled ER originated from RT-PCR with primers specific for the tandem repetition of exon 2. As with the analysis by restriction enzyme digestion (B), the sequences show that total mRNA contained mRNA from both alleles (A and G at nt 158 of exon 17), whereas repetition of exon 2 was restricted to the WKY allele.
Since the sequence of exons 1-14 in COT was the same in the two strains, it could be concluded that the strain-specificity of exon repetition in COT was not determined by differences in the sequence of the repeated exon, nor the adjacent exons in the two strains.

To study the segregation of the COT alleles in the F2 population, one particular SNP was chosen because it allowed discrimination of the two alleles by restriction digestion with the restriction endonuclease AfII (see figure 3.18). COT mRNA was amplified by RT-PCR between exon 16 and exon 17 with primers COT E16F and COT E17R. The use of primer COT E16F avoided amplification from genomic DNA, because there is a large intron between exon 16 and exon 17 (3'UTR). The parental WKY and SHR, and an F1 heterozygote were used as control. An aliquot of the PCR reaction was extensively digested with the restriction endonuclease AfII, and DNA fragments were separated on agarose gel and the result is shown in figure 3.19 A (lanes S, W and H). The DNA fragment from the parental SHR was not digested by AfII as expected, since the SNP in the 3'UTR of the SHR COT allele abrogates the AfII restriction sequence. In contrast, the PCR product obtained from WKY was completely digested by AfII as expected. In the F1 heterozygote, the two alleles were expressed at a comparable level, since the amount of undigested and digested PCR product was quite similar. This differed from what was found in the SA gene. In fact, in that case, the SHR SA allele was over-expressed compared to the WKY SA allele (see figure 3.16 F, “normal mRNA”). All the F2 rats in figure 3.14 were genotyped at the COT gene using the same procedure described before (RT-PCR and restriction enzyme analysis) and the result is shown in figure 3.19 A (animals 1-16). These experiment showed that animals 4, 12, 14 and 15, which did not exhibit exon repetition in COT (see figure 3.14 B and 3.14 C) were SHR COT homozygote. In contrast, the remaining 12 animals that exhibited exon repetition in COT were either WKY COT homozygote (animals 2, 5, and 6), or heterozygote (animals 1, 3, 7, 8-11, 13 and 16). Since the genotype was inferred from COT mRNA, it is possible that the animals that appeared homozygote could actually be heterozygotes that are expressing only one of the two alleles. To investigate this possibility, all the homozygote rats of figure 3.19 A were genotyped again, but this time genomic DNA was used as template. The PCR reaction was carried out with primers COT E17F and COT E17R. An aliquot of the PCR reaction was extensively digested with the restriction endonuclease AfII, and DNA fragments were separated by electrophoresis on agarose gel and the result is shown in figure 3.20. The data are in agreement with those shown in figure 3.19 A. Therefore, all the rats that appeared to be homozygote in figure 3.19 were indeed homozygote.
Figure 3.20. Confirmation of homozygosity in the animals that appeared homozygote in figure 3.18. The 3' UTR of COT was amplified from genomic DNA from the rats that appeared WKY COT homozygote (animals 2, 5 and 6) or SHR homozygote (animals 4, 12, 14 and 15) in figure 3.18 A. Primers COT E17F and COT E17R were used for DNA amplification. An aliquot of the PCR reaction was extensively digested with the restriction endonuclease AfIII. DNA fragments were separated on a 1% agarose gel and visualised by ethidium bromide staining. The undigested PCR fragment derived from the SHR COT allele is indicated by the black arrow whereas the two DNA fragments derived from cleavage of the WKY PCR product are indicated by the dashed arrows.
Figure 3.21. Diagrammatic representation of the primers used to selectively amplify the different COT mRNA isoforms. Amplification is carried out by two rounds of PCR. The second round of PCR is semi-nested, since it uses an internal primer (COT E16R) and one primer that was used in the first amplification (COT E17R). (A) Amplification of total COT mRNA. The first PCR reaction is performed with primers COT E1F and COT E17R which amplify from both the normal COT mRNA isoform and the one that contains the repetition of exon 2. An aliquot of the first PCR reaction is then used as template for the second amplification which is done with primers COT E16F and COT E17R. With these two sets of primers, total COT mRNA is amplified. (B) Selective amplification of the mRNA isoform containing the tandem repetition of exon 2. The first PCR reaction is performed with primer COT 2-2 junction (it amplifies only from the isoform that contains the tandem repetition of exon 2) and primer COT E17R. An aliquot of this reaction is then used as template for a second PCR amplification with primers COT E16F and COT E17R.
Since all the rats possessing at least one WKY COT allele manifest exon repetition of COT, it can be concluded that exon repetition of COT co-segregates with the chromosomal region around the WKY COT gene. As discussed before for the SA gene, exon repetition in COT might be caused by a gene encoding a trans-acting factor that causes exon repetition in COT and that co-segregates with the WKY COT allele. To address this question, exon repetition in COT mRNA was studied in two heterozygotes from the F2 population (rats number 7 and 13 of figure 3.19 A). COT mRNA was amplified by RT-PCR between exon 1 and exon 17 with primers COT E1F and COT E17R. This set of primers amplifies total COT mRNA (with and without exon repetition). A second RT-PCR reaction was performed with primers COT 2-2 junction and COT E17R. Primer COT 2-2 junction was specifically designed to amplify COT cDNA containing the tandem repeat of exon 2. Due to the large size of the PCR products and also to the high annealing temperature that was used to avoid non-specific amplification with primer COT 2-2 junction, the amount of PCR product obtained, even though detectable on agarose gel, was insufficient for further analysis. In order to obtain a suitable amount of PCR product, an aliquot of the previous PCR reaction was used as template for a second PCR reaction (semi-nested) with primers COT E16F and COT E17R. The strategy is outlined in figure 3.21. An aliquot of the PCR reaction was extensively digested with the restriction endonuclease AflII and DNA fragments were separated on agarose gel. The result is shown in figure 3.19 B. When primer COT E1F was used in the first PCR reaction, both the WKY and SHR COT alleles were detected (see lanes 7 T and 13 T). This demonstrated that both the WKY and SHR COT alleles were expressed in the two heterozygotes. The use of primer COT 2-2 junction in the first PCR reaction gave a completely different result. In this case, only the WKY COT allele was detected (lanes 7 ER and 13 ER). These data were confirmed by sequencing the PCR products obtained from the second PCR reaction and the result is shown in figure 3.19 C. When total COT mRNA was amplified using primer COT E1F in the first PCR reaction, both alleles were detected (see “Total mRNA”). In contrast, when primer COT 2-2 junction was used in the first PCR amplification, in a DNA sequence chromatograph of the PCR product from the second amplification only the nucleotide corresponding to the WKY COT allele could be seen (see “Repetition exon 2” in figure 3.19 C). These data clearly demonstrate that exon repetition in COT is allelespecific: the COT mRNA isoform with the tandem repeat of exon 2 is derived solely from the WKY COT allele but not from the SHR COT allele. The same results were obtained from the F1 heterozygote (WKY X SHR, named H) of figure 3.19 A (data not shown).
Figure 3.22. Diagrammatic representation of the primers used to amplify the COT mRNA isoforms. (A) The set of primers COT E1F and COT E7R amplifies a PCR product of different size from each COT mRNA isoform. (B) The set of primers COT E2F and COT E2R amplifies a PCR product of different size from the COT mRNA isoforms in which either exon 2 alone, or exons 2 and 3 are tandemly repeated. (C) The set of primers COT E3F and COT E2R amplifies a PCR product exclusively from the COT mRNA isoform in which exons 2 and 3 are tandemly repeated.
Figure 3.23. Comparison of exon repetition of COT in WKY and Sprague-Dawley by RT-PCR from kidney RNA. (A) Amplification of all COT mRNA isoforms with flanking primers COT E1F and COT E7R. (B) Amplification of repetition of exon 2 and exons 2 and 3 with primers COT E2F and COT E2R. (C) Specific amplification from the COT mRNA isoform with a tandem repeat of exons 2 and 3. This RT-PCR reaction was performed with primers COT E3F and COT E2R. The exon composition of each amplicon is shown by the black boxes and white numbers. W is WKY, SD is Sprague-Dawley and Bl is the blank reaction.
3.2.5. The pattern of exon repetition is allele-specific in COT

As mentioned before, in the Sprague-Dawley rat, three COT mRNA isoforms were identified. One corresponded to the normal COT mRNA, whereas the other two contained a tandem repeat of either exon 2 alone or exons 2 and 3 (Caudevilla C 1998). Interestingly, in WKY, while the mRNA isoform containing the tandem repeat of exon 2 was very abundant, repetition of exons 2 and 3 could not be detected (see figure 3.7). The failure to detect the COT mRNA isoform with the tandem repeat of exons 2 and 3 in WKY could be due to several reasons. The first explanation was that this isoform was absent in WKY. Other possibilities such as a low abundance of this particular isoform in WKY or a problem with the PCR conditions could not be ruled out. To investigate this, exon repetition of COT was evaluated in WKY by RT-PCR with three different sets of primers, which are shown in figure 3.22. Kidney RNA from Sprague-Dawley was included as a control. First of all, primers COT E1F and COT E7R were used. These primers are supposed to amplify all three COT mRNA isoforms. In Sprague-Dawley, all three PCR products were amplified (normal COT mRNA, repetition of exon 2 and also the simultaneous repetition of exons 2 and 3) as shown in figure 3.23 A lane SD. In contrast, from WKY only the PCR products corresponding to the normal COT mRNA and the tandem repeat of exons 2 were amplified (see lane W in figure 3.23 A). This result suggested that the PCR conditions allowed efficient amplification from the COT mRNA isoform with the tandem repeat of exons 2 and 3. In order to determine whether this isoform was absent or just present at a low level in WKY, a second RT-PCR reaction was performed with primers COT E2F and COT E2R which were specifically designed to amplify a PCR product from the COT mRNA isoforms containing exon repetition (repetition of exon 2 alone or simultaneous repetition of exons 2 and 3). From Sprague-Dawley RNA two PCR products were amplified. One corresponded to the repetition of exon 2 alone, whereas the other one contained a repetition of exons 2 and 3 as shown in figure 3.23 B lane SD. In contrast, from WKY only a single PCR product was amplified that corresponded to the repetition of exon 2 (see lane W in figure 3.23 B). Finally, a third RT-PCR was carried out with primers COT E3F and COT E2R, which can only amplify a PCR product from the COT mRNA isoform with the tandem repeat of exons 2 and 3. These primers amplified a PCR product from Sprague-Dawley but not from WKY, thus confirming the absence of repetition of exons 2 and 3 in WKY as shown in figure 3.23 C (compare lane SD and W).

This difference in the pattern of exon repetition between WKY and Sprague-Dawley could be caused by strain-specific differences in trans-acting factors, or by the presence of different COT alleles in the two strains. For example, the WKY COT allele
Figure 3.24. Exon repetition of COT in heterozygotes (WKY X SD). The parental WKY and Sprague-Dawley were also included in the experiment. COT mRNA extracted from kidney was amplified by RT-PCR with three different sets of primers. In each panel, W and SD are the parental WKY and Sprague-Dawley respectively, whereas 1 and 2 are two F1 heterozygotes obtained by crossing a WKY male with a Sprague-Dawley female. (A) Simultaneous amplification of all three COT mRNA isoforms. The RT-PCR was performed with primers COT E1F and COT E7R. (B) Amplification of PCR products from the mRNA isoforms that contain repeated exons. The RT-PCR was performed with primers COT E2F and COT E2R. An amplicon of different size was obtained from the mRNA isoform with the repeat of exon 2 alone or exons 2 and 3, as expected. (C) Specific amplification from the COT mRNA isoform with a tandem repeat of exons 2 and 3. The RT-PCR reaction was performed with primers COT E3F and COT E2R.
might be unable to produce the mRNA isoform with the exon 2-3 repetition. This question could be answered by analysing COT mRNA in a heterozygote (WKY X SD), providing there was at least one SNP between the two strains in one exon of the COT gene. In order to identify a SNP, COT mRNA from Sprague-Dawley and WKY was amplified by RT-PCR and sequenced. The DNA sequence of the Sprague-Dawley and WKY COT 3’ UTR differed at many positions (see figure 3.18) and this makes it possible to determine from which COT allele a particular mRNA had been transcribed.

To obtain heterozygotes carrying one COT allele from Sprague-Dawley and one from WKY, a Sprague-Dawley female was crossed with a WKY male and from the litter, two heterozygotes were randomly chosen and utilised in the following experiments. RNA was extracted from kidney and COT mRNA was amplified by RT-PCR from the two heterozygotes and the parentals WKY and Sprague-Dawley. First of all, the efficiency of exon repetition of COT in the two heterozygotes was compared to that of the parental strains. For this purpose, an RT-PCR was carried out with primers COT E1F and COT E7R, which amplify between exon 1 and exon 7. The result is shown in figure 3.24 A. The ratio between the different COT mRNA isoforms in the two heterozygotes was approximately an average of the ratio in the parental strains. In fact, the isoform with the tandem repeat of exons 2 and 3 seemed less abundant in the two heterozygotes compared to the parental Sprague-Dawley. Moreover, the isoform with the repetition of exon 2 was more abundant in the two heterozygotes than in Sprague-Dawley. These observations were confirmed with a second reaction performed with primers COT E2F and COT E2R, as shown in figure 3.24 B. Finally, a third RT-PCR reaction carried out with primers COT E3F and COT E2R (they can only amplify a PCR product from the COT mRNA isoform containing the tandem repeat of exons 2 and 3) confirmed the presence of the COT mRNA isoform with the tandem repeat of exons 2 and 3 in the two heterozygotes and in the parental Sprague-Dawley, and its absence in WKY (see figure 3.24 C).

In order to determine whether in heterozygotes the COT mRNA isoform containing the tandem repeat of exons 2 and 3 was derived exclusively from the Sprague-Dawley COT allele, or also from the WKY COT allele, a similar strategy to that employed in figure 3.19 B was utilised. COT mRNA was amplified by RT-PCR with three different sets of primers, which are shown in figure 3.25. The first set of primers (COT E1F and COT E17R) was designed to amplify a PCR product from all three COT mRNA isoforms. In contrast, the second set of primers, COT 2-2 junction and COT E17R, was specifically designed to amplify a PCR product only from the COT mRNA isoform with the tandem repeat of exon 2. Finally, a third set of primers, COT 3-2 junction and COT E17R, was specifically
Figure 3.25. Diagrammatic representation of the primers used in the two-step amplification of the different COT mRNA isoforms in the two heterozygotes derived from a cross between WKY and SD. In the first PCR, the antisense primer is always COT E17R. The sense primer is different (COT E1F or COT 2-2 junction, or COT 3-2 junction) and determines which mRNA isoform/s will be amplified. The second amplification is always performed with primers COT E16F and COT E17R. (A) Primers COT E1F and COT E17R amplify a PCR product from all three COT mRNA isoforms. An aliquot of this PCR reaction can be re-amplified with primers COT E16F and COT E17R. With this approach total COT mRNA is analysed. (B) Primer COT 2-2 junction selectively amplifies a PCR product only from the COT mRNA isoform in which exon 2 is tandemly repeated. An aliquot of this PCR reaction can be re-amplified with primers COT E16F and COT E17R. In this way the COT mRNA isoform containing a tandem repeat of exon 2 is analysed. (C) Primer COT 3-2 junction amplifies a PCR product only from the COT mRNA isoform in which exons 2 and 3 are tandemly repeated. An aliquot of this PCR product can be re-amplified with primers COT E16F and COT E17R. This strategy allows to specifically analyse the COT mRNA with the tandem repeat of exons 2 and 3 can be studied.
Figure 3.26. The pattern of exon repetition in COT is allele-specific. COT mRNA isoforms were selectively amplified by RT-PCR from two heterozygotes (WKY X Sprague-Dawley). Amplification was done in two steps. In the first reaction, the forward primer used determined which mRNA isoform was amplified. To analyse total COT mRNA (T), primer COT E1F was used. The COT mRNA isoform with a tandem repeat of exon 2 was amplified with primer COT 2-2 junction. Finally, the isoform containing the tandem repeat of exons 2 and 3 was selectively amplified with primer COT 3-2 junction. The reverse primer was always COT E17R. In the second PCR amplification, primers COT E16F and COT E17R were used and an aliquot from the first amplification reaction was used as template. The strategy was illustrated in figure 3.24. A SNP in exon 17 allows the WKY COT allele to be discriminated from the Sprague-Dawley COT allele. This point mutation creates a sequence that is cleaved by the restriction endonuclease Afl II. (A) Agarose gel analysis of the PCR products from the second amplification, after they had been extensively digested with the restriction endonuclease Afl II. The PCR product derived by amplification from Sprague-Dawley COT mRNA is indicated by the black arrow. In contrast, the PCR product derived from WKY COT mRNA is cleaved into two DNA fragments which are indicated by the red arrows. (B) DNA sequence chromatographs of the PCR products derived from the second amplification. The WKY and Sprague-Dawley COT alleles are represented by the A (in green) and G (in black) nucleotides, respectively.
A

1 2
T 2-2 2-3 T 2-2 2-3

SD
WKY

B

Total COT mRNA

Repetition exon 2

Repetition exons 2-3
designed to amplify a PCR product only from the COT mRNA isoform with the tandem repeat of exons 2 and 3. PCR reactions with the three sets of primers were performed in separate test tubes. Due to the low amount of amplicon obtained from these reactions, an aliquot from this first amplification was used as template for a second PCR with primers COT E16F and COT E17R. An aliquot of this PCR reaction was extensively digested with the restriction endonuclease AfllI. DNA fragments were then separated on agarose gel, and the result is shown in figure 3.26 A. The analysis of total COT mRNA (lanes labelled T) revealed that both alleles were expressed at comparable levels (the Sprague-Dawley and WKY COT allele correspond to the undigested and digested PCR products respectively). In contrast, when the first PCR amplification had been carried out with the primer that specifically amplified the repetition of exon 2 (primer COT 2-2 junction), the amount of digested PCR product (WKY allele) was predominant (compare lanes T with lanes 2-2). Finally, when primer COT 3-2 junction was used in the first round of amplification, only the undigested PCR product (Sprague-Dawley allele) was present. This demonstrated that in the two heterozygotes, the COT mRNA isoform with the simultaneous repeat of exons 2 and 3 had been derived solely from the Sprague-Dawley COT allele (see lanes 2-3 in figure 3.26 A). These data were confirmed by sequencing an aliquot of the second PCR reaction (primers COT E16F and COT E17R) from the three different experiments and the result is shown in figure 3.26 B. When the first round of PCR had been carried out with primer COT E1F (which amplifies all three COT mRNA isoforms), both alleles were detected in approximately equal amounts as shown in figure 3.26 B “Total COT mRNA” (WKY→ A, Sprague-Dawley → G). In contrast, the WKY allele was predominant when the primer specific for the repetition of exon 2 (COT 2-2 junction) was used in the first reaction (see figure 3.26 B “Repetition exon 2”). Finally, when primer COT 3-2 junction was used in the first PCR, only the G nucleotide (Sprague-Dawley COT allele) was detected, as shown in figure 3.26 B “Repetition exons 2 and 3”. All these data taken together clearly showed that in heterozygotes, even though both the WKY and Sprague-Dawley COT alleles were expressed, the COT mRNA isoform containing the simultaneous repetition of exons 2 and 3 had been derived solely from the Sprague-Dawley allele. It can be concluded that the lack of the COT mRNA isoform with the tandem repeat of exons 2 and 3 in the WKY strain results from the inability of the WKY COT allele to produce this isoform, and not from strain-specific differences in trans-acting factors between WKY and Sprague-Dawley.
Figure 3.27. Organisation of the 5’ end of the COT gene. Genomic DNA was extracted from a tail biopsy from SHR (S), WKY (W) and Sprague-Dawley (SD), and extensively reacted with either EcoRI or HindIII. (A) Restriction map of a portion of the COT gene comprised between the promoter (represented as a blue rectangle) and exon 3. The exons are numbered from 1 to 3 and represented by the vertical lines. The probe corresponding to the promoter region is represented by the blue rectangle above the promoter whereas the probe consisting of exon 2 of COT is shown as a green rectangle. The positions of the EcoRI recognition sequences are indicated by the letter E whereas the recognition sequences of HindIII are represented by the letter H. (B) Ethidium bromide stained agarose gel of restricted genomic DNA after electrophoresis. M1 and M2 are molecular markers from MBI Fermentas (λ DNA restricted with HindIII and Gene Ruler 1Kb, respectively). (C) Southern blot hybridisation with a probe corresponding to the COT promoter (on the left) or COT exon 2 (on the right). When genomic DNA was digested with EcoRI, two hybridising bands were detected by both probes in all three rats. In contrast, when genomic DNA was reacted with HindIII, both probes detected a single hybridising band in all three rats.
3.2.6. Genomic organisation of the COT gene

As mentioned in the introduction of this chapter, exon duplication in a gene could account for the presence of mRNA isoforms with repeated exons. In the Sprague-Dawley rat, the duplication of exon 2 in the COT gene had been ruled out by southern blotting (Caudevilla C 1998). In principle, a duplication of exon 2 in the WKY COT allele could explain the presence of the COT mRNA isoform with the tandem repeat of exon 2. In order to investigate this possibility, a southern blot experiment was carried out. Genomic DNA was extracted from a tail biopsy from WKY, SHR, and Sprague-Dawley. About 10 μg of genomic DNA were reacted with the restriction endonucleases EcoRI and HindIII. DNA fragments were then separated on a 0.7% agarose gel and blotted onto a hybond membrane N+. A map of the COT gene with the position of the restriction sites that are cleaved by EcoRI and HindIII is shown in figure 3.27 A. The DNA sequence used to create this restriction map was retrieved from http://www.ncbi.nlm.nih.gov/genome/seq/RnBlast.html. The membrane was then probed with a fragment corresponding to COT exon 2 (which was amplified with primers COT I1 F and COT I2 R) and the result of the hybridisation is shown in figure 3.27 C and indicated as exon 2. From genomic DNA that had been reacted with the restriction endonuclease HindIII, a single fragment of approximately the expected size was detected by the probe from all three rats. This was in agreement with the previously published data (Caudevilla C 1998). Strangely, from genomic DNA that had been reacted with the restriction endonuclease EcoRI, two bands of different size were detected by the probe. This was in disagreement with the published results relative to Sprague-Dawley in which a single band had been detected after digestion with EcoRI. The difference between the two experiments was that a different tissue and a different animal were used. In fact, in the Caudevilla paper, the southern blot had been performed with genomic DNA extracted from liver, instead of tail. Several reasons could explain this discrepancy:

1) There could be variability between different rats. Some might have two different COT alleles and in one of them, the recognition sequence for EcoRI might be abrogated by a point mutation.

2) The EcoRI restriction site in intron 1 could be blocked by methylation. If this were true, the sequence might be methylated in the tissues of tail but not in liver. Furthermore, methylation in tail would not be complete, since a significant fraction of DNA was completely digested.

3) The restriction digestion with EcoRI might have been incomplete in all three samples. However, this is quite unlikely. In fact, on the ethidium bromide stained
Figure 3.28. Restriction enzyme analysis with Eco RI of a PCR product corresponding to intron 1 of COT from SHR (S), WKY (W) and Sprague-Dawley (SD). Amplification of COT intron 1 was performed with primers COT E1F and COT E2R. (A) Agarose gel picture of the PCR products. 5 μl of PCR reaction were separated on a 1% agarose gel. The template used was genomic DNA from SHR, WKY or Sprague-Dawley (SD). Bl is the negative control of the PCR reaction. (B) Restriction enzyme analysis of the PCR products in A. 5μl of undigested PCR reaction (U) were loaded in parallel with 5μl of the same reaction that had been extensively digested with the restriction endonuclease EcoRI (D). The experiment was carried out with the PCR products from SHR, WKY and Sprague-Dawley (SD).
gel, distinct bands representing repetitive elements were visible. This suggested that
the genomic DNA had been completely digested. It would also be unlikely that
approximately 50% of the genomic DNA would be partially digested in all three
samples.

The first hypothesis (a point mutation that abrogated the EcoRI restriction site inside intron
1 of COT in one chromosome) was tested by amplifying intron 1 from the same genomic
DNA that had been used in the southern blot. The PCR products were loaded on agarose
gel and are shown in figure 3.28 A. An aliquot of each PCR reaction was extensively
digested with the restriction endonuclease EcoRI and the DNA fragments were separated
on agarose gel. The undigested PCR products were also loaded on the same gel for a
comparison and the result is shown in figure 3.28 B. The PCR products derived from the
three rats were completely digested after incubation with EcoRI. This ruled out the
possibility of a point mutation of the EcoRI recognition sequence on one chromosome. If
there was a point mutation of the EcoRI sequence in one chromosome, then half the
amount of the PCR product would have been cleaved, while the other half would have not.

The size of the higher band in the southern of figure 3.27 C corresponds to the size of
the genomic region comprised between the EcoRI site just upstream of the promoter, and
the EcoRI site inside exon 2 (see figure 3.27 A). One possibility was that the EcoRI site
inside intron 1 was blocked by methylation. The sensitivity of the restriction endonuclease
EcoRI to DNA methylation has been previously reported (Geahigan KB 2000; del Arco A
1993). Therefore, it is possible that the bigger band could result from the blockage of the
EcoRI restriction site inside intron 1 by DNA methylation. If the presence of the double
band was caused by DNA methylation, then hybridisation of the membrane with the
promoter region should give a band of 7.2 Kb as before and instead of the band of around 4
Kb, a band of approximately 2.8 Kb would be detected. To test this, the promoter of the
COT gene was amplified with primers COT Prom F and COT Prom R, eluted from agarose
gel, labelled with $^{32}$P α-dCTP, and used to probe the same membrane which had been
previously stripped from the COT exon 2 probe and the result is shown in figure 3.27 C
and indicated as promoter. When genomic DNA had been cleaved with HindIII, a single
band was detected after hybridisation. Moreover, the size was the same as that of the band
had been detected by hybridisation with COT exon 2. In theory the digestion with HindIII
should have given two hybridising bands, one that was detected, and another that, due to its
small size, migrated out of the agarose gel when the DNA fragments were being separated
by gel electrophoresis. Again, from the EcoRI digestion, two bands were detected. The
larger had the same size as the one that had been detected by the COT exon 2 probe. The
Figure 3.29. Variation in the extent of DNA methylation in different tissues and different strains. Genomic DNA from kidney (K), liver (L) and tail (T) of three different strains, SHR, Sprague-Dawley (SD) and WKY, was digested with the restriction endonuclease Eco RI, which is sensitive to DNA methylation. DNA fragments were separated on a 0.7 % agarose gel and transferred to a Hybond N+ membrane. The membrane was then hybridised with a probe corresponding to COT exon 2. The lower band (D = digested) is obtained from cleavage of the Eco RI recognition sequence in intron 1 of COT. In contrast, the upper band (U = undigested) is due to the failure of the Eco RI restriction endonuclease to cleave the Eco RI site inside COT intron 1.
Figure 3.30 Percentage of methylation in different tissues and different rat strains. The upper and lower band of figure 3.28 from each sample were quantified with Optiquant and the percentage of methylated DNA was calculated. In each rat, methylation is higher in kidney than in liver. Moreover, in the Sprague-Dawley rat and the WKY strain, the highest level of methylation is observed in genomic DNA extracted from tail.
Figure 3.31. The detection of two hybridising bands with exon 2 of COT is not due to partial digestion. Genomic DNA from a tail biopsy of WKY was extensively reacted with the restriction endonuclease EcoRI. Fragments were separated on a 0.7% agarose gel and transferred to a Hybond N+ membrane. Hybridisation with exon 2 of COT (C) detected a lower band (L) and an upper band (U). In contrast, hybridisation with a probe corresponding to exon 12 of the albumin gene (A) detected a single band, demonstrating that the upper band detected with exon 2 of COT is not due to partial restriction of genomic DNA.
other was about 2.8 Kb in size, and seemed to correspond to the genomic region composed of the promoter and the 5’ end of intron 1. All the data so far seemed to support the hypothesis that a fraction of the EcoRI recognition sequence was blocked by DNA methylation. From the southern blot previously published (Caudevilla C 1998), which had been obtained with genomic DNA from liver, a single band was detected after digestion with EcoRI. This suggested the EcoRI restriction site inside COT intron 1 might be methylated in a tissue-specific manner. To further investigate this, genomic DNA was extracted from kidney, liver and tail of the SHR, Sprague-Dawley and WKY and was reacted with the restriction endonuclease EcoRI. A southern blot was carried out as before, which was then probed with exon 2 of COT. Two bands were detected in all samples except for the SHR liver sample, in which only the lower band of approximately 4 Kb was clearly visible (see figure 3.29). Furthermore, the relative ratio between the upper and lower band varied in different samples. The intensity of all of the bands was quantified, and the intensity of the upper band was divided by the combined intensity of the upper and lower band and the relative plot is shown in figure 3.30. From this data it is possible to see that the percentage of the upper band was higher in kidney than in liver in all three rats. Furthermore, in both the Sprague-Dawley and WKY, the intensity of the upper band was highest in the tail.

All of the data fitted the hypothesis that the upper band in the southern blots resulted from the failure of the restriction endonuclease EcoRI to cleave its recognition sequence inside intron 1. To ensure that the higher band was not due to the incomplete digestion of genomic DNA by EcoRI, a southern blot was repeated with genomic DNA from the WKY tail biopsy. When the membrane was probed with COT exon 2, two bands of similar intensity were detected (see lane C of figure 3.31). The probe was then stripped off and the membrane was re-probed with exon 12 of the albumin gene (a region amplified with primers Alb E12F and Alb E12R). This time, only a single band was detected. If the double band seen with the probe corresponding to exon 2 of COT was due to partial digestion, then more than one band would have also been detected by the albumin exon 12 probe. The other possible explanation, the duplication of exon 2, would have had to have happened only in some cells but not in others, which seems quite unlikely. Furthermore, if the presence of the two bands in the southern blot were due to a duplication of exon 2, two hybridising bands should have been obtained from digestion with HindIII as well. Thus, it can be concluded that the most likely explanation for the presence of the upper band observed after digestion of genomic DNA with EcoRI is the partial methylation of its recognition sequence inside intron 1. The extent of methylation appeared to vary
Figure 3.32 Sequence chromatographs relative to portions of the COT promoter in Sprague-Dawley (SD), SHR and WKY. The promoter region was amplified with primers COT Prom F and COT Prom R in separate tubes. As template, genomic DNA from Sprague-Dawley, SHR or WKY was used. PCR products were separated by electrophoresis on agarose gel. From each sample, a single amplicon of identical size was obtained. The PCR products were separately eluted and sequenced with the primers used for amplification. The chromatographs of two portions of COT promoter are shown. The sequence chromatograph on the left refers to a region of the promoter that includes a single nucleotide polymorphism (SNP). The nucleotide in the middle is A in both the Sprague-Dawley rat and SHR strain. In contrast in WKY, there is the simultaneous presence of the A and C nucleotides. The chromatographs on the right represent a region of the COT promoter further downstream from the SNP. In Sprague-Dawley and SHR this region is identical, and the signal from only one nucleotide is present at each position. In contrast in WKY there are many positions at which the signal from two different nucleotides is present simultaneously (starting from nucleotide at position 9). Due to space limitations only a short sequence is shown but the presence of two signals at the same position carried on from nucleotide at position 9 in the chromatograph to the end of the sequence (data not shown). This suggests that the eluted fragment was a mixture of two PCR products with different sequences. There seem to be many point mutations between the two amplicons. Alternatively, all the differences could be due to a deletion or insertion in one of the two amplicons. Both chromatographs come from the sequence that was performed with primer COT Prom F. Sequencing with primer COT Prom R confirmed the presence of two different PCR products only in WKY (data not shown). Green = A, blue = C, black = G, red = T.
Figure 3.33 Alignment between the SHR COT promoter and the C type COT promoter in WKY. The mutations (point mutations or deletions/insertions) are shown in bold in WKY. Three point mutations (SNPs) are present. Furthermore, in the C type promoter there is an insertion of six nucleotides at position 178, an insertion of an A nucleotide and a deletion of six nucleotides further downstream. All the differences are shown in bold in WKY. The underlined sequence represents the recognition sequence of the restriction endonuclease AvaI [C(T/C)CG(A/G)] which cleaves the C type COT promoter in WKY but not the A type COT promoter.
considerably between different tissues as shown in figure 3.30. Tissue-specific differences in DNA methylation have been already reported (del Arco A 1993).

3.2.7. The promoter of the COT gene

In order to further characterise the COT gene, its promoter region was sequenced in WKY, SHR and Sprague-Dawley. A portion of approximately 1.1 Kb, beginning just upstream of exon 1 of COT was amplified using primers COT Prom F and COT Prom rev. Genomic DNA from SHR, WKY and Sprague-Dawley was used as template. PCR products were separated by gel electrophoresis, eluted and sequenced. An alignment between the SHR and Sprague-Dawley sequences revealed that they were identical (data not shown). In contrast, the sequence obtained from WKY presented a few differences compared to that of SHR. At one nucleotide position, the COT promoter of WKY appeared to be heterozygote and from a certain region onward the sequence seemed to contain two different DNA sequences as shown in figure 3.32. The PCR product obtained from WKY was cloned into pGEM® -T Easy Vector and some of the clones were sequenced. The clones were found to contain two different inserts thus confirming the presence of two types of COT promoter in the WKY strain. The sequences of both types of promoter were aligned with the sequence of the COT promoter in SHR. One type of promoter was identical to that of the SHR and Sprague-Dawley whereas the other sequence had several differences which are shown in figure 3.33. The differences included three SNPs, one short deletion of 6 nucleotides, one insertion of an A nucleotide in a string of A nucleotides in the WKY sequence and a short insertion of 6 nucleotides. The SNP at position 53 creates a recognition sequence for the restriction endonuclease Aval in the WKY promoter (the one that is different from the SHR), but not in that of SHR or Sprague-Dawley (WKY→ C, SHR and Sprague-Dawley → A). The promoter sequence of WKY that is identical to that of SHR was named “A promoter” whereas the other type of promoter, which is shown in figure 3.33 was named “C promoter”. The two COT promoters could be easily distinguished because the SNP at position 53 introduces a restriction site recognised by the enzyme Aval (it cleaves only the DNA molecule with the C nucleotide, the “C type” promoter). Therefore, after restriction with the enzyme Aval, the PCR product corresponding to the A promoter is not restricted whereas the PCR product corresponding to the C promoter is cleaved into two smaller fragments. To investigate whether all WKY rats possessed two types of COT promoter, the SNP that differentiates the A and C type of COT promoters was exploited to genotype the WKY and SHR rats of figure 3.7. The COT promoter in these animals was amplified by PCR with primers COT Prom Aval F and
Figure 3.34 Genotype analysis of the COT promoter in WKY and SHR. The DNA region corresponding to the COT promoter was amplified with primers COT Prom Ava IF and COT Prom R and an aliquot of the PCR reaction was subjected to restriction digestion with the enzyme AvaI. DNA fragments were then separated by agarose gel electrophoresis and visualized by ethidium bromide staining. The "A type promoter" is not cleaved by the restriction endonuclease AvaI and is indicated by the letter A and the black arrows. In contrast, the "C type promoter" is cleaved into two DNA fragments which are indicated by the dashed arrows and the letter C. All WKY rats tested possessed both the A and C type of COT promoter whereas the SHR rats had only the A type promoter.
Figure 3.35 Analysis of segregation of the COT promoter in the F2 population of figure 3.14. The COT promoter was amplified with primers COT prom F and COT Prom R. After separation on agarose gel by electrophoresis, the PCR products were eluted and sequenced. The electropherograms shown are relative to the portion of the sequence at the A/C SNP (position 53 in figure 3.33. (A) Sequence chromatographs relative to the four F2 rats that did not exhibit exon repetition of COT. Only the signal corresponding to the A nucleotide is visible. (B) Sequence chromatographs relative to 11 F2 rats that exhibited exon repetition of COT. In each case, the signals from both the A and C nucleotides are visible. For one rat (rat number 5 of figure 3.14) the amount of PCR product was insufficient for DNA sequence analysis. Restriction digestion with the enzyme Ava I confirmed the presence of both types of COT promoter in this animal (data not shown). The numbers underneath the chromatographs identify the F2 animals and are the same as in figure 3.14. Green = A, blue = C.
COT Prom R. After resolving an aliquot of the PCR reaction by agarose gel electrophoresis, a single PCR product of the expected size was observed in all animals (data not shown). Another aliquot of the PCR reaction was incubated with the restriction endonuclease AvaI and extensively digested. DNA fragments were separated by agarose gel electrophoresis and the result is shown in figure 3.34. While the seven SHR animals had only the COT A promoter (the PCR product is not restricted by AvaI), all the eight WKY rats were found to have both types of promoter (the A promoter is represented by the uncleaved higher band whereas the C promoter is represented by the two smaller fragments).

Next, the segregation of the COT promoter was investigated in the F2 population of figure 3.14. The promoter region of COT from all the F2 rats was amplified by PCR as before. PCR products were separated by gel electrophoresis, eluted and sequenced and the results are shown in figure 3.35. In the four rats that did not show exon repetition (animals number 4, 12, 14 and 15 of figure 3.14), only the signal relative to the A nucleotide was visible (see figure 3.35 A). Therefore it was concluded that in these animals, only the “A type” promoter of COT was present. In contrast, all the other animals, which exhibited exon repetition of COT, had two types of COT promoter, since both nucleotides (A and C) were present simultaneously (shown in figure 3.35 B). Therefore, exon repetition in COT seemed to co-segregate with the “C type” promoter of COT.

The presence of two different types of promoter in WKY was quite puzzling, since there was only one type of COT 3’UTR in WKY (see figure 3.19). It was possible that the WKY strain was heterogeneous. Some WKY rats might have a C/C genotype of the COT promoter, or an A/A genotype, or heterozygote A/C (the WKY rats in figure 3.7 are siblings, therefore the presence of two types of COT promoter could simply be explained by assuming that they were derived from a cross between a parental AA and a parental CC). To address this question, two crosses were made between WKY rats. The parental rats were genotyped and were found to have both types of COT promoter (data not shown). The genotype of the COT promoter of the first filial generation (F1) rats was determined by restriction enzyme analysis as previously explained and the results are shown in figure 3.36. From each animal three DNA fragments were detected after restriction enzyme analysis, therefore it was concluded that all the offspring had the same A/C genotype. None of the rats had a C/C or A/A genotype. If the two COT promoters were present on two different chromosomes, they would segregate independently. Therefore from a cross between heterozygotes, the offspring should be composed of A/A homozygotes, A/C heterozygotes and C/C homozygotes in a 1:2:1 ratio (assuming that all combinations are
Figure 3.36 COT promoter genotype of rats obtained from two independent crosses between two WKY males and two WKY females. Cross 1 produced a litter of 10 rats whereas from cross 2 only three animals were obtained. The four parental WKY were genotyped prior to mating and were found to possess both the A and C type of COT promoter (data not shown). Animals were mated and genomic DNA was extracted from the F1 animals. In order to obtain the genotype, the promoter region of the COT gene was amplified with primers COT Prom Ava IF and COT prom R. An aliquot of the PCR reaction was subjected to restriction analysis with the enzyme Aval. DNA fragments were separated by agarose gel electrophoresis and visualized by ethidium bromide staining. The upper band (indicated by the letter A and the black arrows) represents the “A type” promoter, which is not restricted by Aval. The smaller DNA fragments are derived from the restriction of the “C type” promoter and are indicated by the letter C and the dashed arrows. All rats possess both types of promoter.
Figure 3.37. Schematic representation of the duplication of the COT gene in WKY. The blue box represents the promoter (A type). The red box is the 3' UTR region of the COT gene (the T letter represents a SNP which is found in the WKY UTR). The white rectangle is the portion of gene comprised between the promoter and the 3' UTR. Finally the black rectangles represent chromosomal regions flanking the COT gene. (A) Only one copy of the COT gene is present on the chromosome. (B) A duplication event has created two tandemly-arranged copies of the COT gene. (C) A number of mutations (point mutations, deletions and insertions) converted the A type promoter of the first copy of the gene into the C type promoter which is represented as a yellow box with the letter C inside. Obviously, the mutations could occur in the downstream copy rather than in the upstream one. The final result (presence of two types of COT promoter on the same chromosome) would be the same.
equally viable). In other words, 50% of the rats should be homozygous (either A/A or C/C) and the remaining 50% should be heterozygous (A/C). The likelihood of not finding any homozygous rats among the 13 rats tested is approximately $0.012\% \left(0.5\right)^{13}$. The absence of homozygotes could be explained in two ways:

1) The COT gene in WKY is duplicated. One possibility is that the original gene had an A type promoter and a WKY type 3' UTR (one that is cleaved by AflIII, as shown in figure 3.19). A duplication event would then create two identical COT genes. Later on, mutations (three mutations creating a SNP, one causing a short deletion and two mutations causing short insertions) in the promoter of one of the genes, transformed the “A type” promoter into the “C type” promoter. These hypothetical events are shown in figure 3.37. The presence of two COT genes tandemly arrayed would explain the absence of rats with a CC or AA genotype at the promoter region because each chromosome would provide both types of COT promoter.

2) One chromosome in WKY carries a COT gene with a “C type” promoter whereas the gene with the “A type” promoter is present on the other chromosome. The combination of two “C type” COT genes is lethal as is the combination of two “A type” COT genes. The lethality could depend on the combination of two identical COT genes being lethal, or could depend on other genes which are in close proximity to the COT gene.

To distinguish between these two possibilities, the copy number of the COT gene in SHR, WKY and Sprague-Dawley should be determined. A previous report did not detect exon duplication in the COT gene (Caudevilla C, 1998) but their approach would have not detected large duplications in the gene.

3.3. DISCUSSION

Exon repetition is an extraordinary and unexpected phenomenon that subverted our understanding of how genes functioned. In fact, the presence of tandemly repeated exons in an mRNA molecule that was derived from a gene in which the exons are present in single copy was not explained by the classical knowledge of how genes work. The discovery of exon repetition raised many questions:

1) Is the strain-specificity of exon repetition caused by strain-specific differences in trans-acting factors?
2) Is exon repetition a general aberration of gene expression, or does it happen only in certain genes?

3) If exon repetition is gene-specific, does it have any function?

4) Could exon repetition be dangerous? For example, could a specific exon from a gene that undergoes exon repetition, insert itself into the mRNA derived from several other genes thus disrupting the coding sequence in messenger RNA of those genes?

Strain and tissue-specificity of exon repetition in the SA gene suggested at first that the presence or absence of exon repetition in different strains was determined by strain-specific differences in trans-acting factors. This hypothesis was strengthened by the observation that exon repetition in a second gene, named COT, was also taking place in the WKY strain but not in the SHR (see figure 3.7). But the observation that exon repetition in SA did not co-segregate with exon repetition in COT (see 3.13) demonstrated that exon repetition in the two genes could not be controlled by the same trans-acting factor. This was further strengthened by the presence of exon repetition of COT in liver, from which exon-repetition of SA was not detected. From the F2 population it was also clear that the property of exon repetition in SA had to reside in the chromosomal region around the WKY SA allele. There were two possible explanations for this. The first one was that the WKY and SHR SA alleles were functionally different with regard to exon repetition; only the WKY SA allele was able to produce the SA mRNA isoforms with repeated exons. The second explanation was that exon repetition in SA was determined by a gene close to the WKY SA allele that encoded a trans-acting factor responsible for exon repetition in SA. The answer to this crucial question came from the analysis of SA mRNA in heterozygote rats (carrying a WKY SA allele and an SHR SA allele; thus if exon repetition was due to a trans-acting factor, this presumably would act on both alleles in a heterozygote). These experiments clearly showed that the mRNA isoforms containing repeated exons were derived solely from the WKY SA allele (see figure 3.17), and therefore demonstrated that exon repetition in SA was allele-specific.

Exon repetition in COT had been shown to be species-specific (Caudevilla C 2001). The results in this chapter demonstrated that exon repetition in COT is allele-specific. In fact, it is only observed with the WKY COT allele but not with the SHR COT allele. These findings suggested that determining whether a particular gene in a certain species undergoes exon repetition is not straightforward. For example, there might be several different alleles of a gene in the same species with some alleles (or maybe only one) producing mRNA isoforms with repeated exons, whereas other alleles might not. For
example if the SHR strain had been used to investigate exon repetition of COT instead of Sprague-Dawley, it would have been wrongly concluded that in rat, the COT gene does not exhibit exon repetition.

The fact that exon repetition is allele-specific in both SA and COT raises the possibility that the existence of alleles with different properties (some producing exon repetition, others not) might be quite common. It would certainly be interesting to study other examples of exon repetition in other genes and species to see how common this phenomenon is.

The finding of the allele-specificity of exon repetition provides useful information about the questions raised above. First of all, since exon repetition appears to be restricted to only certain alleles, it can be concluded that it is not a general aberration of gene expression but instead, it is tightly restricted to only certain genes (even certain alleles).

As far as the biological function of exon repetition is concerned, there are no obvious phenotypic differences between SHR and WKY, and the SHR strain does not have any phenotypic abnormalities that might be linked to the absence of exon repetition, nor does the WKY strain show any abnormalities. Hence, it can be concluded that exon repetition in COT and SA is non-essential. The possible effects of exon repetition on translation of the mRNA isoforms containing repeated exons will be addressed in the next chapter.

The insertion of exons from one gene into the mRNA molecules derived from other genes could have deleterious consequences, for instance it could disrupt the coding sequence or create potentially harmful hybrid proteins. Of course, cells would have probably evolved strategies that are specifically intended to prevent this from happening. But even so, it would still be possible that certain exons, from particular genes, could have a strong natural propensity to be ligated to other exons that were part of a different pre-mRNA molecule (transcribed from the same gene or from a different gene). But the observation that exons from one allele are not found in the mRNA derived from the other allele indicated that the possibility that an exon from one gene could attack the RNA molecules derived from other genes was very unlikely indeed. Therefore, it can be concluded that exon repetition in one gene does not interfere with the proper expression of other genes. It is possible though, that the presence of repeated exons in mRNA could disrupt the coding sequence of the gene that undergoes exon repetition. This is especially more likely if the exon being repeated were in the middle of a gene (after the exon that contains the AUG translational start codon).

In this chapter it was shown that exon repetition is restricted to the WKY SA (and COT) alleles. In heterozygotes, the mRNA isoforms with repeated exons are produced only
by the WKY allele. For this reason, it was concluded that the strain-specificity of exon repetition was not due to differences in strain-specific trans-acting factors. This of course, was based on the assumption that both alleles in heterozygotes would have access to the same trans-acting factors. But apparently, the allele-specificity of exon repetition in SA does not explain the tissue-specificity. For example, why does the same allele produce mRNA isoforms with repeated exons in kidney but not in liver? This could be simply explained by assuming that exon repetition in SA requires a trans-acting factor, and this factor is expressed in kidney but not in liver. Other explanations are still possible though. For example, the same allele could undergo changes in chromatin remodelling in a tissue-specific manner. Such changes could include the addition or removal of a methyl, acetyl, phosphate and ubiquitin groups from histone proteins. Other important changes include the positioning of nucleosomes. Some of these modifications could potentially lead to the production of different mRNA isoforms from the same allele in different tissues. An effect of chromatin structure on transcription termination has already been shown (Alen C 2002). Therefore, it is possible that chromatin modifications could affect the types of mRNA isoforms that are produced by a particular gene.
CHAPTER 4
FUNCTIONAL SIGNIFICANCE OF EXON REPETITION

4.1. INTRODUCTION

4.1.1. Exon repetition can create new proteins

As mentioned before, alternative splicing is by far the most important mechanism used by mammals to expand their proteome without increasing the number of genes. From a single gene, through alternative splicing, several different mRNA isoforms can be produced. One of the most extreme examples of how alternative splicing can create many different mRNA isoforms from the same gene is the Drosophila homolog of human Down Syndrome cell adhesion molecule (dscam). This gene, in theory, can produce more than 38,000 mRNA isoforms. Some of the mRNA isoforms produced by alternative splicing contain open reading frames (ORFs) that encode novel protein products.

Exon repetition is another way of producing several mRNA isoforms from a single gene. Hence, exon repetition can potentially contribute to the expansion of the proteome. Some of the mRNA isoforms with tandemly repeated exons contain ORFs that encode new proteins. For example, in the Sprague-Dawley rat, three mRNA isoforms are derived from the COT gene. One isoform corresponds to the normal COT mRNA, in which each exon is present in single copy, while the other two isoforms contain a tandem repeat of exon 2 alone, or exons 2 and 3 simultaneously. The normal COT mRNA isoform encodes the normal COT protein product, whereas the mRNA isoform with the tandem repeat of exons 2 and 3 has a longer open reading frame that encodes a longer protein product differing from the normal one only at the N-terminus as shown in figure 4.1. Interestingly, a polyclonal antibody that was raised against a peptide corresponding to the N-terminus of the COT protein (sequence 43-54, ANEDEYKKTEEI) detected two bands of different size in the Sprague-Dawley rat (Caudevilla C 1998). In contrast, pre-immune serum failed to detect any bands. The smaller protein had an estimated size corresponding to that expected for the normal COT protein. The size of the larger protein was consistent with the size that would be expected for a protein derived from the translation of the COT mRNA in which
Figure 4.1. Comparison of the two COT protein isoforms present in the Sprague-Dawley rat. (A) Amino acid sequence of the protein derived from translation of the normal COT mRNA isoform. The normal protein is composed of 612 amino acid residues. The first methionine is shown in bold. (B) Amino acid sequence of the protein derived from translation of the COT mRNA isoform with a tandem repeat of exons 2 and 3. This protein is composed of 698 amino acid residues. The amino acid residues with a gray background at the N-terminus of the protein, are derived from translation of the first repeat of exons 2 and 3. The underlined amino acid residues are derived from translation of the portion of the second copy of exon 2 upstream of the AUG codon. This short stretch of protein sequence is absent in the normal COT protein. The methionine residue shown in bold is the same as the one in the normal COT protein. Downstream of the second methionine, the two proteins are identical in sequence.
exons 2 and 3 are repeated. This suggests that the COT mRNA isoform that contains the tandem repeat of exons 2 and 3 is actually translated in vivo. It is not known whether this novel protein performs new functions, or if indeed, it has any function at all.

In the case of the voltage-gated sodium channel, exons 12, 13 and 14 are tandemly repeated. This mRNA isoform is predicted to encode a longer protein product containing a partial repeat of domain II (Akopian AN 1999), but its presence in vivo has not been investigated. Finally, the mRNA isoform of the human Sp1 gene that contains a tandem repeat of exons 2 and 3 could encode for a protein with duplicated transcriptional activation domains. This hypothetical protein might possess stronger activation activity since a synergistic effect between activating domains has been shown (Courey AJ 1989; Pascal E 1991).

It should be pointed out that many mRNA isoforms with repeated exons contain short ORFs encoding truncated proteins at their 5' end. This is caused by the presence of premature termination codons (PTCs). It is not known whether these truncated proteins have any function, or even if they are actually translated. It should be remembered that approximately one third of genetic diseases and cancers are caused by the presence of a PTC. Truncated proteins can have a deleterious effect on the cell. For example, by acting in a dominant-negative manner, they can disrupt the function of the wild type protein. For instance, the long QT syndrome is associated with a high risk of sudden death. Mutations in the KCNQ1 gene, which encodes the cardiac KvLQT1 potassium ion (K+) channel, cause the autosomal dominant Romano-Ward (RW) syndrome. It has been shown that an N-terminus-truncated KvLQT1 isoform endogenously expressed in the human heart exerts strong dominant-negative effects on the full-length KvLQT1 protein (Mohammad-Panah R 1999). Another example is the mouse model for Rubinstein-Taybi syndrome (RTS). This was generated by an insertional mutation into the cyclic AMP response element-binding protein (CREB)-binding protein (CBP) gene. Heterozygous CBP-deficient mice, which had truncated CBP protein (residues 1-1084) containing the CREB-binding domain (residues 462-661), showed clinical features of RTS (Oike Y 1999).

4.1.2. Non sense mediated mRNA decay

The synthesis of truncated proteins is the cause of diseases only in a small proportion of cases. The main effect of a PTC is the induction of degradation of the mRNA molecules that contain the premature termination codon. This was discovered many years ago in the
β-globin gene (Kinniburgh AJ 1982) and is now known as nonsense-mediated mRNA decay (NMD). This system is ubiquitous in all eukaryotic organisms and is capable to recognise the presence of PTCs in mRNA molecules and then targeting these molecules for degradation. Of course, cells need a way of discriminating between a normal stop codon and a PTC. In mammals, as a general rule, a PTC is defined by its position relative to the final exon/intron junction in an mRNA molecule. If the position of a stop codon is within 50 nucleotides (5’) of the newly formed exon-exon junction, then it is recognised as a normal stop codon. If the stop codon is positioned further upstream, then it is recognised as a PTC, and the mRNA is subjected to degradation (Zhang J 1998). The involvement of NMD in destroying potentially harmful mRNAs has been demonstrated. For example, in *C. elegans*, there is a mutation that introduces a PTC in the mRNA of the gene UNC-54, which encodes a myosin heavy chain. In heterozygotes that have a functional NMD system, the mutated allele produces only 5% of the total UNC-54 mRNA and the worms appear healthy (Pulak R 1993). In contrast, in heterozygotes that have a deficiency in the NMD system, the expression levels of mRNA from the wild type and the faulty gene are approximately the same and the worms have an aberrant phenotype due to the dominant-negative effect of the truncated form of the myosin protein. Similar examples are also known in humans. For example, β-Thalassemia is caused by mutations in the β-globin gene. This gene is composed of three exons, and the majority of mutations that cause β-Thalassemia introduce a PTC in the first or second exon of this gene. As in *C. elegans*, heterozygote individuals are generally healthy and the amount of mRNA produced from the faulty gene is either absent or very low (Forget BG 1974; Baserga SJ 1988). However one group of mutations behave in a different way. They are located inside the last exon of the β-globin gene, and hence escape NMD. Individuals that are heterozygote for this type of mutation, have high levels of the mutated form of β-globin mRNA and show clinical symptoms (Hall GW 1994).

Quite often, mRNA isoforms with repeated exons have short upstream ORFs, followed downstream by the same ORF that is present in the normal mRNA isoform. These isoforms might not be subjected to NMD, since they contain a normal uninterrupted downstream ORF. Thus, it could be that exon repetition might be detectable only with those particular exons whose repetition does not trigger NMD. Interestingly, it is the exons near the 5’ end of the gene that most often are repeated in mRNA. The corresponding mRNA isoforms normally contain a short upstream ORF encoding a truncated version of
the protein and a downstream longer ORF that encodes the normal protein. In contrast, if exon repetition involved exons located in the middle of the gene, then the main ORF would most probably be interrupted by stop codons (unless the length of the repeated exon/s is a multiple of three) and this in turn would trigger NMD. The only example of exon repetition that involves exons located in the middle of a gene is the voltage-gated sodium channel. In this case though, the repetition of exons 12-13-14 does not introduce any stop codons but it simply increases the length of the ORF. All the above observations could explain why exon repetition appears to be such a rare phenomenon, for example it might be selected against by natural selection, and even if mRNA isoforms with repeated exons are produced, it might be difficult to detect them because they could be efficiently degraded by NMD.

4.1.3. Exon repetition can potentially decrease the production of the normal protein product from a gene

In paragraph 4.1, it was explained that exon repetition can create new ORFs encoding new proteins (either a longer or a truncated protein). However the presence of repeated exons could also have an effect on the efficiency at which the corresponding mRNA is translated into a full length protein product by the ribosome. For example, if the exon containing the AUG translational start codon is repeated, and the AUG in the first copy of the exon is out of frame from the AUG present in the second repeat, then the mRNA might be inefficiently translated into a full length protein. In fact, it is likely that the AUG present in the first copy of the repeated exon would be recognised by the ribosome as translational start codon, therefore decreasing the translation rate of the downstream ORF. This might be the case in the COT mRNA isoform that contains a tandem repeat of exon 2. In fact, the AUG translational start codon is found in exon 2 and the ORF that begins in the first copy of exon 2 is interrupted by a stop codon in the second copy of exon 2. Therefore, in order to synthesise the full length protein, the ribosome must start translation at the AUG in the second copy of exon 2. This is illustrated in figure 4.2. The normal COT mRNA contains a single ORF, the translation of which produces a full length COT protein, as illustrated in figure 4.2 A. In contrast, in the mRNA isoform with a tandem repeat of exon 2, two ORFs are present. The upstream ORF encodes a truncated COT protein whereas the downstream ORF encodes the full length COT protein as shown in figure 4.2 B. Hence, only when the ribosome selects the second AUG as translational start codon is the full length protein
Figure 4.2. Translation of different COT mRNA isoforms. (A) Schematic representation of the normal COT mRNA isoform that contains a single ORF which is translated into the normal COT protein, represented by a red bar. (B) Schematic representation of the COT mRNA isoform containing the tandem repeat of exon 2. This isoform contains two ORFs. The upstream ORF (comprised between the blue AUG start codon and the blue UGA stop codon) encodes a truncated COT protein of 42 amino acid residues which is shown as a blue bar. In contrast, the downstream ORF (comprised between the red AUG start codon and red UAG stop codon) encodes the full length COT protein, again shown as a red bar. From this isoform, the full length protein is synthesized only when the second AUG is selected by the ribosome as translational start site.
synthesised. But the presence of an upstream ORF might pose a problem. This will be discussed in detail in the next paragraph.

4.1.4. Factors that influence the efficiency of translation

The efficiency with which a particular mRNA molecule is translated into protein depends upon many factors. The 5' untranslated region (5'UTR) seems to play a very important role in regulating mRNA translation. In particular, its length (Kozak 1991a), the presence of secondary structures (Kozak 1994), and the presence of AUG codons upstream of the real translational start site (Geballe AP 1994) all seem to affect the efficiency at which an mRNA is translated into protein. The presence of upstream AUG codons generally reduces the rate of translation from the true start codon. Interestingly, in humans, the presence of AUG codons upstream of the actual start codon is encountered in more than 40% of mRNAs analysed (Peri S 2001). Another important factor is the flanking sequence around the AUG start codon. The typical mammalian consensus sequence is (GCC)GCCRCCAUGG (Kozak 1987). The most important positions in the Kozak sequence are the nucleotides at positions –3 and +4 (shown underlined) relative to the A of the AUG. It has been shown that the presence of upstream start codons correlates to a weak recognition sequence (Rogozin IB 2001). For example, of all the human mRNAs that contain AUGs upstream of the real start codon, in 35% of the cases, the upstream codons have a better consensus sequence than the real codon, and in 12% of cases, the upstream AUGs have a similar context to the authentic AUG. It seems as though both factors that lower the efficiency of translation are simultaneously present in these particular mRNAs.

Interestingly, the incorporation of AUG codons in the 5'UTR seems to be selected against. For example, the expected number of AUG codons in 5'UTRs was estimated based upon the length and base composition and it was found that the observed number of AUGs is significantly lower than the expected number (Pesole G 1996), as shown in table 4.1.
Table 4.1. Comparison between the expected (E) and the observed (O) number of AUG codons in 5' UTRs in different organisms (modified from Pesole G 1996).

<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of sequences</th>
<th>Average length</th>
<th>E</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>5459</td>
<td>160</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>Mus sp.</td>
<td>3330</td>
<td>159</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>Rattus sp.</td>
<td>2589</td>
<td>161</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>Arabidopsis sp.</td>
<td>1023</td>
<td>101</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>Aspergillus sp.</td>
<td>52</td>
<td>98</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Drosophila sp.</td>
<td>1134</td>
<td>288</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>Gallus sp.</td>
<td>739</td>
<td>141</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>Saccharomyces sp.</td>
<td>267</td>
<td>136</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>Xenopus sp.</td>
<td>765</td>
<td>136</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>Zea sp.</td>
<td>315</td>
<td>115</td>
<td>12</td>
<td>6</td>
</tr>
</tbody>
</table>

4.1.5. Selection of a downstream AUG as translational initiation codon

In an mRNA molecule containing AUG codons upstream of the true translational start codon, there are several ways by which the true codon can be selected as the starting translation point:

1) Leaky scanning. In this case, the upstream AUG is rather weak, and a substantial fraction of the ribosomes fail to start translation at such codon. Instead, they proceed in scanning the mRNA until they reach the downstream AUG (the one that directs translation of the main ORF) and start translation. An upstream AUG might be weak, for instance because it is in a poor context. Another factor that influences the efficiency of utilisation of an AUG is its distance from the 5' end of the mRNA. If this distance is less than 40 nt, than the AUG is used for translation quite inefficiently (Kozak 1991b). If scanning is obstructed, for example by the presence of a secondary structure downstream of a weak AUG, there is a corresponding increase in the efficiency at which such AUG is used for translation (Kozak 1991a).

2) Presence of an IRES (internal ribosome entry site). If an IRES sequence is present inside an mRNA molecule, then the cap-dependent mechanism of translation is bypassed and translation can initiate directly at an internal AUG codon. (Stoneley M 1998; Coldwell MJ 2000; Rubtsova MP 2003).
3) Translation reinitiation. The 80S ribosome starts to translate the upstream ORF (uORF). When the stop codon is reached, it is believed that the 60S ribosomal subunit is released. Instead, the 40S subunit resumes scanning and initiates translation of the downstream ORF. This mechanism is enhanced in cases in which the uORF is short and the inter-cistronic region is long (Kozak 1991). A long intercistronic region is thought to have a positive effect because it might give the ribosome enough time to recruit factors involved in translation initiation (such as the tertiary complex eIF-2, GTP and Met-tRNAi) after it has stopped translating the uORF. On the other hand, a long uORF is believed to have a negative effect on reinitiation because the ribosome might lose some factors required for initiation of translation. It is believed that some of these factors remain loosely bound to the ribosome and are gradually shed off as the ribosome proceeds to translate an ORF. To strengthen this hypothesis is the observation that a pseudoknot (it slows down elongation of translation) located in the uORF inhibits translation reinitiation at the downstream AUG (Kozak 2001). Reinitiation is used, for example, to modulate translation of the GCN4 gene in yeast.

4.1.6. Effects of exon repetition on translation

As mentioned before, by inserting AUG codons upstream of the ORF that encodes the full length protein, exon repetition could decrease the efficiency at which the mRNA isoform with repeated exons is translated into a full length protein product. Hence, if exon repetition in a particular gene is very efficient (i.e. the majority of mRNA molecules contain repeated exons), it could cause a significant reduction in the expression level of the full-length protein product encoded by that gene, with possible deleterious consequences for the cell. For example, in the COT gene, the AUG translational start codon is located in exon 2. The COT mRNA isoform in which exon 2 is repeated contains two ORFs. The upstream ORF starts from the AUG in the first copy of exon 2 and terminates at the TGA stop codon located in the second copy of exon 2. This ORF encodes a truncated version of the COT protein. The downstream ORF starts at the AUG codon inside the second copy of exon 2 and terminates at the stop codon in exon 17. The nucleotide sequence of the COT mRNA isoform with the tandem repeat of exon 2 is shown in figure 4.3. Hence, if the first AUG is selected by the ribosome as translational start codon, a truncated protein is produced. The full length protein is only synthesised when the ribosome selects the AUG of the downstream ORF as the translational start site. But the presence of the upstream
Figure 4.3. Nucleotide sequence of the cDNA corresponding to the COT mRNA isoform with a tandem repeat of exon 2. The nucleotide sequence corresponding to the first copy of exon 2 is underlined whereas the second copy of exon 2 is shown with a grey background. Two open reading frames are present. The first ORF is comprised between the ATG codon (shown in bold) in the first copy of exon 2 and the TGA stop codon (shown in bold and italic) in the second copy of exon 2. This upstream ORF encodes a truncated COT protein. The second ORF, which encodes the full length COT protein, is comprised between the ATG codon inside the second copy of exon 2 and the TAG stop codon at the end of the sequence. The ATG translational start site of the second ORF is only 7 nucleotides downstream from the stop codon of the first ORF.
AUG codon (which of course is in the same context as the downstream AUG codon) might pose a problem, for the reasons explained before. A similar situation is observed in the SA gene. In this case, repetition of exons 2, 3 and 4 might lead to the production of a truncated protein and a concomitant decrease in the synthesis of the full-length SA protein (the AUG translational start codon is located in exon 3 in the SA gene).

In this chapter, the effect of the presence of repeated exons on the translation of the main ORF were investigated. The data showed that in certain mRNA isoforms with repeated exons, the downstream ORF is translated at a much lower efficiency compared to the same ORF in the normal mRNA (i.e. without repeated exons). Therefore, if exon repetition happens with high efficiency, it might actually disrupt gene function.

4.2. RESULTS

As mentioned in the introduction of this chapter, the presence of a tandem repeat of exon 2 in COT mRNA might impede the synthesis of the full length protein product. As shown in figure 4.2, normal mRNA can only be translated into a full length protein. In contrast, the COT mRNA isoform that contains a tandem repeat of exon 2, can be translated into two distinct proteins depending on which AUG is chosen as translational start codon. If the AUG in the first copy of exon 2 is selected by the ribosome, then a short truncated protein is synthesised, because the ORF that starts from this AUG is interrupted by a stop codon located in the second copy of exon 2, as shown in figure 4.2. From this mRNA, the normal COT protein can only be produced when the AUG in the second copy of exon 2 is selected as translational start codon. In order to investigate the relative frequency at which the two AUG translational start codons are selected by the ribosomes, the COT 5'UTR from both the normal transcript and the transcript containing the tandem repeat of exon 2 were cloned upstream of the reporter gene firefly luciferase, so that they became the 5' UTR of the firefly luciferase gene. For this purpose, plasmid pGL3-Control Vector was used. This vector contains an optimised 5' UTR upstream of an open reading frame that encodes the firefly luciferase open reading frame. Transcription is driven by the SV40 promoter and enhancer. A schematic representation of the pGL3-Control Vector is shown in figure 4.4. The optimised 5' UTR was excised and replaced with the 5' UTR from normal COT mRNA and that of the COT mRNA isoform that contains a tandem repeat of exon 2, as described in materials and methods. The construct containing the 5' UTR of normal COT mRNA was named pGL3-Normal COT 5' UTR, whereas the
Figure 4.4. Schematic representation of pGL3-Control Vector. The firefly luciferase gene is driven by the SV40 Promoter/Enhancer. An SV40 polyadenylation signal is located at the end of the luciferase gene. An optimized 5' UTR is present upstream of the firefly luciferase ORF, between the restriction sites Hind III and Nco I. This 5' UTR allows the firefly-encoding mRNA to be efficiently translated.
construct containing the 5' UTR with the tandem repeat of exon 2 was named pGL3-Rep Exon 2 COT 5' UTR. The two COT 5' UTRs, the normal one and the one containing the tandem repeat of exon 2, are shown in figure 4.5.

In order to determine the efficiency at which the AUG in the second copy of exon 2 is selected for translation, an amount of DNA equivalent to 0.25 µg of pGL3-Basic Vector (a normalisation was made that took into account the different size of the different constructs, so that approximately the same number of DNA molecules were transfected for each construct) of each construct was transfected into Hela cells. Construct pGL3-Basic Vector (Promega), which lacks a promoter, was used as negative control, while pGL3-Control Vector, which has an optimised 5' UTR, was used as positive control. In order to normalise for the possible difference in cell number and efficiency of transfection and protein extraction, 10 ng of a second plasmid, phRL-CMV Vector (Promega), which contains the renilla luciferase gene driven by the CMV promoter was co-transfected with each of the firefly luciferase constructs. The experiment was performed in triplicate for each construct. The ratio between the firefly and renilla luciferase was calculated and the results are shown in figure 4.6. To ensure that there was a linear correlation between the intensity of the signal from firefly luciferase and the amount of transfected DNA, a parallel experiment was carried out in which 0.25 and 0.5 µg of construct pGL3-Normal COT 5' UTR were transfected into Hela cells, while the amount of renilla luciferase plasmid was kept constant. The result is shown in figure 4.7. A two-fold increase in the amount of transfected DNA approximately doubled the intensity of the signal from firefly luciferase, demonstrating that there was indeed a linear correlation between the amount of DNA transfected and the signal from firefly luciferase, at least in the range of concentration used.

As expected, the signal from pGL3-Basic Vector, which lacks a promoter, was very low, approximately 2,000 times less compared to pGL3-Control Vector (see bar A in figure 4.6). The activity of the 5'UTR of the normal COT mRNA isoform, was comparable to that of pGL3-Control Vector (see bars C and B, respectively). In contrast, the value obtained from the construct containing the 5' UTR with the tandem repeat of exon 2, was approximately nine times lower (see bar D in figure 4.6). The reduced amount of firefly luciferase protein obtained from construct pGL3-Rep Exon 2 COT 5' UTR, could be due to two different reasons:

1) The increased length of the 5' UTR that contains the tandem repeat of exon 2, or the creation of a secondary structure due to the tandem repeat of exon 2
Figure 4.5. Nucleotide sequence of the 5' UTR of two different COT mRNA isoforms. (A) Nucleotide sequence of the 5' UTR of the normal COT mRNA isoform. Only one AUG codon is present (shown in bold), which corresponds to the translational start codon of the ORF encoding the full length COT protein. The sequence with a grey background is the 5' end of exon 2 whereas the upstream nucleotides represent exon 1. (B) Nucleotide sequence of the 5' UTR of the COT mRNA isoform containing a tandem repeat of exon 2. The nucleotides of the first copy of exon 2 are underlined whereas the 5' end of the second repeat of exon 2 is shown in grey background. The AUG translational start codon (shown in bold) and located in the first copy of exon 2 is part of an ORF that terminates at the UGA stop codon (shown in bold italic) in the second copy of exon 2. Translation of this ORF produces a short truncated version of the COT protein. The downstream AUG codon in the second copy of exon 2 is the translational start codon of the second ORF that encodes the full length COT protein.
Figure 4.6. Measurement of the efficiency of translation of the COT 5' UTRs by firefly luciferase assay. Approximately the same number of plasmid molecules of each construct were transfected into Hela cells, together with 10 ng of plasmid phRL-CMV. Proteins were extracted 24 hours after transfection and the activity of firefly and renilla luciferase was measured. Each construct was tested in triplicate. The average of the three measurements is shown, together with the standard deviation. The Y axis represents the ratio between the signal from firefly and renilla luciferase, multiplied by a factor of 1,000.
Figure 4.7. Correlation between the intensity of the firefly luciferase signal and the amount of construct pGL3-Normal COT 5' UTR transfected into Hela cells. The Y axis represents the ratio between the signals from firefly and renilla luciferase multiplied by a factor of 1,000. Either 0.25 µg or 0.5 µg of construct pGL3 Normal COT 5' UTR were transfected into Hela cells, together with 10 ng of plasmid phRL-CMV Vector, which encodes the renilla luciferase protein.
2) The AUG in the first copy of exon 2 is selected by the ribosomes as translation start codon, thus reducing the efficiency of translation from the second AUG.

In order to discriminate between these two possibilities, a third construct was prepared. This construct is the same as construct pGL3-Rep Exon 2 COT 5' UTR, except that the AUG codon inside the first copy of exon 2 was mutated to a UUG codon. This was done by overlapping PCR with primers containing the point mutation. The strategy is illustrated in figure 4.8. As shown by bar E in figure 4.6, the mRNA from this new construct was translated with an efficiency comparable to that of the mRNA derived from construct pGL3-Normal COT 5' UTR. This demonstrates that the decreased translatability of the mRNA derived from construct pGL3-Rep Exon 2 COT 5' UTR was due to the presence of the AUG codon in the first copy of exon 2, and not to an increased length of the 5'UTR, or to a particular secondary structure in this 5' UTR.

In order to investigate the effects that the repeated exons in the SA mRNA isoforms have on translation, the optimised 5' UTR of pGL3-Control Vector was replaced with the 5' UTRs of the three SA mRNA isoforms, as described in materials and methods. The construct containing the 5' UTR of normal SA mRNA was named pGL3-Normal SA 5' UTR. The constructs with the 5' UTR of the SA mRNA with a tandem repeat of exon 2 or exons 2, 3 and 4 were named pGL3-Rep Exon 2 SA 5' UTR and pGL3-Rep Exons 2-3-4 SA 5' UTR, respectively. All the SA 5' UTRs are shown in figure 4.9. These three constructs were transfected into Hela cells, together with 10 ng of plasmid phRL-CMV Vector. The experiment was performed in triplicate for each construct. Proteins were extracted 24 hours after transfection and the activity of firefly and renilla luciferases was measured. The ratio between firefly and renilla was calculated for each sample and multiplied by a factor of 1,000. The mean value of the three experiments was then calculated and the result is shown in figure 4.10. Both the 5' UTR from normal SA mRNA and that containing a tandem repeat of exon 2, allowed efficient translation of the firefly protein (see bars B and C respectively). In contrast, the efficiency at which the 5' UTR containing the tandem repeat of exons 2, 3 and 4 was translated into the firefly luciferase protein, was drastically reduced (see bar D). The results with the COT 5' UTRs showed that exon repetition reduced the efficiency of translation because of the upstream AUG codon associated with the tandem repeat of exon 2. In order to investigate the role of the AUG translational start site inside the first copy of exon 3 in SA, this codon was mutated to a UUG codon by overlapping PCR performed which primers carrying the point mutation.
Figure 4.8. Creation of a point mutation in a DNA molecule by overlapping PCR. (A) Schematic representation of the 5' UTR of the COT mRNA isoform with the tandem repeat of exon 2. The upstream ATG will be mutated to a TTG codon. The cDNA is used as template for two distinct PCR reactions (PCR A and PCR B). PCR A is performed with primers 1 and 2 whereas in PCR B, primers 3 and 4 are used. Primers 2 and 3 overlap and each carry a point mutation that is indicated by the star (the A nucleotide of the ATG has been substituted with a T nucleotide). Primers 1 and 4 have a recognition sequence for the appropriate restriction endonucleases at their 5’ end. The two PCR fragments are then purified from agarose gel to eliminate the primers. (B) The two PCR products are mixed together. Since they overlap over a certain region, once melted, the antisense strands can anneal to one another (indicated by the vertical lines). (C) A further amplification is performed with primers 1 and 4. The PCR product contains a TTG codon instead of the original ATG codon. (D) Final PCR product. The DNA molecule can now be digested with the appropriate restriction endonucleases and cloned into a suitable vector.
Figure 4.9. Nucleotide sequence of the 5' UTR of the SA mRNA isoforms. The nucleotides of exon 1 are shown with no background. The nucleotide sequence of exon 2 has a yellow background, exon 3 has a dark blue background, and finally, exon 4 has a grey background. The true translation initiation AUG codon is shown with green letters. All the other AUG codons are shown in purple letters. (A) Sequence of the 5' UTR of the normal SA mRNA isoform. Three AUG codons are located upstream of the true translational initiation codon. (B) Sequence of the 5' UTR of the SA mRNA isoform with a tandem repeat of exon 2. The second copy of exon 2 is shown in a light blue background. Six AUG codons are present upstream from the true translational start codon. One such AUG codon is created by the junction of the two copies of exon 2. (C) Sequence of the 5' UTR of the SA mRNA isoform with a tandem repeat of exons 2, 3 and 4. Eighteen AUG codons are located upstream of the AUG translational start codon associated with the full length ORF encoding the SA protein. The green AUG in the first copy of exon 2 is part of a truncated ORF that terminates with the UGA stop codon (shown in red letters) in the second copy of exon 2. This short ORF encodes a truncated version of the SA protein, the sequence of which is shown in D. (D) Amino acid sequence of the truncated SA protein encoded by the ORF in the SA mRNA isoform with the tandem repeat of exon 2. This truncated protein consists of 154 amino acid residues. The black letters represent the amino acid residues at the N-terminus of the protein that are identical to the normal SA protein, whereas the red letters are the amino acid residues that are unique to this short protein.
Figure 4.10. Measurement of the efficiency of translation of the SA 5' UTRs by firefly luciferase assay. Approximately the same number of plasmid molecules of each construct were transfected into Hela cells, together with 10 ng of plasmid phRL-CMV. Proteins were extracted 24 hours after transfection and the activity of firefly and renilla luciferase was measured. Each construct was tested in triplicate. The average of the three measurements is shown, together with the standard deviation. The Y axis represents the ratio between the signal from firefly and renilla luciferase, multiplied by a factor of 1,000.
This construct was named pGL3-Mut AUG SA 5' UTR and was transfected into Hela cells and tested as before. The result is shown by bar E in figure 4.10. It is evident that the inactivation of the AUG codon in exon 3 did not recover the efficiency of translation of the SA 5' UTR containing the tandem repeat of exons 2, 3 and 4. In fact, the efficiency at which this 5' UTR is translated is only slightly higher than that of construct pGL3-Rep Exon 2 SA 5' UTR.

The normal 5' UTR of COT does not have any AUG codons upstream of the translational start codon. In contrast, the normal 5' UTR of SA contains 3 AUG codons upstream of the translational start codon. The efficiency at which the normal 5' UTR of the SA and COT gene are translated was compared and the result is shown in figure 4.11. It is clear that the 5' UTR of COT is translated at a much higher efficiency (approximately 10-fold) compared to the 5' UTR of the SA gene.

4.3. DISCUSSION

Exon repetition is an unexpected phenomenon that was only recently discovered (Caudevilla C 1998; Akopian AN 1999; Frantz SA 1999). It allows genes to produce new mRNA isoforms, probably through a non-linear way of pre-mRNA processing. One isoform coincides with the normal mRNA, in which each exon is present in single copy. In contrast, the other mRNA isoforms contain tandem repeats of specific exons. Some of these new isoforms can be translated into new proteins. For example, a polyclonal antibody raised against a COT peptide, detected two proteins of different sizes in liver peroxisomes of Sprague-Dawley rats. The larger protein is thought to be derived from translation of the mRNA isoform in which exons 2 and 3 are tandemly repeated. Hence, exon repetition could be another strategy employed by mammals, to increase their proteome complexity. In the mRNA produced from the voltage-gated sodium channel gene in rat, exons 12, 13 and 14 are tandemly repeated and this introduces a partial repeat of domain II in the ORF of this mRNA isoform. The existence of this novel protein has not been confirmed and remains purely speculative. Finally in the human Spl gene, one mRNA isoform contains a tandem repeat of exons 2 and 3. The transcriptional activation domains A and B are encoded by exon 3. Therefore, the mRNA isoform with a repetition of exons 2 and 3 could potentially encode a protein with duplicated transcriptional activation domains.

Other potential new proteins could be derived by translation of the short ORFs at the 5' end of the mRNA that are often created by exon repetition. As mentioned before, these
Figure 4.11. Comparison of the efficiency of translation of the 5' UTRs of the normal mRNA isoform of SA and COT. Approximately the same number of plasmid molecules of each construct were transfected into Hela cells, together with 10 ng of plasmid phRL-CMV. Proteins were extracted 24 hours after transfection and the activity of firefly and renilla luciferase was measured. Each construct was tested in triplicate. The average of the three measurements is shown, together with the standard deviation. The Y axis represents the ratio between the signal from firefly and renilla luciferase, multiplied by a factor of 1,000. A represents the 5' UTR of the normal Sa mRNA isoform whereas B is the 5' UTR of the normal COT mRNA isoform. The amount of firefly luciferase obtained from the COT 5'UTR is approximately 10 times higher than that from the SA 5'UTR.
truncated proteins could have a deleterious effect on the cell but their existence has not been investigated in any of the known cases of exon repetition.

The possible creation of new proteins is not the only effect of exon repetition. For example, the repeated exons might reduce the rate of translation of the full length protein from the downstream ORF. When the mRNA is scanned by the ribosome, usually the first AUG with a good kozak sequence is selected as translational start codon by the ribosome. Hence, if the exon that contains the AUG translational start codon is repeated, the corresponding mRNA isoform might be translated into a full length functional protein, rather inefficiently. This seems to be the case in COT mRNA. In fact, when the 5’UTR from the normal COT mRNA isoform was cloned upstream of the reporter gene firefly luciferase, the corresponding protein was synthesised very efficiently. In contrast, a parallel experiment with the 5’ UTR from the mRNA isoform that contains a tandem repeat of exon 2, revealed a significant reduction (approximately 9 fold) in the synthesis of firefly luciferase. The short intercistronic region (the distance between the AUG translational start codon in the second repeat of exon 2 and the stop codon of the upstream ORF) in the COT mRNA isoform with the tandem repeat of exon 2 is not a favourable context for translation reinitiation. In fact, it is believed that a long inter-cistronic region gives the ribosome enough time to recruit the factors that are required for translation initiation and that had been lost by the ribosome while it was translating the upstream ORF.

The reduced translation of the downstream ORF in the COT mRNA isoform with the tandem repeat of exon 2 is entirely due to the presence of the AUG codon in the first copy of exon 2, and not due to the increased length of the 5’ UTR. In fact, when the first AUG codon was mutated to another codon (UUG), the rate of synthesis of the firefly luciferase was the same as that from the COT 5’ UTR from the normal isoform (compare bars C and E in figure 4.6). The possibility that the mutation of the A nucleotide of the AUG codon to a T nucleotide in construct pGL3-Mut COT 5’ UTR could drastically alter a possible secondary structure of the UTR is quite unlikely. It would be interesting to see if the tandem repeat of exon 2 in COT mRNA decreases the efficiency of translation of full length protein in vivo as well. This could be investigated by determining the amount of COT total mRNA and protein in SHR and WKY. If the repeat of exon 2 lowers the efficiency of translation of the full length ORF in vivo, then the ratio between the amount of COT protein and COT total mRNA should be lower in WKY compared to SHR.

The fact that the AUG codon present in the first repeat of exon 2 decreased the rate of translation of the downstream AUG suggested that the first AUG of the tandem repeat in the
COT mRNA isoform with the tandem repeat of exon 2 is indeed used for translation, at least in Hela cells, and indicated that a truncated protein might be synthesised. Short peptides encoded by uORFs have already been detected (Hackett PB 1986; Raney A 2000). Furthermore, reporter genes have been fused to uORFs and the corresponding protein was synthesised. It would be interesting to see whether a truncated protein product is synthesised in rats that produce the COT mRNA isoform with the tandem repeat of exon 2 (WKY and Sprague-Dawley), but not in rats that do not produce this isoform (SHR).

As far as the SA 5’UTRs are concerned, the tandem repeat of exon 2 alone did not affect the rate of synthesis of the reporter gene (compare bars B and C of figure 4.10). In contrast, the simultaneous repetition of exons 2, 3 and 4 decreased the amount of firefly luciferase by about 15-fold (compare bars B and D in figure 4.10). Surprisingly, mutation of the translational start codon in the first copy of exon 3 did not improve the efficiency of translation of the UTR with the tandem repeat of exons 2, 3 and 4, suggesting that other factors decrease the efficiency of translation. Possible candidates are the AUG codons located between the mutated AUG codon and the AUG encoding the firefly luciferase protein. As can be seen in figure 4.9 C, there are 14 such codons, all out of frame from the AUG translational start codon. This shows that the situation in the SA mRNA 5’ UTRs is more complicated than that in COT mRNA. In fact, with the SA 5’ UTR with the tandem repeat of exons 2, 3 and 4, a ribosome that is scanning the mRNA in a cap-dependent manner will encounter 18 AUG codons before reaching the AUG linked to the firefly ORF instead of only 4 AUG codons as is the case for the normal SA mRNA. Thus, it is likely that one of these 18 upstream AUG codons would be selected for translation initiation. When the AUG in the first copy of exon 3 is mutated to a UUG codon, this leaves another 17 AUG codons upstream of the translational initiation codon of firefly.

All the conclusions about the inhibition of translation of the main ORF by repeated exons have been made on the assumption that the various 5’ UTRs do not alter mRNA stability. For instance, if the mRNA isoforms containing the repeated exons have a shorter half-life, the resulting reduced mRNA level would explain the lower firefly luciferase signal. But the fact that mutating the AUG codon in the first repeat of COT exon 2 restored the level of luciferase, argues against a possible change in mRNA stability. This of course, could still be possible in the case of the SA 5’ UTR with the tandem repeat of exons 2-3-4. One way of investigating this further could be the mutation of other AUG codons in this UTR and the effect that these mutations have on the amount of firefly luciferase.
The simple fact that the presence of repeated exons decreases the synthesis of full length protein is not necessarily harmful. This depends on whether the protein level in the absence of exon repetition is optimal, sub-optimal or excessive. For example, if a particular protein is too abundant in the cell, then exon repetition in that gene could have a useful effect by lowering the protein concentration to a more optimal level. For example, many proto-oncogenes contain uORFs in the 5'UTR. This is thought to tightly regulate the amount of protein that is produced by translation of the mRNA. This of course does not take into account the possible harmful effect of truncated proteins that might be synthesised from mRNA isoforms containing repeated exons, when the exon containing the translational start codon is among the repeated exons.
CHAPTER 5
MECHANISM OF EXON REPETITION

5.1. INTRODUCTION

5.1.1. Characteristics of exon repetition

As mentioned before, the most likely explanation for the presence of repeated exons in mRNA is that the exons are contributed by two pre-mRNA molecules. The exons from two distinct pre-mRNA molecules could be joined by trans-splicing. In higher eukaryotes, pre-mRNA molecules are normally processed through cis-splicing. Hence, if we assume that the mechanism for exon repetition is trans-splicing, the genes that exhibit exon repetition must possess special features that allow trans-splicing to efficiently compete with cis-splicing. Moreover, the mechanism must explain all the characteristics of exon repetition which are:

- Allele-specificity
- Tissue-specificity (in the case of the SA gene)
- Gene-specificity
- Exon-specificity

*Allele-specificity:* exon repetition in COT and SA is only observed with the WKY alleles and the characteristic is determined by cis-acting elements.

*Tissue-specificity:* exon repetition is tissue-specific in the SA gene. It is observed in kidney but not in liver (Frantz SA 1999).

*Gene-specificity:* the vast majority of genes do not seem to exhibit exon repetition. For example out of 545 genes studied by comparing EST sequences with genomic sequences, only one EST from the FBXO7 gene was found to contain a tandem repeat of exon 2 (Hide WA 2001).

*Exon-specificity:* in all the cases of exon repetition which have been discovered so far, not all the exons are repeated. In contrast, in every given gene, only certain exons are repeated whereas all the remaining exons are present in single copy in the mRNA. As shown in chapter 3, in the COT gene the repeated exons are allele-specific (see figure 3.26).
particular, in WKY only exon 2 is repeated whereas in Sprague-Dawley repetition involves either exon 2 alone or exons 2-3 (see figure 3.23).

5.1.2. Possible mechanisms for generating exon repetition

Several mechanisms that can allow the production of mRNA isoforms containing repeated exons can be envisaged. The observation that exon repetition in SA is both allele-specific and tissue-specific seems to suggest that several factors or elements are required. This could also explain why exon repetition seems to be so rare. A detailed description of four hypothetical mechanisms that could cause exon repetition is given below. Obviously, other mechanisms are still possible.

**Mechanism 1. Presence of a cryptic promoter and pause sites**

The most likely explanation for the presence of mRNA isoforms with repeated exons is that they are derived from the contribution of two pre-mRNA molecules. One likely possibility is that one pre-mRNA molecule is spliced in \textit{trans} with the second molecule. \textit{Trans}-splicing is a very common phenomenon in lower eukaryotes, such as trypanosomes and nematodes. For example in trypanosomes, virtually all mRNA molecules have a common sequence at their 5' end. This sequence derives from an RNA molecule called the spliced leader (SL), and is added to all pre-mRNAs by \textit{trans}-splicing (Murphy WJ 1986; Sutton RE 1986). The SL molecule has a 5' splice site but lacks a downstream 3' splice site. In contrast, the pre-mRNAs have a 3' splice site but lack an upstream 5' splice site. Since the 5' splice site is uncoupled from the 3' splice site, the only way for these molecules to undergo splicing is by \textit{trans}-splicing. The SL RNA provides the 5' splice site whereas the 3' splice site is contributed by the mRNA molecule. Of course, exon repetition is very different from the addition of SL to mRNAs in trypanosomes. In fact, if exon repetition is due to \textit{trans}-splicing, then the reaction is between two pre-mRNA molecules that have been transcribed (or are being transcribed) from the same gene whereas in the case of SL, the two RNA molecules that undergo \textit{trans}-splicing are transcribed from different genes. Interestingly, the mod(mdg4) gene in Drosophila produces several mRNA isoforms by \textit{trans}-splicing and it appears that transcription of specific groups of exons is due to putative promoters (Dorn R 2001).

If we assume that exon repetition is caused by intra-allelic \textit{trans}-splicing and that this requires the presence of a 5' splice site uncoupled from a downstream 3' splice site, it has to be explained how pre-mRNA molecules possessing only a 5' splice site or a 3' splice
site are produced. For example in SA, a pre-mRNA molecule with a 3' splice site uncoupled from an upstream 5' splice site could be produced by a cryptic promoter in intron 1 as shown in figure 5.1. Since transcription from such a promoter would start downstream of exon 1 (and its associated 5' splice site), the pre-mRNA molecule being transcribed would possess the 3' splice site preceding exon 2, but would lack the 5' splice site associated with exon 1. The allele-specificity of exon repetition in SA could be explained if, for example, the putative cryptic promoter was disrupted in the SHR SA allele. For example, it could be disrupted by the LINE element which is only present in intron 1 of the SHR SA allele but not in the WKY SA allele (Frantz SA 1996). A mechanism based on a cryptic promoter could also account for the tissue-specificity of exon repetition of SA by assuming that the promoter element is active in the proximal tubules of the kidney and inactive in liver (tissue-specific promoter).

The existence of pre-mRNA molecules with a 5' splice site but lacking a downstream 3' splice site could be explained by the presence of pause sites in the introns positioned downstream of the repeated exons. Pause sites are DNA elements at which RNA polymerase II, the enzyme that transcribes protein-encoding genes, pauses (Keene RG 1999; Yonaha M 1999; Meininghaus M 2000). For example, the hypothetical presence of pause sites in introns 2 and 4 of the SA gene would explain why only exon 2 alone or exons 2-3-4 are repeated.

Mechanism 2. Presence of complementary sequences in pre-mRNA

It has been shown that complementary sequences on two different pre-mRNA molecules favour trans-splicing in vitro (Konarska MM 1985; Solnick 1985). Moreover, trans-splicing between two pre-mRNA molecules designed to base-pair to one another has also been demonstrated in transient co-transfection experiments (Puttaraju M 1999) and in vivo (Puttaraju M 1999; Chao H 2003). This strategy is being investigated as a way of correcting genetic diseases (Liu X 2002; Dallinger G 2003; Mansfield SG 2003). Finally, the same findings were confirmed in vivo (Chao H 2003).

The allele-specificity observed in exon repetition of SA and COT could be explained by the presence of complementary sequences in the WKY alleles but not in the SHR alleles. The lack of exon repetition of SA in liver could be explained by the low level of expression of SA in liver compared to kidney. Since two different pre-mRNA molecules are required for trans-splicing, it is very likely that its efficiency will depend on the concentration of pre-mRNA. The assumption is that the different abundance of SA mRNA in kidney and in liver is due to an increased transcription rate in kidney, and not to a
Figure 5.1. Possible mechanism for exon repetition in SA, based on the presence of a cryptic promoter inside intron 1 (represented by the blue arrow) and pause sites for RNA polymerase II in introns 2 and 4 (represented by the purple ovals). The red arrow is the natural promoter of the SA gene. The black horizontal line represents the first four introns of the SA gene. The exons are numbered and shown as vertical white rectangles. The curved red lines are nascent transcripts that have originated from the natural promoter. The blue curved line is a nascent transcript that was transcribed from the cryptic promoter. The purple boxes represent RNA polymerase II molecules that are stalled at pause sites. The light blue box is an actively elongating RNA polymerase II molecule. The red transcripts contain a 5' splice site but no downstream 3' splice site because it has not been transcribed yet. In contrast, the blue transcript contains a 3' splice site but no upstream 5' splice site because exon 1 with its relative 5' splice site is missing from this transcript. The uncoupled splice sites could be spliced to one another by trans-splicing. If the red transcript on the left is trans-spliced to the blue transcript, an mRNA isoform with a tandem repeat of exon 2 is produced. In contrast, if the red transcript on the right is trans-spliced to the blue transcript, repetition of exons 2, 3 and 4 occurs.
difference in mRNA stability in different tissues. Another explanation for tissue-specificity is the concomitant requirement of complementary sequences and a trans-acting factor that is specifically expressed in kidney but not in liver. Alternatively there could be a liver-specific trans-acting factor that binds one of the complementary sequences and thereby interferes with the base-pairing of the two sequences.

**Mechanism 3. Presence of an exonic splicing enhancer**

An exonic splicing enhancer (ESE) is a sequence inside an exon that stimulates splicing. ESEs are bound by proteins that belong to the SR (serine arginine rich) family. It was previously shown that a mutation in a putative ESE in exon 2 of COT abrogated exon repetition (Caudevilla C 2001).

In the SA gene there is a point mutation at the end of exon 3 that differentiates the WKY and SHR SA alleles. The lack of exon repetition from the SHR SA allele could be due to the abrogation of an ESE by the point mutation in exon 3. A mechanism based on an ESE and its cognate protein could also account for the tissue-specificity by assuming that the SR protein that binds to the ESE is expressed in a tissue-specific manner. For example, this putative protein could be expressed in the proximal tubule cells but not in liver.

**Mechanism 4. Weak 5’ splice site and/or poorly defined exon upstream of the repeated exon/s**

The strength of the splice sites that take part in a splicing reaction influences both splicing efficiency and kinetic. The propensity of the exons to be spliced is also important. If the exon located 5’ to the repeated exon is poorly defined, for example because it is very short, then cis-splicing might be very inefficient. In these conditions, trans-splicing could efficiently compete against cis-splicing. A diagrammatic representation of this mechanism is shown in figure 5.2.

A mechanism based on the presence of a poorly defined exon upstream of the repeated exons has been suggested by Caudevilla and co-workers (Caudevilla C 1998). Furthermore, competition between cis-splicing and trans-splicing has been observed in mammalian cells (Eul J 1995). To support this hypothesis is the observation that exon 1 is short in both SA and COT (67 and 30 nucleotides, respectively). Moreover, neither SA nor COT has a consensus 5’ splice site associated with exon 1.
Figure 5.2 Schematic representation of a possible mechanism for generating mRNA isoforms with repeated exons. It is based on the presence of a weak 5’ splice site upstream of the exon that undergoes repetition and/or a poorly defined exon upstream of the repeated exon. A portion of the gene is shown as a black horizontal line. Exons are shown as green rectangles. Two nascent pre-mRNA molecules are also shown (one red, the other blue) which are attached to the enzyme RNA polymerase II (shown as a light blue square). (A) Diagrammatic representation of the portion of a gene from which only the normal mRNA isoform is produced. Splicing between exon 1 and 2 on the same molecule proceeds very efficiently (represented by the thickness of the arrows) and this prevents the 5’ splice site (associated with exon 2) on one pre-mRNA molecule to trans-splice to the 3’ splice site on the other molecule. (B) Diagrammatic representation of the portion of a gene that produces not only the normal mRNA isoform but also an mRNA isoform containing a tandem repeat of exon 2. Cis-splicing between exons 1 and 2 on the same molecule proceeds very slowly and inefficiently (shown by the thin arrows). This is due to the presence of a poorly defined exon 1 and/or a weak 5’ splice site associated with exon 1. The slow kinetic of cis-splicing allows the 3’ splice site of exon 2 on one molecule, to be trans-spliced to the 3’ splice site of exon 2 from the other nascent pre-mRNA molecule. This trans-splicing reaction leads to the formation of an mRNA isoform in which exon 2 is tandemly repeated.
5.2. RESULTS

5.2.1. Cloning of the SHR SA allele

As shown in chapter 3, exon repetition is allele-specific. This offers the opportunity to pinpoint the determinants that cause exon repetition. In fact, it is quite likely that the inability of one allele to produce mRNA isoforms containing repeated exons is due to mutations (either point mutations or deletions/insertions) that disrupt DNA elements that allow exon repetition. Therefore, it was decided that the first step to unravel the mechanism underlying exon repetition would be to clone and sequence the SHR and WKY SA alleles.

Three λ (EMBL-3) clones that contained the majority of the SHR SA allele were kindly provided by Simon Frantz and Nilesh Samani and are shown in figure 5.3. The insert of λ-GC1 was sub-cloned by Simon Frantz into pBluescript SK (+) that had been previously digested with the restriction endonuclease XhoI. Since the insert of λ-GC1 contained an internal XhoI recognition sequence, two clones were obtained and were named AT6 and AT9. In order to clone the remaining part of the SHR SA allele, DNA was prepared from clones λ-50 and λ-3′4 as described in materials and methods. The inserts were excised with the restriction endonuclease XhoI and sub-cloned into pBluescript SK (+). The gap regions between λ-50 and λ-GC1 and between λ-GC1 and λ-3′4 were amplified by PCR using SHR genomic DNA as template. The amplification reactions were performed with the Expand Long Template PCR System and PCR products were cloned into pGEM® -T Easy Vector. In this way, a number of recombinant plasmids were obtained that contained the entire genomic sequence of the SHR SA allele. The plasmids are shown in figure 5.4 and listed in table 5.1.
Figure 5.3. Diagrammatic representation of the SHR SA allele and the λ clones that contain genomic portions of the gene. The exons are represented by vertical lines and are numbered from 1 to 15. The white box in intron 1 represents the LINE element. The rectangles under the SA gene represent the inserts of three lambda clones (named λ-50, λ-GC1 and λ-3'4) that contain genomic portions of the SHR SA allele.
Figure 5.4. Schematic representation of the plasmid clones that span the entire SHR SA allele. The contig starts 3,867 bp upstream of exon 1 and ends 2,278 bp downstream of the TAA stop codon which is located in exon 15 (the last exon of the SA gene). Clones SHR 50 and SHR 3’4 were subcloned from \( \lambda \) clones. Clones PCR 1-5 were derived by cloning PCR products obtained by amplification from SHR genomic DNA. Clones AT6 and AT9 were kindly provided by Simon Frantz. The SA gene is composed of 15 exons which are represented by the vertical black lines. The introns are represented by the black horizontal line. The white box inside intron 1 is the LINE element which is found only in intron 1 of the SHR SA allele but not in the WKY SA allele.
### Table 5.1 List of the plasmid clones whose inserts form a contig that spans the entire genomic sequence of the SHR SA allele.

<table>
<thead>
<tr>
<th>NAME OF CLONE</th>
<th>CLONING VECTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHR 50</td>
<td>pBlue Script KS</td>
</tr>
<tr>
<td>SHR PCR 1</td>
<td>pGEM-T Easy</td>
</tr>
<tr>
<td>SHR AT6</td>
<td>pBlue Script KS</td>
</tr>
<tr>
<td>SHR AT9</td>
<td>pBlue Script KS</td>
</tr>
<tr>
<td>SHR PCR 2</td>
<td>pGEM-T Easy</td>
</tr>
<tr>
<td>SHR PCR 3</td>
<td>pGEM-T Easy</td>
</tr>
<tr>
<td>SHR PCR 4</td>
<td>pGEM-T Easy</td>
</tr>
<tr>
<td>SHR PCR 5</td>
<td>pGEM-T Easy</td>
</tr>
<tr>
<td>SHR 3'4</td>
<td>pBlue Script KS</td>
</tr>
</tbody>
</table>

All the inserts were sequenced by a biotechnology company (Genterprise) with the Big Dye Terminator system. The contig is 32,311 in length. It starts 3,845 bp upstream of exon 1 and ends 2,281 bp downstream of the TAA stop codon which is located in exon 15. The gene is composed of 15 exons. The ATG translational start codon is located in exon 3. All the splice sites confirm to the GT-AG rule. The entire sequence of the SHR SA allele is shown in appendix A of this chapter.

### 5.2.2. Comparison between the DNA sequence of the SHR and WKY SA alleles

The sequence of the SHR SA allele was aligned with that of the WKY SA allele (the WKY SA allele was cloned by Jian-hua Jia using the Expand Long template PCR system, Roche) using the Gene Jockey program. The two sequences, with the exception of a large insertion represented by a LINE element (long interspersed sequence) in intron 1 of the SHR SA allele were strikingly similar, having an identity of over 99%. Except for the LINE element there were only minor differences represented by a few point mutations. In particular there were 39 differences in intron 1 (25 were SNPs whereas the remaining were differences in the number of di or tri-nucleotide repeats at microsatellite sequences). The only difference in the coding sequence was the point mutation located at the end of exon 3 (A in SHR, G in WKY), which was the same as the one used for the experiments relative to the SA heterozygotes described in chapter 3 (see figure 3.16). This is a sense mutation that does not change the amino acid sequence of the SA protein.
In a 1.2 Kb region upstream of exon 1 there are six single nucleotide polymorphisms (SNPs). Another difference is represented by 5 A nucleotides in an oligo(A) tract.

The differences in the intronic sequences are shown in figure 5.5. They consist of SNPs or deletions and insertions of only 1-4 nucleotides.

5.2.3. Mechanism 1. Cryptic promoter and pause sites.

Is there a cryptic promoter in intron 1 of the WKY SA gene?

As mentioned in the introduction of this chapter, one possible mechanism that could elicit exon repetition in SA involves the presence of a cryptic promoter in intron 1 of the WKY SA allele and pause sites in introns 2 and 4. The hypothetical promoter element might be disrupted in the SHR SA allele by the LINE element and this would account for the allele-specificity of exon repetition of SA. Furthermore, by assuming that the promoter in intron 1 exhibits transcriptional activity only in the proximal tubule cells, this mechanism could also explain why exon repetition of SA is found in the proximal tubule cells but not in the liver. To investigate the presence of a promoter region, the entire intron 1 of the WKY SA allele was PCR amplified, cloned into the promoter-less vector pECFP-1 and the construct was named WKY SA intron 1 pECFP-1. The pECFP-1 vector contains an open reading frame (ORF) encoding the enhanced cyan fluorescent protein (ECFP). This gene is an enhanced cyan fluorescent variant of the Aequorea victoria green fluorescent protein gene (GFP). The ECFP gene contains six amino acid substitutions. One substitution changes the fluorescent properties (excitation and emission wavelengths). The other five substitutions enhance the brightness and solubility of the protein. In order to compare the transcriptional activity of WKY intron 1 with that of the natural promoter of the WKY SA allele, the latter was also cloned into the same vector and the construct was named WKY SA NP pECFP-1. One µg of DNA of each construct was separately transfected into 293 cells with calcium phosphate method. Since the inserts had a different size, the amount of DNA transfected for each clone was normalized, so that approximately the same numbers of plasmid molecules were transfected for each construct. 293 cells were chosen because it was known that a minigene spanning the region comprised between exons 1 and 3 of the WKY SA allele could recapitulate exon repetition in transient transfection experiments (Janhua Jia, personal communication). Therefore, if a putative promoter located in intron 1 was involved in exon repetition, it would have to be active in this cell line. The vector pECFP-
Figure 5.5. Result of the alignment between the SHR and WKY SA alleles. The entire SA gene is shown excluding the promoter region. The exons are represented by the vertical bars and the exon’s number is reported underneath. The vertical arrows above each intron indicate the number of differences present between SHR and WKY in each intron. The first number represents the number of SNPs whereas the number in brackets, when present, indicates the number of short (1-4 nt) deletions and insertions.
N1 in which the ECFP gene is under the control of the strong CMV promoter was used as a positive control. First of all, cells were observed under a fluorescence microscope 48 hours after transfection. Cells that have been transfected and that are expressing sufficient levels of ECFP protein emit fluorescence upon irradiation at the appropriate wavelength. Only the cells that were transfected with pECFP-N1 were green. In contrast, cells that had been transfected with constructs WKY SA intron 1 pECFP-1 and WKY SA NP pECFP-1 tested negative. In an attempt to detect the ECFP protein by western blot, total proteins were extracted with RIPA buffer and separated by SDS-PAGE. Proteins were transferred to a membrane which was probed with a monoclonal antibody specific for GFP (which cross-reacts with ECFP) to detect the amount of ECFP. A positive signal was obtained after 5 seconds of exposure only from the cells that had been transfected with plasmid pECFP-N1 (CMV promoter) as shown in figure 5.6. Even after ten minutes of exposure, there was no signal from the other two samples (data not shown). These data could be explained in two ways. First of all, it was possible that both intron 1 and the natural promoter of SA were inactive in 293 cells. On the other hand, their transcriptional activity could have been too low to be detected by fluorescence microscopy or western blot. In order to discriminate between these two possibilities, a much more sensitive promoter assay, based on the reporter gene firefly luciferase was utilised. This assay can detect less than 1 femtogram of firefly luciferase, corresponding to approximately 10⁻²⁰ moles of protein. The natural promoter of the WKY SA allele was excised from construct WKY SA NP pECFP-1 and sub-cloned into the promoter-less pGL3-Basic vector and the construct was named WKY SA NP pGL3. pGL3-Basic vector contains the reporter gene firefly luciferase. Intron 1 of the WKY SA gene was also cloned into pGL3-Basic vector and the construct was named WKY SA intron 1 pGL3. Either 0.5 or 1.0 μg of pGL3 based constructs was used to transfet 293 cells. In order to normalise the assay for the efficiency of transfection and protein extraction, 2 ng of plasmid phRL-CMV, which contains the renilla luciferase gene driven by the CMV promoter were co-transfected together with the firefly luciferase constructs. The experiment was performed in duplicate for each construct and for each concentration and the results are shown in table 5.2.
Figure 5.6 Evaluation of the promoter activity of intron 1 of the WKY SA allele. Cells were transfected with construct WKY SA intron 1 pECFP-1, or construct WKY SA NP pECFP-1 (transcription is driven by the WKY SA natural promoter) or construct pECFP-N1 (transcription is driven by the CMV promoter). 48 hours after transfection, total proteins were extracted and resolved by SDS-PAGE. A western blot was performed with an anti-GFP antibody that cross-reacts with ECFP.
Table 5.2 Transcriptional activity of intron 1 (construct WKY SA intron 1 pGL3) and the natural promoter of the WKY SA gene (construct WKY SA NP pGL3). The values represent the ratio between the signals obtained from firefly luciferase and renilla luciferases, multiplied by a factor of 1,000. The data relative to pGL3-Basic vector, which was used as negative control, are also shown.

<table>
<thead>
<tr>
<th>CONSTRUCT</th>
<th>Amount of DNA (µg)</th>
<th>F.fly/Ren 1st sample X1000</th>
<th>F.fly/Ren 2nd sample X1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>pGL3-Basic vector</td>
<td>0.5</td>
<td>12.6</td>
<td>11.6</td>
</tr>
<tr>
<td>WKY SA intron 1 pGL3</td>
<td>0.5</td>
<td>2.36</td>
<td>2.26</td>
</tr>
<tr>
<td>WKY SA NP pGL3</td>
<td>0.5</td>
<td>923</td>
<td>983</td>
</tr>
<tr>
<td>pGL3-Basic vector</td>
<td>1.0</td>
<td>26.8</td>
<td>24.2</td>
</tr>
<tr>
<td>WKY SA intron 1 pGL3</td>
<td>1.0</td>
<td>4.33</td>
<td>3.93</td>
</tr>
<tr>
<td>WKY SA NP pGL3</td>
<td>1.0</td>
<td>2106</td>
<td>1803</td>
</tr>
</tbody>
</table>

The average of each sample was then plotted and is shown in figure 5.7. From construct WKY SA NP pGL3, in which transcription is fired from the SA natural promoter, approximately an 80-fold increase in firefly luciferase activity compared to pGL3-Basic vector was obtained. In contrast, the firefly luciferase activity from construct WKY SA intron 1 pGL3, which contains intron 1 of the WKY Sa gene, was less than that of pGL3-Basic vector. The signal obtained from the promoter-less plasmid pGL3-Basic vector is probably due to a low transcriptional activity from the vector backbone. In fact, to reduce it, the vector was designed to contain a synthetic poly(A) signal/ transcriptional pause site.

Due to the considerable length of intron 1 (more than 5 Kbp), the detection of promoter activity by measuring a protein product might be problematic. In fact, an RNA molecule transcribed from intron 1 would probably have a long 5’ UTR, unless the putative promoter element is located at the very end of intron 1. It is quite likely that a long 5’ UTR would contain many AUG codons. Hence, it is likely that a ribosome scanning along this mRNA would start translation from an AUG codon located upstream of the initiation codon of the firefly luciferase ORF. Therefore, even if an mRNA containing the ORF for firefly luciferase was produced, the RNA molecule might not be efficiently translated into the firefly luciferase protein. To avoid this problem, the DNA sequence corresponding to intron 1 of the WKY SA gene was scanned with a promoter finder (http://bimas.dccn.nih.gov/molbio/proscan/) for possible promoter regions. The promoter finder identified a putative promoter element and its sequence and position inside intron 1 are shown in figure 5.8. The putative promoter was also present in intron 1 of the SHR SA...
Figure 5.7 Evaluation of transcriptional activity of intron 1 of the WKY SA gene. Promoter activity was measured indirectly by determining the luciferase activity of protein extracts from cells that had been transfected with construct WKY SA intron 1 pGL3. The promoter-less cloning vector pGL3-Basic vector was also transfected as a negative control. In order to compare the transcriptional activity of intron 1 with that of the natural promoter of the SA gene, construct WKY SA NP pGL3 was also transfected. Either 0.5 μg or 1.0 μg of firefly luciferase constructs were transfected in 293 cells. In order to normalise the assay for the efficiency of transfection and protein extraction, 2 ng of a plasmid phRL-CMV which contains the renilla luciferase gene driven by the CMV promoter were co-transfected with the firefly luciferase constructs. Transfection was carried out with the calcium phosphate method. Total proteins were extracted 24 hours after transfection and the activities of firefly and renilla luciferases were measured. The experiment was performed in duplicate for each construct and each concentration and the average value was plotted. The Y axis represents the ratio between the activity of firefly and renilla luciferases, multiplied by a factor of 1,000. (A) The experiment was carried out with 0.5 μg of firefly luciferase constructs. (B) The experiment was carried out with 1.0 μg of firefly luciferase constructs.
allele. The DNA sequence corresponding to the natural promoter of the SA gene was scanned with the promoter finder. Strangely, it failed to identify a promoter region. To test whether the putative promoter element possessed transcriptional activity, it was PCR amplified, cloned into pGL3-Basic vector and assayed as previously described. Again, pGL3-Basic vector and the natural promoter of the WKY SA allele were included in the experiment as negative and positive control respectively. Two μg of DNA of each construct were transfected together with 200 ng of plasmid phRL-CMV Vector for normalization. The intensity of the signal obtained from the putative promoter present in intron 1 was comparable to that of the vector alone, indicating that this DNA region was did not initiate transcription in 293 cells. The data are shown in table 5.3 and plotted in figure 5.9.

<table>
<thead>
<tr>
<th>CONSTRUCT</th>
<th>FIREFLY/RENILLA X 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>pGL3-Basic vector</td>
<td>8.6</td>
</tr>
<tr>
<td>WKY CP pGL3</td>
<td>4.6</td>
</tr>
<tr>
<td>WKY SA NP pGL3</td>
<td>382</td>
</tr>
</tbody>
</table>

Table 5.3. Transcriptional activity of the putative promoter region (construct WKY CP pGL3) in intron 1 and the natural promoter of the WKY SA gene (construct WKY SA NP pGL3). The values represent the ratio between the signals obtained from firefly luciferase and renilla luciferases, multiplied by a factor of 1,000. The data relative to pGL3-Basic vector, which was used as negative control, are also shown.

It was possible that a putative promoter was present in intron 1 but had not been detected by the promoter finder. To investigate this, a second approach was used. This method was based on the direct detection of RNA by RT-PCR and therefore was not affected by the obstacles due to translation of mRNA. For this assay, a genomic portion of the SA gene between exons 1 and 3 was cloned from both WKY and SHR into the promoter-less vector pECFP-1. The two constructs were named WKY 1-3 pECFP-1 and SHR 1-3 pECFP-1 respectively. To monitor the transcriptional background from the cloning vector backbone, a portion of the WKY SA gene comprised between exons 2 and 3 was also cloned into the same vector and used as negative control. This construct was named WKY 2-3 pECFP-1. All the constructs are depicted in figure 5.10. Each plasmid was separately transfected into Hela cells. These cells were suitable because exon repetition could be recapitulated by SA minigenes transiently transfected inside these cells (Jian-hua Jia personal communication). Transfection was carried out using the calcium phosphate
Figure 5.8 Putative promoter region identified inside intron 1 of the SA gene by the promoter finder: http://bimas.dcrt.nih.gov/molbio/proscan

(A) Diagrammatic representation of intron 1 of the SA gene of WKY. The putative promoter element is represented by the white arrow. Exons 1 and 2 are shown as white vertical rectangles. (B) Nucleotide sequence of the putative promoter.
Figure 5.9 Evaluation of transcriptional activity of a putative promoter inside intron 1 of the WKY SA gene. Promoter activity was measured indirectly by determining luciferase activity. 293 cells were transfected by the calcium phosphate method with 2 μg of firefly luciferase constructs and 200 ng of phRL-CMV Vector which contains an ORF encoding renilla luciferase driven by the CMV promoter. The promoter-less plasmid pGL3-Basic vector was used as negative control. Construct WKY SA NP pGL3 in which transcription of the luciferase gene is driven by the WKY SA natural promoter was used as positive control. Total proteins were extracted 24 hours after transfection and the activities of firefly and renilla luciferases were measured. The Y axis represents the ratio between the activity of firefly and renilla luciferases, multiplied by a factor of 1,000.
Figure 5.10 Evaluation of the transcriptional activity of SA intron 1. A genomic region comprised between exons 1 and 3 of SHR and WKY was cloned into the promoter-less vector pECFP-1 to yield constructs SHR 1-3 pECFP-1 and WKY 1-3 pECFP-1 respectively. The region comprised between exons 2 and 3 in WKY was also cloned into the same plasmid and used as negative control. (A) Diagrammatic representation of the cloning vector pECFP-1. (B) Schematic representation of the inserts that were cloned into the pECFP-1 vector. The names on the left refer to the names of the constructs. The exons are represented by the vertical white numbered rectangles. The introns are shown as black horizontal lines. The blue horizontal box inside intron 1 of SHR 1-3 pECFP-1 represents the LINE element.
method as described in materials and methods. RNA was extracted and cDNA synthesised as described in materials and methods. PCR amplification was performed with primers SA E2F and SA E3R and the result is shown in figure 5.11 A. The amplification of a PCR product from the RNA extracted from the cells that had been transfected with construct WKY 2-3 pECFP-1 (see lane 3 in figure 5.11 A) suggested that a significant amount of transcripts were transcribed from the vector backbone. To investigate this, a second PCR amplification was carried out in which primer pECFP-1F (it primes on the vector backbone just upstream of the insert) was used as the forward primer instead of primer SA E2F and the result is shown in figure 5.11 B. The amount of PCR product obtained with primer pECFP-1F was similar to that obtained with primer SA E2F (compare figure 5.11 A with figure 5.11 B, in which amplification was carried out with the same number of cycles). Overall, the data suggest that there is transcription from the vector backbone.

5.2.4. Mechanism 2. Complementary sequences

As mentioned in the introduction, the presence of complementary sequences on two separate pre-mRNA molecules could facilitate their interaction through base-pairing. Since the mRNA isoforms with repeated exons are likely to arise from the contribution of two pre-mRNA molecules, base-pairing would significantly increase the chance of the two pre-mRNA molecules to react with one another by bringing them in close proximity. From the genomic sequence of the WKY SA gene, complementary regions were found. In particular, a microsatellite in intron 1 (dinucleotide AC repeated 20 times in WKY) could potentially base-pair with a microsatellite in intron 4 (dinucleotide TG repeated 21 times in WKY). The same microsatellites were also present in the SHR SA allele. In particular, the AC dinucleotide was repeated 24 times instead of 20, whereas the TG dinucleotide was present with 21 copies like in WKY. Since the two microsatellite sequences are present in both the WKY and SHR SA alleles, a mechanism based exclusively on complementary sequences in the pre-mRNA molecules cannot account for the allele-specificity of exon repetition in SA. Moreover, this mechanism could not account for the tissue-specificity of exon repetition in SA, unless it was assumed that there is a threshold concentration of pre-mRNA above which the mRNA isoforms with repeated exons can be produced. Since SA mRNA is more abundant in kidney than in liver, it is possible that the concentration of SA pre-mRNA is also higher in kidney than in liver. Alternatively, there could be a tissue-specific trans-acting factor expressed only in the proximal tubule cells and that binds to one of the microsatellite regions, thus impeding base-pairing.
Figure 5.11. Analysis of promoter activity by RT-PCR. 293 cells were transfected with 0.5 μg of constructs WKY 1-3 pECFP-1, SHR 1-3 pECFP-1 and WKY 2-3 pECFP-1. Total RNA was extracted 48 hours after transfection and cDNA was synthesised by RT. SA cDNA was then amplified by PCR. (A) PCR performed with primers SA E2F and SA E3R. 1, WKY 1-3 pECFP-1. 2, SHR 1-3 pECFP-1. 3, WKY 2-3 pECFP-1. 4, untransfected cells. 5, PCR blank. (B) PCR performed with primers pECFP-1 and SA E3R. 1, WKY 1-3 pECFP-1. 2, SHR 1-3 pECFP-1. 3, WKY 2-3 pECFP-1. 4, untransfected cells. 5, PCR blank.
In conclusion, since complementary sequences are present in both the WKY and SHR SA alleles, their involvement in exon repetition is very unlikely.

5.2.5. Mechanism 3. Exonic splicing enhancers

As mentioned before, exon repetition in COT was proposed to be dependent on the presence of an exonic splicing enhancer in exon 2 (Caudevilla C 2001). First of all, the authors analysed COT mRNA in several different mammalian species (human, bovine, pig, mouse, Cos7 cells derived from monkey) and found exon repetition of COT only in rat. They aligned the COT cDNA sequences from all the species investigated and found that a sequence corresponding to a putative exonic splicing enhancer was present in rat but absent from all the other species examined (GAAGAAG in rat, AAAAAAA in the other species). This sequence was positioned 117 nt downstream of the 5' end of exon 2 in rat. They showed that the putative exonic splicing enhancer stimulated trans-splicing in vitro. They then constructed a COT sub-minigene consisting of exon 2 with flanking intronic sequences (125 nt of intron 1 and the first 25 nt of intron 2), followed by the last 115 nt of intron 2 and exon 3. They transiently transfected the sub-minigene into Hela cells and harvested total RNA, 24 hours post-transfection. RT-PCR experiments with primers specific for the COT exonic sequences of rat amplified two PCR products. One amplicon contained exon 2 joined to exon 3 (normal exon organisation). The other PCR product contained two copies of exon 2 which were joined to exon 3 (repetition of exon 2). The putative exonic splicing enhancer in exon 2 of the sub-minigene was then mutated from GAAGAAG to the sequence AAAAAA and the corresponding minigene failed to recapitulate exon repetition in transient transfection experiments. It was therefore concluded that the ESE in exon 2 was required for exon repetition in COT mRNA. The DNA sequence of exon 2 in the COT gene of SHR was shown to be identical to that of exon 2 of COT from WKY and Sprague-Dawley and therefore the putative exonic splicing enhancer sequence (GAAGAAG) is present in all three rats. The DNA sequence chromatograph relative to the ESE in exon 2 of the SHR COT gene is shown in figure 5.12. This demonstrates that the presence of the putative ESE sequence is not sufficient to elicit exon repetition in COT mRNA in vivo. It is not known whether the SHR COT allele is unable to produce exon repetition because it has a DNA element that inhibits exon repetition, or because it is lacking a second DNA element required for exon repetition. The assumption that in order to have exon repetition, several DNA elements are required could explain why exon repetition seems to be so rare in genes.
Figure 5.12 Exon 2 of the SHR COT allele contains the putative exonic splicing enhancer (ESE). This sequence was shown to be required for exon repetition of COT (Caudevilla et al., 2001). COT mRNA was amplified by RT-PCR from SHR kidney cDNA using primers COT E1F and COT E7R. The PCR product was purified from agarose gel and sequenced with primer COT E1F. A portion of the sequence chromatograph is shown and the putative ESE sequence (GAAGAAG) is highlighted by the brown rectangle.
5.2.6. Mechanism 4. Weak exon and 5' splice site of the exon 5' of the repeated exons

As mentioned before, the most likely explanation for the existence of the mRNA isoforms that contain tandemly repeated exons is that their exons are derived from two pre-mRNA molecules. One possibility is that two pre-mRNAs could be joined together by trans-splicing. Trans-splicing is simply a splicing reaction in which the 5' and 3' splice sites are provided by two different pre-mRNA molecules. If trans-splicing is the mechanism that generates the mRNA isoforms with repeated exons, then the genes that exhibit exon repetition must possess peculiar characteristics so that trans-splicing can efficiently compete against cis-splicing. Competition between trans and cis-splicing has already been shown in vitro (Solnick 1985; Puttaraju M 1999) and also in transient transfection assays with cell lines (Puttaraju M 1999). For instance, the presence of a weak 5'splice site upstream of the repeated exon/s might thwart cis-splicing and in these conditions a strong 5' splice site located on a different pre-mRNA molecule could be spliced in trans to the 3' splice site. Interestingly, the sequence of the 5' splice site associated with exon 1 in COT and SA does not match the consensus sequence. It should be pointed out that in most cases of exon repetition, the repeated exon/s are spliced to the second exon (Caudevilla C 1998; Frantz SA 1999; Takahara T 2000; Flouriot G 2002; Takahara T 2002). In this case, if the 5’ splice site associated with exon 1 is weak, there are no other 5’splice sites from the same gene that can be spliced to the 3’ splice site associated with exon 2. This might explain why in most instances exon 2 is repeated.

The length of the upstream exon could also play an important role. For example it was shown that a 3’ splice site that normally undergoes cis-splicing can undergo trans-splicing if it is placed in the right context, more precisely in an outron which is an intron that lacks a well defined 5’ splice site. When an untranslated 27 nt exon was placed upstream of the 3’ splice site, cis-splicing occurred with very low efficiency whereas trans-splicing remained dominant (Conrad R 1993). The authors speculated that the short length of the exon might cause it to splice inefficiently to the downstream exon. Interestingly, exon 1 is rather short in both SA and COT (66 bp and 30 bp respectively). A short and poorly defined exon might be spliced very weakly, for example because it is unlikely to contain exonic splicing enhancers which stimulate splicing and exon inclusion in mRNA. This mechanism was postulated in the discussion of one paper dealing with exon repetition (Caudevilla C 1998). The low splicing efficiency and/or kinetic could probably be enhanced by the concomitant
presence of a weak 5' splice site associated with a short exon 1. In order to determine if the 5' end of the SA gene (exon 1 and relative 5' splice site) played a role in exon repetition, a series of mutant constructs were created from construct WKY 1-3 ECFP-N1 which was kindly provided by Janhua Jia. This construct contains a genomic portion of the WKY SA gene comprised between exons 1 and 3. To investigate a possible effect of exon 1 of SA, both its length and sequence were changed. In one mutant, the length of exon 1 was increased by fusing the human β-globin gene exon 1, to exon 1 of SA. This construct was named WKY 1-3 Glob 1 ECFP-N1. In a second mutant, the entire exon 1 of SA was removed (except for the most 3' 6 nucleotides) and replaced with exon 1 of the human β-globin gene. This construct was named WΔ 1-3 Glob 1 ECFP-N1. As mentioned before a weak 5' splice site associated with exon 1 might allow trans-splicing to compete efficiently against cis-splicing. To investigate this, a mutant was constructed in which the sequence of the 5' splice site associated with exon 1 of SA was mutated to a consensus 5' splice site (canonical sequence CAG/GTAAGT) and the construct was named WKY 1-3 5' SS mutant ECFP-N1. All the constructs are shown in figure 5.13. First of all, the effect of exon 1 of SA (both its sequence and length) was studied. For this purpose, clones WKY 1-3 Glob 1 ECFP-N1 and WΔ 1-3 Glob 1 ECFP-N1 were transiently transfected in Hela cells. The experiment was performed in duplicate for each construct with 1 µg of DNA. Total RNA was extracted 72 hours after transfection and cDNA was synthesised as described in materials and methods. To evaluate the amount of total SA mRNA and that of the mRNA isoform with a tandem repeat of exon 2, two different PCR reactions were performed. The first PCR reaction was performed with primers SA 2-2 junction and SA E3R. This set of primers amplifies a PCR product exclusively from the mRNA isoform that contains a tandem repeat of exon 2. DNA amplification was carried out with 35 cycles and the result is shown in figure 5.14 A. The expected PCR product was amplified from the mRNA produced by both clones. The second PCR reaction was performed with primers β Glob E1F and SA E3R and amplification was carried out for 30 cycles. Since the efficiency of exon repetition from minigenes is quite low, only the PCR product corresponding to the mRNA isoform with the normal exon organisation was detected as shown in figure 5.14 B. The difference in the size of the PCR products obtained from clones WKY 1-3 Glob 1 ECFP-N1 and WΔ 1-3 Glob 1 ECFP-N1 is due to the different length of exon 1 in the two clones (the sequence corresponding to SA exon 1 is missing in clone WΔ 1-3 Glob 1 ECFP-N1). From these results it can be concluded that neither the length of exon 1 of SA nor its
Figure 5.13 Diagrammatic representation of the SA minigene constructs that contain mutations at the 5' end. Construct WKY 1-3 ECFP-N1 was kindly provided by Jian-hua Jia. In construct WKY 1-3 5' SS mutant ECFP-N1, the 5' splice site associated with exon 1 was mutated to a consensus splice site (CAG/GTAAGT) and is shown as a star. In construct WKY 1-3 Glob 1 ECFP-N1, exon 1 was lengthened by fusing exon 1 from the human β-globin gene upstream of exon 1 of SA. Finally, in construct WΔ 1-3 Glob 1 ECFP-N1, SA exon 1 was removed (except for the last 5 nucleotides) and replaced with the exon 1 of the human β-globin gene. The ORF encoding the cyano fluorescent protein is represented by the black box. The SA exons are shown as white boxes, β-globin exon 1 is shown as a grey box. The CMV promoter is shown as a black arrow. The polylinker region is represented by the dotted line. Finally, the SA introns are shown as black horizontal lines.
Figure 5.14. Analysis of the mRNA isoforms produced by constructs WKY 1-3 Glob 1 ECFP-N1 and WΔ 1-3 Glob 1 ECFP-N1 in a transient transfection experiment in Hela cells. Plasmid DNA was transfected into Hela cells. The experiment was performed in duplicate for each of the two constructs. Lanes 1 and 2 refer to construct WKY 1-3 Glob 1 ECFP-N1; lanes 3 and 4 refer to construct WΔ 1-3 Glob 1 ECFP-N1. Total RNA was extracted 72 hours after transfection. cDNA was synthesised by RT and amplified with two different sets of primers. (A) Detection of the isoform with a tandem repeat of exon 2. 35 cycles of amplification were carried out with primers SA 2-2 junction and SA E3R. (B) Detection of the normal isoform. 30 cycles of amplification were carried out with primers β Glob E1F and SA E3R. The difference in size is due to the shorter length of exon 1 in construct WΔ 1-3 Glob 1 ECFP-N1.
sequence are required for exon repetition in the context of the minigene. In fact both clones produced the mRNA isoform with the tandem repeat of exon 2 with an apparently similar efficiency.

As mentioned before, the efficiency of exon repetition obtained from minigenes is much lower compared to the efficiency observed \textit{in vivo}. The length of exon 1 in the minigene WKY 1-3 ECFP-N1 is 35 nucleotides longer compared to exon 1 of the natural SA gene and this is due to the presence of the polylinker sequence comprised between the transcription start site and exon 1. In order to investigate whether this additional sequence was the cause for the lower efficiency of exon repetition obtained from the minigene WKY 1-3 ECFP-N1 compared to rat kidney, a new construct was prepared in which most of the vector sequence comprised between the transcription start site and exon 1 of SA was removed and the new construct was named WKY 1-3 Nhe I ECFP-N1. In this construct only 13 nucleotides are present between the transcription start site and exon 1 of SA compared to 35 nucleotides of clone WKY 1-3 ECFP-N1. To evaluate the efficiency of exon repetition in this construct, a transient transfection experiment was performed with Hela cells. For a comparison, construct WKY 1-3 ECFP-N1 was separately transfected. Construct WKY 1-3 5' SS mutant ECFP-N1 was also tested in order to evaluate the role that the 5' splice site associated with exon 1 had on exon repetition. The experiment was performed in duplicate for construct WKY 1-3 5' SS mutant ECFP-N1 and WKY 1-3 ECFP-N1 and in triplicate for construct WKY 1-3 Nhe I ECFP-N1. Total RNA was extracted 72 hours after transfection and the SA mRNA isoforms were analysed by RT-PCR. Again, the mRNA isoform with the tandem repeat of exon 2 was amplified with primers SA 2-2 junction and SA E3R for 35 cycles and the result is shown in figure 5.15 A. To evaluate the presence of the normal SA mRNA isoform, a second amplification with 30 cycles was performed with primers SA E2F and SA E3R and the result is shown in figure 5.15 B. From all three constructs, the expected PCR product was amplified. All these data suggest that neither the sequence nor the short length of exon 1 is essential requirements for exon repetition. Furthermore, the same conclusion can be made for the 5' splice site sequence. These conclusions are just qualitative. To determine whether exon 1 and its adjacent 5' splice site influence the efficiency of exon repetition, the amount of each mRNA isoform should be carefully quantified.
Figure 5.15. Analysis of the mRNA isoforms produced by constructs WKY 1-3 ECFP-N1, WKY 1-3 5' SS mutant ECFP-N1 and WKY 1-3 NheI ECFP-N1 in a transient transfection experiment in Hela cells. Plasmid DNA was transfected into Hela cells. The experiment was performed in duplicate for constructs WKY 1-3 ECFP-N1, WKY 1-3 5' SS mutant ECFP-N1 and in triplicate for construct WKY 1-3 NheI ECFP-N1 (lanes 1 and 2 refer to construct WKY 1-3 ECFP-N1; lanes 3 and 4 and 5 refer to construct WKY 1-3 NheI ECFP-N1 and lanes 6 and 7 refer to construct WKY 1-3 5' SS mutant ECFP-N1). Total RNA was extracted 72 hours after transfection. cDNA was synthesised by reverse transcription and amplified with two different sets of primers. (A) Detection of the normal isoform. 30 cycles of amplification were carried out with primers SA E2F and SA E3R. (B) Detection of the isoform with a tandem repeat of exon 2. 35 cycles of amplification were carried out with primers SA 2-2 junction and SA E3R.
5.3. DISCUSSION

For many years it was believed that pre-mRNA molecules in mammals were exclusively processed in a linear manner through cis-splicing to yield messenger RNA molecules. But in recent years, some unusual mRNAs were discovered whose existence could not be explained by cis-splicing. Some of these molecules contained coding sequences that had been transcribed from different chromosomes (Kawasaki T 1999; Zhang C 2003). It was speculated that these mRNAs had been produced by trans-splicing. Another example of mRNA molecules that could not have been derived by a linear processing of a pre-mRNA molecule were those that contained tandemly repeated exons and that had been produced by a gene that did not have any duplicated exons. In all the cases of exon repetition discovered so far, the trivial explanation of exon duplication in genomic DNA was ruled out by southern blot analysis. The last example consisted of mRNA molecules that contained scrambled exons. In this case, the order of the exons was different from the exon arrangement in the gene (Caldas C 1998; Finta C 2000).

The most likely explanation for all these unusual mRNA molecules is that they arise by trans-splicing. In this case, two different pre-mRNA molecules are joined together by a splicing reaction. For example in the case of the SA gene, the 5' splice site associated with exon 2 or exon 4 in one molecule could be spliced in trans to the 3' splice site associated with exon 2 of another SA pre-mRNA molecule thus yielding mRNA isoforms with a tandem repeat of one or more exons. If trans-splicing is indeed the mechanism underlying exon repetition, the genes that produce the mRNA isoforms with repeated exons must have some distinct characteristics that favour trans-splicing. As mentioned before in the introduction of this chapter, one possible mechanism would be the presence of pre-mRNA molecules in which the 5' (or 3') splice site is uncoupled from the corresponding 3' (or 5') splice site. In particular, one pre-mRNA molecule could have a 5' splice site but not the corresponding downstream 3' splice site. One possible way of producing this kind of molecule is the presence of pause sites at specific positions (see figure 5.1). In this way, the 5' splice site on the nascent pre-mRNA molecule could splice in trans with another molecule that possesses an uncoupled 3' splice site. The latter could be generated from a promoter inside the intron upstream of the repeated exon (intron 1 in the case of SA). The experimental evidence gathered so far argues against the presence of a cryptic promoter in intron 1 of the SA gene. It is still possible though, that the presence of pause sites on their own might cause the formation of mRNA isoforms containing repeated exons. In this case,
splicing of exon 1 to exon 2 in the same nascent molecule should be blocked or delayed, or for some reason, splicing between two different pre-mRNA molecules (trans-splicing) is capable of efficiently competing with intra-molecular splicing (cis-splicing). Interestingly, it was shown that the time that RNA polymerase II spends at pause sites can be influenced by specific proteins (Renner DB 2001). An involvement of a trans-acting factor in the mechanism of exon repetition could account for the tissue-specificity in the case of SA, by assuming that the protein is expressed in the proximal tubule cells but not in liver.

As mentioned before, the presence of complementary sequences in pre-mRNA molecules was shown to promote trans-splicing both in vitro and in vivo (Chao H 2003). A mechanism based on the presence of complementary sequences at least in the SA gene is rather unlikely. In fact, even though complementary sequences are present in the SA gene, they are found in both the WKY and SHR SA alleles. Therefore, this mechanism would not explain the allele-specificity of exon repetition in SA. Moreover, a mechanism based on complementary sequences could not account for the tissue-specificity of exon repetition in SA unless one sequence is blocked by the binding of a tissue-specific trans-acting factor (for example an RNA-binding protein).

In the case of the COT gene, exon repetition was attributed to the presence of an exonic splicing enhancer sequence (ESE) in exon 2 of the COT gene (Caudevilla C 2001). Strangely, even though the SHR COT allele possesses the ESE sequence, it is unable to produce mRNA isoforms with repeated exons. It could be argued that the SHR strain lacks the cognate trans-acting factor that binds to the ESE sequence. This question was addressed by analysing COT mRNA in heterozygote rats that were obtained by crossing SHR with WKY. The analysis of total COT mRNA showed that both alleles were expressed at similar levels (see figure 3.19 C “Total mRNA”). In contrast, analysis of the COT mRNA isoform with the tandem repeat of exon 2 showed that it was derived exclusively from the WKY COT allele (see figure 3.19 C “Repetition exon 2). These data demonstrated that the absence of exon repetition in SHR could not be explained by the absence in this strain of a putative cognate protein that binds the exonic splicing enhancer in exon 2. The absence of exon repetition of COT in SHR could be explained in two ways:

1) The SHR COT allele contains a cis-acting element that blocks exon repetition.
2) Exon repetition of COT requires the presence of other cis-acting elements that are not present in the SHR COT allele.

It is not known which of the two hypothesis is correct.
The last possible mechanism that was investigated depended on the presence of a poorly defined exon and/or a weak 5' splice site upstream of the repeated exon/s. This was evaluated by mutating a WKY SA minigene. Some mutations were aimed at changing the sequence or the length of exon 1 (clones WA 1-3 Glob 1 ECFP-N1 and WKY 1-3 Glob 1 ECFP-N1 respectively). In another mutant, the 5' splice site associated with exon 1 was changed to a consensus 5' splice site (clone WKY 1-3 5' SS mutant ECFP-N1). The purpose of constructing all these mutants was to determine whether a presumed upregulation in the efficiency and/or kinetic of cis-splicing could abrogate the production of the isoform with the tandem repeat of exon 2. It was possible that in these conditions, trans-splicing would no longer be able to efficiently compete against cis-splicing. From preliminary experiments, none of the mutants were able to block or significantly reduce exon repetition. Therefore exon 1 and its associated 5' splice site do not seem to play a role in exon repetition.

The allele-specificity of exon repetition in COT and SA offers the possibility of identifying the region/s in these genes that cause the inclusion of repeated exons in the mRNA. From the DNA sequence comparison between the WKY and SHR SA alleles several differences were found. A possible candidate that could account for the allele-specificity of exon repetition in SA is the LINE element, which is found only in intron 1 of the SHR SA allele. Surprisingly, an SHR SA minigene consisting of the genomic sequence from exons 1 to 3 could recapitulate exon repetition in transient transfection experiments (Jian-hua Jia, personal communication). This suggested that elements of the SA gene outside this might block exon repetition of SA in the SHR SA allele. Another possible explanation is that there are differences in chromatin structure or epigenetic modifications between the WKY and SHR SA alleles. The difference between the WKY and SHR alleles that is responsible for exon repetition must be inheritable.

In most of the reported cases of intergenic trans-splicing, the efficiency is very low, approximately 1% of the normal isoforms (those derived by cis-splicing). This suggested that trans-splicing could be a "noise" of pre-mRNA splicing rather than a highly specific process with a biological function. But even if the efficiency is low, there seems to be a specificity in intergenic trans-splicing since it is observed only with some exon combinations but not with others. In the case of SA and COT, the isoforms with repeated exons are very abundant, representing more than 50% of the mRNA produced by these genes (see figure 3.11). Furthermore, a substantial amount of the mRNA isoform
containing a tandem repeat of exons 2 and 3 is also observed in the human Sp1 gene (Takahara T 2000). These findings demonstrate that exon repetition is not just a "noise" of gene expression but instead it is produced by a very efficient mechanism.

Another characteristic of exon repetition is that it is tightly restricted. For example in heterozygotes it is observed only in mRNA derived from one allele but not from the other. The most likely explanation for this is that exon repetition is a co-transcriptional process. Pre-mRNA molecules are likely to be processed before they become detached from the DNA template and therefore pre-mRNAs transcribed from one allele could not interact with pre-mRNAs transcribed from the other allele (unless the two alleles were in very close proximity in the nucleus). The cell would probably benefit by this. For instance, the co-transcriptional processing of pre-mRNA molecules containing exons that have a tendency to be ligated to exons present on other pre-mRNA molecules would reduce the risk of producing hybrid mRNA molecules that would probably encode chimeric proteins with potentially dangerous effects for the cell.

In this chapter, four possible mechanisms were investigated and none were supported by results. Of course, other mechanisms are still possible. For example, the enzyme RNA polymerase II could jump or slide backward once it has transcribed the exons that are found repeated in mRNA. In this way, a single pre-mRNA molecule would be produced that contains a tandem repeat of specific exons. After cis-splicing, these precursors would yield mRNA isoforms with repeated exons.

Another possibility is if exon 1 became disengaged from the CTD domain of the RNA polymerase II. This would probably thwart cis-splicing and in these circumstances trans-splicing might take place between two nascent pre-mRNA molecules. This was also proposed by Tasic and collaborators (Tasic B 2002).
Appendix A. Genomic sequence of the SHR SA allele. The exonic sequences are shown in bold. The dinucleotides corresponding to the GT and AG sequences of the 5’ and 3’ splice sites are shown with a yellow background. The sequence with a red background inside intron 1 is the LINE element. The translational start site situated in exon 3 is represented by the green ATG codon. The TAA stop codon located in exon 15 is shown in red. The microsatellite downstream of the LINE element (shown with a light blue background) is complementary to the microsatellite sequence shown in light blue background and located in intron 4.
CONCLUSIONS

Exon repetition was discovered a few years ago and since then very little progress has been made. An intriguing question was the reason why exon repetition in SA is strain-specific (Frantz SA 1999). In particular, it is observed in the WKY strain of rat but not in the SHR. Considering that exon repetition of SA is also tissue-specific (it is detected in the proximal tubule cells of the kidney but not in liver), the most likely explanation for strain-specificity (and also tissue-specificity) seemed to be the presence of a trans-acting factor in the proximal tubule cells of WKY that caused exon repetition in SA mRNA. This idea was initially supported by this work which showed that exon repetition of a second gene named COT, was present in the same strain that exhibited exon repetition of SA (WKY strain) but not in the strain that lacked it (SHR strain). These findings suggested that the same trans-acting factor might cause exon repetition in several different genes. Interestingly, the analysis of segregation of exon repetition in the two genes in an F2 population obtained from a cross between WKY and SHR showed that the property of exon repetition in the two genes segregate independently. This ruled out the possibility that exon repetition in SA and COT could be determined by a single trans-acting factor. It was still possible that two different trans-acting factors were present in WKY, one causing exon repetition in SA and the other one in COT. Alternatively, exon repetition could be specifically associated with certain alleles. It was known that WKY and SHR possessed two different SA alleles (Frantz SA 1996). To discriminate between these two possibilities, SA mRNA was analysed in heterozygotes carrying one SA allele from WKY and one from SHR. A single nucleotide polymorphism (SNP) in exon 3 allowed to determine from which allele a particular mRNA had been transcribed. These experiments showed that while the majority of the normal mRNA isoform is derived from the SHR SA allele, the isoforms containing the repeated exons are solely produced from the WKY SA allele. Therefore, the strain-specificity of exon repetition in SA is not determined by a strain-specific trans-acting factor. The determinants of exon repetition in SA are cis-acting elements which are present only in the WKY SA allele. A similar analysis of COT mRNA in heterozygote rats showed that the COT mRNA isoform with the tandem repeat of exon 2 is derived exclusively from the WKY COT allele. Therefore, as in SA, exon repetition of COT is also allele-specific. The inability of the SHR SA and COT alleles to give exon repetition might be explained by the lack in these genes of one or more cis-acting sequences that are required for exon repetition. Alternatively, one or more elements that inhibit exon repetition might be present in the SHR alleles. The second explanation is in agreement with results obtained from an
SHR SA minigene comprised between exons 1 and 3. Upon transfection into mammalian cells, the SHR SA minigene could recapitulate exon repetition (Jian-hua Jia, unpublished data). Similarly, a COT minigene from Sprague-Dawley (the DNA sequences comprised in the minigene are identical in Sprague-Dawley and SHR) could recapitulate exon repetition in transfection experiments (Caudevilla C 2001).

The allele-specificity of exon repetition indicates that the phenomenon happens while the pre-mRNA molecules are still attached to the DNA template and it is restricted to one chromosome. Similarly, the protochaderin mRNA isoforms that are believed to result from trans-splicing are produced only intrachromosomally (Tasic B 2002). If exon repetition is indeed caused by trans-splicing, the reaction is probably carried out cotranscriptionally. This would be in agreement with the many observations that indicate that pre-mRNA processing is a co-transcriptional event (Beyer AL 1988; Eperon LP 1988; Bauren G 1994). The restriction of exon repetition to only one allele is in contrast with the findings from a Drosophila gene named lola (longitudinals lacking). In this case, trans-splicing appears to be inter-allelic and requires chromosome pairing suggesting that close proximity is needed (Horiuchi T 2003). It has been suggested that the discrepancy between mammals and insects might be due to a lack of chromosome pairing during transcription in mammals (Horiuchi T 2003).

One key question about exon repetition is its biological significance. Does it have a specific function, or is it just a by-product of an error-prone pre-mRNA processing reaction? The fact that more than 50% of the mRNA in both SA and COT contains repeated exons suggests that exon repetition is a highly efficient process. The tissue-specificity of exon repetition in SA (Frantz SA 1999), the regulation of the efficiency of exon repetition in a voltage-gated sodium channel gene by the nerve growth factor (Akopian AN 1999) and the ability of exon repetition to produce new proteins (Caudevilla C 1998) all seem to suggest that exon repetition might have a biological function. On the other hand, the absence of exon repetition of SA and COT in SHR indicates that at least in these two genes, exon repetition is not essential. Therefore, even if the novel COT protein product has a biological function, it is not essential even in rat.

Potentially, exon repetition might have a negative effect on translation. In particular, if the exon containing the AUG translational start site is repeated and the two AUG codons are out of frame, an upstream ORF encoding a truncated protein is created. The presence of the upstream AUG translational start codon might inhibit translation of the downstream codon, which encodes the full length protein. This aspect was investigated in
chapter 4. The 5' UTRs from normal and exon repetition mRNA isoforms of COT and SA were fused to a reporter gene (firefly luciferase). The results relative to the repetition of exon 2 of COT indicate that it significantly reduces translation efficiency from the downstream AUG translational start codon (compare C and D in figure 4.6). Furthermore, since mutation of the AUG codon to a UUG codon in the first copy of exon 2 increased the efficiency of translation to a level comparable with that obtained from the normal COT 5'UTR (compare C and E in figure 4.6), it can be concluded that the reduced level of translation from the downstream AUG codon in the exon repetition isoform is due to the presence of the upstream AUG and not to a change in the secondary structure of the 5'UTR. A similar result was obtained from the 5'UTRs of the SA gene. Again, repetition of the exon containing the AUG translational start codon (exon 3) significantly reduced the efficiency of translation from the downstream AUG. In this case, mutation of the AUG to a UUG codon in the first copy of exon 3 had little effect on translation of the AUG in the second copy of exon 3. This is probably due to the presence of several AUG codons. It is likely that one or a few of these codons are used for translation and therefore inhibit synthesis of the luciferase reporter protein. As a general rule, it can be concluded that repetition of the exon containing the AUG translational start codon causes a reduction in translation of the downstream AUG that directs synthesis of the full length protein if the two codons are out of frame. Whether this has beneficial or deleterious consequences would ultimately depend on the level of protein expression in the absence of exon repetition. For example if an excessive amount of protein was being produced, exon repetition might have a beneficial effect by decreasing the amount of protein to a more optimal level. On the other hand, if an optimal amount of protein was being synthesised in the absence of exon repetition, than the effect would be deleterious. This does not take into account the potential production of truncated proteins (in principle, they could act as dominant negative) that can be derived from the translation of the upstream ORFs. It is interesting to note that translation is not completely abolished by exon repetition, at least with the reporter assay used in chapter 4. The residual translation could be due to either leaky scanning or translation reinitiation. If translation reinitiation is involved, then it is expected that the amount of protein produced is going to be influenced by the concentration of initiation factors of translation. In particular, the higher the concentration of initiation factors, the more protein would be expected to be produced.

Another aspect of exon repetition which is not well understood is its mechanism. An exonic splicing enhancer was shown to be required for exon repetition in the COT
gene (Caudevilla C 2001). Interestingly, this work demonstrated that the presence of the ESE in exon 2 is not sufficient to cause exon repetition of COT. In fact, even though the SHR COT allele possesses the putative exonic splicing enhancer sequence in exon 2, it does not produce mRNA isoforms with repeated exons. It could be argued that the SHR strain might lack the cognate protein that supposedly binds the ESE sequence, but this was ruled out by the results obtained from the heterozygote rats (cross between SHR and WKY). In these animals, the mRNA isoform with the tandem repeat of exon 2 is only derived from the WKY COT allele but not from the SHR COT allele, even though both alleles are expressed (see figure 3.19 B and C). It can be concluded that the presence of the ESE in exon 2 of COT, even though required, is not sufficient to confer exon repetition. The most likely explanation for the absence of exon repetition of COT in SHR is that other sequences are required but are absent in the SHR COT allele. Alternatively, it is possible that the SHR COT allele possesses DNA elements inside or near the COT gene that inhibit exon repetition. The second explanation is supported by the ability of a COT minigene to produce the mRNA isoform with the tandem repeat of exon 2 (Caudevilla C 2001).

It has recently been shown that the promoter of a gene can have a profound effect on pre-mRNA processing (Cramer P 1997; Kadener S 2001; Kadener S 2002). Therefore, the promoter region would appear to be a good candidate as an element that is involved in the allele-specificity of exon repetition. But since the COT promoter of SHR and Sprague-Dawley are identical in sequence (at least a region of 1.2 Kb upstream of exon 1), it can be concluded that the absence of exon repetition of COT in SHR is not due to the promoter region.

A possible mechanism to explain exon repetition was originally proposed by Caudevilla and co-workers (Caudevilla C 1998) but had never been tested. They speculated that a short exon 1 would splice inefficiently to the downstream exon 2 and this might allow the 5' splice site adjacent to exon 2 of a second pre-mRNA molecule to efficiently compete against cis-splicing. Interestingly, exon 1 is relatively short in both COT and SA (27 and 67 nt respectively). This hypothesis was tested in chapter 5. The length of exon 1 in a WKY SA minigene was substantially increased by fusing upstream of exon 1 of SA, 133 nt of exon 1 from the human β-globin gene. The corresponding minigene named WKY 1-3 Glob 1 ECFP-N1 recapitulated exon repetition in transient transfection experiments (see figure 5.14, lanes 1 and 2). These data suggest that a long exon 1 in SA does not block exon repetition. To test whether the sequence of exon 1 played an important role, a second mutant minigene was constructed. In this case, exon 1 of SA was replaced with exon 1 of the
human β-globin gene (except for the 5' splice site) and this construct was named WΔ 1-3 Glob 1 ECFP-N1. Since this new mutant produced exon repetition in transient transfection experiments (see figure 5.14, lanes 3 and 4), it can be concluded that exon repetition in SA is not determined by the sequence of exon 1. In contrast, replacement of exon 2 with exon 8 of SA or exon 2 of β-globin abolished exon repetition (Jian-hua Jia, unpublished data).

It could be speculated that an increase in the efficiency of splicing between exons 1 and 2 would thwart trans-splicing. To test this hypothesis, the 5' splice site adjacent to exon 1 was mutated to a consensus 5' splice site. The resulting minigene was still able to produce exon repetition in mammalian cells (see figure 5.15, lanes 6 and 7).

The mechanism for exon repetition of SA involving a cryptic promoter inside intron 1 was also tested. The data obtained suggest that intron 1 does not contain a promoter element. However, due to the inherent problems with the experiments (presence of AUG codons for the luciferase assay), the presence of a putative promoter in intron 1 of the SA gene cannot be ruled out. The presence of pause sites in introns 1 and 4 was also part of the mechanism. Interestingly, insertion of a pause site in intron 2 of an SA minigene increased the efficiency of exon repetition (Jian-hua Jia, unpublished data). This result suggests that pause sites might be part of the mechanism for the production of mRNA isoforms with repeated exons.
REFERENCES


Fong N, Bentley DL. (2001). “Capping, splicing, and 3’ processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD.” Genes Dev. 15(14): 1783-1795.


Exon repetition: a major pathway for processing mRNA of some genes is allele-specific

Roberto Rigatti, Jian-Hua Jia, Nilesh J. Samani and Ian C. Eperon*

Department of Biochemistry, University of Leicester, Leicester LE1 7RH, UK and Department of Cardiology, University of Leicester, Glenfield Hospital, Leicester LE3 9QP, UK

Received November 5, 2003; Revised and Accepted December 8, 2003

ABSTRACT

Exon repetition describes the presence of tandemly repeated exons in mRNA in the absence of duplications in the genome. Its existence challenges our understanding of gene expression, because the linear organization of sequences in apparently normal genes must be subverted during RNA synthesis or processing. It is restricted to a small number of genes in some of which over half of the mRNA contains specific patterns of repetition. Although it is sometimes assumed to arise by trans-splicing, there is no evidence of this and the efficiency is very much higher than for examples of bona fide trans-splicing in mammals. Furthermore, a potentially ubiquitous reaction such as trans-splicing is not consistent with a phenomenon that involves such a high proportion of the products of so few genes. Instead, it seems more probable that exon repetition is caused by a specific trans-acting factor. We have tested this and demonstrate for the two best characterized examples that the property is restricted to specific alleles of the affected genes and is determined in cis. It is not determined by exonic splicing signals, as had been suggested previously. In heterozygotes, RNA transcribed from the two alleles of an affected gene can have fundamentally different fates.

INTRODUCTION

The discovery of exon repetition suggested that something extraordinary can happen during RNA processing (1,2). Tandem repeats of specific exons were found originally in a majority of the mRNA from two rat genes, COT (carnitine octanoyl transferase) and Sa (a medium-chain acyl-CoA synthetase) (1,2), although in the case of Sa the repeats were seen only in mRNA from the kidney of specific rat strains (2). Since these reports, exon repetition has been observed in mRNA from a small number of other rat and human genes (3–7). In most cases, the repeats were detected by RT–PCR and confirmed by direct analysis of the mRNA (1,2,4,5,7). In principle, exon repetition might arise from duplications in the genome or unusual RNA processing reactions. Duplications of the repeated exons in the genome have been excluded (1,2,4,5,7). Tandem duplication of the entire gene might allow run-through transcription and splicing, but this too was excluded in the case of Sa (2). Hence, it was inferred that exon repetition must take place at the level of the RNA, and it was the first example of heterogeneity in natural mammalian mRNA that did not arise from the use of alternative signals in linear, contiguous sequences.

The only reaction known at present that might explain exon repetition is intragenic trans-splicing. However, the precursors in mammals do not support this. Most examples of trans-splicing in mammalian cells have involved pairs of partial genes in which transcripts began or ended within the introns (which therefore became "outruns"), such that there were unpaired splice sites (8–10). There is no evidence of appropriate internal promoters in the Sa or COT genes. Trans-splicing of intact genes has been reported very rarely, perhaps in part because it could be detected only if it were interallelic, meaning that the mRNA contains sequences originating from two distinguishable alleles, or intergenic, containing sequences from two different genes. A comprehensive analysis of protocadherin gene expression excluded the existence of trans-splicing between transcripts originating from different alleles, although there was evidence of a very low level of probable trans-splicing between tandem gene clusters on the same chromosome (11). The only precedent for efficient trans-splicing between transcripts from one intact gene comes from Drosophila, in which splicing involves alleles of lola on paired homologous chromosomes (12). However, it seems that lola splicing might depend upon the use of alternative promoters to produce transcripts with partial terminal introns (12). We infer that, even if trans-splicing of intact genes were a ubiquitous side-reaction, it could account for neither the high level of exon repetition nor its restriction to so few genes. Exon repetition is clearly special.

If exon repetition is not an occasional aberration of RNA processing, but efficient, it is unclear why it should be restricted to so few genes. The absence of exon repetition in Sa mRNA from one strain of rat suggests that there is a strain-specific trans-acting factor that promotes exon repetition in the mRNA from specific genes. The existence of such a factor would be a strong indication that exon repetition had a biological function. We have investigated exon repetition of the Sa and COT genes, and conclude that it is tightly restricted and not determined by a common trans-acting factor;

*To whom correspondence should be addressed. Tel: +44 116 252 3482; Fax: +44 116 252 3369; Email: eci@le.ac.uk

Nucleic Acids Research, Vol. 32 No. 2 © Oxford University Press 2004; all rights reserved
furthermore, we show that it is a unique property of specific alleles acting in cis.

MATERIALS AND METHODS

Analysis of exon repetition by RT-PCR

RNA extraction, reverse transcription and PCR were as described (2). In all figures, the mRNA was extracted from kidneys. RT-PCR showed that the proportion of COT mRNA with exon repetition was similar in all tissues tested. This proportion was quantified by amplification in the presence of [α-32P]dCTP and measurement of the intensities of the signals after various numbers of cycles. The efficiencies of amplification for each product were determined by fitting a straight line to each plot of log10 signal versus cycle number, and used to calculate the starting ratios after cycle 1 (13). The ratios were corrected for the numbers of dCMP moieties incorporated on each strand during amplification. Reverse transcription was done with 2 μg of total RNA and oligo(dT) primers. PCR products were isolated for sequence analysis by electrophoresis on agarose gels, followed by elution using centrifugation through Spin-X filters (Costar). Sequencing was done using Big Dye terminators (Applied BioSystems). PCR conditions were as follows:

- Figure 2 panel A: 94°C, 2 min, then 25 cycles of 94°C, 10 s, 55°C, 45 s, 72°C, 1 min 30 s, 72°C, 10 min.
- Figure 2 panel B: 94°C, 2 min, then 28 cycles of 94°C, 45 s, 55°C, 45 s, 72°C, 45 s, 72°C, 10 min.
- Figure 3 panels A and B: 94°C, 2 min, then 28 cycles of 94°C, 10 s, 55°C, 30 s, 72°C, 1 min 72°C, 10 min.
- Figure 4 panel A and C: 94°C, 2 min, then 30 cycles of 94°C, 10 s, 55°C, 45 s, 72°C, 45 s, 72°C, 10 min.
- Figure 4 panel B: 94°C, 2 min, then 30 cycles of 94°C, 10 s, 60°C, 30 s, 72°C, 1 min 72°C, 10 min.
- Figure 5 panel A: 94°C, 2 min, then 35 cycles of 94°C, 10 s, 55°C, 45 s, 72°C, 1 min 72°C, 10 min.
- Figure 5 panel B: first round PCR with primers COT E1F and COT E17R: 94°C, 2 min, then five cycles of 94°C, 10 s, 65°C, 30 s, 72°C, 3 min, followed by five cycles of 94°C, 10 s, 63°C, 30 s, 72°C, 3 min, then 25 cycles of 94°C, 10 s, 60°C, 30 s, 72°C, 3 min, 72°C, 10 min.
- Second round of PCR, primers COT E16F and COT E17R: 94°C, 2 min, then 30 cycles of 94°C, 10 s, 55°C, 30 s, 72°C, 1 min 72°C, 10 min.

Primer sequences

COT E1F, GAGTGCAAGAGGCGAAGGCGGCTGC
COT E3R, TGATGCAATGTCTGGCCAC
COT E7R, GGGCCGACAGGTTCATCCAG
COT E2F, GGAATGTTCGTTCTTCAATTGAC
COT E2R, GGAATGTTCGTTCTTCAATTGAC
COT E16F, TTCCCATGGTACATAATGGATAC
COT E17R, CATGAAATCTCATAGGCTTG
SA E1F, GGCTTTCTCTCCCATTAAGCAGGTCTA
SA E2R, ATGGATTTGGTGGTTAGGCATCG
SA E4R, CACCACTTGGGATCTTGCGGAAGTTGAAAAAC
SA E4R, CACCACTTGGGATCTTGCGGAAGTTGAAAAAC
SA E4R, CACCACTTGGGATCTTGCGGAAGTTGAAAAAC

Figure 1 shows the sites of annealing of the primers on COT and SA mRNA.

Gene sequences

The sequences of cloned Sa genes from WKY and SHR rats have been deposited with accession numbers AY455861 and AY456695, respectively.

RESULTS

The rat Sa gene is expressed in the liver and in kidney proximal tubules, and it is expressed more abundantly in the spontaneously hypertensive (SHR) rat than in the normotensive Wistar-Kyoto (WKY) rat (14,15). However, exon repetition in Sa is restricted to the WKY rat, in which either exon 2 or exons 2 and 3 are tandemly repeated (2). Exon repetition in COT mRNA has been investigated in Sprague-Dawley (SD) rats, in which either exon 2 or exons 2 and 3 are
tandemly repeated, and the detection of a protein product suggested that repetition might have a biological role (1). If exon repetition was determined by a common trans-acting factor, then it would be predicted that SHR rats would not show repetition in COT mRNA. To test this, COT mRNA from WKY and SHR rats was examined by RT-PCR. Amplification between exons 1 and 7 (Fig. 2A) in mRNA from eight WKY rats showed a very strong signal from exon repetition-containing mRNA compared with normal mRNA, confirmed by RT-PCR between the two copies of exon 2 (Fig. 2B). Quantitative PCR showed that the fraction of cDNA molecules with repetition of exon 2 was 54% (see Supplementary Material). However, the parallel experiments with mRNA from seven SHR rats showed no evidence of exon repetition. This shows that COT exon repetition occurs in the same rats as Sa exon repetition, consistent with the hypothesis of a common trans-acting factor. The results also show that the exon repetition of COT mRNA is non-essential, even though at least one exon repetition isoform appears to be translated (1).

It has been suggested previously that exon repetition in COT mRNA depended on a purine-rich enhancer sequence in exon 2, based on splicing reactions with exogenous RNA and experiments with transfected mini-genes (16). To examine whether the failure of the SHR COT mRNA to produce repetition isoforms was caused by a difference in the exon sequence, COT cDNA was amplified by PCR from WKY, SD and SHR rat strains and sequenced. Exon 2 is identical in all
The point mutation creates a recognition sequence for the restriction studied in the F2 population of Figure 3. A SNP in the COT 3'UTR was heterozygotes. The segregation of the WKY and SHR COT alleles was Figure 5. Analysis of COT gene alleles involved in exon repetition in parental lines (S, W) and an FI heterozygote (H).

animals are numbered as in Figure 3. The results are also shown for the total mRNA from F2 animals derived from a WKY X SHR cross. The from the COT 3'UTR can be cleaved into two fragments by Aflll only if endonuclease Aflll in the WKY COT allele. Thus, amplification products used to analyze the distribution of the COT alleles in the F2 population. The point mutation creates a recognition sequence for the restriction endonuclease Aflll.

amplification products; normal mRNA amplified by primers in exons 1 and 17 represent the alleles present in the specific lines of rats to exon repetition was tested further by examining whether the property of exon repetition in the Sa and COT genes co-segregate when WKY and SHR rats are crossed. First, the Sa genotypes of rats in the second filial (F2) generation from the cross were determined, and RT–PCR was used to determine their pattern of exon repetition. Rats homozygous for the WKY allele of the Sa gene all expressed mRNA with repetition of either exon 2 or exons 2, 3 and 4 (animals 1–7; Fig. 3A); rats homozygous for the SHR allele showed no exon repetition (animals 8–16; Fig. 3A). In both cases, the result matched the phenotype of the parental rats with the same Sa genotype. If exon repetition of the Sa mRNA depended upon a trans-acting factor, then it is likely to be linked to the Sa WKY allele.

When the same rats were screened by RT–PCR of COT RNA, it was evident that exon repetition in COT mRNA did not co-segregate with exon repetition in Sa mRNA (animal 4 showed exon repetition in Sa but not COT mRNA, and animals 8, 9, 10, 11, 13 and 16 showed exon repetition in COT but not Sa mRNA; Fig. 3B and C, cf. Fig. 3A). The level of exon repetition appeared to vary somewhat in the positive animals (Fig. 3B; see below). We conclude that Sa and COT exon repetition are not determined by a common trans-acting factor but are caused either by separate trans-acting factors or by cis-acting properties of the WKY and SHR alleles of each gene.

The contributions of trans-acting and cis-acting factors in exon repetition were tested by determining which alleles produced the mRNA isoforms containing repeated exons in Sa heterozygotes. The allele from which the RNA originates can be identified via a polymorphism in exon 3. Genotype analysis of the F2 population revealed a number of rats that contained the WKY and SHR alleles of the Sa gene. Two of these were analyzed. Amplification between exons 1 and 2 of Sa mRNA produced bands indicative of exon repetition (animals 17 and 18; Fig. 4A), but they were weaker than in rats homozygous for the WKY allele (animals 1 and 2). This conclusion was confirmed by two further analyses with different sets of primers (Fig. 4B and C). The contribution of each allele to the different mRNA isoforms was determined by sequence analysis of the PCR products. To test whether sequence analysis would give semi-quantitative data, PCR products from normal mRNA from the two homozygous parent lines were derived by amplification between exons 1 and 4 (Fig. 4D), mixed in various ratios and subjected to sequencing. Figure 4E shows that the heights of the peaks at the position of the single nucleotide polymorphism (SNP) in exon 3 were correlated with the starting proportions of the alleles in the mixture of PCR products, although the signal from the WKY allele was a little weaker. The mRNA isoforms in the F2 population were analyzed by sequencing three RT–PCR products: normal mRNA amplified by primers in exons 1 and 4, exon 2 repetition mRNA amplified by a primer specific for the junction of exon 2 with exon 4 with exon 2 and a reverse primer in exon 4 (as in Fig. 4B), and exon 2-3-4 repetition mRNA amplified by a primer specific for the junction of exon 4 with exon 4 and a reverse primer in exon 4 (as in Fig. 4C). Sequencing across the SNP in exon 3 showed that in two heterozygotes tested (Fig. 4F, animals 17 and 18), the normal mRNA derived predominantly from the SHR allele (Fig. 4F, three strains. We conclude that, even if the purine-rich sequence in exon 2 is required for exon repetition, it is not sufficient. Furthermore, all of the exons in SD and SHR rats are identical, even though a high proportion of COT mRNA from the SD rats contains repeated exons whereas mRNA from SHR contains no repetition. Thus, exonic sequences do not determine the property of exon repetition in the SD rat strain.

Figure 5. Analysis of COT gene alleles involved in exon repetition in heterozygotes. The segregation of the WKY and SHR COT alleles was studied in the F2 population of Figure 3. A SNP in the COT 3'UTR was used to analyze the distribution of the COT alleles in the F2 population. The point mutation creates a recognition sequence for the restriction endonuclease Aflll.

This procedure is similar to that in (A), with prior amplification between exon 1 and 17. ER. Exon repetition-containing COT mRNA was amplified between the exons 2-3-jet and exon 17, and then re-amplified and analyzed as for T. (C) Sequence analysis of products of amplification between exons 16 and 17. As in (B), PCR products derived initially by RT–PCR between exons 1 and 17 represent the alleles present in total mRNA, whereas those derived initially by RT–PCR with a primer specific for the tandem repetition of exon 2 represent the alleles present in exon repetition isoforms.
normal mRNA), whereas mRNA containing repetition of exon 2 (Fig. 4F, repetition exon 2) and mRNA containing the exon 2-3-4 repetition (Fig. 4F, repetition exons 2, 3 and 4) gave signals identical to those from these isoforms in the WKY homozygote (Fig. 4F, W). We conclude that mRNA isoforms with exon repetition in the heterozygote are derived from the WKY allele alone, and therefore that exon repetition is a cis-acting property of that allele. The lack of a contribution to the repetition isoforms from the SHR allele is consistent with the relatively faint level of exon repetition in animals 17 and 18 in Figure 4A.

The analysis was repeated for COT mRNA. To characterize the COT genotype in the F2 population, mRNA from the F2 animals was tested for the presence of a WKY-specific AflIII site in exon 17. The assignments were confirmed by genomic PCR (not shown). The results in Figure 5A show that the digestion products (lower bands) associated with the presence of a WKY allele are seen in animals 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 13 and 16, all of which showed exon repetition in COT mRNA in Figure 3. We conclude that COT exon repetition segregates with the presence of a WKY allele of the COT gene. Hence, exon repetition is a cis-acting property of the allele or it is determined by a trans-acting factor closely linked to the WKY allele.

The two possibilities could be distinguished by determining whether in heterozygotes the RNA showing exon repetition was derived from both alleles or only the WKY allele, as was the case for Sa mRNA. However, the analysis of exon repetition in heterozygotes could not be done directly, because the SNP was distant from the repeated exons. Instead, mRNA with a repetition of exon 2 was amplified by a primer specific for the junction of exon 2 with exon 2 and a reverse primer in exon 17, and re-amplified in a second round of PCR between exons 16 and 17, as above. Both digestion with AflIII (Fig. 5B) and sequence analysis of a second SNP in exon 17 (Fig. 5C) showed that the exon repetition-containing mRNA was derived from the WKY allele. As a control, total COT mRNA was amplified between exons 1 and 17, and then analyzed by re-amplification and both digestion and sequencing. The two alleles were present in almost equal concentrations. Hence, just as with the Sa gene, the property of exon repetition in the COT gene is restricted to the WKY allele even in a heterozygote. This explains the variation in the apparent level of exon repetition seen in Figure 3B: in the left-hand panel, animals 2, 5 and 6 were homozygous for the WKY allele, whereas animals 1, 3 and 7 were heterozygotes and the SHR allele would not have contributed to the exon repetition signal. The positive animals in the right-hand panel were all heterozygotes and gave a more uniform apparent level of repetition.

**DISCUSSION**

Exon repetition was an extraordinary and unexpected phenomenon that appeared to subvert our understanding that genes were expressed via a linear series of signals on contiguous sequences of DNA or RNA. However, it has never been clear whether it was essential. It had been shown that mRNA containing repetition of exons 2 and 3 of COT is translated to produce a longer protein, from which it was inferred that exon repetition might have functional significance (1), whereas we had found that the exon repetition in Sa mRNA from WKY rats was absent in SHR rats (2). In this work, we have demonstrated that SHR rats do not show exon repetition in COT mRNA either. Thus, the best characterized and most efficient examples of exon repetition are inessential. In addition, we have examined the effect of the repetition of COT exon 2 (which contains the initiation codon) upon expression of a reporter function, and the results show that the upstream, out of frame initiation codon reduces expression substantially (R. Rigatti, unpublished data). Since this is the major isoform in WKY rats, the implication is that exon repetition is actually detrimental to COT expression in these rats. These results only deepen the mystery surrounding exon repetition. There is no obvious reason why such an extraordinary property should have arisen, one that does not arise from a background level of imprecision in splicing but is emphatically specific and efficient.

The allele specificity of exon repetition in the heterozygotes demonstrates that it is not determined by diffusible factors, whether encoded on the same chromosome or elsewhere; the determinants are allele-specific. Allele-specific expression is usually associated with quantitative considerations of transcription efficiency or the proportion of two different splicing isoforms. In the case of exon repetition, it appears that allelic differences acting strictly in cis determine whether or not a transcript will undergo qualitatively different and extraordinary reactions.

The differences in sequence between the two alleles of the Sa gene suggest few candidate determinants. Within the 1.2 kb preceding exon 1, there are six single nucleotide differences (SNPs) between the WKY and SHR alleles, and a difference of five nucleotides in the length of an oligo(A) tract. The first five exons differ at only one position. Most of the introns contain only a few differences, but intron 1 contains 39 differences in 5.3 kb (25 SNPs and differences in tracts of di- or trinucleotide repeats), and there is a LINE element of 1.4 kb in SHR that is missing in the WKY allele. Since the LINE element is found only in the strain that does not display exon repetition, its only potential role would be as a suppressor rather than a determinant of exon repetition. In the case of the COT gene, our analysis of the genomic sequences is incomplete; we have not yet found any differences between the SHR allele and the SD allele, and we cannot exclude the possibility that the determinants of exon repetition are epigenetic.

Allele specificity excludes a number of possible explanations for exon repetition, including the simplest form of the idea that it arises by trans-splicing. It is not consistent with the idea that freely diffusing pre-mRNA sequences undergoing processing are brought into close proximity by specific RNA-binding proteins and that trans-splicing takes place. There is evidence from studies of intergenic trans-splicing that chromosomal proximity is required. For example, it was observed in protocadherin that there was no trans-splicing between alleles, but there was an extremely low level (less than 0.5% of mRNA) of trans-splicing between gene clusters on the same chromosome (11). In the *lola* gene of *Drosophila*, there is a high level of interallelic trans-splicing (possibly resulting from internal promoters), but this requires the proper pairing of chromosome homologues (12). This suggests that intragenic trans-splicing might be highly efficient.
If trans-splicing is involved in exon repetition, then the most plausible explanation of allele specificity is that the reaction is required to take place while the transcripts are associated with the gene. The differences between the alleles may well affect signals involved in transcription, with very remarkable consequences, and it will be of great interest to determine what the signals are and why the vast majority of genes lack them.

**SUPPLEMENTARY MATERIAL**

Supplementary Material is available at NAR Online.

**ACKNOWLEDGEMENTS**

This work was supported by the Medical Research Council, UK, and the Wellcome Trust, UK.

**REFERENCES**


