INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Université d'Ottawa · University of Ottawa

CHARACTERIZATION OF THE Mpv 20 PROVIRAL INSERTION SITE AND OF THE UNP GENE

© Marco Di Fruscio

Thesis submitted to the
School of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Department of Biochemistry,
University of Ottawa,
Ottawa, Ontario
CANADA

National Library of Canada

Acquisitions and Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada

Acquisitions et services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-28334-8

Canada

ABSTRACT

Mpv 20 is a transgenic mouse line generated by retroviral infection of embryos in the laboratory of R. Jaenisch. Of the seventy lines generated, only four displayed a recessive lethal phenotype. Three of the four are early embryonic lethal; these include Mov 13 (Schnieke et al. 1983), Mov 34 (Soriano et al. 1987), and Mpv 20 which is the focus of this thesis. The fourth line Mpv 17 (Weiher et al. 1990) displays an adult lethal phenotype.

The proviral integration site was cloned and a series of genomic clones were isolated from DNA libraries generated from both heterozygote and wild type animals. Using a combination of exon trapping and genomic sequencing of a 3.5 kb region surrounding the proviral integration site, it was determined that the virus integrated into an intron of the murine homologue of the human *Npat* (Nuclear Protein at the AT locus) gene (accession D83243, U58852, X97186). The virus integrated in the opposite transcriptional orientation relative to the cellular gene.

Northern analysis using a partial murine lung *Npat* cDNA, generated by RT-PCR, demonstrated on average a reduction of 1.6 fold *Npat* message in RNA obtained from heterozygote tissues relative to a wild type littermate control. Analysis of the closely linked *Atm* gene showed no detectable difference in the level of brain RNA between wild type and heterozygote littermates.

A 5' RACE protocol was used to isolate a longer cDNA which was used on Southern blots to examine the possibility of methylation differences between the different animals. Under the experimental conditions utilized, no observable differences were detected between wild type and heterozygote animals. RACE product did however detect a truncated message

on a Northern blot present in heterozygote animals but not in the wild type controls. This band contained proviral sequences as demonstrated by probing of the same Northern with an MuLV LTR probe.

During a search for candidate genes present at the Mpv 20 locus, a cDNA named Unp was isolated (Gupta et al. 1993). It was subsequently determined that this gene product was a member of a large family of deubiquitinating enzymes and displayed similar enzymatic activity using a series of assays (Baker and Gray unpublished data). The second part of the thesis dealt with the determination of the genomic structure of the *Unp* gene. To this end, cloned DNA from the gene was sequenced. The *Unp* gene was found to be present in 22 exons spanning 47.3 kb and has been assigned the Genbank accession AF026469. All exons, with the exception of exon 1 which contains the transcriptional start site and exon 22 which contains the complete untranslated region, were flanked by the consensus splice acceptor/donor dinucleotides AG/GT (Mount 1982). Compilation of exonic sequences identified the presence of sequencing errors present in the initially published cDNA sequence (Gupta et al. 1993). The modified cDNA sequence encodes an additional 169 amino acids 5' to what was believed to be the initiating Met.

A 1.1 kb Eco RI /Sac I region upstream of the transcriptional start site was examined for promoter activity using a CAT reporter construct. Low activity was observed in this construct. Surprisingly, a substantial increase in CAT activity was observed when a 556 bp segment at the 5' end was deleted from the reporter construct leading us to believe that some repressor element must be present in the deleted region. The promoter fragment lacking the repressor area was found to display activity in both orientations in our experimental assay.

To the memory of Luigi Di Fruscio
1925-1987.

Acknowledgements

Dr. Douglas Gray (Doug) my supervisor who had to put up with me during my stay. For his role in the genomic cloning of both the *Unp* and Mpv 20 genes and for knowing when not to approach me.

Dr. Hans Weiher for the initial generation of the Mpv 20 line. Dr. T. Imai for supplying us with the human *NPAT* and mouse *ATM* cDNA probes. Yvonne Avis for her role in performing the CAT assays. Doris Leopold, Suzanna Dramnic , and especially Scott Kelso for their roles in operating the automated sequencer and who were extremely happy to see me leave.

Drs. Suzette Delaloge, Dominique Dubois, and Richa Sharan for all the late night intellectually stimulating philosophical, theological, and psychological discussions. Manon Dubé morning coffee break sessions and for the apple picking. Helen Tai and David Stojdl for the numerous evenings of sitting around eating and drinking. Sandra Jones for always having something up her sleeve. Giampaolo De Biase for resparking my interest in theatre, and for all the 1989s that we shared.

To the spirits of Bob Marley and Ludwig Van Beethoven whose music carried me through some tense moments.

To my wife Cécile Robard who kept me sane for the past three years and was a constant support and companion, and who never tried to organize the tons of articles floating through the appartment. And finally to all the members of my family who were a constant support during my graduate studies.

LIST OF ABBREVIATIONS

ATCC	American type culture collection
ATM	Ataxia-telangiectasia
βgal	Beta-galactosidase
CAT	Chloramphenicol acetyl transferase
cM	Centi Morgan
CTP	Cytosine triphosphate
DEPC	Diethylpyrocarbonate
DNA	Deoxyribose nucleic acid
EC	Embryonal carcinoma
EDTA	Ethylene-diamine tetra-acetic acid disodium salt
ES	Embryonic stem
HEPES	N-2-Hydroxyethylpiperazine-N'2-ethanesulphonic acid
HMG	High mobility group
HS	Hypersensitive site
I.U.	International unit
LTR	Long terminal repeat
MOPS	3-(N-Morpholino)propane-sulfonic acid
MPSV	Myeloproliferative sarcoma virus
MuLV	Murine leukemia virus
Npat	Nuclear protein at the AT locus
ONPG	2-Nitro-phenyl-β-D-galactopyranoside
OPC	Oligonuleotide purification cartridge
PBS	Phosphate buffered saline
PEG	Polyethylene Glycol
PCR	Polymerase chain reaction
RACE	Rapid amplification of cDNA ends
RNA	Ribose nucleic acid
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
TE	Tris EDTA
SDS	Sodium lauryl sulphate
SM	Suspension media
SSC	Sodium Citrate/Sodium Chloride
STE	Sucrose Tris EDTA
STET	Sucrose Tris EDTA Triton
Sry	Sex determining region on the Y chromosome
TLC	Thin layer chromatography
Tris	Tris(hydroxymethyl)methylamine
UV	Ultra-violet

X-gal	5-Bromo-4-Chloro-3-Indolyl-β-D-Galactopyranoside
YAC	Yeast artificial chromosome

Table of contents

Abstract i
Dedication iii
Acknowledgements iv
List of abbreviations v
List of tables and figures ix
Preface 1
Chapter 1
Background of the Mpv 20 line and the identification of *Unp* 2
Introduction (I) 4
    Transgenic animals generated by Microinjection of DNA 5
    Retroviral lifecycle 8
    Mov 1 to Mov 13 lines of transgenic mice 9
    The Mov 13 locus 11
    Mov 15 to 45 and Glob 1 to 3 lines of transgenic mice 15
    The Mpv 17 line of transgenic mice 17
    Activation of cellular genes following retroviral infection 18
    Integration at the Evi-1 and Cb1/fim-3 loci 19
    Npat and *ATM* as candidate genes at the Mpv 20 locus 22
Introduction (II)
    Unp and the ubiquitin pathway 25
Chapter 2
Materials and Methods 29
    Generation of transgenic mouse lines 29
    In vitro culture of embryos 29
    Molecular cloning 30
    Screening of libraries 31
    Phage minipreps 31
        Primary infection 31
        Secondary infection 32
        DNA preparation 32
    RNA isolation using the Guanidinium/cesium chloride method 33
    RNA isolation, LiCl Method 33
    Northern blotting 34
    Restriction digests 34
    DNA isolation, Southern blotting 35
        Isolation of genomic DNA 35
        Mini preparation of plasmid DNA, Boiling lysis method 35
        Large preparation of plasmid DNA, Triton lysis method 36
    Subcloning of 4 kb fragment 37
    Exon trapping 37
        Preparation of plasmid DNA 37
        Tissue culture 38

Primary amplification 38
Secondary amplification 39
Synthesis and purification of oligonucleotides 39
DNA sequencing 40
Subcloning of PCR products 40
Probes 41
5' Rapid amplification of cDNA ends (RACE) 42
HMG box PCR 42
Genomic PCR 42
Construction of CAT plasmids 43
CAT assays 44
β-galactosidase assay 44
Chapter 3
Results (I) 46
Generation and identification of a mutant transgenic mouse line 46
Molecular cloning of the Mpv 20 proviral insertion site 49
Identification of a transcriptional unit at the Mpv 20 locus 52
Identification of the disrupted *Npat* gene 62
Northern and Southern blotting analysis 66
Determination of the mechanism of *Npat* inactivation 69
Results (II) 76
Mapping intron/exon borders 76
CAT assays 83
Chapter 4
Discussion (I) 87
Discussion (II) 95
Conclusion 98
References 100
Appendix 1. Nucleotide sequence of the *Unp* gene 111
Appendix 2. MatInspector analysis of 1.1 kb Eco RI/Sac I fragment containing the *Unp* 126

List of Tables and Figures

Table 1. Progeny of 20/wt X 20/wt mating 47

Table 2. Development of embryos to blastocysts *in vitro* 47

Figure 1. Appearance of cultured embryos from a Mpv 20 heterozygous cross 48

Figure 2. Genomic map surrounding the Mpv 20 locus....50

Figure 3. Southern blot analysis of heterozygous (h) vs wild type (w) CFW DNA 51

Figure 4. Northern and Southern analysis of cDNA clone EK10 53

Figure 5. Sequence analysis of a 4kb Eco RI fragment originating from phage clone M2017 56

Figure 6. "Zoo blot" using the GRAIL fragment as a probe 57

Figure 7. Sequence of trapped exon and nucleotide alignment to *Sry* (Sex determining region on the Y chromosome) 58

Figure 8. Southern blot analysis of *Sry* PCR products 60

Figure 9. Sequence of the *Sry* PCR product and homology to the human *Npat* cDNA 61

Figure 10. Sequence of genomic DNA at the Mpv 20 proviral insertion site 63

Figure 11. Nucleotide and predicted translation product alignments between mouse and the human *NPAT* 64

Figure 12. Sequence alignment confirming the presence of a duplicated region in the mouse genome 65

Figure 13. Southern blot analysis of heterozygous (h) vs wild type (w) CFW DNA using partial Npat cDNA as a probe 67

Figure 14. Northern blot analysis of tissue RNAs from heterozygote (h) and a wild type (w) littermate 68

Figure 15. Southern analysis of 5' RACE product 72

Figure 16. Nucleotide sequence and alignment of the RACE9 product isolated 73

Figure 17. Southern analysis of methylation patterns in heterozygote (h) vs wild type (w) CFW DNA 74

Figure 18. Northern blot analysis of heterozygote (h) versus wild type (w) RNA 75

Figure 19. Mapping of exons present in phage RC9A by Southern blot analysis 78

Figure 20. Genomic map of *Unp* locus showing intron/exon positions 79

Figure 21. Splice acceptor/donor sites of the sequenced exons 80

Figure 22. Corrected Unp cDNA sequence and its predicted protein product 81

Figure 23. Sequence alignment of Intron 3 with S2 ribosomal protein pseudogene 82

Figure 24. Identification of promoter activity present in a cloned 1.1 kb Eco RI/Sac I genomic fragment of the *Unp* gene 85

Figure 25. Sequence of the 1.1 kb genomic Eco RI/Sac I fragment containing the *Unp* promoter 86

Figure 26. Proposed mechanisms for the inactivation of the *Npat* gene 94

PREFACE

The work in this thesis is composed of two unrelated projects. All headings bearing the Roman numeral (I) will describe experiments carried out on the Mpv 20 mutant mouse strain. This strain was generated by retroviral infection of early embryos. Work with the heading (II) will describe experiments related to the murine deubiquitinating enzyme Unp.

The work presented in this thesis has been submitted for publication as described below:

(I) Di Fruscio M, Weiher H, Vanderhyden BC, Imai T, Shiomi T, Hori TA, Jaenisch R, and Gray DA. (1997) Proviral inactivation of the Npat gene of Mpv 20 mice results in early embryonic arrest. *Mol Cell Biol* **17**:4080-4086

(II) Di Fruscio M, Gilchrist CA, Baker RT, and Gray DA. Genomic structure of Unp, a murine gene encoding a ubiquitin-specific protease. Manuscript submitted.

CHAPTER 1

BACKGROUND OF THE Mpv 20 LINE AND THE IDENTIFICATION OF *UNP*

Of the seventy lines of transgenic animals generated by Jaenisch and co-workers using retroviral infections, only four have displayed a recessive lethal phenotype. Three of these include Mov 13 (Schnieke et al. 1983), Mov 34 (Soriano et al. 1987, Gridley et al. 1990), and Mpv 17 (Weiher et al. 1990). The fourth strain obtained was Mpv 20 which will be the focus of the first part of this thesis. The mutation results in a homozygous recessive lethal phenotype apparently arising at the eight-cell stage of embryogenesis. The phenotype is believed to arise due to the failure of compaction of the eight cell morulae.

Animals were previously infected using a replication deficient recombinant retrovirus MPSVneo (Seliger et al. 1986), by H. Weiher, in the same screen that yielded the Mpv 17 line (Weiher et al. 1990). Mating experiments failed to generate homozygote animals which carried the disrupted allele suggesting an embryonic lethal phenotype. Embryos were examined at 11 days postcoitum , but no homozygote animals were recovered. *In vitro* culturing of one cell zygotes demonstrated that a fraction of the embryos never proceeded past the eight cell stage of development. The locus was cloned by screening a genomic library constructed from DNA isolated from heterozygote animals carrying the integrated provirus (DA Gray). Initial experiments were carried out using "zoo blots" to identify regions of the cloned DNA which would be conserved among species pointing to the presence of coding sequences. One genomic fragment was found to be conserved and was used to screen an undifferentiated F9 embryonal carcinoma cDNA

library. This initial screen identified a cDNA, designated F9, which was believed to arise from the Mpv 20 locus. The F9 cDNA was used to rescreen a testis cDNA library in an attempt to obtain a full length transcript. One cDNA clone designated T4 was initially isolated. Northern blots hybridized with the T4 clone showed a complex pattern of expression which included both ubiquitiously expressed and tissue specific transcripts which seemed to be developmentally regulated.

My initial project after joining the laboratory was to isolate and characterize some of these tissue specific clones. Library screens were carried out using the genomic fragment which identified the original F9 clone. None of the cDNAs isolated gave similar expression as that obtained for T4 on Northern blot analysis. Southern blot analysis showed that all the cDNAs obtained were highly repetitive in nature. Sequencing of a genomic region containing the area used to generate the original probe, demonstrated that F9 was isolated by virtue of the presence of a polypurine stretch which was homologous to the genomic DNA. Comparison of the F9 with the T4 cDNA showed that F9 was a chimeric cDNA which contained both genomic and cDNA sequences. The T4 cDNA was excluded as a possible candidate gene from the Mpv 20 locus. A database analysis I carried out determined that the T4 cDNA (called Unp) showed high homology to the *tre* oncogene (Huebner et al. 1988). The goal my project was then to try and identify which gene was disrupted during the proviral integration event.

The first part of this thesis will deal with the description of the mutant phenotype and the previous cloning of the genomic region surrounding the proviral integration site. Next it will describe the experiments performed that lead to the identification of the murine *Npat*

gene (see below) as a candidate gene which may be responsible for the recessive phenotype. It will also provide some evidence for a mechanism of the proviral inactivation which does not seem to involve a methylation dependent event.

INTRODUCTION

(I)

Classical genetics has depended on the analysis of induced or spontaneous mutations for the identification of genes involved in normal development. These events are rare and it is sometimes difficult to identify the gene or gene product in question on a molecular level. As a result of the lack of models and developmental mutations available, significant advances in the understanding of the molecular event involved in mammalian development have been hindered.

One way to circumvent the above problem is to generate mutant animals which display defects in their normal development. Analysis of the mutated genes in these animals may enable the assignment of a role for the gene products in the mutant phenotypes.

In the past number of years, new technologies in the field of insertional mutagenesis have been developed which may generate transgenic animals that carry one or more disrupted genes. These technologies involve the introduction of exogenous DNA into either the embryo proper or into embryonic stem (ES) cells which are subsequently introduced into the animal. These model animals may be achieved by direct introduction of foreign DNA, as in the case of microinjection into the zygotic pronucleus, transfection of ES cells with foreign DNA, or by infection with a retrovirus which may carry a selectable marker aiding in the cloning of the disrupted site. A number of review articles have been published which

describe the advantages and disadvantages of the different techniques, and describe different transgenic mice obtained (Jaenisch 1988, Gridley et al. 1987, Reith and Bernstein 1991).

Breeding the transgenic animals obtained to the homozygote state is carried out to identify recessive mutant phenotypes. Some of these phenotypes may be subtle and not easily detectable whereas others may display severe defects in development or viability. A number of phenotypes have been described as a result of an insertional mutagenesis event, but only in a limited number of cases have the actual genes responsible been isolated. Examples which were generated by proviral insertions include the Mov 13 locus which encodes the α1 (I) collagen gene (Schnieke et al. 1983), the *Mov 34* gene (Soriano et al. 1987, Gridley et al 1990) in which the provirus integration disrupted the *p40* gene encoding a regulatory subunit of the 26S proteosome (Tsurumi et al. 1995), the *Mpv 17* gene (Weiher et al. 1990) which encodes a peroxisomal protein (Zwacka et al. 1994), and the Mpv 20 line in which the virus disrupts the murine *Npat* gene (Di Fruscio et al. 1997).

This section will give a general overview of the field and describe a few examples in which genes have been identified using insertional mutagenesis. In particular it will focus on the mutational events which have identified genes essential for normal development or differentiation. Where possible, I will describe some experiments which have identified a possible mechanism of action.

TRANSGENIC ANIMALS GENERATED BY MICROINJECTION OF DNA

Microinjection results in the integration of the foreign DNA as concatamers in the genome which may be accompanied by rearrangements (Covarrubias et al. 1986), deletions

(Karls et al. 1992), duplication, or translocation of the host sequences near the integration site. In the HUGH 3 insertional mutant (Covarrubias et al. 1986, 1987), five copies of the transgene are present in which the transgene array is interrupted at least twice by mouse genomic sequences. This line also carries a 10 kb deletion of host genomic sequence at the site of the integration. These large changes in the host sequence make it difficult to interpret the mutation data as well as identify the gene or genes responsible for the observed early postimplantation embryonic lethal phenotype.

A more detailed example in which the mutational analysis is complicated by the loss of a large section of the host genome is in the case of the cloning of the dystonia musculorum (dt) locus. Dt is a hereditary neurological disorder of the mouse transmitted as an autosomal recessive mutation. Homozygote animals carrying the mutation display degeneration of axons in the sensory neurons of both the peripheral and central nervous system.

In a series of experiments carried out by Kothary et al. (1988, 1989), a construct containing an _Escherichia coli (E. coli)_ *lacZ* transgene under the control of the mouse heat-shock gene (hsp68) promoter was microinjected into a pronuclei of one-cell zygotes. Southern blot analysis revealed that seven of the 63 animals born harboured the transgene. All seven of the founder animals were able to transmit the transgene through their germ-line.

The seven lines generated showed inducible expression of the hsp68 promoter, as assessed by X-gal staining, when subjected either to heat-shock or arsenite insults. This expression was observed in a number of different tissues and at different points in development. In addition one of the above lines (designated Tg4) also showed constitutive expression of the *lacZ* gene in unstressed developing neural tubes. LacZ activity was first

detected in fetus neural tube at day 9.5 of gestation and staining was most intense at day 13.5. This activity was not detected in neonatal mice.

No apparent phenotype was observed in the heterozygous Tg4 mutant mice. In crosses of two heterozygous mice, one quarter of the pups generated began to show signs of limb incoordination at approximately day 10-12 following birth. The homozygote mice progressively lost coordination and most died before weaning or shortly thereafter. The above phenotype resembled that of a spontaneous mutation in the neural tissue-specific gene dystonia musculorum (dt). Crosses between heterozygote animals of both series, showed that the *Tg4* gene was allelic to the *dt* gene.

The identification of the genomic structure surrounding the integration site in the Tg4 line was carried out by Brown et al. (1994). A Southern blot was generated which contained Tg4/Tg4 DNA digested with enzymes which did not cut in the transgene. When lacZ was used as a probe, a band for the integration complex of approximately 70 kb in length was observed. This confirmed previous estimates (approximately 15-20) of the number of copies of the transgene present which was inserted in a single head-to-tail concatamer. Since only a single band hybridized with the lacZ probe, it was concluded that only one integration site was present in Tg4.

The cloning of the integration site was carried out using the lacZ probe and screening a Tg4/+ genomic library. Seven clones were identified which contained the transgene sequence. When total mouse DNA was used to reprobe these clones, all seven contained mouse genomic sequences. Isolation of a non-repetitive 1.8 kb Eco RI fragment from one of the clones was used to screen a wild-type genomic library. The initial clone obtained was

used as a starting point for a chromosome walk. A 75 kb contig was generated which spanned the integration complex. Southern blot analysis demonstrated that the integration of the transgene was accompanied by a deletion of 45 kb of host genomic sequence. Pulse field gel electrophoresis was used to generate a long range restriction map of the locus. It was observed that no other rearrangements occurred within 500 kb of the integration complex.

Subsequent experiments using a combination of exon trapping and zoo blot analysis on DNA from rat, human, monkey, sheep, rabbit, hamster or pig, lead to the subsequent identification of the disrupted gene responsible for the observed mutant phenotype as the hemidesmosomal protein bullous pemphigoid antigen 1 (Brown et al. 1995a & b).

RETROVIRAL LIFECYCLE

Retroviral vectors have circumvented some of the problems associated with the loss or rearrangement of host DNA described above. Retroviruses are viruses whose genome are composed of ribose nucleic acids (RNA). These viruses enter a permissive host cell by interacting with a cell surface receptor. The murine leukemia virus receptor has been identified as the cationic amino acid transporter (Kim et al. 1991, Wang et al. 1991). Using a reverse transcriptase present in the virion core, retroviruses reverse transcribe their genome into deoxyribose nucleic acid (DNA). The cytoplasmic double stranded DNA produced enters the nucleus of the host cell by a process which is poorly understood but may require cell division. Once present in the nucleus, a viral encoded integrase enzyme catalyses the ligation of the viral DNA into the host DNA. This "proviral" form of DNA is flanked on

both ends by long terminal repeats (LTRs) which have been identified to contain the viral promoter elements. The result is a stable integration of the provirus into the host genome and the proviral DNA subsequently replicates along with the host DNA and segregates as a Mendelian locus. No large rearrangements or deletions have been reported to occur at or near the integration site. However, a duplication of 4 to 6 base pairs is present on both sides of the integration site, which is characteristic of the particular viral integrase enzyme and not of the cell type being infected (Panganiban 1985). If retroviral infection occurs early, the foreign DNA may be stably incorporated in the germ line resulting in the generation of a transmitting founder animal.

The ability for the virus to replicate and spread is not a prerequisite for integration as was demonstrated in the generation of the Mov 1 to Mov 13 lines of transgenic mice described below (Jaenisch 1976). Integration however has been found to occur to a large degree in DNase I-hypersensitive areas (Rohdewohld et al. 1987, Vijaya et al. 1986) of the genome, indicating that an open chromatin structure may be required. This structural feature is common to actively transcribing genes and therefore may result in these genes being disrupted preferentially. The provirus may integrate in the promoter, intron, or exon of a given gene and result either in the activation or inactivation of the disrupted gene or of neighbouring cellular genes (see below for examples). These activation/inactivation events may occur at the proviral integration site or at a considerable distance from the actual integration event (e.g. cb-1/fm3 locus described below).

Mov 1 TO Mov 13 LINES OF TRANSGENIC MICE

Jaenisch and co-workers (Jaenisch 1976, Jähner and Jaenisch 1980, Jaenisch et al.

1981) have infected embryos with the murine Moloney leukemia virus (M-MuLV) at different stages of development: at the premorula and morula stage by overnight cocultivation of 4 to 16 cell embryos with subconfluent Cl1-1a cells which were producing M-MuLV; at the blastocyst stage by microinjecting a Cl1-1a cell into the blastocoel; and at the mid-gestation stage by microinjection of M-MuLV into day 8 embryos *in utero*. They were subsequently able to isolate 13 transgenic strains of mice each carrying a single provirus integrated into distinct chromosomal locations (Mov 1 to Mov 13). All the strains generated were able to transmit the viral genomes through their germ line in a Mendelian fashion. In contrast, animals which were infected as newborns never transmitted the viral genome to the next generation. With the exception of Mov 4 and Mov 6 all other members showed no large rearrangement or deletions in the viral or host genomes using a combination of restriction digests and Southern blot analysis . Mov 4 and Mov 6 appeared to have suffered a deletion or rearrangement at the 3' end of the proviral DNA.

Proviral integration was observed in animals that were infected at all three of the above stages of embryogenesis. Preimplantation embryos were susceptible to virus integration but the virus itself was unable to express its functions. In midgestation, M-MuLV was able to integrate as well as express functions in all organs.

Subsequent experiments (Jaenisch et al. 1983) were carried out in an effort to determine if any genes were disrupted by the proviral integrations. Heterozygous animals were crossed in an attempt to produce homozygote animals which carried both disrupted alleles. Homozygote animals were obtained for strains Mov 1 to Mov 12. None of these animals showed a distinct phenotype from that observed in the parental heterozygotes. These

results could arise from the proviruses integrating into silent regions of the mouse genome or the integrating events may have occurred in genes which have redundant functions and therefore may be nonessential. Alternatively, the provirus may have integrated into an intron of a gene which upon RNA processing was able to produce a correctly spliced product. Preliminary results obtained by Rohdewohld et al. (1987) have observed expression at the Mov 10 locus in undifferentiated F9 embryonal carcinoma cells, which are closely related to early embryonic cells, but not in their differentiated counterpart (Martin 1980). This would imply that the provirus in Mov 10 integrated into a region of DNA that is transiently expressed in early embryogenesis and its disruption is not sufficient to display a detectable early embryonic mutant phenotype. Further studies carried out by Mooslehner et al. (1991) have identified the integration event occurring into the first exon of a gene encoding a putative GTP-binding protein. In this mouse strain, the proviral integration does not disrupt expression of the gene indicating that the integration event in a transcription unit does not necessarily affect transcription therefore resulting in no phenotype being observed.

THE Mov 13 LOCUS

The Mov 13 locus is the best characterized retroviral insertion mutant. It will be described in detail in an attempt to demonstrate how retroviruses may disrupt and subsequently inactivate a cellular gene.

In contrast to the breeding results obtained above, Mov 13 mice generated small litters. No homozygote animals were identified when 14 adults and 55 embryos at day 15 to 19 of gestation were examined by Southern blot analysis (Jaenisch et al. 1983). Thus it

was assumed that homozygotes die prior to day 15. Subsequent examination of embryos *in utero* at different stages of development demonstrated that homozygotes die between 11 and 14 days of gestation. The Mov 13 proviral integration therefore suggested that an essential gene function was being disrupted resulting in a recessive lethal phenotype.

In an attempt to characterize the locus in more detail, a Northern blot containing poly $(A)^+$ RNA isolated from both Mov 13 homozygote embryos and wild-type embryos was probed using cellular flanking regions from the Mov 13 locus (Schnieke et al. 1983). Mov 13 locus specific transcripts were identified in normal wild-type embryos. In contrast, no message was detected from Mov 13 homozygote embryos. Expression at the locus began at day 12 of gestation and increased during fetal development. No or low levels of expression was observed in adult animals. Mov 13 homozygote embryos did not express the gene and were arrested in development between days 11 and 12 of gestation which coincided with the time of the Mov 13 locus activation.

The expression data taken together with several other pieces of evidence led the authors to speculate that the gene product was possibly a secreted protein which was part of the extracellular matrix. A number of cDNA clones available for those proteins were tested and it was determined that the provirus integration had occurred at the 5' end of the $\alpha 1$(I) collagen gene completely blocking its transcription.

A 14 kb fragment which represented the Mov 13 preintegration site was cloned and identified as the $\alpha 1$(I) collagen gene (Harbers et al. 1984). This fragment was determined to contain the 5' end of the $\alpha 1$(I) collagen gene. Comparing the sequence between the wild type and from the equivalent mutant allele, it was determined that the provirus had

integrated within the first intron. The orientation of the M-MuLV genome transcription was in the opposite orientation relative to that of the cellular gene. Preliminary data also suggested that the provirus integrated close to a DNase I-hypersensitive site and that differences existed between the chromatin conformation before and after the integration event.

That work was expanded upon by Breindl et al. (1984) using limited digestion with DNase I. Experiments were carried out on both collagen producing (3T3 fibroblasts and day 15 embryos) and nonproducing (embryonal carcinoma cells, adult liver and brain) cells. Cells which expressed the α1(I) collagen RNA showed the presence of three DNase I-hypersensitive sites present in the 14 kb genomic fragment described above, whereas in nonexpressing cells only two sites were detected.

Hypersensitive site 1 (HS 1) was found to be in the body of the gene. HS 2 was present in the intron in the vicinity of the Mov 13 proviral integration, and HS 3, using S1 mapping analysis, was located to the promoter region of the α1(I) collagen gene. Site HS 1 and HS 2 were detected in the chromatin of all the mouse cells examined, the HS 3 site was only detected in cells which produced collagen. The appearance of the HS 3 site was believed to be developmentally regulated and its appearance coincided with the point in development when transcription of the α1(I) collagen gene began.

It followed that the loss of the HS 3 site in Mov 13 homozygote embryos was believed to arise as a consequence of the gene no longer producing a mRNA transcript. S1 analysis was used to confirm whether a transcript was being produced in the Mov 13 homozygote embryos or whether the presence of the provirus was causing the transcript to

be intrinsically unstable. In both situations the gene product would not be present which would coincide with the observed phenotype. As mentioned the provirus integrated in the first intron of the collagen gene therefore a probe was generated which spanned the first exon. This probe would be able to detect transcripts, if any were present, that were initiated in both wild type and mutant cells. The results showed a 112 bp fragment was protected in both wild-type and heterozygote Mov 13 mice and in a cell line derived from wild-type embryos but that no protected fragment was present in Mov 13 homozygotes or a cell line derived from a homozygous embryo. In the Mov 13 heterozygote mice the 112 bp fragment had half the intensity as the corresponding wild-type. Thus HS 3 appears during embryonic development in cells that activate the collagen I gene. Subsequent work confirmed both the block in initiation of transcription and a change in the methylation pattern of the mutant allele as being present (Hartung et al. 1986, Jähner and Jaenisch 1985).

Finally, the mechanism of the mutation was examined by Barker et al. (1991). From all the data that was obtained to date they reasoned that two possible mechanisms were possible to explain the Mov 13 mutation. First, the possibility that the insertion event displaced cis-acting regulatory sequences resulting in both the *de novo* methylation and the alteration in the $\alpha 1(I)$ collagen gene chromatin structure subsequently leading to the inactivation of the gene. Alternatively, the proviral integration may have disrupted or displaced DNA sequence elements which were essential for the transcription of the gene and thus the resulting change in the methylation status of the Mov 13 locus may have arisen as a consequence of the inactivation as opposed to its cause.

To distinguish between the two mechanisms cloned versions of the mutant and wild-

type allele which were propagated in _E. coli_, and therefore were demethylated, were introduced into fibroblast cell lines derived from day 12 Mov 13 homozygous embryos. In the above experiment, the wild-type allele was expressed whereas the mutant allele still showed no expression implying that the methylation status of the gene was not the mechanism at work in the Mov 13 mutant. To determine the difference between the displacement or disruption of an element responsible for the correct initiation of transcription of the gene a construct containing only a single LTR was generated and used in transfection experiments. This construct was also unable to activate transcription therefore adding support to the possible disruption of a cis-action element present in the first intron as opposed to its displacement from the promoter. This element is possibly different in tissues which overproduce or exclusively produce the type I collagen as was demonstrated by Kratochwil et. al. (1989) in which teeth derived from homozygote Mov 13 mice were able to produce normal levels of $\alpha 1$ (I) collagen.

Mov 15 TO 45 AND Glob 1 TO 3 LINES OF TRANSGENIC MICE

Using a similar protocol as above Soriano et al. (1987) an additional 34 transgenic lines of animals, Mov 15-45 and Glob 1-3, were generated. Heterozygote crosses of each line, with the exceptions of Mov 34 and Mov 24, generated the expected 1:2:1 Mendelian ratio of offspring. Male mice from the Mov 24 line only passed on the transgene to their male offspring indicating that the provirus integrated into the Y chromosome. As a result no homozygote animals were obtained for this line. Similar to the Mov 13 line described above, no Mov 34 homozygote animals were detected using Southern blot analysis of tail biopsies.

Examination of embryos at day 8 to 13 of gestation also failed to produce the appearance of homozygote animals. However a large number of conceptuses with reabsorbed embryos were seen when two heterozygote animals were crossed relative to wild type or wild type/heterozygote crosses. This suggested that homozygote embryos had implanted in the uterus but had failed to develop normally. Four to eight cell embryos were isolated from double heterozygote and control crosses. The same fraction of embryos developed to the blastocyst stage, indicating that homozygote embryos develop normally up to the blastocyst stage but died prior to day 8 of gestation. Closer examination revealed that homozygote embryos died before reaching the egg cylinder stage (day 6.5).

A 12 kb Eco RI fragment containing the provirus and flanking genomic region was cloned into a phage vector. Restriction analysis showed that the 5' LTR was 140 bp from one Eco RI site and that the 3' LTR was 2.3 kb from the second Eco RI site. This 2.3 kb fragment showed the presence of a single band in wild type DNA and an additional band in heterozygote samples using Southern blot analysis confirming that it in fact was derived from the proviral integration site. A unique 4.8 kb fragment 4-9 kb downstream of the 3' LTR detected a single 1.7 kb transcript using Northern blot analysis in all tissues examined. Quantitation of the amount of message in day 10 littermate embryos, relative to a α-tubulin, showed a 1.6 fold decrease in message in the heterozygote embryos and a 2.2 fold decrease in adult animals. This data indicates that the proviral integration interferes with the transcription of the gene. Subsequent mapping placed the provirus integration at a position 2.5 kb 5' to the Mov 34 transcriptional unit and indicated that the provirus was orientated in the same transcription direction as the gene (Soriano et al. 1987).

The identification of a possible role of the Mov 34 gene product was determined by Tsurumi et al. 1995. They were interested in isolating a regulatory subunit (p40) of the 26 S proteasome. The 40 kDa protein was purified and subjected to Lys-C protease and six purified peptides were subjected to partial protein sequencing. All six peptides showed sequence similarity to the mouse Mov 34 gene product. Using a PCR based strategy a human cDNA clone of p40 was obtained from human hepatoblastoma cells HepG2. The translated sequence showed 94.4% amino acid identity to the translated mouse protein thus providing a possible function for the Mov 34 product.

THE Mpv 17 LINE OF TRANSGENIC MICE

In contrast to the above, the Mpv 17 and Mpv 20 lines were generated by infection with the replication deficient recombinant retrovirus MPSVneo (Seliger et al. 1986, Weiher et al. 1987). Mpv 20 will be the focus of the first part of the thesis and will be examined in detail below. In contrast to the Mov 13, Mov 34 and Mpv 20 lines which display an early embryonic lethal phenotype, Mpv 17 mice die as adults as a result of chronic renal failure (Weiher et al. 1990).

Homozygote mutant mice are indistinguishable from littermates until 2-3 months of age. At the onset of disease, Mpv 17 animals displayed inactivity, weight loss, and pallor. Animals started to die by 9 weeks of age and over 90% of animals were dead by 18 weeks of age. Serum analysis at 6-8 weeks of age showed abnormal levels of creatinine, cholesterol and albumin. Blood samples showed a marked decrease in red blood cell count and hemoglobin consistent with the pallor observed as a result of anemia. Animals at the terminal stage of the disease displayed high amounts of protein in their urine consistent with

general kidney failure.

An LTR probe from MPSV was used to screen a genomic library generated from DNA obtained from an Mpv 17 mouse. Initially 3 clones were isolated each containing sequence 3' of the provirus. In order to obtain DNA 5' to the proviral integration site, a wild-type library was screened and the preintegration site clone was obtained. Southern blot analysis confirmed that the clone arose from the Mpv 17 locus. A conserved DNA fragment near the site of integration was used to probe Northern blots containing RNA derived from different tissues of both wild type and Mpv 17 mice. The results showed the presence of a 1.7 kb transcript present in all tissues examined in wild type animals. This transcript was also present in EC and ES cell lines indicating that the Mpv 17 message is ubiquitously expressed throughout embryonic development and in adult animals. In contrast, the transcript was completely absent in Mpv 17 mutant animals regardless of age or stage of disease. The above findings point to the provirus playing some role in disrupting the Mpv 17 locus. This disruption interferes with the expression of a stable transcript.

A database search of the Mpv 17 predicted protein product by Zawacka et al. (1994) identified homology to the peroxisomal membrane protein pmp22 (Kaldi et al. 1993). Immunofluorescence using an antibody directed against the Mpv 17 protein showed colocalization of the gene product with the peroxisomal protein catalase. This data taken together with functional data has led to the assignment of the Mpv 17 protein product as being a peroxisomal protein.

ACTIVATION OF CELLULAR GENES FOLLOWING RETROVIRAL INFECTION

The above experiments described animals in which the integration of a provirus into

the genome results in the loss of function of the disrupted gene. The following series of experiments will describe proviral insertional events which result in the activation of cellular genes.

INTEGRATION AT THE Evi-1 AND Cb-1/fim-3 LOCI

Morishita et al. (1988) were interested in identifying genes capable of transforming haematopoietic cells. They were interested in cells that still maintained their dependence on growth factors for viability, namely Interleukin-3 (IL-3), but were no longer capable of terminal differentiation or of being induced to differentiate. To this end, they examined cell lines which were isolated from retrovirus induced myeloid leukemias. These cell lines still displayed IL-3 dependence yet they had undergone an event which rendered them refractory to differentiation. To identify which genes were affected they examined sites of viral insertion into known proto-oncogenes. They also determined if any common viral integration sites could be identified. Five of the 37 cell lines examined contained proviruses in a common integration site designated the ecotropic virus integration 1 site (Evi-1). The cell lines examined further were designated NFS-60, NFS-78, NFS-48, and DA-1.

To identify whether viral integrations affected transcription, a Northern blot containing poly $(A)^+$ RNA from the NFS-78 cell line was probed with a genomic fragment spanning the preintegration sites of the four above mentioned cell lines. The result showed a weak and diffuse hybridization signal in the 1-7 kb range. A cDNA library was subsequently constructed and the same probe was used to isolate two independent clones (78-5, 78-8) which contained 2.2 and 2.5 kb inserts. Both cDNA clones contained CasBrM MuLV sequences, which was the retrovirus used in the infection of the parental cells, and cellular

sequences. A cDNA library was also constructed from RNA originating from NFS-58 cells. Screening using the 2.5 kb insert above resulted in the isolation of two independent clones (58-1, 58-2). No viral sequences were present in clone 58-2 and it was then used as a probe in Northern blots. Transcripts were detectable in cell lines containing *Evi*-1 rearrangements (NFS-58, NFS-60, NFS-78, DA-1) but not in cell lines without the *Evi*-1 rearrangements. Thus the provirus was promoting the transcription of a cellular gene which would not normally be expressed. Neither the *Evi*-1 protein nor the transcript have been detected in normal haematopoietic cells. The gene has been found to be expressed normally in kidney and developing oocytes in the ovary (Morishita et. al. 1990). The integration occurred near or in 5' noncoding exons of a novel gene. The cDNA sequence predicted the gene product to contain zinc fingers. These data demonstrated that retroviral activation of a gene encoding a putative zinc-finger protein may be implicated in the transformation of haematopoietic cells.

In dealing with the same system as described above, Bartholomew and Ihle (1991) examined a series of myeloid leukemia cell lines which demonstrated the expression of *Evi*-1 gene product but in which the viral integration was in a genetically linked locus *cb-1/fim-3* (DA-3 and DA-34). Pulse field gel electrophoresis demonstrated that the two loci were a maximum distance of 120 kilobases apart. Using probes from the 5' end of *Evi*-1 and from both ends of *cb-1/fim-3*, a chromosome walk was carried out. Restriction mapping placed the two loci 90 kb apart with the *cb-1/fim-3* locus oriented 5' to *Evi*-1. Northern analysis of RNA isolated from both DA-3 and DA-34 cell lines showed a 6 kb transcript when an *evi*-1 probe was utilized, which is the same size transcript seen in kidney and ovary. In order to confirm that the transcripts observed in the above lines were derived from the *Evi*-1 gene, cDNAs

were isolated from a DA-3 cell library with a *Evi*-1 probe. Thirteen independent clones were isolated and the largest clone (designated 21,1) was determined to contain the entire *Evi*-1 coding region by restriction digest mapping. S1 nuclease protection assays on DA-3 and DA-34 cells suggest that the *Evi*-1 transcripts were initiated at the start sites normally used in tissues that express *Evi*-1. The authors had previously demonstrated by Southern blotting that the majority of the retrovirus, integrated in DA-3 cells was deleted and possibly only a single LTR remained in the *cb-1/fim-3* locus (Bartholomew et al. 1989). They subsequently reasoned that if the LTR in DA-3 cells was functioning as a promoter, a stable transcript should be observed on Northerns when genomic fragments near the viral integration site are used as probes. Twenty different fragments spanning 40 kb of the integration site were employed and none showed any transcripts being present. The above data that the single LTR present at the *cb-1/fim-3* locus in DA-3 cells is either inactive or does not give rise to a stable transcript.

The activation of the *Evi*-1 gene by the LTR in *cb-1/fim-3* could arise from long range changes in the chromatin structure between the two loci thereby resulting in possible changes in the methylation state of the *Evi*-1 gene. This change might in turn influence the expression of the *evi-1* transcript. Changes due to the LTR presence at the *cb-1/fim-3* locus should be seen on the allele containing the integrated provirus. When a 5.5 kb Eco RI fragment which was C + G rich and contained the 5' region of *Evi-1* and included exon 1 was examined with both methylation sensitive and insensitive enzymes. The results showed that there was an increase in methylation on the allele harbouring the *cb-1/fim-3* provirus relative to the unrearranged allele. Increased methylation is normally associated with the inactivation of

speculated that an increase in methylation around the *Evi-1* gene promoter resulted in either the enhanced binding of transcription factors required for proper expression or alternatively prevented the binding of negative factors that would normally inhibit expression. This may be a possibility since a protein which binds methylated CpG islands has been discovered to be present in mammals (Lewis et al. 1992).

NPAT AND *ATM* AS CANDIDATE GENES AT THE MPV 20 LOCUS

Experiments which will be presented in the thesis will demonstrated that the provirus present in the Mpv 20 line integrated into the murine homolog of the human *NPAT* gene. The following is a brief description of the *NPAT* locus and how it was initially identified.

Ataxia-telangiectasia (ATM) is a human autosomal recessive disorder that manifests a wide range of clinical symptoms which include degeneration of the cerebellum, dilation of blood vessels in the eyes, ears, and face, atrophy of the thymus, elevated incident of cancer especially leukemia and lymphomas, and premature aging (reviewed in Harnden 1994, Shiloh 1995). The *ATM* gene has been cloned and mapped to human chromosomal position 11q22-23.

In a search to identify genes that may have a modifying effect on the ATM phenotype, a novel gene designated *NPAT* (Nuclear protein at the AT locus) has been identified which lies approximately 0.5 kb upstream of the *ATM* gene (Imai et al. 1996). A PCR strategy was employed to isolate coding sequences which lay in a YAC (yeast artificial chromosome) contig of ~ 2 Mb of DNA spanning human chromosome 11q22.3-q23.1. The strategy involved digesting the YACs with either BssHII or Sac II restriction enzymes both of which contain only G and C in their recognition sequence. It was reasoned that since the CpG

contain only G and C in their recognition sequence. It was reasoned that since the CpG dinucleotides are not scattered randomly in the genome but are instead clustered in CpG islands which are located 5' of genes (Bird 1986, 1987), therefore only regions that are 5' to genes of interest would be cut. Following the restriction digests synthetic linkers of known sequence were ligated to the ends of the DNA. A PCR step was employed using an oligonucleotide directed against the linker region and a second oligonucleotide against the human Alu repeat which is scattered throughout the genome. Amplified fragments were used to screen either a testis or a human placenta cDNA library resulting in the isolation of three overlapping clones. All three clones hybridized back to the YACs when used as probes on Southern blots. Sequence analysis revealed no similarities to any known entry in the databases. Southern blots indicated that clone T4-41 was located at the centre of the AT locus and thus was further characterized. Northern blot analysis using the above clone showed the presence of two hybridizing bands of 6.2 and 5.4 kb in all tissues examined. Southern blot analysis revealed that the T4-41 sequence was highly conserved throughout evolution with signals being detected in yeast, and seven vertebrate samples analysed. T4-41 was used as a probe to rescreen cDNA libraries resulting in the isolation of three additional clones which span 5.9 kb. Sequence of the composite cDNA identified 66 nucleotides in the 5' untranslated region followed by an open reading frame of 4281 nucleotides and a long 3' untranslated region of 1553 nucleotides. The translated sequence of the predicted protein of 1427 amino acids was subjected to PSORT analysis which identified a potential nuclear localization signal of four continuous basic amino acids in the carboxyl terminus. Database search of the complete *Npat* protein identified weak homologies with yeast *cdc24* (Bender and Pringle

1992), human protein-tyrosine phosphatase zeta precursor (Krueger and Saito 1992), and human nucleoporin NUP214 (Von Lindern et al. 1992). Also present were consensus sequences for phosphorylation with CDK4/E2F and CDK6/E2F complexes (Nigg 1993). Genomic analysis has demonstrated that the *NPAT* and *ATM* genes lie 0.5 kb apart and that they are transcribed in opposite directions using the same promoter (Imai et al. 1996). Comparative genomic analysis has shown that both murine genes are also closely linked and map to mouse chromosome 9 (Matsuda et al. 1996).

Atm knockout mice have been described (Barlow et al. 1996, Xu et al. 1996, Xu and Baltimore 1996). These mice are viable in the homozygote state, as opposed to the *Mpv 20* strain, and display all the hallmarks of having ATM. The viability of the *Atm* knockout mice excludes the locus as encoding a candidate gene for the observed recessive lethal phenotype in the Mpv20 strain.

(II)

Unp AND THE UBIQUITIN PATHWAY

Ubiquitin is a 76 amino acid peptide which is highly conserved across species lines and is best known for its role in tagging proteins for proteolytic degradation (reviewed in Hershko 1988 and Jentsch 1992). Briefly, the degradation pathway requires the activation of ubiquitin using an ATP dependent process. This links the terminal Gly of ubiquitin to the sulfhydryl group of a member of the ubiquitin-activating enzyme (E_1) family. The activated ubiquitin is subsequently transferred to the sulfhydryl group of a member of ubiquitin-conjugating enzyme (E_2). E_2 catalyses the transfer of activated ubiquitin to the ϵ-amino group of lysine residues in the target protein. In some instances a branched ubiquitin chain is added to the target protein by the action of a ubiquitin ligase (E_3) which transfers an activated ubiquitin, which is present on E_2, to the lysine 48 residue of the ubiquitin moiety already bound to the target protein. The targeted protein is then degraded by the 26S proteasome (reviewed by Rechsteiner et al. 1993) and the ubiquitin molecules are subsequently recycled.

In addition to its degradation role, ubiquitin has also been shown to be involved in a number of different processes. Palombella et al. (1994) have demonstrated that the ubiquitin-proteasome pathway is essential for limited proteolytic cleavage of the NF-κB1 precursor protein to yield the activated form. Ubiquitin also displays a chaperone function in aiding the proper assembly of the 40S ribosomal subunit (Finley et al. 1989). Ubiquitin has also been shown to be involved in the cell cycle at a number of different points which include the destruction of cyclins (Zachariae and Nasmyth 1996), regulating the abundance of the cyclin-

dependent kinase inhibitor p27 (Pagano et al. 1995), and in the separation of sister-chromatids in fission yeast (Funabiki et al. 1996).

The genes coding for ubiquitin are also well conserved in eukaryotic cells. The free 76 amino acid ubiquitin molecule is not synthesised in cells, but rather is synthesised as a linear polypeptide from four genes which fall into two separate classes. Class I genes encode a polyprotein of tandemly repeated ubiquitins which are subsequently released by the cleavage of the Gly-Met bonds that join the repeats. Class II genes encode a ubiquitin moiety which is covalently attached to one of two sequences of 52 or 76-80 amino acids. The 52 amino acid protein is found in the 60S ribosomal subunit whereas the yeast 76 amino acid protein is found in the 40S ribosomal subunit and was identified as being ribosomal protein S37 (Finley et al. 1989).

Since no free ubiquitin is present, an enzymatic activity must be present in cells which cleaves the ubiquitin precursors generating a pool of ubiquitin essential for the survival of cells. A family of deubiquitinating enzymes was identified which are capable of cleaving both the natural and engineered substrates thus generating a pool of the peptide. The Ubp proteins are prototypical members of the family of deubiquitinating enzymes and will be discussed below. Although all these enzymes have been shown functionally to have the ability to remove ubiquitin from both synthetic substrates as well as the Class II ubiquitin gene products, no other natural substrates have been identified.

During the search for candidate genes present at the *Mpv 20* locus, a cDNA was isolated from an undifferentiated F9 embryonal carcinoma library. Sequence analysis precluded this cDNA as originating from the *Mpv 20* locus. The cDNA sequence predicted

the possible presence of a nuclear localization signal. Northern analysis using a series of different tissues demonstrated the message as being ubiquitously expressed. From the above two pieces of data, the cDNA was designated *Unp* (*u*biquitous *n*uclear *p*rotein) (Gupta et al. 1993). A database search of the predicted translated product identified significant homology between *Unp* and the human *TRE* oncogene (Huebner et al. 1988) which is believed to have arisen from the fusion of DNA derived from chromosomes 5, 17, and 18. This homology was seen in the portion of the gene derived from chromosome 17, designated *TRE*-2 (Nakamura et al. 1992), and was the only region that was identified to be absolutely necessary for transformation. Chromosome mapping of *Unp* placed it on mouse chromosome 9 which is syntenic with human chromosome 3. From the mapping data, it was concluded that *Unp* was not the murine homolog of *TRE-2* but rather both genes were members of a larger family present on different chromosomes.

Two separate lines of evidence have lead to the assignment of the *Unp* gene product as being the mammalian prototype of deubiquitinating enzymes. First, family members of these enzymes have been identified by their ability to cleave ubiquitin from linear ubiquitin-protein fusions. Gilchrist et. al. (manuscript in preparation) have demonstrated that *Unp* possesses the ability to cleave ubiquitin from a linear substrate. *Unp* like the *UBP* family members of deubquitinating enzymes share homology in the conserved cysteine and histidine domains which are essential for catalytic activity (Baker et al. 1992, Tobias and Varshavsky 1991). Second, Papa and Hochstrasser (1993) identified that the yeast *DOA4* gene encodes a deubiqitinating enzyme which is related to the human *TRE-2* oncogene (renamed to *TRE-17* to reflect its chromosomal location).

Some of these enzymes may have interchangeable target specificity. Baker et al. (1992) demonstrated that a yeast strain which lacks four genes coding for deubiquitinating enzymes (*YUH1*, *UBP1*, *UBP2*, and *UBP3*) grows normally and still displays high deubiquitinating activity. These data suggest that other unidentified family members are functional in the above strain. Alternatively, a yeast strain lacking the *DOA4* gene (Papa and Hochstrasser 1993) results in the yeast displaying poor growth. The data point to an accumulation of ubiquitinated substrates and the authors speculate that the gene product may be functioning late in the deubiquitinating pathway and may be associated with the 26S proteasome.

Since *Unp* is a member of a larger family of mammalian deubiquitinating enzymes, we were interested in generating a mouse which carries the disrupted gene to examine the possible role played by the gene product in a murine system. This would enable us to determine if *Unp* has redundant functions in a cell or whether it is essential for cell survival. To achieve this end we needed to clone the gene and determine its structure i.e. the positions of the intron/exon borders.

The second part of the thesis will deal with the genomic sequencing of the *Unp* gene and some of the studies that were carried out in order to initially identify and partially characterize its promoter.

CHAPTER 2

MATERIALS AND METHODS

GENERATION OF TRANSGENIC MOUSE LINES

Retroviral infection of 4 to 16 cell CFW mouse embryos was previously carried out in the laboratory of R. Jaenisch using a published protocol (Jähner and Jaenisch, 1980; Weiher et al., 1990). Briefly, Rat 1 cells infected with MPSVneo and F-MuLV (provided by W. Ostertag, University of Hamburg) as helper-virus were used as producer cells. Embryos were co-cultured with producer cells, then transferred to pseudopregnant C57B1/6J x CBA females. Progeny mice were genotyped by Southern analysis of tail biopsy DNA using a neo-specific probe. From 42 mice examined, 11 contained one or more copies of the neo gene. Eighteen transgenic strains were obtained by breeding founder animals to wild-type CFW mice.

IN VITRO CULTURE OF EMBRYOS

Five week old Mpv 20 heterozygous females were superovulated by intraperitoneal injection of 5 I.U. pregnant mares' serum gonadotropin followed 44 h later by 5 I.U. human chorionic gonadotropin (hCG). Following the second injection the females were paired with either heterozygous or normal CFW male mice. The following morning the females were checked for vaginal plugs. Mated females were killed 48-64 h after hCG injection and their oviducts and uteri were gently flushed with CZ-B embryo culture medium (Chatot et al., 1989) supplemented with 875 mg/l taurine (CZ-B/T). Embryos were collected, washed, and transferred to 0.8 ml CZ-B/T supplemented with 1 mg/ml glucose in embryo culture dishes

(Falcon; Becton Dickinson, Lincoln Park, New Jersey) for subsequent culture. Embryo development was assessed at 12 h intervals monitoring the degree of compaction.

MOLECULAR CLONING

The Mpv 20 integration site was cloned by established methods as described in Sambrook et al. (1989). Briefly, Bgl II digested liver DNA from a heterozygous animal was fractionated on a 0.7% agarose gel. DNA in the size range of 10 to 20 kb was purified from a gel slice using Geneclean (Bio101), and ligated into the phage vector λEMBL3 (Stratagene). Ligated DNA was packaged using Gigapack reagents (Stratagene). Phage were plated on the *E. coli* K802 strain. Filter lifts were hybridized (5X SSC, 5X Denhardt's, 0.5% SDS, 100 μg/ml herring sperm DNA, 50% formamide) to a radiolabelled 754 bp Nhe I/Spe I fragment derived from the MPSV LTR. Filters were hybridized in standard solution (5X SSC, 50% formamide, 42°C) and washed using high stringency conditions (0.1X SSC/0.1% SDS, 65°C), under which the probe does not detect endogenous proviral sequences on Southern blots. Five positive clones were identified from approximately 10^6 phage, one of which was λBg4.1. Additional clones were obtained using a 2.5 kb genomic Bam HI fragment from λBg4.1 to screen a phage library constructed from DNA partially digested with Mbo I, fractionated into 15-20 kb lengths by sucrose gradient centrifugation, and ligated into λEMBL3 arms as above.

The Unp gene was cloned from D3 (mouse embryonic stem cells) DNA partially digested with Mbo I, fractionated as above, and ligated into λDASH II (Stratagene) arms. Initial clones were isolated using a radiolabelled Unp cDNA probe. End fragments of the isolated clones were used to perform a chromosome walk in both directions.

SCREENING OF LIBRARIES

Titers of libraries were determined by mixing 100 μl serial dilutions of phage in suspension media (SM) (100 mM NaCl, 10 mM $MgSO_4$, 50 mM Tris pH 7.5, 0.01% gelatin) to 100 μl of overnight *E. coli.* grown in LB/Mg (10g bacto-tryptone, 5g yeast extract, 10g NaCl and 10 mM $MgSO_4$ per liter of water). Mixtures were incubated at 37°C for 15 min and 5 ml of melted top agarose (0.75% agarose in LB/Mg) at 42°C was added and spread on 1.5% LB/Mg agar plates. Agarose was allowed to solidify at room temperature prior to overnight incubation at 37°C. Strains of *E. coli* employed were as follows: for genomic libraries K802, and for cDNA libraries XL-1 Blue strain.

Following determination of titers, 500 000 pfu were plated onto 150 mm agar plates as described above. Following overnight incubation, plates were chilled at 4°C for 1 hr and lifts were performed, in duplicates, by placing nylon membranes over the plates and allowing DNA to transfer. DNA was denatured by autoclaving membranes at 100°C for two min and subsequently cross-linked in a UV chamber (BioRad). Membranes were prehybridized for greater than two hours at 42°C and hybridized overnight (42°C) in the presence of radiolabelled probes.

Positive clones were picked as agarose plugs and transferred to 1 ml of SM media containing two drops of chloroform ($CHCl_3$). Phage were allowed to elute from plugs either for 4 hr at room temperature or overnight at 4°C. Phage were subjected to two more rounds of purification prior to preparation of DNA.

PHAGE MINIPREPS

PRIMARY INFECTION

A phage plaque was picked and added to one drop of overnight bacterial culture. Following a 10 min incubation at 37°C, 2 ml of LB media (containing 10 mM Mg^{+2}) was added and culture was incubated until it cleared (4-6 hr). Following incubation, 2 drops of $CHCl_3$ was added and debris pelleted by centrifugation (4000g, 5 min).

SECONDARY INFECTION

Fifteen µl of primary lysate was added to 1 ml of overnight bacterial culture. Following absorption for 10 min at 37°C, culture was added to 50 ml of LB media (Mg^{+2}) and incubated at 37°C with vigorous shaking overnight. Ten drops of $CHCl_3$ and 1.5 g NaCl are added and debris was pelleted.

DNA PREPARATION

Phage lysate was added to 5 g of PEG 8000 (polyethyleneglycol) and placed on ice for 1 hr. Following incubation, phage were collected by centrifugation for 30 min (3000g, 4°C). Supernatant was removed and pellet resuspended in 0.6 ml SM solution. DNase I (2.5 µl of 10 mg/ml) was added and incubated at 37°C for 1 hr. Following incubation, 2.5 µl diethlypyrocarbonate (DEPC) was added. The sample was subsequently divided into two 0.3 ml aliquots and 0.6 ml 1M Tris pH 8.5, 1% SDS, 0.1M EDTA was added to each aliquot. Samples were heated to 65°C for 5 min. A of 5M solution of potassium acetate (0.3 ml) was added to each tube and incubated on ice for 30 min. Debris was remove by centrifugation in a microfuge (15 min, 4°C). Supernatants were transferred to fresh tubes and filled with isopropanol. DNA was collected and pellets were resuspended in 200 µl of TE. Samples were pooled, phenol extracted once and ethanol precipitated. Pellets were resuspended in

200 µl of TE.

RNA ISOLATION USING THE GUANIDINIUM/CESIUM CHLORIDE METHOD

Mice were sacrificed by cervical dislocation. Tissues were removed and used fresh or frozen in liquid nitrogen and stored at -80°C. The procedure was according to Turpen and Griffith (1986). Briefly, tissues were homogenized in 5 volumes (g of tissue/ml solution) of 4 M guanidinium isothiocyanate, 5 mM sodium citrate (pH 7.0), 0.1 M β-mercaptoethanol, and 0.5% Sarkosyl. One g of cesium chloride (CsCl) was added for each 2.5 ml of homogenate. Homogenates were layered onto 2.0 ml cushions of 5.7 M CsCl in 0.1 M EDTA. Samples were centrifuged at 20 000 rpm in a Beckman SW41 rotor for 24 hr at 4°C. Following centrifugation, RNA pellets were dissolved in 10 mM Tris (pH 7.4), 5 mM EDTA, and 1% SDS and extracted with chloroform. To the aqueous phase, 0.1 volume of 3 M sodium acetate and 2.2 volumes of ethanol were added and stored at -20°C for at least 2 hr. RNA was collected by centrifugation at 10 000 rpm for 15 min in a Beckman JA-20 rotor. Pellet was resuspended in 1 ml of H_2O and stored at -80°C. Samples were quantitated using UV spectroscopy using an O.D. at 260 nm of 1 to be equivalent to 40 µg/ml of RNA (Sambrook et al. 1989).

RNA ISOLATION, LiCl METHOD

Fresh or frozen tissues were placed in approximately 3 ml of 6 M urea 3 M LiCl. Samples are subsequently homogenized using a polytron (setting 7) for one min. RNA was precipitated by incubating the sample overnight at 4°C. Following overnight incubation, RNA was collected by centrifugation for 30 min (4°C) at 10 000 rpm in a Beckman JA-20 rotor. Pellet was resuspended in 2.5 ml of urea/LiCl solution and spun as described above.

Following centrifugation, pellet was resuspended in 2 ml of 10 mM Tris (pH 7.5), 0.5% SDS, and 50 µg/ml of proteinase K. The resulting solution was incubated for 20 min at 37°C. The solution was subsequently extracted twice with phenol followed by two $CHCl_3$:isoamyl alcohol (24:1) extractions. NaCl was added to a final concentration of 0.1 M, followed by the addition of 2.5 volumes of ethanol, and precipitated overnight at -20°C. RNA was collected by centrifugation and resuspended in 0.2 ml of water.

NORTHERN BLOTTING

Tissues were flash frozen in liquid nitrogen, and processed for RNA extraction by the LiCl/urea method. RNA samples were resuspended in water, and optical densities determined spectrophotometrically. Equal quantities of RNAs were pipetted into a loading dye containing ethidium bromide and separated on a 1% formaldehyde-agarose gel (Sambrook et al. 1989). The integrity and quantity of RNA samples were confirmed by UV transillumination. RNAs were transferred to Hybond N (Amersham) by capillary transfer using 10 X SSC. RNAs were crosslinked to the membrane in a UV chamber and hybridized using cDNA fragments labelled with ^{32}P dCTP using a random oligonucleotide system (Boehringer Mannheim). Hybridizing signals were quantitated by densitometer scanning using a Molecular Dynamics phosphoimager.

RESTRICTION DIGESTS

Digests were carried out using enzymes obtained from New England Biolabs (NEB), Gibco (BRL), and Boehringer Mannheim (BM) following incubation conditions of the manufacturer. For plasmid DNA 2 units of enzyme was used per μg of DNA and incubated for 2 hrs. Genomic DNA was digested overnight with 10 units of enzyme per μg DNA.

DNA ISOLATION, SOUTHERN BLOTTING

ISOLATION OF GENOMIC DNA

Approximately 0.5 g of fresh or frozen tissues were minced. Tissues were added to 3 ml of extraction buffer (75 mM NaCl, 25 mM EDTA, 10 mM Tris (pH 7.6), 1% SDS, and 400 μg/ml proteinase K) and incubated either at 50°C for 3 hr, or overnight at 37°C with periodic swirling. Following incubations, samples were gently extracted three times with an equal volume of phenol, and dialyzed overnight against TE (10 mM Tris, pH 7.5, 1 mM EDTA) at 4 C. Following dialysis, samples were quantitated using UV spectroscopy using an O.D. at 260 nm of 1 to be equivalent to 50 μg/ml of DNA (Sanbrook et al. 1986).

DNA was digested with the appropriate enzymes and fractionated on agarose gels. Gels were denatured for a minimum of 30 min in 0.5 N NaOH, 1.5 M NaCl and subsequently neutralized for 30 min in 1 M Tris pH 8.0, 1.5 M NaOH prior to transfer. Transfers were by capillary onto Hybond N (Amersham) membranes using 10 X SSC. DNA was crosslinked prior to hybridization with radiolabelled ^{32}P dCTP probes.

MINI PREPARATION OF PLASMID DNA, BOILING LYSIS METHOD

Bacteria were inoculated in 5 ml of LB/Amp media and incubated overnight at 37°C. DNA was isolated following the protocol of Holmes and Quigley (1981). Cells were transferred to 1.5 ml tubes and spun for 2 min. Cells were resuspended in 350 μl of STET (8% sucrose, 0.5% Triton X-100, 50 mM EDTA, and 10 mM Tris, pH 8). Twenty-five μl of lysozyme solution (10 mg/ml in 10 mM Tris, pH 8) was added, vortexed for 1 sec, and boiled for 3 min. Samples are spun for 15 min and supernatant was collected . Isopropanol (350 μl) was added to the supernatant and DNA was recovered by centrifugation.

LARGE PREPARATION OF PLASMID DNA, TRITON LYSIS METHOD

Bacteria were inoculated into 500 ml of LB/Amp media (ampicillin 150 μg/ml) and incubated at 37°C overnight. Cells were harvested by centrifugation at 10 000 rpm in a JA-14 rotor at 4°C. Supernatant was discarded and cells were resuspended in 5 ml of STE solution (50 mM sucrose, 25 mM Tris, pH 8, and 10 mM EDTA). One ml of 0.5 M EDTA and 1 ml lysozyme solution (hen egg white lysozyme, freshly dissolved at 5 mg/ml in 0.25 M Tris, pH 8) and incubated for 5 min. An additional ml of 0.5 M EDTA was added and incubated for 5 min. Solution was subsequently placed on ice for 10 min. Eight ml Triton solution (0.1% Triton, 50 mM Tris, pH 8, and 60 mM EDTA) was added and incubated on ice for 15 min. Following incubation, samples were centrifuged for 1 hr at 20 000 rpm in a Beckman JA-20 rotor (4°C). Solutions were transferred to fresh centrifuge tubes, 50% volume of PEG solution added and incubated on ice for 30 min (30% polyethylene glycol (PEG) 8 000, 1.5 M NaCl). DNA was collected by centrifugation for 10 min at 10 000 rpm in a Beckman JA-20 rotor (4°C) and pellet redissolved in 2 ml TE. CsCl (2.26 g) and 50 μl of 10 mg/ml ethidium bromide were added. Solutions were transferred to 3 ml ultracentrifuge tubes, sealed and spun at 95 000 rpm for 2 hr or 60 000 rpm overnight in a Beckman TLV-100 rotor. Plasmid band was removed and extracted with water saturated butanol until pink colour was completely removed. Two volumes of TE buffer and 6 volumes of ethanol were added and plasmid collected by centrifugation at 7 000 rpm for 10 min (Beckman JA-20 rotor). Plasmids were resuspended twice in 0.5 ml of 0.3 M sodium acetate and precipitated with 1 ml ethanol. Final pellets were resuspended in desired volume of TE buffer.

SUBCLONING OF 4 KB FRAGMENT

Genomic clone M2017 was digested with Eco RI and fractionated on a 0.8% preparatory gel. A 4 kilobase (kb) fragment was excised from the gel and purified by Geneclean (Bio101). The fragment was cloned into the Eco RI site of plasmid pGEM-4 (Promega) and introduced into DH5α strain *E. coli* which was made competent using a RbCl protocol (Hanahan 1983). Plasmid was incubated in the presence of *E. coli* for 30 min on ice and subsequently subjected to heat shock at 42°C for 2 min. One ml of LB media was added and bacteria was allowed to recover for 1 hr at 37°C. Aliquots of recovered bacteria were plated onto LB/Ampicillin plates (150 μg/ml ampicillin) and incubated at 37°C overnight. Colonies were screened by boiling lysis DNA mini preps.

EXON TRAPPING

Phage spanning the integration site were digested with Sal I to release inserts. Inserts were isolated using Geneclean (Bio101) and either subjected to a partial Mbo I digest, or to complete digest with Bam HI. DNA was subsequently cloned into the Bam HI site of the exon trapping vector pSPL1 (Buckler et al. 1991)(Gibco BRL).

PREPARATION OF PLASMID DNA

DNA was prepared using the CsCl protocol described previously. Twenty μg of DNA was used per transfection. In control transfections, 50 ng of control plasmid was mixed with herring sperm DNA to bring the final amount of DNA to 20 μg. DNA was ethanol precipitated and resuspended to a final concentration of 40 μg/ml in sterile 0.1X TE (1 mM Tris pH 8.0, 0.1 mM EDTA pH 8.0). Four-hundred and forty-four μl of the above DNA was mixed with 500 μl of sterile 2X HEPES buffered saline (HBS) (280 mM NaCl, 10 mM KCl,

1.5 mM $Na_2HPO_4 \cdot 2 H_20$, 12 mM dextrose, and 50 mM HEPES pH 7.05). Sixty-two µl of 2 M $CaCl_2$ was slowly added to the DNA and the resulting mixture was incubated at room temperature for 20-30 min.

TISSUE CULTURE

Cos-1 cells (African green monkey kidney cells)(ATCC) were plated out at a density of $8x10^5$ to $16x10^5$ cells per 100mm culture dish in the presence of 10 ml Dulbecco's modified Eagle's medium supplemented with 10% fetal calf serum. Cells were incubated at 37° C in a 5% CO_2 atmosphere until 40-60% confluence. Following incubation, media was removed and $CaCl_2$ precipitated DNA added. Cells were rocked gently and incubated at room temperature for 15 min. Following incubation, 9 ml of media containing 100 µM chloroquine was added and incubated for 7-8 hrs at 37°C. Media was removed, cells were washed with phosphate buffered saline (PBS), and 10 ml of fresh media was added. Cells were incubated at 37 °C for 24 to 48 hrs prior to isolation of RNA. RNA was isolated using the guanidinium /cesium chloride method.

cDNA was generated using manufacturer's conditions (BRL). Briefly, first strand synthesis was primed using 20 µM of the SA4 oligonucleotide 5' CACCTGAGGAGTGAATTGGTCG 3' in the presence of 1-3 µg of total RNA. Two-hundred units of Superscript II reverse transcriptase was added to each reaction and incubated at 42°C for 30 min. Reactions were heated to 55°C and 3 units of RNase H was introduced followed by a 10 min incubation.

PRIMARY AMPLIFICATION

One half of each reverse transciptase (RT) reaction was used in the primary PCR

reaction in the presence of 100 μM of each oligonucleotide SA4 and SD3 5' GTGAACTGCACTGTGACAAGCTGC 3' in the presence of 2 units of Vent exo- polymerase (New England Biolabs). Cycling conditions were as follows: 35 cycles at 94°C for 45 sec, 58°C for 30 sec, and 72°C for 1 min, followed by an additional hold for 10 min at 72°C. One tenth of the PCR reactions were subsequently fractionated on 2% preparative agarose gels and bands greater than 379 bp were isolated and purified by Geneclean.

SECONDARY AMPLIFICATION

Purified primary PCR product was reamplified using 100 μM of each of the following nested oligonucleotide primers dUSA1 5' CUACUACUACUACACCTGAGGAGTGAATTGGTCG 3' and dUSD1 5' CUACUACUACUAGTGAACTGCACTGTGACAAGCTGC 3' using the same cycling conditions as above. Secondary PCR products were treated with uracil DNA glycosylase to remove the uracil residues, producing abasic sites that do not form base pairs. The resulting products contain single-stranded termini and are subsequently subcloned into pAMP10 (Buckler et al. 1991).

SYNTHESIS AND PURIFICATION OF OLIGONUCLEOTIDES

Oligonucleotides were synthesized on a Applied Biosystems (ABS) model 392 synthesizer using phosphoamidite monomers. Oligonucleotides were purified using ABS oligonucleotide purification cartridges (OPC) following the supplier's protocol. All reagents used were of HPLC grade. Briefly, oligonucleotides were cleaved from the synthesis cartridge with 3 ml of ammonium hydroxide (NH_4OH). Solutions were subsequently incubated at 55°C overnight prior to purification. Following incubation 1 ml of H_2O was

added and oligonucleotides were bound to an OPC column, which was primed by flushing the column with 5 ml of acetonitrile (CH_3CN) followed by 5 ml of 2 M triethylamine. Columns were subsequently washed with 3x5 ml of 1.5 M NH_4OH and rinsed with 2x5 ml H_2O. Trytl blocking groups were removed by addition of 5 ml 2% trifluoroacetic acid (TFA). One ml of the above solution was passed through the column and allowed to remain therein for 5 min at room temperature. Subsequently, the remaining TFA solution was passed through the column and the column was rinsed with 2x5 ml of H_2O. Oligonucleotides were eluted with 1 ml of a 20% CH_3CN solution and dried in vacuo. Oligonucleotides were quantitated by UV spectra using an optical density (O.D.) of 1 at 260 nm to be equivalent to 33μg/ml of oligonucleotide (Applied Biosystems 1987).

DNA SEQUENCING

Sequencing was performed either manually using the Sanger dideoxy method (Sanger et al. 1977) or using an automated system. Manual DNA sequencing was performed using a T7 dideoxy sequencing kit (Pharmacia). Briefly, 2-3 μg of plasmid DNA was denatured for five minutes at room temperature in the presence of NaOH. Samples were neutralized using Tris pH 4.5, and ethanol precipitated. Twenty ng of primer was added to each template and sequencing reaction was carried out in the presence of ^{35}S-dATP. Reactions were run on a 6% acrylamide/50% urea denaturing gel.

Automated sequencing was performed on an ABI 377 DNA Sequencer using dye-terminator chemistry using AmpliTaq sequencing kits (Perkin-Elmer 1995).

SUBCLONING OF PCR PRODUCTS

All PCR fragments generated were gel purified using Geneclean (Bio101). Inserts

were subsequently phosphorylated using 2 units of T4 polynucleotide kinase in 15 μl of 1X T4 ligation buffer for 30 min at 37°C. Following incubation, the kinase was inactivated by heating the reaction for 20 min at 70°C. Additional ligase buffer was added either in the presence of pGEM-4 (Promega) digested with Sma I or pLitmus 28 (New England Biolabs) digested with Eco RV and 2 units of T4 DNA ligase in a final reaction volume of 25 μl.

PROBES

All PCR generated probes were amplified with Vent exo- polymerase using 2 units of enzyme per 100 μl reaction volume.

The GRAIL probe was amplified from the 4kb Eco RI genomic fragment originating from phage M2017 using the following oligonucleotide primers: G-for 5'GTGCGCCCTGCTCGAGG 3' and G-rev 5'GTGCTCTGCATCTTGTG 3'. Cycling conditions were for 20 cycles at 94°C for 40 sec, 60°C for 1 min, and 72°C for 1 min, followed by a 10 min hold at 72°C.

The *Npat* RT-PCR probe was amplified from a murine testis cDNA library using the following oligonucleotide primers: 5 mnpat 5' TTTAATGTCTCCTGGTAGACG 3' and 3 mnpat 5' GGCCTAGCAGTTCATCAATTG 3'. Cycling conditions were 40 cycles at 94°C for 30 sec, 55°C for 30 sec, 72°C for 1 min , followed by a 10 min cycle at 72°C.

A 2.4 kb *ATM* probe was generated by Eco RI digestion of the mouse mATM 6 cDNA (courtesy of T. Shiomi , National Institute of Radiological Sciences, Inage-ku, Japan) followed by purification from low melt agarose.

A 5.0 kb Sac I fragment containing most of the coding region was obtained from wild-type Moloney virus. An 754 bp LTR probe from MuLV was isolated from a

Nhe I/Spe I digest.

HMG box probe was isolated as a 380 bp Bgl II/Pst I fragment of plasmid p422 (ATCC) which contained a nonrepetitive 3.5 kb Eco RI fragment of the mouse *Sry* gene (Gubbay et al. 1990).

5' RAPID AMPLIFICATION OF cDNA ENDS (RACE)

5' RACE was carried out using a random primed/oligo dT lung library (Stratagene, lambda ZAP II) as a template. Two μl of the cDNA was mixed with 10 pmoles of each oligo 3 mnpat (see above) and KS 5' CGAGGTCGACGGTATCG 3' from pBluescript KS (Stratagene) and subjected to the following cycling conditions in the presence of 2 U of Vent exo-: 94°C for 4 min , 80°C for 2 min, followed by 40 cycles of 94°C for 30 sec, 58°C for 30 sec, and 72°C for 1.5 min followed by one additional cycle at 72°C for 10 min.

HMG BOX PCR

HMG PCR was essentially as described in Denny et al. (1992b). Briefly phage DNA spanning the Mpv 20 locus were digested with Sal I to release the inserts. Inserts were purified from low melt agarose gels by Geneclean and 50 ng of each were used as templates. Oligonucleotides were from a conserved region among HMG box containg proteins and were as follows: Soxfor 5' ATGAA(C/T)GC(A/C/T)TT(C/T)AT(A/G/T)GT(A/G/C/T)TGG 3' and Soxrev 5' GG(A/G/C/T)(C/T)(G/T)(A/G)TA(C/T)TT(A/G)TA(A/G)T(C/T)(A/G/T)GG 3'. Cycling conditions were as follows: 94°C for 5 min, 80°C for two min, followed by 30 cycles at 94°C for 30 sec, 48 or 52°C for 30 sec, and 72°C for 1 min. PCR products were gel purified and sequenced directly using the above oligonucleotides.

GENOMIC PCR

Genomic PCR was carried out using 2 μg of CFW mouse DNA as a template. Oligonucleotides were as follows: LR 5' CTGCTGGCTTGAACACTG 3' and 2041 5' TCACATCATCTGGACCGC 3'. PCR was carried out using the expand high fidelity PCR system (Boehringer Mannheim). Cycling conditions were an initial 94°C for 4 min, 80°C for 2 min, followed by 10 cycles at 94°C for 15 sec, 55° C for 30 sec, and 72°C for 2 min, an additional 15 cycles in which the extension time increased by 20 sec per cycle with a final 10 min hold at 72°C. Insert was sequenced using the LR oligonucleotide above and the following oligonucleotides: Bgl1 5' GAGAGCAAGAGGATTGGC 3', Bgl 2 5' CAATCTGGAACACAGGAC 3', Ambanti 5' GTCTGTCTGCTTAGGTAC 3', and Ambsense 5'GTACCTAAGCAGACAGAC 3'.

CONSTRUCTION OF CAT PLASMIDS

A 1.1 kb Eco RI/Sac I genomic fragment containing the first 108 bases of the cDNA and 1 kb of sequence immendiately 5' to the cDNA start, was ligated into the Eco RI/Sac I site of pLitmus 28 (New England Biolabs) generating pDG75. The insert was subsequently removed by a Nsi I/Kpn I digest and cloned into the Pst I site of the pCAT (Promega) reporter plasmid generating pDG83 and pDG84. pDG89 was generated by first digesting pDG83 with Bgl II followed by purification on a 0.8% low melt agarose gel by Geneclean and subsequently religated. pDG101 and pDG102 were generated using a PCR protocol. Oligonucleotides corresponding to positions 570-588 and 970-987 (numbering from genomic sequence) were synthesized carrying a Pst I site on their 5' end. PCR was performed using 50 ng of pDG75 as template and 10 pmoles each of oligonucleotide 5' promoter 5' AAAAACTGCAGCTCAATGCGTCCGTGTGC 3' and oligonucleotide 3' promoter

5' AAAAACTGCAGGCGCACGGCGTGCTATGGG 3'. Cycling conditions were 94 °C for 4 min. 80 °C for 2 min followed by 30 cycles of 94 °C for 30 sec, 55 °C for 30 sec. and 72 °C for 1 min and one 10 min extension at 72 °C using 2 units of Vent exo- DNA polymerase. PCR product was subsequently digested with Pst I and cloned into the Pst I site of pCAT.

CAT ASSAYS

Transiently transfected 293T primary human embryonal kidney cells (ATCC) were harvested 24 hrs following transfection. Dishes were trypsinized and recovered cells were washed with ice cold PBS. Cells were resuspended in 100 μl of 0.25 M Tris pH 7.5 and lysed by three rounds of freeze/thaw (ethanol/dry ice bath/ 37 °C bath) cycles. Debris was removed by a 10 min centrifugation in a microfuge at 4 °C. Soluble proteins were quantified using the BioRad micro assay. 100 μg of each sample was used to perform β-galactosidase assay (see below). Remaining samples were heated for 10 min at 65 °C and spun as above. Five μg of soluble proteins were used for CAT assays using a Green Fast fluorescence kit (Molecular Probes). A 1 μl aliquot was spotted onto a thin layer chromatography (TLC) plate (VWR) and separated using 15% methanol/ 85% $CHCl_3$ as the mobile phase. TLC plates were analyzed on a Storm (Molecular Dynamics) flouroimager using "Image Quant" software. As an internal control cells were cotransfected with pDM2 (gift from Dr McBurney) in which the pgk-1 promoter (3-phosphoglycerate kinase) is driving the expression of lacZ.

β-GALACTOSIDASE ASSAY

100 μg of protein was mixed with 200 μl of ONPG (2-Nitro-phenyl-β-D-galactopyranoside) (4 mg/ml) (BDH) in assay buffer (100 mM $NaPO_4$ pH $_{7.0}$, 1 mM $MgSO_4$, and 10 mM β-mercaptoethanol). The absorbance was measured at 420 nm in a Beckman DU

640 spectrophotometer and the data was analyzed using the manufacturer's "Kinetic/Time" program. The slope was used to compare the efficiency of transfection between samples.

CHAPTER 3

RESULTS

(I)

GENERATION AND IDENTIFICATION OF A MUTANT TRANSGENIC MOUSE LINE

Mpv 20 transgenic mice were generated during the same screen for recessive lethal mutations that generated the Mpv 17 strain, whose adult kidney phenotype has been described previously (Weiher et al., 1990). Of the eighteen transgenic mouse lines tested, only Mpv 17 and Mpv 20 were found to harbour recessive lethal phenotypes. In the Mpv 20 strain, no animals were identified among the progeny of heterozygous crosses that were homozygous for the MPSV-neo provirus (the results of tail DNA biopsies of test crosses summarized in Table 1). No homozygotes were identified among day 11 midgestation embryos from similar crosses (Table 1). These data suggested an early recessive lethal phenotype, possibly occurring prior to implantation. To examine this possibility preimplantation embryos were cultured from heterozygous x heterozygous or heterozygous x wild type (wt) crosses. Approximately one quarter of embryos from crosses of heterozygous x heterozygous animals failed to proceed from the morula to the blastocyst stage (Table 2). Many of the arrested embryos had the appearance of uncompacted 8 cell morulae (Figure 1).

Table 1. Progeny of 20/wt x 20/wt matings.
DNA was isolated from the Mpv 20 heterozygote (20/wt) animals and subjected to slot-blot analysis. A neo probe (Beck et al. 1982) was used to identify the presence of the transgenic allele.

Table 2. Development of embryos to blastocysts *in vitro.*
Matings were established as noted. Fertilized oocytes were flushed from the mice ovaducts and cultured *in vitro* at 37°C. Progression of development was monitored every 12 hours.

Table 1

Progeny of 20/wt x 20/wt Matings

	Total	wt/wt	20/wt	20/20
adults	199	67	132	0
day 11 embryos	55	21	34	0

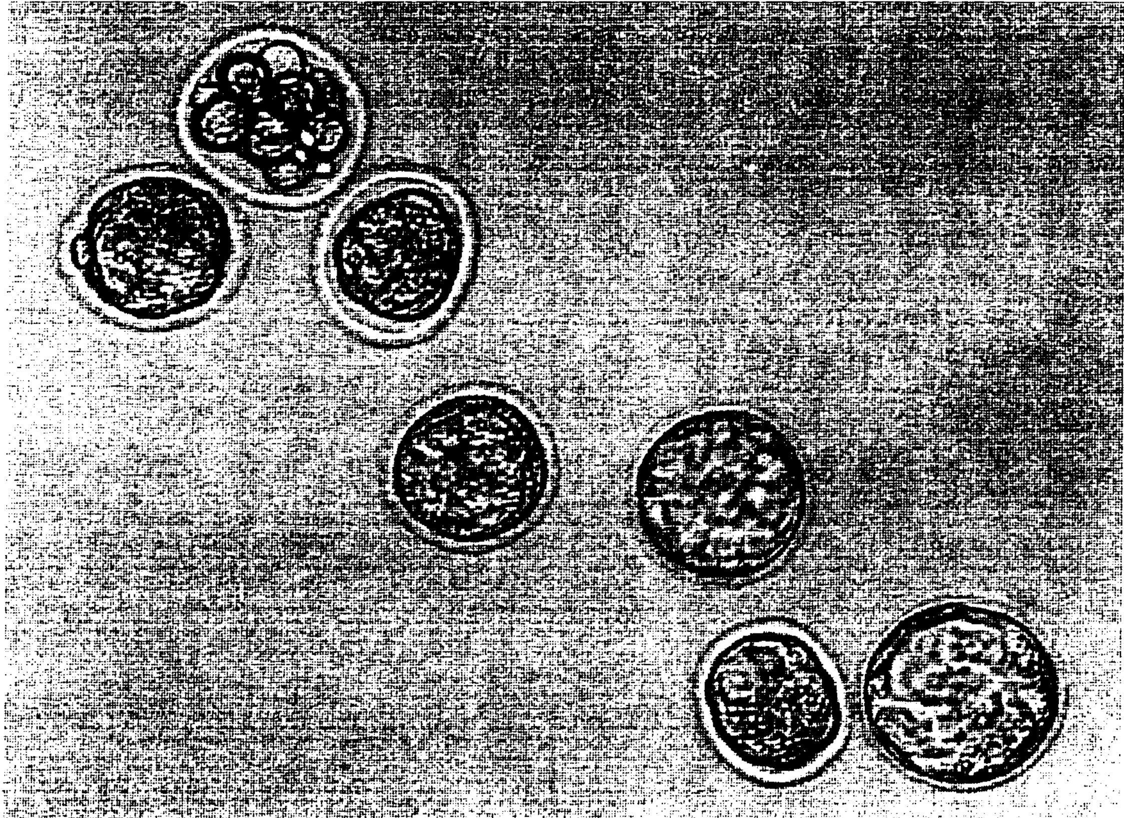
Table 2

Development of Embryos to Blastocysts *In Vitro*

mating	blastocysts	percent
wt xwt	49/54	91
wt x 20/wt	43/50	86
20/wt x 20/wt	110/155	71*

Significant difference by Z-test
(α=.0006)

Figure 1. Appearance of cultured embryos from an Mpv 20 heterozygous cross. Photograph was taken using phase contrast microscopy at 3.5 days postcoitum. The uppermost embryo, indicated by the open arrowhead, arrested at the uncompacted 8 cell stage, while the remainder compacted and in some cases formed blastocysts.



MOLECULAR CLONING OF THE MPV 20 PROVIRAL INSERTION SITE

To identify the gene or genes affected by the insertion of proviral sequences, the region of the insertion was cloned into bacteriophage vector EMBL3. A proviral LTR fragment was used as a probe to screen a genomic library constructed from Mpv 20 heterozygous liver DNA. This library consisted of gel purified 13 kb Bgl II fragments ligated in to the Bam HI site of the lambda vector. The clone isolated by this strategy (designated Bg4.1 on Figure 2) was found to contain a 13 kilobase proviral junction fragment. From clone Bg4.1, a mouse genomic 2.5 kb Bam HI fragment was chosen as a probe to screen an Mbo I partial-digest library of wild type DNA. This screen yielded clone M2041. Using the same strategy, genomic probes were obtained which detected phage clones with overlapping inserts spanning approximately 50 kb, centred around the site of integration of the MPSV-neo provirus (Figure 2). To confirm that these phage clones were in fact derived from the insertion site, a Southern blot of DNA from wild type and Mpv 20 heterozygous mice was hybridized with a probe derived from phage M2041 (Figure 3). As expected, insertion of proviral sequences resulted in the appearance of additional bands in the Mpv 20 heterozygous lanes that were in complete agreement with the previously deduced restriction map of the locus (D.A. Gray, unpublished data). Digests with additional enzymes demonstrated that integration was not accompanied by any gross genomic rearrangements near the integration site (D.A. Gray, unpublished data).

Figure 2. Genomic map surrounding the *Mpv 20* locus.
The initial clone Bg4.1 was isolated from a heterozygote genomic library using a 754 bp Nhe I/Spe I fragment derived from the MPSV LTR. Hatched box shows the position of the probe used to isolate the equivalent wild type clone phage clone M2041. End fragments were subsequently used to initiate a chromosome walk in both directions (refer to Materials and methods section molecular cloning of the proviral insertion site). Arrow shows the transcriptional orientation of the integrated provirus. B=Bam HI, Bc=Bcl I, Bg=Bgl II, E=Eco RI, E5=Eco RV, H=Hind III, K=Kpn I, S=Sac I.

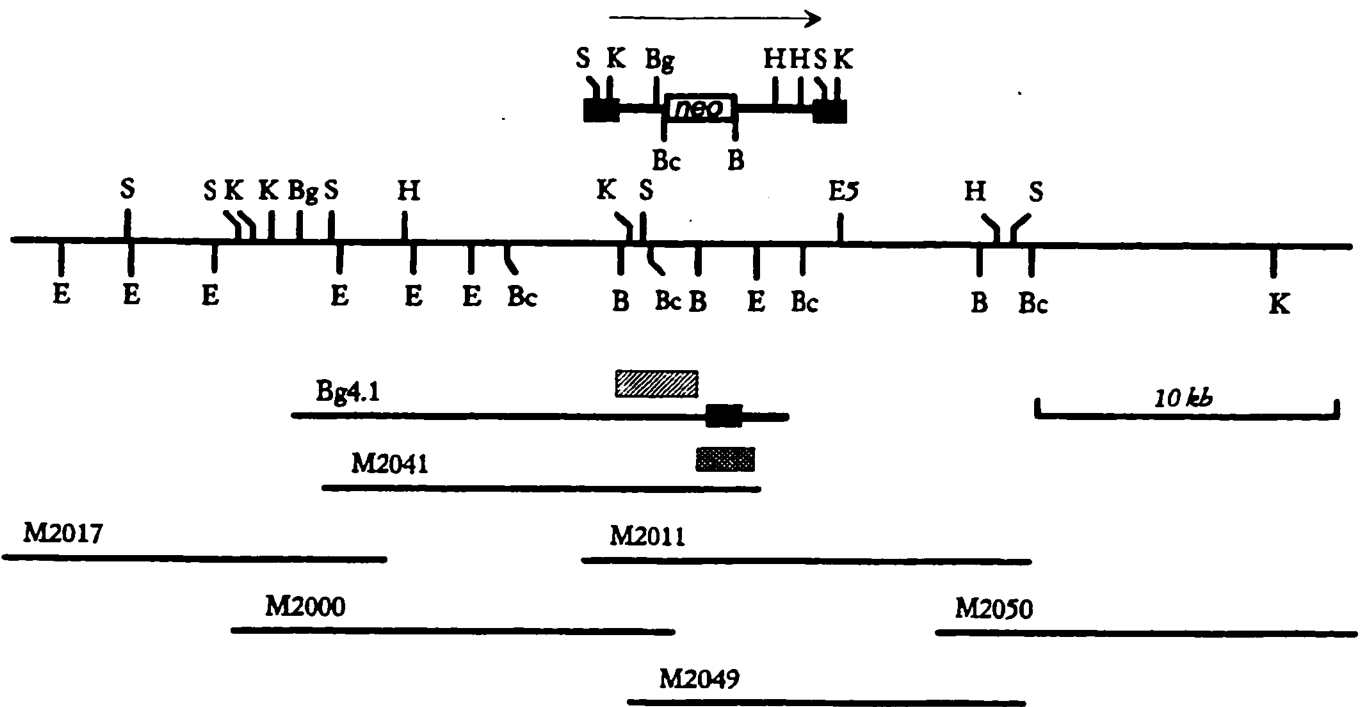

S K Bg
neo
H H S K
Bc B
S
S K K Bg S
H
K S
E5
H S
E E E E E E Bc B Bc B E Bc B Bc K
Bg4.1
10 kb
M2041
M2017
M2011
M2000
M2050
M2049

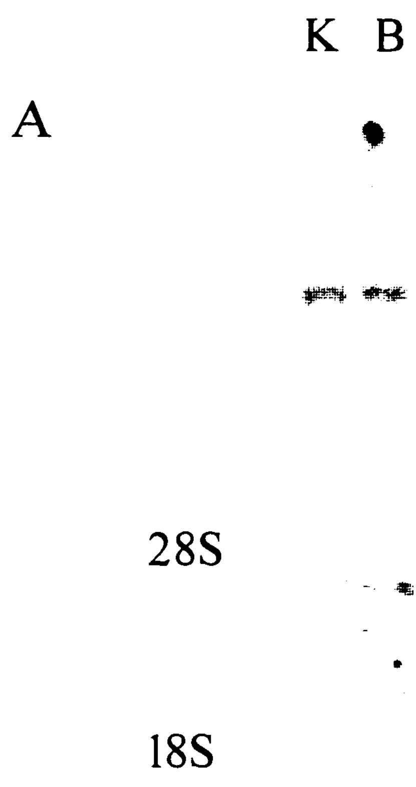
Figure 3. Southern blot analysis of heterozygote (h) versus wild type (w) CFW DNA. A genomic fragment originating from clone M2041 spanning the preintegration site (cross hatched box in Figure 2) was used to probe a Southern blot containing both heterozygote and wild type DNA digested with either Eco RI or Kpn I. Blot was washed at high stringency (0.1X SSC/0.1% SDS 65°C). Both enzymes show an additional band appearing in the heterozygote lanes confirming that clone M2041 is derived from the proviral integration site.

Eco RI
Kpn I
w h w h
kb
12
10
8
6

IDENTIFICATION OF A TRANSCRIPTIONAL UNIT AT THE MPV 20 LOCUS

The initial strategy used was to identify regions, in the cloned genomic DNA, which were evolutionarily conserved. Small restriction fragments (approximately 1 kb) were obtained by digestion of cloned phage DNA with enzymes having 4-base recognition sites (Alu I, Rsa I and Sau 3AI). Nonrepetitive fragments were chosen which were distributed over the entire 50 kb cloned region. These were hybridized to Southern blots of DNA from various metazoan species. A 1kb Rsa I fragment from clone M2017 detected single copy sequences in the genomes of several mammals including rabbits, rats, and humans (D. A. Gray, unpublished data). This fragment was used to screen a L1210 (murine lymphocytic leukemia cell line) cDNA library (provided by P. Duncan) and a murine kidney library (provided by T. Marshall). Six positive clones were identified from over one million plaques screened. These six clones were brought through to a tertiary screen and phage DNA was subsequently isolated. Following digestion of the phage, with the appropriate restriction enzymes, inserts were purified from a preparative gel by Geneclean. Each insert was used to probe Northern blots containing total RNA from both kidney and heart. In all six cases, none of the inserts identified a similar size message as the original Rsa I fragment (see Figure 4A for representative sample of clone EK10). These inserts were then used to probe Southern blots of total mouse genomic DNA digested with a series of enzymes and in all cases examined, each insert demonstrated either the presence of repetitive sequences or alternatively the inserts were hybridizing to pseudogenes (see Figure 4B for representative sample).

Figure 4. Northern and Southern analysis of cDNA clone EK10.
(A) Northern blot analysis of clone EK10. cDNA was used to probe both kidney (K) and brain (B) total RNA. One very large transcript (approximately 12 kb) is observed in both tissues examined. **(B)** Southern analysis of clone EK10. The above probe was used on a Southern blot containing genomic CFW DNA digested with Xho I (X), Kpn I (K), and Eco RI (E). Data showed the presence of a repetitive element present in the probe cDNA. Both blots were washed in 0.1X SSC/0.1% SDS at 65°C. Southern blot was exposed for 5 min to X-ray film.


A
K B
28S
18S

B
X K E

Next, the presence of exons in the region of the genome of interest were sought by use of sequence data and computer analysis. A 4 kb Eco RI fragment from clone M2017 which contained the above Rsa I fragment was subcloned into the pGEM-4 sequencing vector and sequenced (Figure 5). This sequence was analysed using the GRAIL (Gene Recognition Analysis Internet Link) program which resulted in the identification of a putative exon being present between positions 502 and 782 with an excellent probability rating in the open reading frame spanning positions 484 to 1011. The program assesses this rating based on the mammalian codon usage preference and the presence of consensus splice donor/splice acceptor sites. PCR primers complimentary to this region were designed and used to amplify a probe from the cloned genomic fragment. This probe was used on a "zoo blot" containing mouse, rat, and pig DNA and was found to be conserved (Figure 6). The fragment however never hybridized to any message on Northern blots containing total (10 μg of sample) or poly(A)$^+$ selected RNA (2 μg of sample) obtained from a series of different tissues.

An exon trapping strategy was then employed in an attempt to identify putative exons in the cloned phage DNA. This strategy relies on the presence of functional splice donors/splice acceptor sites being present in the region of DNA being examined. The splicing vector contains exonic sequences of the HIV-1 *Tat* gene separated by an intron which contains a multiple cloning site. If the cloned DNA contains functional elements, a correctly spliced chimeric molecule will be generated that can be amplified using primers complimentary to the cloning vector.

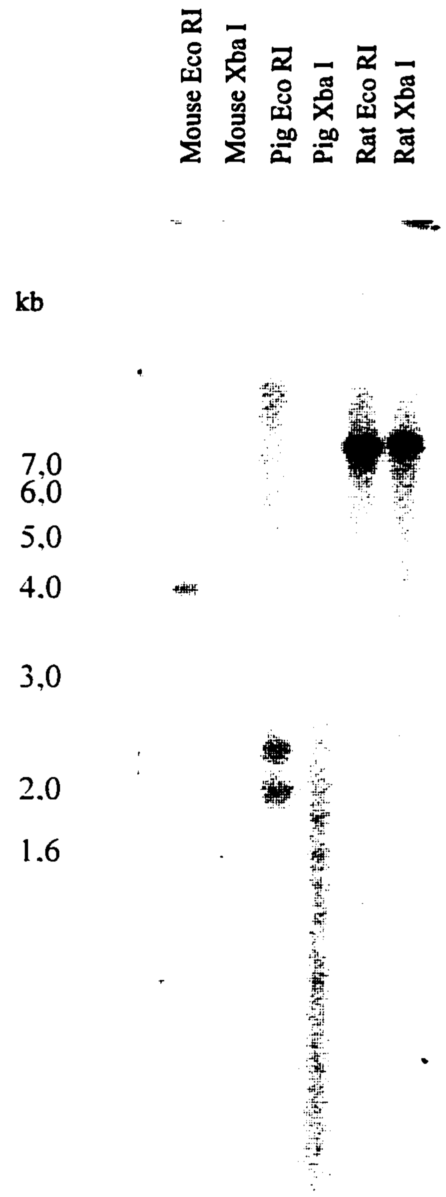
Five Mbo I clones from phage M2017 as well as the 2.5 kb Bam HI fragment from

phage M2041 (present in both orientations) which lies next to the proviral integration site were used to transfect Cos-1 cells. One putative exon was identified in this screen and was sequenced (Figure 7A). This exon was then employed as a probe to identify the genomic fragment from which it originated. The trapped exon mapped back to the pSPL-1 vector which was believed to contain the 2.5 kb Bam HI fragment from clone M2041 (data not shown).

Figure 5. Sequence analysis of a 4kb Eco RI fragment originating from phage clone M2017. 4 kb Eco RI fragment from clone M2017 was cloned into the pGEM-4 sequencing vector. Database analysis using the Gene Recognition and Analysis Internet Link (GRAIL) identified one region between positions 502-782 (bold uppercase) lying in the reading frame spanning positions 484-1011 (uppercase) as harbouring a potential exon. Underlined regions are the sequences used in generating oligonucleotides in order to amplify a probe from the region of interest.

```
   1 gaattctgca  gcagcatgcc  ctccaggcca  gcctaagtta  gaacttctgt  ccaacatgaa    60
  61 aagagaaaga  agaatgctct  tgagtggcca  gacagcttgg  taactctctt  tctgccctcc   120
 121 cctcagagaa  gagacttgcc  ccaaatatga  ttatccaagg  gcagggcagc  tgtaattcct   180
 181 gaagagaaca  tgccacagat  ggttcaagtt  ggcatccgtc  ctccatggct  tgcagggata   240
 241 gacccagttc  cttctcccta  agagcttggt  taaacaggtt  agttgggcag  tggaaaccca   300
 341 agcttgattg  acagagctca  ttgaggttca  gagtagccta  caagatgagc  agaaggacca   360
 361 ctgcctggct  gaccacagag  acctcgctgg  aaagggtagt  ggatatagac  acaaatgact   420
 421 ctcccactcc  tccctggcct  ctggaaccac  cctttcccca  gggcccgccc  actctccata   480
 481 ctcACTCCCA  CTCCTTGCGC  CGGGCCTGCG  GGGTGCTCTG  CATCTTGTGT  GCCTCGTGGG   540
 541 CCTTGATGAT  GTCCCGCTGC  TCGATGAGCC  CCCCTCGCCT  CCGCTTCATG  ACCTCTGGGC   600
 601 TCACAGGAAG  GAAGGGGCTG  GAGCCTGCAT  CCAGCGTGGT  GCTGCCCTCC  CAGCCCGGGC   660
 661 CCTCGCCTTC  CTGCACCGGG  AAGAAGTTGT  CGAAGCCGGC  CACCGCGGAG  CGCGGGGAAG   720
 721 CAGGCAACAT  GCCCACCTCG  TCGGAGGCAT  TGGCCTCGAG  CAGGGCGCAC  TGCTGCGGCC   780
 781 GCCATGCCTC  CTCATCAGGC  AGAGCCAGGA  CGTGGCGGCG  CCTGCCAGCA  GCCCCGGGCT   840
 841 CAGGGCTCCC  TCCTGTAGCC  GGCACAGGGT  GGCTCGGCAG  CTTCGGCCTG  GGCTCAGTCC   900
 901 ACGCATCTCG  GGCACCATCA  CAGAACATTG  CTCTGCCAGG  AAAAAGGGTG  GGCAGAGACA   960
 961 AAGGGTGGGG  AGGAGAAGGA  GGAAGAGGAG  GAGGAGGGAA  TGAAGGATCA  Tgggcacggg  1020
1021 gaaatagaga  agagaaaaag  attaatgaca  cggcccctca  aagctcaata  atgcttccta  1080
1081 ggcctggcca  agaaactggc  acataggccg  agcttggtaa  aagatgctaa  aatcagtagt  1140
1141 gttgactcta  aactagtctg  ggtccgattc  acagagttgc  agagtttaag  caaagaagat  1200
1201 tcactcctac  catgagattc  aagtccagtg  cagatgacag  aggttccatg  attacagata  1260
1261 ctcctcctct  aacgacgact  ggactaccac  ccaataaacc  cattgttagg  agaaaatatt  1320
1321 gcaaagtaaa  atagcattta  aagtctaagt  cactgaacat  cgcggcctaa  caacacagta  1380
1381 cactgtacag  tgagacaggt  ttcccacatg  atttggggc   tgggaagtga  ggctcactca  1440
1441 ctgcccagaa  ccataagagg  aacataccaa  gagccctgga  aagatcaaaa  ttctagaaat  1500
1501 gaagtacatt  ttctacaaaa  tatgtgtggc  tttcacacca  ccatcaaaag  aggccatcag  1560
1561 tatgtacaca  ggagacaaac  attccagtac  cttctataag  gagcaccgcc  atctcaggtg  1620
1621 ggcagcaatg  tggctaaatg  aaaagaatcc  tcatcacccc  tattccaggt  tcccagctcc  1680
1681 tctgccctca  aggccacctg  actccctggc  cagcacaaag  taggcatgtc  ttttgtctag  1740
1741 gttcctgcct  ttcttatgat  acacacatcc  tcctcaacac  acagagaaca  ttatagatcc  1800
1801 agatgctcta  atgactccct  gaagctggat  ttcctattcc  agaccaagga  gaatatagtc  1860
1861 aaattcactt  tagaaagagt  gcttgccata  tgccaaccac  caggtaccaa  gcatctcata  1920
1921 atgagccaac  tggtccatgt  catacccatt  gggacacaga  ggaacactga  ctcttgtcac  1980
1981 acggcatgat  gggctgggac  tgatgtaagt  ctgatgtctg  taagttctca  gaggagggcc  2040
2041 acattatgta  cccggtgtaa  ccaggagagt  tgtatagaag  ggcaggacct  tgaagctaag  2100
2101 catggacaga  tctctctggg  taagcaccag  gagatggggg  gcaggagata  aggataggga  2160
2161 gattataggt  aaaaggaaga  acatatatgc  acagaggagg  actcaaggaa  cccagaggat  2220
2221 gtgtgtggtg  gtttgaatgt  gcttggccca  cgggaagtgg  caccataagg  agctatggct  2280
2281 ttgttggagt  aggtgtggcc  ctgttggagg  aagtgtgtta  ctgtgtgggt  gggctttgag  2340
2341 ggctccttgt  ggccaaactc  ctcccattgg  ggaagagagc  ctccttctgg  ctgcctgcag  2400
2401 atgccagtct  cttcctggct  tccttcagat  caagatgcag  aactcttggc  atctccagca  2460
2461 ccaagtctgc  ccacacaatg  tcgtgcttcc  caccacggat  aatgggctaa  acttctgaaa  2520
2521 ctgtaaccca  gccccaatta  aatgttgtcc  ttataagagt  tgccttggtc  atggtgtctc  2580
2581 ttcacagcag  taaaacccta  actaagagaa  tgtggaaaac  acttcttcca  taagcctggg  2640
2641 gtcctaatag  tcctttatac  ttcagaggac  ctcagctttc  cacctcaaga  tactacttta  2700
2701 ggtatatagg  aatcaaaaag  gatgctttcc  cagtgctccc  agtgtgcttc  cagtgtgctc  2760
2761 ccagtgtgct  cccagtgtgg  gatagagcct  tgacccacac  cacttgaatg  atttgaatgc  2820
2821 caaggttcct  ttgggcacac  agagcacatg  ggaggaaaaa  tgtccaggct  aagcttaatg  2880
2881 aggaggagaa  atgaaggaga  aatattttca  gtctgcaaga  ctgtctaccc  tttagggtgg  2940
2941 gatcttccac  aggtgatgta  gcacgatctt  gggctagaca  cagccctggg  ccagagactg  3000
3001 catcaatacg  caggtggcaa  gccgacgagg  ctctatccca  cacagggagc  acactgggaa  3060
3061 cactagggta  gcatcctttt  tgatccctgc  acacacctga  cccctagcta  tcttctgcct  3120
3121 tccctaatat  gagccacaaa  ccaaaccgta  aaccagaaaa  gcaaagaaca  tggaatggtt  3180
3181 ataaatttta  atgttcaacc  atattatttt  atagatgagg  aggataagat  ccaaaagtta  3240
3241 aagagaccca  cccagggcca  cactgctagt  aaaccagact  tagatttcaa  gacttagctc  3300
3301 ccatttttgt  ttttatcaca  ctctttgtgc  ttcatgaagc  agaatccagc  gatgccacgc  3360
3361 ctatatcctc  gtcactaatt  aacagagtta  ttgtaaggaa  aagcctagat  tccctgcctt  3420
3421 ccttttccat  cacctggact  catctttcca  gatatgatac  ccaagccttc  cgggagctga  3480
3481 gttgagcagt  ggtcagcaag  gttgagtcaa  gccaatgggg  taggggg agg gaaaatgtcc  3540
3541 aggctaagct  taatgaggaa  gagaaatgaa  ggagaaaata  tttttcaagt  cttgccaaca  3600
3601 gctgggtggg  cctttattcc  tacaatgttt  aaatgtctct  cccccaccat  acacatgcgt  3660
3661 gcacccctcc  cctgacagct  cccactccct  ctccagactt  tagcacctcc  agcaattcca  3720
3721 agtattcttg  gactctggtc  tacagtctcc  tccccgcttc  taagcccatt  gtagacaagg  3780
3781 ctcccaggcc  actagctctc  ccagggcaac  cagaagccag  ggtctggctt  tgtatatctg  3840
3841 tttcccagct  atccccaccc  ccatccttcc  tggatttccc  cttccctgac  actttgccaa  3900
3901 tctaaatcac  gtctttgcca  gaattacaac  agcaataata  tgcacgacag  agctctggct  3960
3961 tcaacagaat  caatttacaa  atatcaaaac  atgaaattat  ctcatctttc  ctttaagaat  4040
4041 tc                                                                    4042
```

Figure 6. "Zoo blot" Southern using GRAIL fragment as a probe.
DNA from CFW mouse liver, rat liver, and pig ovary was digested with either Eco RI or Xba I and probed with the GRAIL PCR fragment (see Figure 5). A positive signal was obtained in all three species examined under the experimental conditions utilized. The probe recognized the 4 kb mouse Eco RI fragment from which it was originally derived. Blot was washed in 1X SSC/0.1% SDS at 25°C.



Mouse Eco RI
Mouse Xba I
Pig Eco RI
Pig Xba I
Rat Eco RI
Rat Xba I
kb
7,0
6,0
5,0
4.0
3,0
2.0
1.6

Figure 7. Sequence of trapped exon and nucleotide alignment to *Sry* (Sex determining region on the Y chromosome).
(A) Sequence of genomic region surrounding the putative trapped exon. Exonic sequences are shown in uppercase. The exon is flanked by the consensus splice acceptor/splice donor dinucleotides AG/GT which are underlined (Mount 1982). (B) Sequence alignment of figure 7A (trap) with the mouse DNA sequence of the *Sry* locus (Genbank accession X67204).

A

```
  1 ccctgaaccagctgtaggtctgtgagccacccatctgcaaggaggaagacaatctgagac   60
 61 agccttttcatttttgtttctttttctgggaatggctttctgtgtctacctgtatgtc  120
121 tgtctgtgtgtactctgtgctttttgtggtctagtacaaagttttctttggttcctagGA  180
181 TGGAACAGAGACAGAGCACCCCACTTGGCCTGGGCCTGAACCATTTTAAGGACATTCACA  240
241 GAAGACTTGGATCTGGTTATTCTTGCAAACTAACCATCTTTTGCAGAAATGAATGGCCCA  300
301 CTTGTAAAGTAGAGTGGCTTGAGgtgggaaattttcatctctctatcatttacagggaga  360
361 aggctataatttatgagggtaggggccacccagaccaagtgccacatggcaattcttggt  420
421 caagaagcccactgtttggatgagaccatttttgtt                         456
```

B

```
 456 AACAAAAATGGTCTCATCCAAACAGTGGGCTTCTTGACCAAGAATTGCCA   407 trap
     || ||||| | |||||||||||||| |||| |||  |||| | |||||||
1073 AAGAAAAAAGTTCTCATCCAAACAGAGGGCCTCTCCACCAGGTATTGCCA   1123 Sry

 406 TGTGGCACTTGGTCTGGGTGGCCC 383 trap
     | ||||||||||||  ||||||||
1133 TATGGCACTTGGTCCAGGTGGCCC 1156 Sry

 307 TTACAAGTGGGCCATTCATTTCTGCAAAAGATGGTTAGTTTGCAAGAATAACCAG  253 trap
     | | | |||||||| ||||| ||||||||||||||||||||| | | | | | ||
1205 TCATAGGTGGGCCACTCATTCCTGCAAAAGATGGTTAGTTTGTAGGGACACCAAG  1259 Sry

 236 AATGTCCTTAAAATGGTTCAGGCCCAGGCCAAGTGGGGTGCTCTGTCTCTG  186 trap
     || |||||||||||||||||||  |||| | ||||  ||||| | |  ||
1290 AAAGTCCTTAAAATGGTTCAGGATCAGGTCCAGTGCAGTGCTTTATGCCTT  1340 Sry

 185 TTCCATCCTAGGAA  172 trap
     ||||||||||||||
1340 TTCCATCCTAGGAA  1353 Sry

 173 AACCAAAGAAAACTTTGTACTAGACCACAAAAAGCACAGAGTACACACAGACAGACA 117 trap
     ||  ||  ||||  |||||| |||  |||||||||||| |  | | ||| ||| |||
1356 AAAAAAGAAAAAAATTGTACAAGAGGACAAAAAGCACAAACAAAATACACACAAACA 1412 Sry

 116 TACAGGTAGA 107 trap
      | | | | |
1413 CAAAAGGACA 1422 Sry
```

Sequence analysis of this genomic fragment using oligonucleotides present within the trapped exon confirmed the presence of a splice donor and acceptor site present at the start and end of the sequenced exon. A data base search on this fragment as well as the above 4 kb Eco RI (Figure 5) fragment revealed some homology to the inverted repeat of the *Sry* (sex determining region on the Y chromosome) locus in the mouse (Gubbay et al. 1992). For the region surrounding the trapped exon the DNA homology translated to 204 of 261 identical positions (78%). Since both male and female transgenic animals were present in the colony, this eliminated *Sry* as a potential candidate gene. However there have been reports of a related family (Sox) of genes identified whose members share homology in the HMG box and were mapped to autosomes (Denny et al. 1992a&b, Wagner et al. 1994, Connor et al. 1994, 1995). A similar PCR strategy as that of Denny et al. (1992b) was attempted to determine if a HMG box was present in the 50 kb of genomic DNA surrounding the Mpv 20 proviral integration. A visible band in the proper size range was observed when the inserts derived from phage M2041 and M2017 were used as a templates. Probing the transferred gel with a 380 bp Bgl II/Pst I fragment from clone p422 (ATCC), containing the HMG box of *Sry*, identified a positively hybridizing band when the membrane was washed at low stringency (Figure 8). The positive fragments were purified and sequenced directly using the same degenerate oligonucleotides to prime DNA synthesis. The sequence of the PCR product (Figure 9 A) surprisingly did not identify any HMG containing proteins when the Genbank database was searched, instead it displayed high homology to the human *Npat* cDNA (Imai et al. 1996). This homology translated to 85 out of 91 identical nucleotides (93.4%) (Figure 9 B). The predicted translated protein showed 100% identity to *Npat* (Figure 9 C).

Figure 8. Southern blot analysis of *Sry* PCR products.
One-fifth of each PCR reaction was run on a 1.5% agarose gel and transferred to a nylon membrane. Blot was probed with a 380 bp Bgl II/Pst I fragment from plasmid p422 (ATCC) which contains the HMG box of the mouse *Sry* gene (Gubbay et al. 1990). Hybridization signals are observed in both lanes which contain phage templates M2017 and M2041. p422 plasmid was used as a positive control. Negative (- ve) control was a PCR reaction containing all reagents except for a template. The annealing temperature for the PCR reaction was 52°C. Blot was washed in 0.5X SSC/0.1% SDS at 42°C.

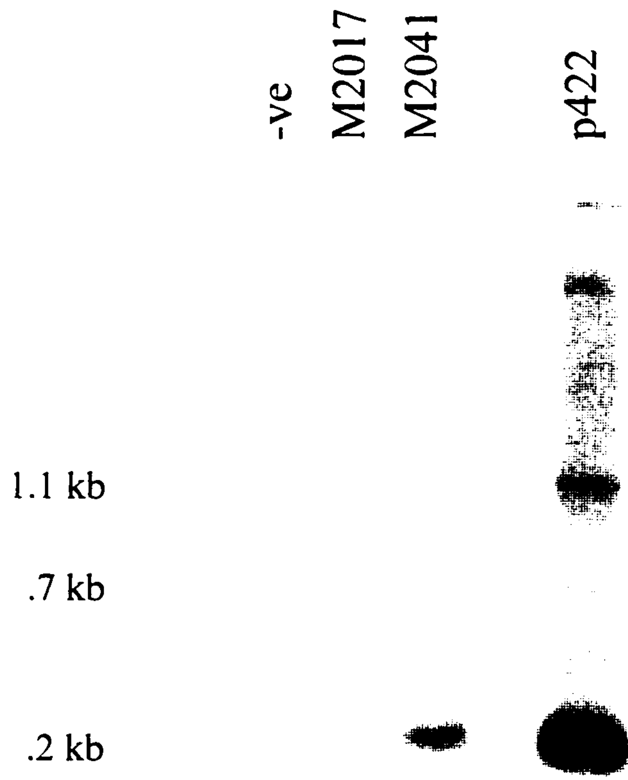

-ve
M2017
M2041
p422
1.1 kb
.7 kb
.2 kb

Figure 9. Sequence of the *Sry* PCR product and homology to the human *Npat* cDNA . PCR product was generated using degenerate oligonucleotides to HMG box (Denny et al. 1992b)(A) The nucleotide sequence of the isolated PCR product. (B) Nucleotide homology to the *Npat* cDNA (Genbank accession D83243, U58852, X97186). (C) Translated product of the PCR sequence displaying identity to a region of the human *Npat* cDNA.

A

```
  1  ggaatgacgtaatatgctttcctatatcttttactaatcttgattttctttttcagCAAA   60
 61  TGGTTATTGAAAATGCAAGAGAGAAAATATTAAGCAACAAATCTCTTCAAGAAAAACTAG  120
121  CAGAAAACATAAATAAGTTTTTGACTAGgtgagctgtat                      159
```

B

```
 57  CAAATGGTTATTGAAAATGCAAGAGAGAAAATATTAAGCAACAAATCTCTTCAAGA  112 SryPCR
     ||||||||||||||||||||| |||| |||||| ||||||||||||||||||||||
793  CAAATGGTTATTGAAAATGCACGAGAAAAAATACTAAGCAACAAATCTCTTCAAGA  848 h Npat

113  AAAACTAGCAGAAAACATAAATAAGTTTTTGACTAG                      148 SryPCR
     ||| |||||||||||||||||||| ||||| |||||
849  AAAGCTAGCAGAAAACATAAATAAATTTTTAACTAG                      884 h Npat
```

C

```
 57  QMVIENAREKILSNKSLQEKLAENINKFLT  146 SryPCR
     QMVIENAREKILSNKSLQEKLAENINKFLT
243  QMVIENAREKILSNKSLQEKLAENINKFLT  272 h Npat
```

IDENTIFICATION OF THE DISRUPTED NPAT GENE

To confirm that the trapped exon arose from the integration site, a Sst I fragment from clone Bg4.1, which contained the proviral LTR/genomic DNA junction, and a 9 kb Eco RI fragment from clone M2041 which contained the preintegration site were cloned into the sequencing vector pGEM-4. Figure 10 shows the composite sequence of the Eco RI/Sac I region surrounding the proviral integration site (refer to Figure 2 for position on genomic map). The translated sequence of the region surrounding the provirus was used in a Genbank database search and also resulted in the identification of the human *Npat* cDNA (Figure 11). The nucleotide homology is present in areas of the genome that are flanked by the consensus splice acceptor/splice donor AG/GT dinucleotides (Mount 1982) defining the position of four of the murine exons present in the sequenced region of DNA. Of 353 exonic nucleotides, 297 were identical (84.1%) which translated to 85.3% (99/116 amino acids) identity on the amino acid level. When conserved amino acid substitutions were included, the homology increased to 91.4% (106/116 amino acids).

Closer examination of the genomic DNA sequence showed that the provirus integrated into an intron in the opposite transcriptional orientation of the gene. Also present was a duplicated region of DNA, the function of which is presently unknown. This duplication was confirmed by sequencing a genomic DNA PCR product which was generated using oligonucleotides outside of the duplicated region (Figure 12).

Figure 10. Sequence of genomic DNA at the Mpv 20 proviral insertion site. Bold uppercase letters denote translated sequences that align to the *Npat* (Genbank accession D83243) translated protein. Putative introns (denoted by lowercase letters) all harbor the conserved splice donor/splice acceptor sequences (Mount 1982). Solid triangle represents the position of the proviral insertion. This was determined by sequencing a Sst I fragment originating from clone Bg 4.1 which contained the provirus/mouse junction fragment. A duplicated region was also observed, the significance of which is unclear (underlined region). Italics indicate position of oligonucleotides used to amplify genomic DNA in order to confirm the presence of a duplication. Sequence is a compilation of multiple reads of both strands of cloned phage DNA.

1 gaattcagggttcatctcttatttctgagtcgagtaattttgccacccattaagactgctc 61
62 atttaggcatagtggagatgccattaatcccggcactggggaagcagaagcatgtaggtc 122
123 tctgacttcgaggccagctgggtctttagagagagtttcaggacagctgaggctacaggca 183
184 gaaaccttgccttgaaaaataacaagaagaagattgttcatttctgtctttctccaaatcc 244
245 tgtatactaagagcattcagttgattgtttttatactttttttcctttatgtttcctttg 305
306 tctttgtgtcaagtgtccaaccttaagtacatagggaccgtattgattcatcttacaatctt 366
367 gtacactgctttgctcatagtggtattcagaaaatgtttttctaggtttaatatacaataa 427
428 tagtatctcattacaagttaataaaatttctgtttctttggcttactctgttttcttaag**A** 488
489 **TACAGCTGGAGAGTCTTTGAATATCATTCCTGATCCTCAGGAAAGGAAAACTCAAACCAGT** 549
550 **TTAATGTCTCCTGGTAGACGTAAAA**ggtaattactttatataatcagcacattgtcaatag 610
611 ccacttaaaaacagacaaacgataactttagtaagtgtttgcattgtgaataaattaataag 671
672 actagattaagtttttgtgacatatttcaaacacatcttttccaatattgtaacttcactg 732
733 aatttgagttgtgtctttaatcaatggtgtgtaacttgctgttcagtaactttgtgtgat 793
794 gatgtgcagaataatggtagagcttatgacagatcttagctcatctgtcgagcgtttcttg 854
855 aatgaatataatgttaatgttataatccagaagagtacatggagctaagtcatgacagcct 915
916 gtttttaggtgtgctatggtccataagaatgcatgaagtttaaatcatgaaagcatgttc 976
977 tcgtttttattttcatatgtatatttttcctgttctagttaagctttctatttcttaaaaa 1037
1038 aaatcagcggtcaggattctaaatggtttttatctgattcttcatctgccaacaatatttt 1098
1099 tttcttttttg▼attagaatttttaactttcttaaacatatacaatgataaaacttatta 1157
1158 ataacatttcagaaactttataataaaaagtctataaatacttaaaactttttcataaa 1218
1219 gccactaataaaatggttcacatttcattataagataggtggtggagtgtagattgtagct 1279
1280 tactggaaaaacacatgcctatcatatatgaggtcttgggataaattcccattaccactaa 1340
1341 agaaagaaagaaagaaaaaggtataaaagtacagagctttgtacacttaggaccacagagct 1401
1402 aatttttgtttgggcgtgatgtttttttgattgtttgtttttccattcacttgtttttctt 1462
1463 tctcttttcatctttatttttgttgtttttaatagaattcattacatctgaattcatgct 1523
1524 tttctgttcttctggctttgtttcttaagaaaagagggggaaaaatctgcctactaccatg 1584
1585 atgatcttaaaatgtgtacttcatacttatcattgatttgttagatacttaacaaacacgcttt 1645
1646 ataatttatttcag**TGAATCACAAAAGAAGAGTC*TCACATCATCTGGACCGC*ATTCATCAA** 1706
1707 **GGAATTTTCAGGATCCAAATGCATTTGCAGTAGAAAAA**gtaagtgacctttctcaaaattca 1767
1768 gtttgaagtattctcacatttaaaaacactgatgtgccgggcgtggtggcgcacgcctttta 1828
1829 atcccagcacttgggaggcagaggcaggcagatttctgagttcgaggccagcctggtctac 1889
1890 aaagtgagttccaggacagccagggctacacagagaaaccctgtctcgggaaaaccaaac 1950
1951 caaaacaaaacaaaaaaacccactgatgtatttatagttaggtgtgtttatgtgtttggtat 2011
2012 gaatctgagagcaagaggattggttatggagcaggggacagtaactgaaagtagcagttag 2072
2073 agaacccgacagagccagaaacgtcaattgatgaactgctaggccttccggtatggtaga 2133
2134 tgtgtgtggggatgagtgtgagagagtgtgagccgaagtagagtcagattgaaacaagtgtt 2194
2195 ttggttttgttttaaaatcattgccaaaaaattacatgggtataatctaatttaaagtaat 2255
2256 taaaatagactgttcattcatcacaaaatatgatctgtttctgaacacatttggatagtgac 2316
2317 aaatgcaccaagtaaaagctatttgtgattgccagctcctggggtttagctgaacatttga 2377
2378 ccaggatcaggtcattgaagaccttgtgtaaaacgtgtagaactgcattcaggatggcgtg 2438
2439 aagatcaagttggtgatctggacaacactttgaattgcaaaatattttgttcagacctga 2499
2500 atgcagatggaggcaccagaacttgcatttctttctactaatgtttatcaaaactagccac 2560
2561 ttggtttttgtgacattttgggggtttttttggtttgtttttggtttttatggtttgtttttgtt 2621
2622 ttgttttcgagacagggtttttttgtatagcctggctgtcctggaactcactttgtagacca 2682
2683 ggctggcctcgaactcagaaatccgcctgcctgtgcctttgagtgctgggattaaaggcg 2743
2744 tgcgccaccacgcatggctgacatttgtttttttaattatttatttagaattgaggtctt 2804
2805 gtcctgtgttccagattgtacataaaatcctgactacaatttataatttatctcagctcc 2865
2866 ccaaattgctggaacaaaaaatgtgatcacaaaaagtgatttacgtttggtttggcgggtt 2926
2927 gggttttgggaatgacgtaatatgcttcctatctttactaatcttgatttctttttc 2987
2988 ag**CAAATGGTTATTGAAAATGCAAGAGAGAAAATATTAAGCAACAAATCTCTTCAAGAAAA** 3048
3049 **ACTAGCAGAAAACATAAATAAGTTTTTGACTAG**gtgagctgtatgattcaaatttctgatt 3109
3110 gtttttacctaatcccccaactgcactaacagtttcttggttttcctttctgttttacag**T** 3170
3171 **GATAGTAGTGTTGCTCAAGTACCTAAGCAGACAGACAGCAACCCGACAGAGCCAGAAACGT** 3231
3232 **CAATTGATGAACTGCTAGGCCTTCCG**gtatggtagatgtgtgtggggatgagtgtgagag 3292
3393 agtgtgagccgaactagagtcagattgaaacagtgtttttggttttgttttaaaatcattgcc 3353
3354 aaaaaattacatgggtataatctaatttaaagtaattaaaatagactgttcattcatcaca 3414
3415 aatatgatctgtttctgaacacatttggtagtgaatacagtgttcaagccagcagatgtc 3475
3476 ttggagtctgttgggagtttgggaagcacacatggtactgagattggtctaagtgcactag 3536
3537 cctccataacactgaaaattctatagacattaaataccttggcaaagggtgtggcagtgga 3597
3598 catgaagtaatttgctgaaacttcaggcttctcagttcttacaacagtgaacttaaaagca 3658
3659 gacagatggttccagagagctc 3680

Figure 11. Nucleotide and translated product alignments between mouse and human *NPAT*. (A) Alignment of exonic sequences surrounding the Mpv 20 insertion site and the published human *NPAT* sequence (Imai et al. 1996). Vertical lines denote identical nucleotide positions. The homology over 353 nucleotides was found to be 84.1%. Numbering for the mouse sequences (mNPAT = murine *Npat*) is relative to the positions in Figure 9, human numbering (hNPAT = human *NPAT*) is to the published cDNA sequence. (B)Alignment of the predicted translation products of the exons flanking the Mpv 20 insertion site and the published human *NPAT* sequence. Identical nucleotides are indicated, whereas + indicates conserved substitutions. Numbering of the human sequence is relative to the amino acid position of the published predicted translated product (Imai et al 1996). m cDNA denotes the protein generated from the translated sequence of exons present in Figure 9. The exonic translated sequences spanning the integration site is denoted by the arrowhead between the Arg (R) and Lys (K) residues.

A

```
488  ATACAGCTGGAGAGTCTTTGAATATCATTCCTGATCCTCAGGAAAGGAAAACTCAAA 544 mNPAT
      |||  |||||||  |||| |||  ||||||||| | ||||||||| |||| | ||
617  GTACCACTGGAGAAGCTTTAAATGTCATTCCTGGTGCTCAGGAAAAGAAAGCACATG 673 hNPAT

545  CCAGTTTAATGTCTCCTGGTAGACGTAAAAG 575 mNPAT
     ||||||||||||||||  |||||||||  |||||
674  CCAGTTTAATGTCTCCCGGTAGACGCAAAAG 704 hNPAT


1660 TGAATCACAAAAGAAGAGTCTCACATCATCTGGACCGCATT---CATCAAGGAATT 1712 mNPAT
     |||||| || |   || |||   ||| |   ||||| || ||||    ||  | ||||||
705  TGAATCTCAGAGAAAAAGTACCACTTTGTCTGGCCCTCATTCAACAATACGGAATT  760 hNPAT

1713 TTCAGGATCCAAATGCATTTGCAGTAGAAAAA 1744 mNPAT
     | || |||||||| || |||||||||||||||
761  TCCAAGATCCAAACGCTTTTGCAGTAGAAAAA  792 hNPAT


2990 CAAATGGTTATTGAAAATGCAAGAGAGAAAATATTAAGCAACAAATCTCTTCA 3042 mNPAT
     ||||||||||||||||||||| |||| |||||| ||||||||||||||||||
793  CAAATGGTTATTGAAAATGCACGAGAAAAAATACTAAGCAACAAATCTCTTCA  845 hNPAT

3043 AGAAAAACTAGCAGAAAACATAAATAAGTTTTTGACTAG 3081 mNPAT
     |||||| |||||||||||||||||||||| ||||| |||||
851  AGAAAAGCTAGCAGAAAACATAAATAAATTTTTAACTAG  884 hNPAT


3170 TGATAGTAGTGTTGCTCAAGTACCTAAGCAGACAGACAGCAACCCGACAGAGC 3222 mNPAT
     |||||  | | |||| |||||||||||||| ||||| | |||||| || ||||
885  TGATAACAATATTGCCCAAGTACCTAAGCAAACAGATAACAACCCTACGGAGC  937 hNPAT

3223 CAGAAACGTCAATTGATGAACTGCTAGGCCTTCCG 3257 mNPAT
     |||| || |||||||||||| | ||||| ||||||
943  CAGAGACTTCAATTGATGAATTCCTAGGACTTCCG  972 hNPAT
```

B

```
    TAGESLNIIPDPQERKTQTSLMSPGRRKSQSQLKSLTSSGPH-SSRNFQDPNAFA   m cDNA
    T GE LN+IP  QE+K    SLMSPGRRKSQSQ KSLTSSGPH + RNFQDPNAFA
185 TTGEALNVIPGAQEKKAHASLMSPGRRKSQSQRKSTTLSGPHSTIRNFQDPNAFA   239  h NPAT

    VEKQMVIENAREKILSNKSLQEKLAENINKFLTSDSSVAQVPKQTDSNPTEPETS   m cDNA
    VEKQMVIENAREKILSNKSLQEKLAENINKFLTSD+++AQVPKQTD+NPTEPETS
240 VEKQMVIENAREKILSNKSLQEKLAENINKFLTSDNNIAQVPKQTDNNPTEPETS   294  h NPAT


    IDELLGLP  m cDNA
    IDE LGLP
295 IDEFLGLP  302 h NPAT
```

Figure 12. Sequence alignment confirming the presence of a duplicated region in the mouse genome. Genomic PCR was carried out using oligonucleotides LR and 2041 which lie outside of the duplicated region (see Figure 10 for exact position). Top nucleotides arose from the sequenced phage DNA given in Figure 10, bottom sequence is the partial sequence obtained from the genomic PCR product which was subcloned in the Eco RV site of pLitmus 28. Repetitive elements were identified to lie between nucleotide positions 2075-2308 and 3210-3442. Underlined are the positions of the oligonucleotides that were used to prime the sequencing reactions. Arrows above the oligonucleotides indicate their 3' end orientation. Oligonucleotide with two arrows indicate the same sequence complimentary to both strands.

```
       →
2012 GAATCTGAGAGCAAGAGGATTGGCTATGGAGCAGGGGACAGTAACTGAAAGTAGCAGTTAG   2072
                                  |    |||||||||||||||||||||||||||
                                  GGAGGGGACAGTAACTGAAAGTAGCAGTTAG

2073 AGAACCCGACAGAGCCAGAAACGTCAATTGATGAACTGCTAGGCCTTCCGGTATGGGTAGA   2133
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     AGAACCCGACAGAGCCAGAAACGTCAATTGATGAACTGCTAGGCCTTCCGGTATGGGTAGA

2134 TGTGTGTGGGGATGAGTGTGAGAGAGTGTGAGCGAAGTAGAGTCAGATTGAAACAAGTGTT   2194
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     TGTGTGTGGGGATGAGTGTGAGAGAGTGTGAGCGAAGTAGAGTCAGATTGAAACAAGTGTT

2195 TTGGTTTTGTTTTAAAATCATTGCCAAAAAATTACATGGGTATAATCTAATTTAAAGTAAT   2255
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     TTGGTTTTGTTTTAAAATCATTGCCAAAAAATTACATGGGTATAATCTAATTTAAAGTAAT

2256 TAAAATAGACTGTTCATTCATCACAAATATGATCTGTTTCTGAACACATTTGGATAGTGAC   2316
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     TAAAATAGACTGTTCATTCATCACAAATATGATCTGTTTCTGAACACATTTGGATAGTGAC

2317 AAATGCACCAAGTAAAAGCTATTTGTGATTGCCAGCTCCTGGGGTTTAGCTGAACATTTGA   2377
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     AAATGCACCAAGTAAAAGCTATTTGTGATTGCCAGCTCCTGGGGTTTAGCTGAACATTTGA

2378 CCAGGATCAGGTCATTGAAGACCTTGTGTAAAACGTGTAGAACTGCATTCAGGATGGCGTG   2438
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     CCAGGATCAGGTCATTGAAGACCTTGTGTAAAACGTGTAGAACTGCATTCAGGATGGCGTG

2439 AAGACCAAGTTGGTGATCTGGACAACACTTCTGAATTGCAAAATATTTTGTTCAGACCTGA   2499
     |||||||||||||||||||||||||||||||||||||||||||||||||
     AAGACCAAGTTGGTGATCTGGACAACACTTCTGAATTGCAAAATATTTT

2500 ATGCAGATGGAGGCACCAGAACTTGCATTTCTTTCTACTAATGTTTATCAAAACTAGCCAC   2560
                                              |||||||||||||||||||
                                              TGTTTATCAAAACTAGCCAG

2561 TTGGTTTTGTGACATTTTGGGGGTTTTTTGGTTTGTTTTGGTTTTTATGGTTTGTTTTGTT   2621
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     TTGGTTTTGTGACATTTTGGGGGTTTTGTGGTTTGTTTTGGTTTTTTTGGTTTGTTTTGTT

2622 TTGTTTCGAGACAGGGTTTTT-TGTATAGCCTGGCTGTCCTGGAACT-CACTTTGTAGACCA 2682
     ||||||||||||||||||| | |||||||||||||||||||||||||| ||||||||||||||
     TTGTTTCGAGACAGGGTTTCTCTGTATAGCCTGGCTGTCCTGGAACTACACTTTGTAGACCA

2683 GGCTGGCCTCGAACTCAGAAATCTGCCTGCCTGTGCCTTTGAGTGCTGGGATTAAAGCCG   2743
     ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     GGCTGGCCTCGAACTCAGAAATCTGCCTGCCTCTGCCTTCTGAGTGCTGGGATTAAAGCCG

2744 TGCGCCACCACGCATGGCTGACATTTTGTTTTTTAATTATTTATTTAGAATTGAGGTCTT   2804
     ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     TGCGCCACCACGCATGGCTGACATTTTGTTTTTTAATTATTTATTTAGAATTGAGGTCTT

          ←
2805 GTCCTGTGTTCCAGATTGTACATAAAATCCTGACTACAATTTATAATTTTATCTCAGCTCC   2865
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     GTCCTGTGTTCCAGATTGTACATAAAATCCTGACTACAATTTATAATTTTATCTCAGCTCC

2866 CCAAATTGCTGGAACAAAAAATGTGATCACAAAAAGTGATTTACGTTTGGTTTGGCGGGTT   2926
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     CCAAATTGCTGGAACAAAAAATGTGATCACAAAAAGTGATTTACGTTTGGTTTGGCGGGTT
```

```
2927 GGGTTTTGGGGAATGACGTAATATGCTTTCCTATCTTTTACTAATCTTGATTTTCTTTTTC 2987
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     GGGTTTTGGGGAATGACGTAATATGCTTTCCTATCTTTTACTAATCTTGATTTTCTTTTTC

2988 AGCAAATGGTTATTGAAAATGCAAGAGAGAAAATATTAAGCAACAAATCTCTTCAAGAAAA 3048
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     AGCAAATGGTTATTGAAAATGCAAGAGAGAAAATATTAAGCAACAAATCTCTTCAAGAAAA

3049 ACTAGCAGAAAACATAAATAAGTTTTTGACTAGGTGAGCTGTATGATTCAAATTTCTGATT 3109
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     ACTAGCAGAAAACATAAATAAGTTTTTGACTAGGTGAGCTGTATGATTCAAATTTCTGATT

3110 GTTTTTACCTAATCCCCCAACTGCACTAACAGTTTCTTGGTTTTCCTTTCTGTTTTACAGT 3170
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     GTTTTTACCTAATCCCCCAACTGCACTAACAGTTTCTTGGTTTTCCTTTCTGTTTTACAGT

                              ←       →
3171 GATAGTAGTGTTGCTCAAGTACCTAAGCAGACAGACAGCAACCCGACAGAGCCAGAAACGT 3231
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     GATAGTAGTGTTGCTCAAGTACCTAAGCAGACAGACAGCAACCCGACAGAGCCAGAAACGT

3232 CAATTGATGAACTGCTAGGCCTTCCGGTATGGGTAGATGTGTGGGGATGAGTGTGAGAG 3292
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     CAATTGATGAACTGCTAGGCCTTCCGGTATGGGTAGATGTGTGGGGATGAGTGTGAGAG

3393 AGTGTGAGCGAAGTAGAGTCAGAATGAAACAGTGTTTTGGTTTTGTTTTAAAATCATTGCC 3353
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     AGTGTGAGCGAAGTAGAGTCAGAATGAAACAGTGTTTTGGTTTTGTTTTAAAATCATTGCC

3354 AAAAAATTACATGGGTATAATCTAATTTAAAGTAATTAAAATAGACTGTTCATTCATCACA 3414
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     AAAAAATTACATGGGTATAATCTAATTTAAAGTAATTAAAATAGACTGTTCATTCATCACA

                                                         ←
3415 AATATGATCTGTTTCTGAACACATTTGGGTAGTGAATACAGTGTTCAAGCCAGCAGATGTC 3475
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||
     AATATGATCTGTTTCTGAACACATTTGGGTAGTGAATACAGTGTTCAAGCCAGCA
```

Mapping data from somatic cell hybrids, using a Bam HI/Eco RI preintegration site DNA fragment as a probe, had placed the Mpv 20 provirus on chromosome 9 (U. Francke, unpublished), which is also the site of the murine *Npat* gene (Matsuda et al. 1996). Based on the sequencing, mapping and expression data (see below), the MPSV-neo provirus in Mpv 20 mice had probably inserted into the *Npat* gene.

NORTHERN AND SOUTHERN BLOTTING ANALYSIS

Oligonucleotides were chosen corresponding to a region of the mouse protein product which had 100% identity to the human protein in order to generate a partial mouse cDNA from a testis library present in the laboratory (constructed by DA Gray). A 291 bp PCR cDNA fragment, which spans amino acids 281-302 (numbering according to the published human sequence) was obtained and sequenced to confirm its integrity (data not shown). This probe was subsequently used in both Northern and Southern blot analysis.

When the above fragment was used to probe a Southern blot containing both wild type and heterozygote DNA, both the wild type as well as the transgenic alleles were detected (Figure 13).

Northern blots of either total or poly(A)$^+$ RNA from mouse brain, lung, and kidney demonstrated that there was a decrease in the amount of message in a heterozygote transgenic animal relative to a wild type littermate (Figure 14 A). Using the hybridization signal from the control to correct for loading, the differences translated to a 1.6 fold, a 1.5 fold and a 1.7 fold decrease in the brain, kidneys and lungs respectively. This is in agreement with the supposition that recessive lethal mutant phenotype is a result of a loss of function of the *Npat* gene product in the Mpv 20 homozygotes.

Figure 13. Southern blot of heterozygote (h) versus wild type (w) CFW DNA using a partial Npat cDNA probe. 291 bp RT-PCR probe was generated from a testis cDNA library using oligonucleotides which span amino acid positions 281-302 (numbering according to the human cDNA). Probe was used on a Southern blot containing heterozygote (h) or wild type (w) CFW DNA digested with Bam HI, Eco RI or Hind III. All three enzymes showed an additional band present in the heterozygote samples confirming that the murine *Npat* cDNA spans the Mpv 20 proviral integration site. Blot was washed in 0.5 X SSC/0.1 % SDS at 55°C.

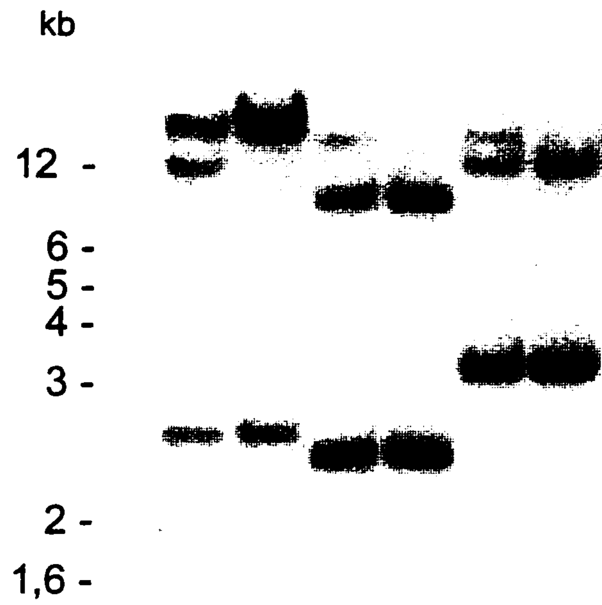
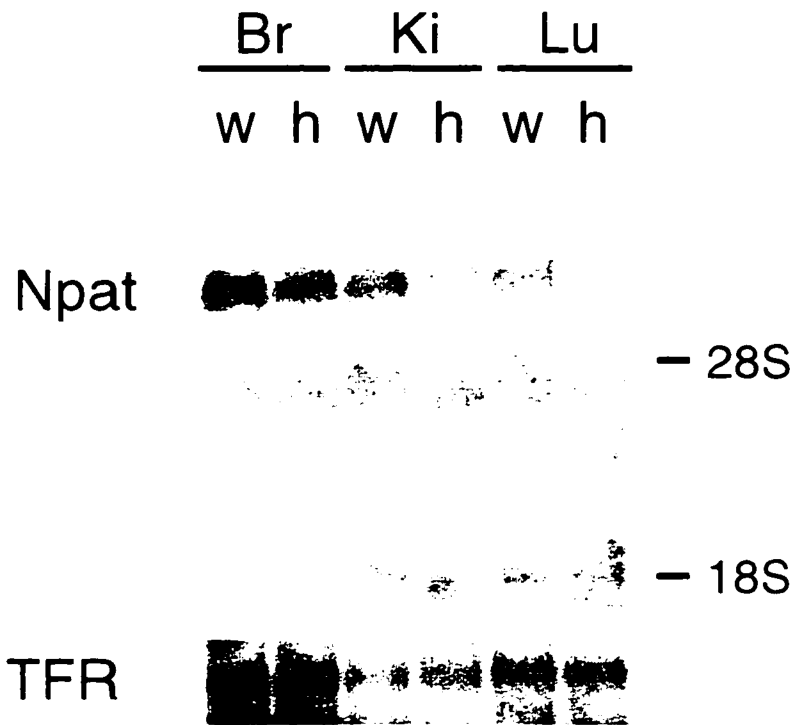
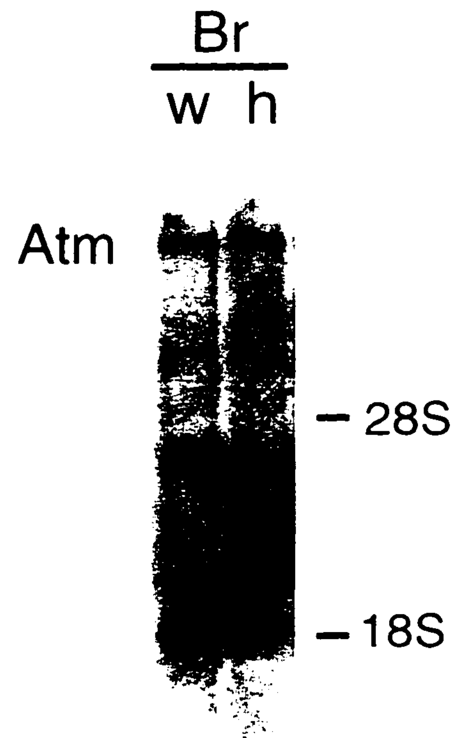

Bam HI
Eco RI
Hind III
h
w
h
w
h
w
kb
12 -
6 -
5 -
4 -
3 -
2 -
1,6 -

Figure 14. Northern blot analysis of tissue RNAs from a heterozygote (h) and a wild type littermate (w). (a) Matched samples of total RNA from brain (Br), kidney (Ki), and lung (Lu) were probed with the RT-PCR *Npat* cDNA probe. In every case examined we observed a reduction in the amount of *Npat* transcript in the heterozygote lane as opposed to the wild type littermate. The bottom panel shows the transferrin receptor cDNA (TFR) (obtained from ATCC) loading control. (b) The Northern blot from 14a was stripped and reprobed with the linked murine *Atm* cDNA. No observable difference in RNA expression was observed between samples.



a)
Br
Ki
Lu
w
h
w
h
w
h
Npat
28S
18S
TFR



b)
Br
w
h
Atm
28S
18S

Northern blots were stripped and reprobed with a mouse *Atm* cDNA fragment probe to determine if the expression of this closely-linked gene (Imai et al. 1996) was also affected by the integration event. There was no difference observed in *Atm* expression between Mpv 20 heterozygous and wild type brain RNAs (Figure 14 b).

DETERMINATION OF THE MECHANISM OF *NPAT* INACTIVATION

We were next interested in determining the mechanism by which the provirus was disrupting the expression of the *Npat* gene. It has been previously demonstrated that C-type murine retroviruses have the ability to cause changes in the methylation patterns of adjacent DNA (Stewart et al. 1982, Jähner and Jaenisch 1985). We employed a RACE protocol (rapid amplification of cDNA ends) in an attempt to clone the 5' end of the cDNA. The same 3 mnpat oligonucleotide which was used to generate the initial 291 bp partial cDNA clone was employed. The 5' oligonucleotide was targeted to the pBluescript region of the lambda Zap II vector into which the cDNA library was initially cloned. Two products were observed when the RACE PCR reaction was fractionated on a agarose gel. These migrated between 600 and 800 bp. To confirm that these were in fact bona fide RACE products, the gel was transferred and probed with the 291 bp partial *Npat* cDNA generated. Both bands gave a strong hybridization signal (Figure 15) and were subsequently isolated and subcloned into the pLitmus 28 vector and sequenced to confirm their identity. The two products had identical sequence with the large fragment extending further upstream. The sequence of this RACE product (RACE 9) is shown in Figure 16 A. The predicted murine protein has 79.5% identity and 84.3% homology to the human protein. Interestingly the murine protein has an additional 9 amino acids of which 8 are present in a single stretch having the sequence

GWYLTFHQ (underlined in Figure 16 B) of which the significance or role if any has not been determined.

The RACE9 product was used as a probe on a Southern blot containing both wild type and heterozygote DNA doubly digested with a series of enzymes and either Msp I or its methylation sensitive isoschizomer Hpa II which both contain CpG sequences in their recognition sites. It was reasoned that since CpG islands are found to a large degree at the 5' end of genes (Bird 1987), these enzymes should be able to distinguish if any changes in the methylation status of the gene or it's promoter was affected following the proviral integration event. These changes in methylation state may have resulted in a change in expression on the affected allele (Bird 1986). In all cases examined DNA fragments that hybridized to the probe were either uncut by the Msp I enzyme or displayed no difference in the ability for Hpa II to cleave the Msp I site present in the given fragment (see Figure 17 for a representative blot). No methylation differences were detected between the different animals examined using the above strategy.

The RT-PCR probe used in Figure 14 A was derived mostly from the region 3' of the provirus. This probe would therefore not be able to detect any truncated message being produced at the integration site. Northern blots containing both wild-type and heterozygote RNA were probed with the RACE 9 product. As can be seen in Figure 18 A, a full length *Npat* message was detected in all tissues examined. A second smaller sized band was also observed in samples derived from heterozygote animals. This band was present in both total and poly(A)$^+$ RNA Northerns indicating that it was polyadenylated. Examination of proviral LTR sequences when read on the opposite strand indicate the possibility of a

polyadenylation signal being used generating a stable polyadenylated truncated message. In order to confirm that the truncated message did in fact contain proviral sequences, a Northern blot was probed with an 754 bp Nde I/Spe I LTR fragment. As expected the probe identified an additional band present in the heterozygote samples which seemed to comigrate with the truncated *Npat* transcript (Figure 18 B). Endogenous proviral sequences were also detected in both wild type and heterozygote animals with the same probe. Northerns were stripped and probed with an internal 5 kb Sac I fragment containing most of the coding sequence from MuLV. Northern blots confirmed that the two upper bands observed (Figure 18 B) in all the lanes were in fact derived from endogenous MuLV provirus (data not shown).

Figure 15. Southern analysis of 5' RACE PCR reactions.
One tenth of RACE PCR reaction was fractionated on a 1.5% agarose gel. Samples were transferred and probed with the initial partial RT-PCR murine *Npat* cDNA product isolated. A doublet is observed in the 5' RACE lane whereas no hybridization signal was detected in the negative (-ve) control lane. Indicated on the left are the position of the DNA size markers. Blot was washed in 0.5X SSC/0.1% SDS at 65°C and exposed for 40 min on a phosphoimager screen.

5' RACE
-ve
2 kb
.9 kb
.6 kb

Figure 16. Nucleotide sequence and alignment of the RACE 9 product isolated.
(A) Nucleotide sequence of RACE 9 product obtained from an oligo dT/random primed lung cDNA library. Product was generated using 3' mnpat and KS oligonucleotides. Sequence shown is a compilation of multiple reads on both strands. (B) Translated product of the above product showing amino acid homology to the human *NPAT* (hNpat) cDNA. Vertical lines represent identical matches whereas + indicate conserved substitutions. Numbering of *hNpat* is according to the published translated sequence (Imai et al. 1996). Additional 8 amino acids present in the translated mouse sequence are underlined.

A

```
  1 CAATTTTGAATGAGTATGTAGCTATGAAAGCAAAAGAAACATCAAATGATGTCCCAACAA  60
 61 TAATGTCATCTCTATGGAAGAAGCTAGACCACACACTCTCTCAGATCAGGAGCATGCACA 120
121 GTTCTCCAGGGTTTGCTGCCCATCAGAGAGGTTGGTATTTAACTTTTCATCAAGCTCGAA 180
181 CGAGAAATGGAATTGCAGAAATCAAGCGGCAGCGATGGCTAGCATCTCAGGCAGCTCCCG 240
241 TCAGTGCAGAGTTACTGGTTTTACCTTATGCTTCAGGACAGTTTACCACTTATCCTTTGG 300
301 TAGCCACACAAGCTGTTAAGCCAACTGGCCCAATTTCAACTCCAGTGAGGTCAAATATTG 360
361 TGGTGGTCAACCAGTCACAGCCACAAAGTACTGTAACTAATACAGCTGGAGAGTCTTTGA 420
421 ATATCATTCCTGGTCCCCAGGAAAGGAAAACTCAAACCAGTTTAATGTCTCCTGGTAGAC 480
481 GTAAAAGTGAATCACAAAAGAAGAGTCTCACATCATCTGGACCGCATTCATCAAGGAATT 540
541 TTCA GGATCCAAATGCATTTGCAGTAGAAAAACAAATGGTTATTGAAAATGCAAGAGAG 600
601 AAAATATTAAGCAACAAATCTCTTCAAGAAAAACTAGCAGAAAACATAAATAAGTTTTTG 660
661 ACTAGTGATAGTACTGTTGCTCAAGTACCTAAGCAGACAGACAGCAACCCGACAGAGCCA 720
721 GAAACGTCAATTGATGAACTGCTAGGCC                                 748
```

B

```
    ILNEYVAMKAKETSNDVPTIMSSLWKKLDHTLSQIRSMESSPGFAAHQRGWYLTFHQ  RACE9
    ||||||||| |+|||+|| ||||||||||||||||||| ||| ||  ||
 62 ILNEYVAMKTKKTSNNVPAIMSSLWKKLDHTLSQIRSMQSSPRFAGSQR--------  110 hNpat

    ARTRNGIAEIKRQRWLASQAAPVSAELLVLPYASGQFTTYPLVATQAVKPTGPISTP  RACE9
    |||| ||||||||| |||| ||+|||||||||||||||| |   ||+ +|+| ||||
111 ARTRTGIAEIKRQRKLASQTAPASAELLVLPYASGQFTTPPSTGTQVTRPSGQISTP  167 hNpat

    VRSNIVVVNQSQPQSTVTNTAGESLNIIPGPQERKTQTSLMSPGRRKSESQKKSLTS  RACE9
     ||  |||| || | ||| | |||||+||| ||+|   |||||||||||||+|| |
168 SRSYFVVVNHSQSQDTVT-T-GESLNVIPGAQEKKAHASLMSPGRRKSESQRKSTTL  222 hNpat

    SGPHSS-RNFQDPNAFAVEKQMVIENAREKILSNKSLQEKLAENINKFLTSDSTVAQ  RACE9
    |||||| |||||||||||||||||||||||||||||||||||||||||||||+ +||
223 SGPHSSIRNFQDPNAFAVEKQMVIENAREKILSNKSLQEKLAENINKFLTSDNNIAQ  279 hNpat

    VPKQTDSNPTEPETSIDELLG  RACE9
    ||||||+||||||||||| ||
280 VPKQTDNNPTEPETSIDEFLG  300 hNpat
```

Figure 17. Southern analysis of methylation patterns in heterozygote (h) versus wild type (w) CFW DNA. Matched DNA samples were digested with either Kpn I or Eco RI in combination with Hpa II or its methylation insensitive isoschizomer Msp I. Blot was probed with the 5' RACE *Npat* product. In all cases examined there were either no Msp I sites in the region of DNA examined (first four lanes) or when a Msp I site was present, there were no observable changes in Hpa II restriction site methylation status(last four lanes). Blot was washed in 0.5 X SSC/0.1% SDS at 55°C.

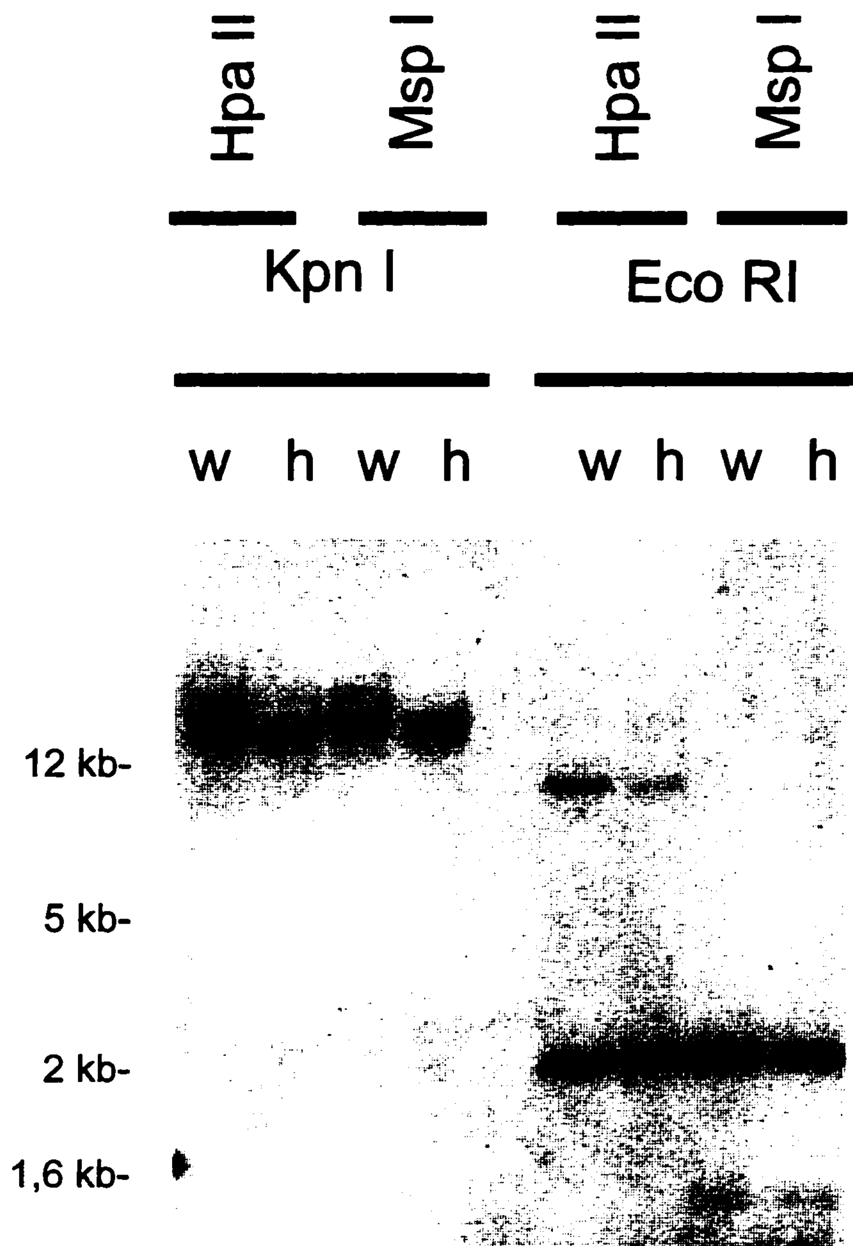
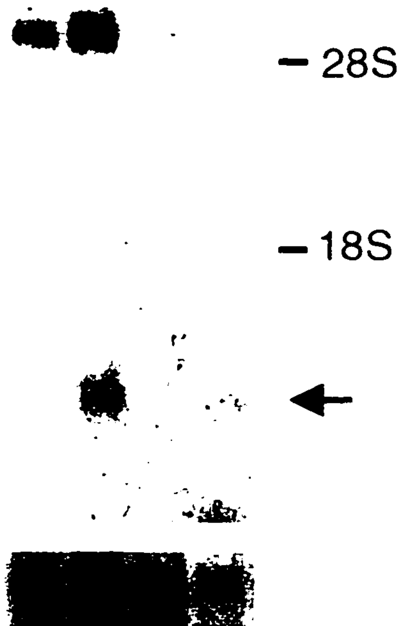

Hpa II
Msp I
Hpa II
Msp I
Kpn I
Eco RI
w h w h
w h w h
12 kb-
5 kb-
2 kb-
1,6 kb-

Figure 18: Northern blot analysis of heterozygote (h) versus wild type (w) RNA .
(A) Poly(A)$^+$ Northern blot was hybridized with a 5' RACE product obtained from a lung cDNA library. This probe contains sequences further upstream of the provirus. The probe hybridizes to the full length message in both heterozygous and wild type samples. an additional band is also observed in the heterozygote samples (see arrow) migrating below the 18S ribosomal RNA. (B) Poly(A)$^+$ Northern was stripped and reprobed with a 754 bp Nhe I/Spe I fragment from the MPSV LTR. Two bands are observed in all samples examined representing endogenous retrovirus, whereas an additional band is once again present in the heterozygote lanes migrating below the 18S ribosomal RNA. Transferrin receptor cDNA was used as a loading control.

a)

Br Ki

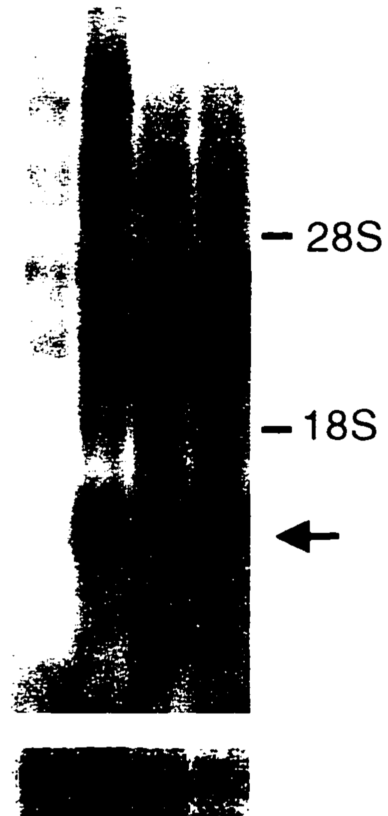
w h w h



b)

Br Ki

w h w h



RESULTS

(II)

MAPPING INTRON/EXON BORDERS

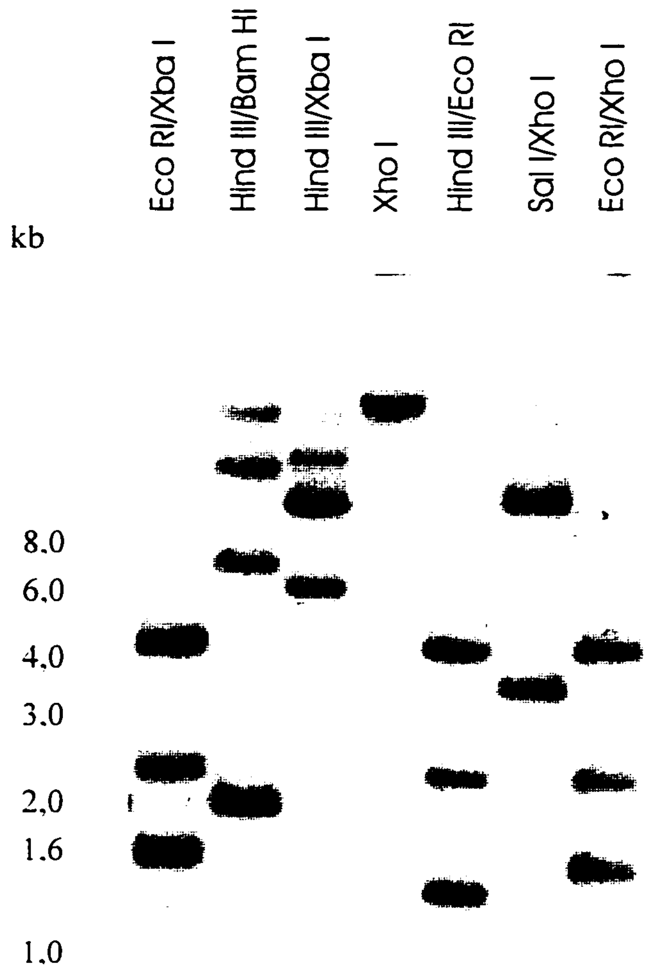
The initial strategy to map the intron/exon borders of *Unp* was to isolate genomic DNA fragments cloned in the λ DASH II vector. Individual phage clones were digested with enzymes present in the phage multiple cloning site and the DNA was transferred by Southern blotting. The blots were subsequently probed with either the full length or small fragments of the *Unp* cDNA (Gupta et al. 1993). This resulted in the identification of DNA fragments which contained exonic sequences, a representative blot is shown in Figure 19 for phage clone λRC9A. The positive fragments were subsequently subcloned into the pGEM-4 sequencing plasmid and sequenced. When the sequencing data (using the T7 and Sp6 oligonucleotides) (Promega) was compiled for the initial 18 fragments subcloned, it was found that only part of 13 exons were identified in the 14000 bases sequenced. The remaining fragments were subcloned and a series of synthetic oligonucleotides were employed as sequencing primers. Figure 20 gives the genomic structure of the gene and shows the positions of the 22 exons. With the exception of exon one, which includes the start of transcription, and exon 22 which contains the end of the transcript, all the intron/exons are flanked by the consensus splice acceptor/donor dinucleotides AG/GT (Mount 1982) (Figure 21). The gene is spread out over 47.3 kb and has been completely sequenced (Genbank accession AF026469). The sequence of which can be found in Appendix 1. The first exon contains a short 5' untranslated region (UTR) of 51 bp followed

by the initiating Met. Exon 8 contains the entire Cys domain which is essential for the catalytic activity of the enzyme, whereas the second domain containing two essential histidine residues was split between exons 20 and 21. Exon 22 was found to contain the entire 3' UTR.

The published cDNA sequence predicted the presence of twelve upstream open reading frames prior to the authentic start AUG (Gupta et. al. 1993). These short open reading frames were believed to play some role in regulating the transcription of the gene as was demonstrated for TGF-β (Arrick et al. 1991). Gilchrist et. al (unpublished data) has demonstrated through N-terminal sequencing of the protein that in fact there is only one upstream AUG present in the cDNA and that the protein starts at the second AUG. The sequence data from the genomic DNA also confirms the above finding (see Figure 22 for modified cDNA sequence) and has identified a few areas in the body of the cDNA where amino acid sequence changes have been inserted or replaced. These changes result in an additional 169 amino acids present at the N-terminal region of the protein. Also the DA substitution present within the body of the protein coding sequence is identical to the residues present in the human cDNA (Gray et al. 1995).

A database search of introns identified embedded in intron 3 a cDNA sequence of the S2 ribosomal protein (Heller et al. 1988, Suzuki et al. 1991). The cDNA is colinear with the intron and is inserted in the opposite transcriptional orientation. Closer examination of the nucleotide homology (Figure 23) reveals a number of regions in which gaps have been introduced. A poly A stretch is also present just downstream of the termination of homology indicating that the S2 sequence is a processed pseudogene.

Figure 19. Mapping of exons present in phage RC9A by Southern blot analysis.
Phage DNA from clone RC9A was digested with the enzyme combination shown above. Southern blot was probed with the entire Unp cDNA and positive hybridizing bands indicate the presence of exonic sequences. Blot was washed in 0.5X SSC/0.1% SDS at 65°C.



Eco RI/Xba I
Hind III/Bam HI
Hind III/Xba I
Xho I
Hind III/Eco RI
Sal I/Xho I
Eco RI/Xho I
kb
8.0
6.0
4.0
3.0
2.0
1.6
1.0

Figure 20. Genomic map of the *Unp* locus showing intron/exon positions.
Schematic representation of the *Unp* locus. The gene is spread out over 47.3 kb. . All the sequence has been compiled from multiple reads of both strands of DNA cloned as overlapping fragments into the sequencing vector pGEM-4. Filled boxes denote the position of the sequenced exons. Names, positions and sizes of the cloned genomic phages are displayed below the map.

2 kb

Eco RI
Bam HI
Eco RI
Eco RI
Eco RI
Eco RI
Bam HI
Eco RI
Bam HI
Bam HI
Eco RI
Eco RI
Bam HI
Bam HI
Eco RI
Bam HI
Eco RI
Eco RI
Bam HI
Eco RI
Eco RI
Bam HI
Bam HI
Bam HI

UG 7-3
UG 8-1
RC 9A
UG 9-3

Figure 21. Splice acceptor /donor sites of the sequenced exons.
Summary of exons present in the *Unp* gene. Uppercase letters denote exonic sequences, the numbering above the exons indicates their position relative to the cDNA positions given in Figure 22.

```
1                    151
tagcacgCCGTGCG.......CGCAGTGgtgagcg
152                  279
tccctagGTATCTT.......TTCTCAGgtggagt
280                  410
ttttcagATCCTGA.......CAGAAAAgtgagtt
411                  537
cttacagGTTGTGG.......ACCATTGgtaagca
538                  683
ttcttagCAACTAT.......GGGTCAGgtaagcg
684                  745
cactcagGTGCTAG.......AGTCAAAgtaagtg
746                  886
cttccagATCAAGC.......GCAGGGGgtaagaa
887                  1004
ttttcagTGGATCT.......TCTGCAGgtagagg
1005                 1178
gtttcagTGCCTGA.......GTTCAAGgtaggtt
1179                 1337
gttgtagACGCAAG.......AGATGCGgtatgat
1338                 1563
cgtgtagGTGGTAG.......TATCCAGgtaccgg
1564                 1646
tttgcagTACCGTG.......AGAAAATgtaagaa
1647                 1741
ttcttagATGGTGG.......TTTTTGTgtaagta
1742                 1933
catgcagGTATGAG.......GTATCAGgttggta
1934                 2022
ttttaagCCGATAC.......TATGAAGgtaagca
2023                 2247
gtatcagGAGATGA.......CTCAACTgtaagtg
2248                 2318
tccacagCTCGATC.......ATCTGAGgtatgtc
2319                 2437
tccgcagGCCTGTG.......ACCCCTGgtatgca
2438                 2587
tctttagGTACTGT.......CAGTCAGgtgggtg
2588                 2691
cactcagAGCTCTG.......GGTCACTgtaagta
2692                 2780
ttggcagACACTGC.......GATAGTGgtgagta
2781                 3629
gttttagACGAAAG.......CATAAAA
```

Figure 22. Corrected *Unp* cDNA sequence and its predicted protein product.
Initially the translation of the cDNA was believed to begin at nucleotide position 556. Sequencing of the gene identified a number of errors which resulted in the addition of an additional 169 amino acids prior to the above position. Bold amino acids show changes which include a few modified bases in the body of the cDNA.

```
   1 CCGTGCGCTAGGCGGATGTCCGGGCCGGCTGGGCCGGGGCGGCGGACGAGATGGCGGAAGGCCGGGGCAGCCGTGAG   78
                                                       M  A  E  G  R  G  S  R  E
  79 CGACCGGATGTGGAGACCCAGAAGACGGAGCTCGGAGCCTTGATGGGGACCACGCTCCAACGTGGGGCGCAGTGGTAT  156
      R  P  D  V  E  T  Q  K  T  E  L  G  A  L  M  G  T  T  L  Q  R  G  A  Q  W  Y
 157 CTTATTGACAGCCGGTGGTTCAAGCAGTGGAAGAAGTATGTGGGCTTTGACAGCTGGGACATGTACAATGTCGGGGAG  234
      L  I  D  S  R  W  F  K  Q  W  K  K  Y  V  G  F  D  S  W  D  M  Y  N  V  G  E
 235 CATAACCTGTTTCCTGGACCTATTGACAACTCCGGACTCTTCTCAGATCCTGAGAGTCAGACCTTGAAGGAGCACTTA  312
      H  N  L  F  P  Q  P  I  D  N  S  G  L  F  S  D  P  E  S  Q  T  L  K  E  H  L
 313 ATCGATGAGCTGGACTATGTGCTGGTCCCAGCCGAAGCCTGGAATAAATTGCTGAATTGGTATGGCTGTGTGGAGGGC  390
      I  D  E  L  D  Y  V  L  V  P  A  E  A  W  N  K  L  L  N  W  Y  G  C  V  E  G
 391 CAGCAGCCTATTGTCAGAAAAGTTGTGGAGCATGGCCTGTTTGTCAAGCACTGCAAAGTGGAAGTGTATTTGCTGGAG  468
      Q  Q  P  I  V  R  K  V  V  E  H  G  L  F  V  K  H  C  K  V  E  V  Y  L  L  E
 469 CTGAAGCTCTGTGAGAACAGTGACCCCACCAATGTGCTAAGTTGCCATTTTAGCAAAGCAGACACCATTGCAACTATT  546
      L  K  L  C  E  N  S  D  P  T  N  V  L  S  C  H  F  S  K  A  D  T  I  A  T  I
 547 GAGAAGGAGATGAGGAAGCTCTTCAACATCCCTGCAGAACGTGAAACACGGCTTTGGAACAAATACATGAGCAACACC  624
      E  K  E  M  R  K  L  F  N  I  P  A  E  R  E  T  R  L  W  N  K  Y  M  S  N  T
 625 TATGAGCAGTTGAGCAAGCTAGACAACACTATCCAGGATGCTGGGCTGTACCAGGGTCAGGTGCTAGTAATTGAGCCC  702
      Y  E  Q  L  S  K  L  D  N  T  I  Q  D  A  G  L  Y  Q  Q  Q  V  L  V  I  E  P
 703 CAAAATGAAGATGGCACATGGCCCCGGCAGAGCCTGCAGTCAAAATCAAGCACTGCACCTAGCAGAAATTTCACTACC  780
      Q  N  E  D  G  T  W  P  R  Q  S  L  Q  S  K  S  S  T  A  P  S  R  N  F  T  T
 781 TCTTCAAAACCATCCGCAAGTCCCTATTGCTCAGTGTCTGCCTCTCTCATTGCAAATGGTGATAGCACTAACAGCTCT  858
      S  S  K  P  S  A  S  P  Y  C  S  V  S  A  S  L  I  A  N  G  D  S  T  N  S  S
 859 GGGATGCACAGCTCCGGTGTCAGCAGGGGTGGATCTGGCTTCTCTGCTTCGTATAATTGCCAGGAGCCCCCATCACCT  936
      G  M  H  S  S  G  V  S  R  G  G  S  G  F  S  A  S  Y  N  C  Q  E  P  P  S  P
 937 CATATACAGCCTGGCCTCTGTGGACTTGGAAACCTGGGGAACACTTGCTTCATGAACTCTGCTCTGCAGTGCCTGAGC 1014
      H  I  Q  P  G  L  C  G  L  G  N  L  G  N  T  C  F  M  N  S  A  L  Q  C  L  S
1015 AACACTGCCCCACTGACTGAGTACTTTCTCAAAGATGAGTATGAGGCCGAGATCAACCGAGACAACCCTCTGGGGATG 1092
      N  T  A  P  L  T  E  Y  F  L  K  D  E  Y  E  A  E  I  N  R  D  N  P  L  G  M
1093 AAAGGGGAGATTGCAGAGGCCTATGCAGAGCTCATCAAGCAGATGTGGTCTGGAAGGGACACTCACGTGGCACCACGG 1170
      K  G  E  I  A  E  A  Y  A  E  L  I  K  Q  M  W  S  G  R  D  T  H  V  A  P  R
1171 ATGTTCAAGACGCAAGTGGGACGTTTTGCCCCTCAGTTTTCTGGCTACCAACAGCAAGACTCCCAGGAGCTGCTAGCC 1248
      M  F  K  T  Q  V  G  R  F  A  P  Q  F  S  G  Y  Q  Q  Q  D  S  Q  E  L  L  A
1249 TTTATTCTAGACGGACTGCACGAGGACCTGAACCGCGTAAAGAAGAAGCCTTACCTGGAGCCCAAGGACGCCAATGGG 1326
      F  I  L  D  G  L  H  E  D  L  N  R  V  K  K  K  P  Y  L  E  P  K  D  A  N  G
1327 CGACCAGATGCGGTGGTAGCAAAGGAAGCCTGGGAAAACCACAGGCTGAGGAATGATTCTGTGATTGTGGATACTTTC 1404
      R  P  D  A  V  V  A  K  E  A  W  E  N  H  R  L  R  N  D  S  V  I  V  D  T  F
1405 CATGGCCTTTTCAAATCGACTTTGGTTTGCCCAGAATGTGCTAAAGTTTCTGTGACCTTTGACCCATTTTGCTATCTA 1482
      H  G  L  F  K  S  T  L  V  C  P  E  C  A  K  V  S  V  T  F  D  P  F  C  Y  L
1483 ACTCTCCCACTGCCTTTGAAGAAGGATCGGATTATGGAGGTCTTCCTGGTTCCTGCTGACCCTCAGTGCAGACCTATC 1560
      T  L  P  L  P  L  K  K  D  R  I  M  E  V  F  L  V  P  A  D  P  Q  C  R  P  I
1561 CAGTACCGTGTGACTGTGCCATTGATGGGGGCCATTTCTGACCTGTGTGAAGCACTCTCCAAGCTGTCTGGCATTGCT 1638
      Q  Y  R  V  T  V  P  L  M  G  A  I  S  D  L  C  E  A  L  S  K  L  S  G  I  A
1639 GCAGAAAATATGGTGGTCACTGATGTATATAATCACCGCTTCCACAAAATTTTCCAAATGGATGAAGGTTTAAGCCAC 1716
      A  E  N  M  V  V  T  D  V  Y  N  H  R  F  H  K  I  F  Q  M  D  E  G  L  S  H
1719 ATCACGCCTCGAGATGACATTTTTGTGTATGAGGTCTGCAACACGTCCATGGATGGCTCAGAGTGTATCACTCTTCCA 1794
      I  T  P  R  D  D  I  F  V  Y  E  V  C  N  T  S  M  D  G  S  E  C  I  T  L  P
1795 GTCTACTTCAGAGAGAAGAAGTCCAGGCCGTCGAGTGCTTCCTCGGGGGCTGTGCTCTATGGACAGCCCCTTCTTGTC 1872
      V  Y  F  R  E  K  K  S  R  P  S  S  A  S  S  G  A  V  L  Y  G  Q  P  L  L  V
1873 TCTGTCCCTAAGCATAAGCTAACCCTAGAGTCTTTGTACCAGGCTGTTTGTGATCGTATCAGCCGCTACATAAAACAG 1950
      S  V  P  K  H  K  L  T  L  E  S  L  Y  Q  A  V  C  D  R  I  S  R  Y  I  K  Q
1951 CCTTTGCCTGATGAGTTTCTCAGCTCACCCTTAGAGCCTGGGGCCTGCAATGGCTCTAGGAGTAGCTATGAAGGAGAT 2028
      P  L  P  D  E  F  L  S  S  P  L  E  P  G  A  C  N  G  S  R  S  S  Y  E  G  D
2029 GAAGAGGAAGAAATGGATCATCAAGAGGAAGGAAAAGAGCAGCTTTCCGAAGTGGAAGGCAGTGGTGAGGACGATCAG 2106
      E  E  E  E  M  D  H  Q  E  E  G  K  E  Q  L  S  E  V  E  G  S  G  E  D  D  Q
2107 GGAGATGACCATAGCGAGAGCGCCCAAAAGGTGAAAGGCCAGCCGAGGCACAAGAGGCTCTTTACCTTCAGCCTCGTG 2184
      G  D  D  H  S  E  S  A  Q  K  V  K  G  Q  P  R  H  K  R  L  F  T  F  S  L  V
2185 AACTCCTGTGGAACTGCTGATATCAATTCACTGGCAACTGACGGGAAACTCCTCAAACTCAACTCTCGATCCACACTG 2262
      N  S  C  G  T  A  D  I  N  S  L  A  T  D  G  K  L  L  K  L  N  S  R  S  T  L
2263 GCCATTGACTGGGACAGTGAGACCCGAAGCCTTTACTTTGATGAGCAAGAATCTGAGGCCTGTGAGAAGCACCTGAGC 2340
      A  I  D  W  D  S  E  T  R  S  L  Y  F  D  E  Q  E  S  E  A  C  E  K  H  L  S
2341 ATGTCACAGCCGCAGAAGAAGAAGAAGGCGGCAGTAGCCCTGCGAGAGTGCATCGAGCTCTTCACCACCATGGAGACC 2418
      M  S  Q  P  Q  K  K  K  K  A  A  V  A  L  R  E  C  I  E  L  F  T  T  M  E  T
2419 CTTGGGGAGCATGACCCCTGGTACTGTCCCACCTGTAAGAAGCACCAGCAGGCAACAAAGAAGTTTGACTTGTGGTCT 2496
      L  G  E  H  D  P  W  Y  C  P  T  C  K  K  H  Q  Q  A  T  K  K  F  D  L  W  S
2497 TTGCCCAAGATCCTGGTGGTTCACCTCAAGCGTTTCTCCTATAACAGATACTGGCGGGATAAACTTGACACCGTGGTG 2574
      L  P  K  I  L  V  V  H  L  K  R  F  S  Y  N  R  Y  W  R  D  K  L  D  T  V  V
2575 GAGTTCCCAGTCAGAGCTCTGAACATGTCCGAGTTTGTCTGTGACCGGTCAGCAAGGCCTTATGTTTATGACCTAATT 2652
      E  F  P  V  R  A  L  N  M  S  E  F  V  C  D  R  S  A  R  P  Y  V  Y  D  L  I
2653 GCTGTGTCCAATCACTATGGAGCCATGGGGGTTGGTCACTACACTGCATATGCGAAGAACAGACTGAACGGAAAATGG 2730
      A  V  S  N  H  Y  G  A  M  G  V  G  H  Y  T  A  Y  A  K  N  R  L  N  G  K  W
2731 TATTACTTTGATGATAGCAGTGTGTCCCTGGCCTCTGAGGACCAGATAGTGACGAAAGCCGCCTACGTGTTGTTCTAT 2808
      Y  Y  F  D  D  S  S  V  S  L  A  S  E  D  Q  I  V  T  K  A  A  Y  V  L  F  Y
2809 CAGCGTCGGGATGACGAATGCTCCAGCACATCTTCGCTCGGCAGTTTCCCCGGTTCTGATGGAGGGGTGAAGCTGAGC 2886
      Q  R  R  D  D  E  C  S  S  T  S  S  L  G  S  F  P  G  S  D  G  G  V  K  L  S
2887 AGCTCACACCAGGGCATGGGGGATGAGGAGGCTTACAACATGGACACCAACTGATGTGGCCTCTGTGACCCTGCCACC 2964
      S  S  H  Q  G  M  G  D  E  E  A  Y  N  M  D  T  N  .
2965 CTCCCAACACCAGCATCCTCCCCCAGGAGAATATCTGTGGCACTCTTGAAGACTGCTCATATCAATCTCAAACCAGAT 3042
3043 GAGGACATTACCTCCTTTTATGAGAAGAGGAAGGAAAGAGAAACCCTGCTTTCTGAGAAGGGCTCTATGTTAAGAGGC 3120
3121 TGGATTATTTTTTCTTTTAAACGGGTGGTGAGAGAAAAACCTGTGGAGCTCAATTGTGGGCTGGGGTTTGGTGCACAG 3198
3199 TGGCATGTGTCTCATGGATTCCAAAGAGTGGAGAGGAAGACCCTGCAATGAGCATGGTGCAGCCATGAAGACCCACCT 3276
3277 GCAGGGCTGCTCCGAACTGGGTTTTGGATTCTGGTGCCCACACTGCTCACCCCAGTAGTCTGGCCACAGTAAACCAAG 3354
3355 CCTCCAGCAAATTTGATGTTGTCCTCTAGTTCTTTGCTGTTCATTGAGATGGGCTTATAGTTTTCTAAACTGTCTCTG 3432
3433 AAACCCGTTTCTATTCTCTTGTTCTCTCAGAATTTTCCAGCCGCCTTCCCGGTTATGAAATGAAGTTAGTATTGGGCT 3510
3511 ATTGTTTTCAGTTTCTCACAGACTCCTAGAAGTTCTGTTTTGTCCCCCACTGCCCTTGATAAAGTGATTCTATGTGGA 3588
3589 ACTTTCATTAGCTCTCCCTTAAAAGAAGTCTATTCATAAAA                                      3631
```

Figure 23. Sequence alignment of intron 3 with the *S2* ribosomal protein pseudogene. Genbank database search identified the rat *S2* protein (Genbank accession X57432) present in intron 3. The cDNA is present in the opposite transcriptional orientation relative to the *Unp* gene. It is colinear with the exon and a number of gaps were introduced to obtain an optimal alignment. Numbering of *S2* is relative to the published sequence (Suzuki et al. 1991). Numbering of intron 3 is relative to the nucleotide position in intron 3 itself.

```
2540 TCTTCTTCTTCGGGAAAACACCAAATGGTGGATGACGCCGGTGCAGCGAGAGGGC 2486 INTRON 3
     |||| ||||||||||||||||||||||| ||||||||||||||||||| ||||||
   1 TCTTTTTCTTCGGGAAAACACCAAATGGCGGATGACGCCGGTGCAGCGGGAGGGC   55  RAT S2

2485 CCGGAGGACCCGGGGCCC 2468 INTRON 3
     ||||||||||||||| ||
  56 CCGGAGGACCCGGGGGCC   73  RAT S2


2467  AGGATTAGGAGGCCGCGGTGGCTTCCGCGGAGGATTCGGCAGCGGTCTTAGGGG 2414 INTRON 3
      |||||||||||||||||| |||||||||||||||||||||||||| ||||||||
  74 CAGGATTAGGAGGCCGCGGCGGCTTCCGCGGAGGATTCGGCAGCGGCCTTAGGGG  128  RAT S2

2413 CCGTGGTCATGGCCGAGGCCGTGGTCAAGG       2384  INTRON 3
     ||| |||| ||||||||||||||| |  ||
 129 CCGCGGTCGTGGCCGAGGCCGTGGCCGCGGTCGAGG 164   RAT S2


2383 CCCCGGGGCTCGTGGAGGTAAAGCTGAAGACAAGGAGTGGATCCCCGTCACCAAG 2329 INTRON 3
     || ||||||||||||||||||||||||||||||||||||||||||||| ||||||
 165 CCGCGGGGCTCGTGGAGGTAAAGCTGAAGACAAGGAGTGGATCCCCGTTACCAAG  219  RAT S2

2328 CTGGGCCGCCTGGTTAAGGACATGAAGATCAAGTCCTTGGAGGAGATCTACCTGT 2274 INTRON 3
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||
 220 CTGGGCCGCCTGGTTAAGGACATGAAGATCAAGTCCTTGGAGGAGATCTACCTGT  274  RAT S2

2273 TCTTCCTGCCTATTAAGGAGTCTGAGATCATTGACTTCTTCCTGGGTGCGTCCCT 2219 INTRON 3
     ||| |||||| ||||||||||||||||| |||||||| ||||||||||| || ||
 275 TCTCCCTGCCCATTAAGGAGTCTGAGATTATTGACTTTTTCCTGGGTGCATCTCT  329  RAT S2

2218 AAAGGATGAGGTTCTGAAAATCATGCCAGTGCAGAAGCAGACTCAGGCTGGCCAG 2164 INTRON 3
     ||||||||| ||||||||||||||||||||||||||||||||||| ||||||||||
 330 AAAGGATGAAGTTCTGAAAATCATGCCAGTGCAGAAGCAGACTCGGGCTGGCCAG  384  RAT S2

2163 CGGACCAGGTTCAAGGCTTTCGTCACTATTGGGGAC-ACAACGGTCACGTTGGTC 2110 INTRON 3
     ||||||||||||||||| || ||| ||||||||||| |||| || ||||||||||
 385 CGGACCAGGTTCAAGGCCTTTGTCGCTATTGGGGACTACAATGGCCACGTTGGTC  439  RAT S2

2109 TTGGTGTTAAGTGCTCCACAGGAGGTTGCCACTGCCATCCGAGGGGCCATCATCT 2055 INTRON 3
     |||||||||||||||||| ||||||| |||||||||||||||||||||||||||
 440 TTGGTGTTAAGTGCTCCA-AGGAGGTAGCCACTGCCATCCGAGGGGCCATCATCC  493  RAT S2

2054 TGGCCAAGCTTTCCATTGTCCCTATGAGGAGAGGCTACTGGGGGAACAAGATTGG 2000 INTRON 3
     ||||||||||||| || |||||  || ||||||||||||||||||||||||||||
 494 TGGCCAAGCTTTCGATCGTCCCAGTGCGGAGAGGCTACTGGGGGAACAAGATTGG  548  RAT S2

1999 CAAGCCCCACACTGTTCCATGCAATGTGACAGGCTGCTGTGGCTCTGTGSTGGTG 1945 INTRON 3
     ||||||||||||||| |||||||| ||||||||| |||||||||||||| ||||
 549 CAAGCCCCACACTGTGCCATGCAAGGTGACAGGCCGCTGTGGCTCTGTGCTGGTT  603  RAT S2

1944 CATCTCATCCCTGCCCCCACAGGCACTGGCATTGTCTCTGCTCCTGTGCCCAAGA 1890 INTRON 3
     | ||||||||||||||||| |||||||||||||||||||||||||||||||||||
 604 CGTCTCATCCCTGCCCCCAGAGGCACTGGCATTGTCTCTGCTCCTGTGCCCAAGA  658  RAT S2

1889 AGCTCCTGATGATGGCCGGTATAGATGACTGMTACACTTCAGCCAGAGGCTGCGC 1835 INTRON 3
     |||| |||||||||||||||||||||||||| |||||||||||||| |||||| |
 659 AGCTACTGATGATGGCCGGTATAGATGACTGCTACACTTCAGCCAGGGGCTGCAC  713  RAT S2

1834 TGCCACCCTGGGCAACTTTGCTAAGGCCACCTTTGATGCCATCTCCAAGACTTAC 1780 INTRON 3
     ||||||||||||||||||||| ||||||||||||||||||||||||||||| |||
 714 TGCCACCCTGGGCAACTTTGCCAAGGCCACCTTTGATGCCATCTCCAAGACCTAC  768  RAT S2

1779 AGCTACCTGACCCCCGACCTVTGGAAAGAGACTGTCTTCACCAAGTCTCCTTATC 1725 INTRON 3
     |||||||||||||| ||||| |||||||||||||| |||||||||||||||||||
 769 AGCTACCTGACCCCTGACCTCTGGAAAGAGACTGTGTTCACCAAGTCTCCTTATC  823  RAT S2

1724 AGGAATTCACTGATCATCTTGTGAAAACCCACACCAGAGTCTCTGTTCAGAGGAC 1670 INTRON 3
     ||||||||||||| |||||||||||||||||||||||||||||||||||||||||
 824 AGGAATTCACTGACCATCTTGTGAAAACCCACACCAGAGTCTCTGTTCAGAGGAC  878  RAT S2

1669 CCAGGCTCCAGCTGTGGCTACCACATAAGGGTTTTTATACAAGA 1626 INTRON 3
     ||||||||||||||||||||||||||||||||||||||||||||
 879 CCAGGCTCCAGCTGTGGCTACCACATAAGGGTTTTTATACAAGA  922  RAT S2
```

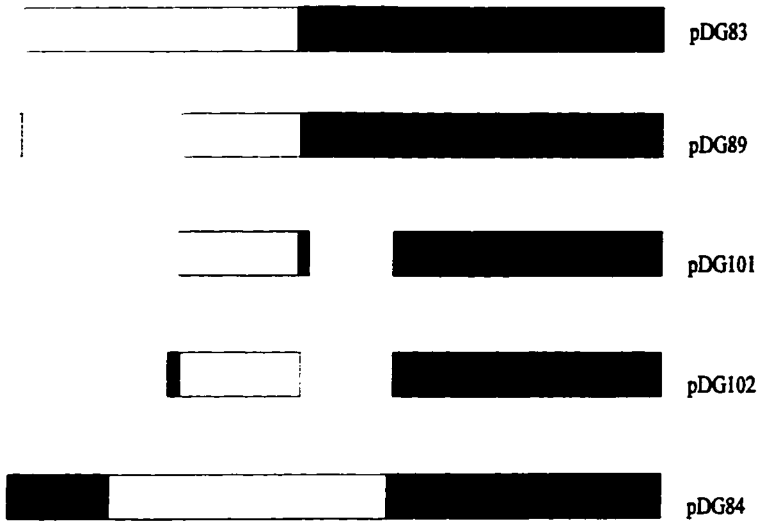
CAT ASSAYS

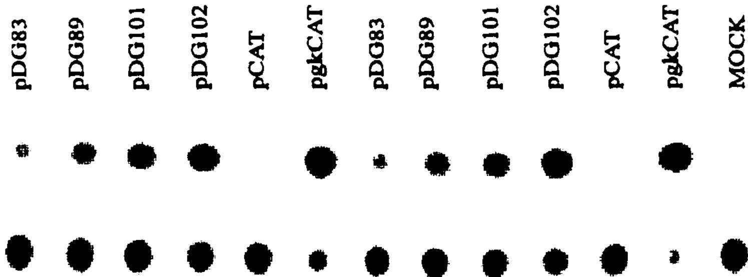
The 5' end of the sequenced *Unp* gene contained a 1.1 kb Eco RI/Sac I fragment which included the first 3 AUGs of the cDNA. Being upstream of the cDNA, it was reasoned that this fragment should contain some or all of the *Unp* promoter. The above fragment was cloned into the Pst I site of promoterless pCAT (Promega) reporter plasmid. Initially constructs pDG83, pDG84, and pDG89 (see Figure 23 A and refer to Materials and Methods, construction of CAT plasmids) were co-transfected with pDM2 (pgk-1 beta-galactosidase construct) as internal control into the 293 T (primary human embryonal kidney) cell line. Results demonstrated that the pDG84 had background levels of promoter activity, whereas pDG83 showed low activity. Surprisingly the deletion construct, pDG89, showed between two and eight fold higher CAT activity compared to the full length pDG83 construct (data not shown, see below for subsequent experimental data). Experiments were repeated in triplicates and the same trend was consistently observed. This data suggested the possible existence of a binding site for a repressor element being present in the deleted region of the DNA. The initial constructs also contained the first 3 AUGs of the cDNA and it was possible that this region in addition to the above putative repressor element was sufficient to decrease promoter activity by the formation of secondary structure in the 5' untranslated region (UTR) of the mRNA. Constructs pDG101 and pDG102 were subsequently generated which did not contain the putative repressor element sequence and in which most of the 5' UTR was removed. All five constructs were retransfected in duplicate and subjected to CAT assays (see Figure 24 B and C). As was the case in previous experiments, CAT activity increased for pDG89 relative to the parental construct, however, CAT activity increased in both the

pDG101 and pDG102 constructs indicating that the 5' UTR was also inhibiting CAT activity, furthermore the highest activity was observed for the pDG102 construct demonstrating that the promoter was functional in both orientations under the experimental conditions utilized. Analysis of the DNA sequence prior to the start of the CAT gene in pDG83 and pDG89 shows that translation would occur at the second AUG of the region and continue until 3 in frame stop codons are found resulting in the loss of CAT product.

Figure 25 shows the sequence of the initial DNA fragment used to construct pDG 83 and 84. *Unp* shows the typical hallmarks of a housekeeping gene (Williams and Fried 1986, Dush et al. 1985, Reynolds et al. 1984, Melton et al. 1984, Williams et al. 1988). It is constitutively expressed in all tissues examined (Gupta et al. 1994), the promoter is GC rich and does not have a TATA box or a CAAT motif present at approximately 30 and 80 bases upstream of the cap site. Also present are potential binding sites for the transcription factors Sp1, AP1, and AP2 determined by a database search using the MatInspector program (Quandt et al. 1995) (see Figure 25 and Appendix 2). These include 12 AP1 sites six of which are on the positive strand and the remainder on the negative strand. Results of the entire MatInspector analysis can be found in Appendix 2.

Figure 24. Identification of promoter activity present in a cloned 1.1 kb Eco RI/Sac I genomic fragment of the Unp gene. (A) Constructs utilized for CAT assays (not drawn to scale). Open box denotes genomic sequences, black box shows presence of cDNA sequences. gray box represents the CAT reporter gene. Initially Constructs pDG83, pDG84, and pDG89 were designed to contain what was believed to be the first three AUGs present in the 5' untranslated region. pDG89 contains a 557 bp deletion between the Bgl II sites present in the gene at positions 9 and 567 (numbering according to the genomic sequence). pDG 101 and pDG102 were generated using oligonucleotides designed after the Bgl II site at position 567 and prior to the first AUG in the cDNA. (B) Thin-layer chromatogram of CAT activity from the transfection of 293 T cells. pCAT represents the vector alone, pgkCAT (pgk-1 promoter driving CAT, gift from Dr McBurney) was used as a positive control, and Mock represents cells which were subjected to the same protocol but in the absence of DNA. (C) Graphical representation of the CAT activity of the above constructs. Data is plotted as per cent of substrate converted to the acetylated form. All values are normalized to β-galactosidase activity present from the cotransfected internal control vector (n=2). Error bars indicate the difference between the average values obtained.


A
pDG83
pDG89
pDG101
pDG102
pDG84


B
pDG83
pDG89
pDG101
pDG102
pCAT
pgkCAT
pDG83
pDG89
pDG101
pDG102
pCAT
pgkCAT
MOCK

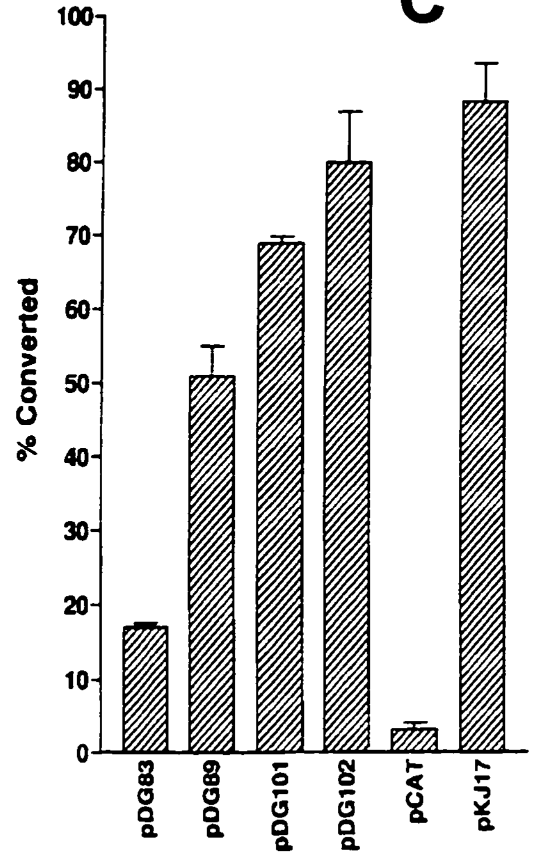

C
% Converted
0
10
20
30
40
50
60
70
80
90
100
pDG83
pDG89
pDG101
pDG102
pCAT
pKJ17

Figure 25. Sequence of the 1.1 kb genomic Eco RI/Sac I fragment containing the Unp promoter. Uppercase letters indicate the position of cDNA sequence and * denotes the initiating methionine. The complete fragment was used to create both pDG 83 and pDG 84 constructs. Italics show the positions of the Bgl II sites that were used to generate construct pDG 89. Bold sequences show the positions of the 5' and 3' promoter oligonucleotides used to generate plasmids pDG101 and pDG102. Indicated are a few potential binding sites for transcription factors identified using the MatInspector program (Quandt et al. 1995). (+) denotes the consensus sequence is present in the same orientation of the sequence, (-) denotes that the sequence is complimentary to the one given in the figure. Sequence was obtained from multiple reads on both strands.

1 gaattcag*agatct*gcctacctctgcctcctgagtgctgggttaaaagcctgtgccatca 60

61 ccgcccggaagaaaacggaatcattgtttcaggaaaaaattaagatgcctggacctctgt 120
SP1(-) OCT-1(+)

121 cggcttcggtggttgctactctctccttgacctggatggctgacaggatacaaagttgtt 180
AP1(+) AP1(+)

181 taagacaactttgtttttgcgtctatttatgactttattttgtttttgcaggtgtgg 240
AP1(+) E47(+)

241 tgatgtatgcttttaatcccagcactctggagaaagggcaaatgggcctaaattcgaga 300

301 ctagcctagattacataaacagctccaggacagcagctagaagaaacctgtcccaaacaa 360

361 acaacaaatttcactattgtgtgcctgtttggcgaacgcgtgccacagtgtctgtgtcga 420
OCT-1(-) E2F(+)

421 gaacaaaggacaactttaaagagtcagtttctttttccacttttacacgggttctgggg 480
AP1(-)

481 atcaaacccagttgtcacccacgtagccatcgcgccggcccttgtttgaaacaactatta 540
AP1(-)

541 atggctggaccacaccggacagtc*caga**tc****tcaatgcgtccgtgtgc**acttggctacagg 600
OCT-1(-)

601 ggctgccgaggcagtaaattcaatctcgcatttaggcagtcttgctgaaagccgcatgga 660

661 accgagtggaaacaagaaaagcaagcggcaggctcctcctgcacggatttagcatcaggc 720
SP1(+)

721 taacctcacaggccaggatggagcgagtgactgatgggtcgagggcgggccagcagtcag 780
AP1(+) SP1(+) AP1(-)

781 atcgcaggagccgctcagacacgggtggatgtggccgccgcctctgggcgtccagcctgc 840

841 ttctgtccgcttcggctccggaggggcgggggggggttaacagcgcgtggggcgggga 900
SP1(+) SP1(+)

901 gaagaggcgaaagtcgacggcggtgccgggcgcgcgcacgtgcgtggttacgtgcgtgcg 960

961 tgggcgcgg**cccatagcacgCCGTG**CGCTAGGCGGATGTCCGGGCCGGCTGGGCCGGGGC 1020
SP1(+)

1021 GGCGGACGAGATGGCGGAAGGCCGGGGCAGCCGTGAGCGACCGGATGTGGAGACCCAGAA 1080
*

1081 GACGGAGCTC 1090

CHAPTER 4

DISCUSSION

(I)

The Mpv 20 mouse strain was one of four recessive lethal mutant strains resulting from a screen of 70 transgenic lines obtained by retroviral infection of embryos (along with Mov 13, Mov 34, and Mpv 17). The mutation frequency was determined to be approximately 5-6%. Although the sample size is obviously quite small, the proviral insertions in these strains appear to favour noncoding regions. Insertion of an MPSV-neo provirus into an intron of the *Npat* gene was found to interfere with expression of the gene (Figure 14). In the Mov 13 strain proviral insertion occurred in the α(1)I collagen gene (Schnieke et al. 1983; Jaenisch et al., 1983; Harbers et al. 1984) resulting in inactivation of a cis-acting transcriptional element (Barker et al. 1991). In the Mov 34 strain, insertion into the 5' region of a gene encoding a proteosomal subunit (Soriano et al. 1987; Gridley et al. 1990 Tsurumi et al. 1995) was found to inactivate RNA expression from the gene. In addition, a separately derived line, A6, also identified the proviral integration event occurring in the first intron of the UbcM4 gene (Harbers et al. 1996) using a different infection protocol. In some cases where the provirus is resident within an intron the mechanism of gene inactivation is still not fully understood.

Retroviral insertion can occur at many sites in the mammalian genome, and there is some evidence of preferential insertion into active chromatin (Vijaya et al. 1986; Rohdewohld et al. 1987). It may be that the noncoding sequences located at the sites of

proviral insertion in the Mov and Mpv strains have a conformation that is particularly accessible to the viral integrase, or that disruptions of coding sequences are merely under-represented in the current small sample size. Alternatively, the presence of additional cellular factors may be present at the above integration sites which may associate with the viral integrase aiding in the integration reaction itself. Kalpana et al. (1994) have demonstrated, using a yeast two hybrid system, that the HIV-1 integrase enzyme associates with high affinity to the human homolog of the yeast transcription factor SNF5. This association was also found to increases the activity of the integrase in an in vitro integration assay. Related factors or binding sites for control elements may be present at the above integration sites. The proviral insertions may allow for their future identification.

There are in fact few reported examples of mutations affecting the development of the mouse morula. The earliest of these is a 5 cM deletion spanning the albino locus on chromosome 7 (Lewis, 1978), in which embryonic arrest occurs at the 2 to 6 cell stage. Insertion of exogenous DNA into and/or the deletion of 2 cM of chromosome 1 in the ßS12 transgenic line resulted in the generation of a morula decompaction (mdn) phenotype (Cheng and Costantini, 1993). Due to the large deletions present in the genomes, the gene or genes responsible for the early embryonic arrest have not been identified. The absence of homozygous midgestation embryos and the preimplantation arrest of one quarter of embryos from an Mpv 20 heterozygous cross suggests a recessive lethal mutation whose phenotype is amongst the earliest known.

Formal proof that the Mpv 20 phenotype is in fact occurring at the 8 cell stage needs to be obtained. It is possible that the Mpv 20 homozygote embryos are more sensitive to

culture effects then their wild type counterparts. This sensitivity may translate into the growth arrest observed *in vitro*. To circumvent this problem, a systematic approach of embryos at all stages of development must be carried out. A PCR based approach can be developed to distinguish between the mutant and/or wild type alleles present in the isolated embryos. Embryos would be isolated at different preimplantation time points between the zygote and the blastcyst stage and after implantation between days 4.5 and 11 postcoitum.

The function of *Npat* is not currently known. Based on the presence of protein motifs in the predicted translation product, it has been speculated that the human *Npat* gene encodes a nuclear protein that may be a substrate of cyclin-dependent kinases (Imai et al., 1996). Others have noted similarities between *Npat* and transcription factors such as *Oct-1* (Chen et al., Mammalian Genome, in press). Until *Npat* is studied further and its function known, it will be difficult to interpret how its disruption leads to embryonic arrest at or around the uncompacted 8 cell stage. Two possibilities must be considered: either the lack of zygotic *Npat* is of no consequence until maternal stores are depleted (which must occur at or around the 8 cell stage), or the *Npat* gene product is not required by the embryo until the 8 cell stage. In the former scenario, the *Npat* gene product could be involved in any number of structural, metabolic, or cell cycle-related activities, whereas the latter scenario implies that the *Npat* product has a direct or indirect role in events occurring at the 8 cell stage (notably, cell polarization or compaction). To the distinguish between the two possibilities, antisense oligonucleotide primers could be microinjected into one cell zygotes prior to the onset of embryonic transcription. If the presence of the maternal message is essential for the embryos' ability to divide, the embryos should be arrested at the one cell

stage of development. This approach has been successfully employed to demonstrate that the mouse *Mos*-encoded protein is necessary for normal meiotic maturation (Paules et al. 1989). In addition concanavalin A was used as a marker to monitor cell surface polarity events that occur at the 8 cell stage of development (Ziomek and Johnson 1980). This protocol could be employed on homozygote embryos to determine if lack of the *Npat* gene product inhibits the normal processes which occurs at this stage.

The above microinjection experiments would demonstrate the essential nature of the *Npat* gene product but would not address whether the mutation is cell autonomous or not. To address this question, an experiment can be designed in which embryos are flushed at the 2-4 cell stage of development, physically disrupted and subsequently mixed with wild type embryos. Resulting embryos are reintroduced into pseudopregnant female mice. Examination of all the tissues in the chimeric animals that are born would be carried out to determine if the homozygote mutant cells contributed to the development of the animal. If this approach is not feasible, early embryos can be transplanted to extrauterine sites to generate teratocarcinomas followed by the isolation of embryonic carcinoma (EC) cells from these tumours (reviewed in Martin 1980). The EC cells are isolated and can be introduced into wild type blastocysts. Contribution of the EC cells to some or all the tissues would indicate that the *Npat* product can behave in a cell autonomous fashion and therefore the mutant cells can be "rescued" by placing them in a wild type environment. This would suggest that the *Npat* gene product can freely diffuse between cells and behave in a paracrine fashion. The feasibility of this approach was demonstrated by the microinjection of ES cells carrying the disrupted 413.d locus, which results in early embryonic lethality at

approximately 8 days postcoitium, into host blastocysts. The recovered midgestation conceptuses showed the ES cells contribution in all extraembryonic and somatic tissues (Conlon et al. 1991). One of the advantages in generating Mpv 20 EC cells is that subsequent work could be carried out in a tissue culture system eliminating the need to resort to the collection of large numbers of embryos.

Upon passage through embryonic cells, C-type proviruses can initiate a general inactivation of flanking genomic regions by a mechanism that involves methylation (Jähner et al., 1982; Jähner and Jaenisch, 1985). If this was the case in the Mpv 20 strain, one might expect the adjacent *Atm* gene to be affected by the insertion of the provirus, but this does not seem to be the case since no change in its expression was found (Figure 14). The presence of a truncated *Npat* message points to an inactivation of the gene through a methylation independent mechanism. The promoter of the homozygote mutant mouse would still initiate the production of a transcript but this would in turn not create a full length message resulting in a functional null mutation. This is consistent with the data seen in the heterozygotes in which no change in expression is seen from the *Atm* gene which uses the same promoter but on the opposite strand. In addition, no changes in the methylation status of the *Npat* gene was observed (Figure 17). It cannot be ruled out that the RACE product obtained was not close enough to the promoter and thus a restriction fragment which might show a difference in sensitivity could not be identified. In order to overcome this difficulty a 5' RACE product that extends to the start of the cDNA must be isolated to confirm that there are truly no changes in the methylation patterns between heterozygote and wild type animals. The independent isolation from a second group of the *Npat* cDNA (designated E14) (Byrd et al.

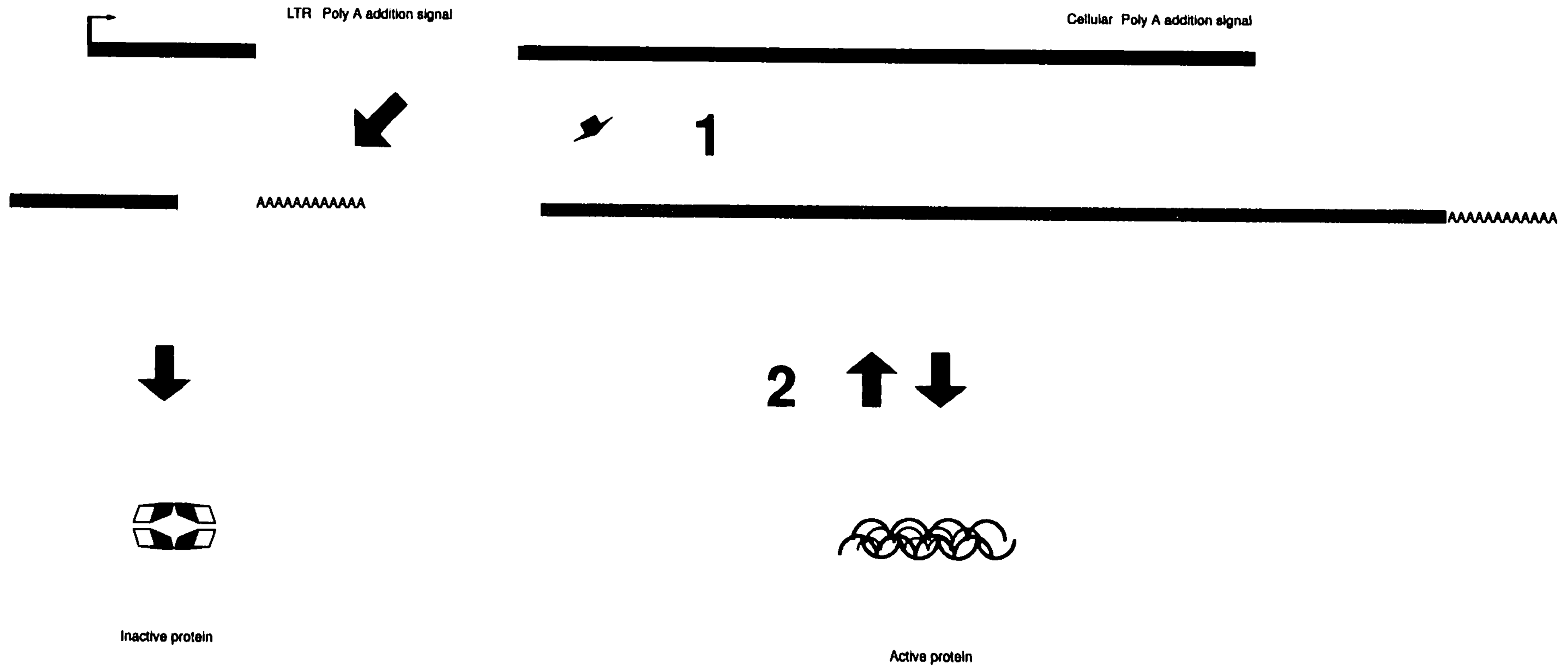
1996) allowed the mapping of the above RACE product. Examination of the genomic structure surrounding the *E14* gene places the RACE product in exon 3. The intron/exon borders of *E14* (*Npat*) was determined and it was found that exon 3 lies a minimum of 17 kb downstream of the first exon. This large distance from the promoter may account for the lack of differences observed in the methylation studies.

Mpv 20 heterozygotes did not display a 2 fold decrease in *Npat* expression for a simple dosage effect phenomena in which the mutated allele does not contribute to the amount of message present in the cell, instead an average of 1.6 fold decrease in *Npat* message was observed in the tissues examine. In the Mov 34 line, Soriano et al. (1987) observed a similar decrease in mRNA in heterozygote embryos. Further proof that a simple dosage effect is not occurring can be seen in the A6 line (Harbers et al. 1996). The homozygote animals die of an apparent placental defect, surprisingly, homozygote animals show a 70±8% decrease in message whereas the heterozygote animals display a 30±5% decrease in message. This points to the possibility that in this mutation the proviral insertion does not completely shut down expression but that some residual activity is present. This could imply that all three genes in the heterozygote animals (i.e. Mov 34, A6 and Mpv 20) are being regulated at some level in the cell in order to maintain a threshold level of protein which is essential to survival (Figure 26). This regulation may be either in an increased stability of the mRNA by factors binding to the long 3' UTR present in the *Npat* cDNA, or the establishment of a positive feedback loop with the protein products maintaining their critical level of expression. If this level must be maintained for the animal's viability, examination of the protein level present in heterozygote as opposed to wild type animals

should demonstrate similar levels of protein. Once a suitable antibody to the *Npat* gene product is available the above possibility can be examined by Western blot analysis of proteins obtained from whole cell lysates of different tissues. Alternatively in a certain subset of cells, the provirus may be efficiently spliced out of the pre-mRNA generating a partial compensation for the lost allele. A final possibility is that only a subset of cells require the *Npat* gene product for their normal development, and only in these cells is the expression increased by some mechanism from the wild type allele.

We favour a model in which the inactivation of the gene arises as a consequence of a truncated message being generated. This message may still initiate from the *Atm/Npat* promoter, the *Npat* message would be processed generating a message which lacks most of the coding region of the protein. This truncated protein would therefore be inactive and unable to perform its normal role resulting in a functional null mutation.

Figure 26. Proposed mechanisms for the inactivation of the *Npat* gene.
Initially a pre mRNA transcript which is initiated at the bent arrow is generated (shown in black) which includes the complete proviral genome (depicted as a white box). In the majority of instances, the slicing machinery will process the message and use the cryptic polyadenylation signal present in the proviral LTR resulting in the generation of a truncated message and inactive protein (depicted on the left). In some instances, the proviral sequence is efficiently spliced out of the pre mRNA generating a full length message which in turn produces an active protein (1). The active protein generated in a second scenario may in fact upregulate its own expression to maintain a threshold level in the cell (2).


LTR Poly A addition signal
Cellular Poly A addition signal
1
AAAAAAAAAAAA
AAAAAAAAAAAA
2
Inactive protein
Active protein

(II)

In a survey of 699 vertebrate sequences Kozak (1987) identified the presence of an upstream AUG codon in fewer than 10 % of mRNAs. Exceptions to the above were seen in 2/3 of proto-oncogene transcripts where these sequences are believed to play a role in regulating the translation of the gene product.

Previously the *Unp* cDNA was believed to contain twelve upstream open reading frames prior to the authentic AUG. These AUGs were speculated to control the translation of the *Unp* message in a similar fashion to that of TGF-β3 (Arrick et al. 1991). Sequencing of the *Unp* gene identified a number of miscalled bases in the published cDNA sequence (Gupta et al. 1993). These changes result in an open reading frame which codes an additional 169 amino acids and eliminates all but one of the upstream AUGs leaving a 5' untranslated region of 51 nucleotides (Figure 22).

The significance of the complete cysteine domain being present in a single exon is unclear. We can speculate that this region may have been present in an ancestral cysteine protease gene which may have undergone a duplication event. This would lead to the generation of new enzymes which were capable of the same catalytic function, but which in turn possess different target specificity due to the presence of the new additional sequences being present in the protein. The only other deubiquitinating enzyme in which the genomic structure is known is *Drosophila fat facets*. Similar to *Unp*, the entire cysteine domain is present in a single exon, whereas the his domain is also split between two exons. This would imply that the ancestral gene must have existed prior to the divergence of insects

and vertebrates. It will be interesting to compare the genomic organization of other family as they become available.

The presence of a promoter which is functional in both orientations may point to the existance of other genes being present upstream of the sequenced gene raising the possibility that the *Unp* gene is in an area of the mouse genome which is highly active. This can be examined using genomic fragments upstream of the sequenced gene and probing Northern blots to examine the possible presence of additional transcripts. Other examples of bidirectional promoters include the *NPAT/ATM* promoter previously mentioned (Imai et. al. 1996). This promoter can drive the simultaneous expression of two reporter constructs when transfected into cells (Imai et. al. unpublished). The H2A and H2B genes of Xenopus xlh3 histone gene cluster are transcribed in opposite directions from initiation points which lie approximately 235 bp apart (el-Hodiri and Perry 1995). The dihydrofolate reductase (*Dhrf*) and *Rep3* genes are also transcribed in opposite directions from a bidirectional promoter region (Wells et al. 1996). Upon closer examination of the above two clusters, two independent promoters are present which share common elements. This arrangement may be useful in regulating the expression of both genes in a cell cycle dependent manner. Whereas the above examples point to the presence of two closely juxtaposed promoters the *Surf 1* and *Surf 2* genes have been shown to be a true bidirectional promoter (Lennard and Fried 1991). This was demonstrated by generating mutations in each of the three elements that were protected in DNase I digestion assays individually. Each mutation individually decreased the activity of both promoters at the same time. Whether *Unp* truly possess a bidirectional promoter or whether there exists two promoters which are in close proximity

to each other and share common functional elements needs to be determined. It cannot be ruled out that the CAT activity observed in both orientations may result as an artifact of the experimental system since the promoter contains no TATA or CAAT boxes and is highly G+C rich. The high G+C rich content of the region upstream of the start of transcription may present to the cell, in the opposite orientation, a string of nucleotides in a similar context as that normally seen. Thus the position of other elements normally found adjacent to the region would in fact be responsible for the correct orientation of the activation of the cellular gene.

The low promoter activity observed for pDG83 possibly arose as a consequence of coding sequences for the *Unp* gene being present in the initial cloned genomic fragment utilized. Examination of the sequence (data not shown) shows that a product would be translated prior to the start of the CAT gene. This product would subsequently terminate by the presence of three in frame stop codons the first of which lies 47 bp upstream of the start of the CAT gene. The residual activity observed possibly arose through a mechanism of termination-reinitiation which has been previously described by Peabody and Berg (1986). In their assay, a bicistronic message was driven off the same promoter. The termination position of the first reading frame relative to the position of the second start of initiation was the critical factor in observing expression of the second reporter. When the distances were relatively small, efficient reinitiation occurred but when the distance between the two transcriptional units reached 50 nucleotides, the level of the second reporter was decreased by two fold.

CAT assays also point to the presence of a repressor element(s) in the region of DNA that was deleted in the pDG89 construct. Smaller deletions must be constructed in

order to delineate the area responsible. Once this area has been identified. the DNA can then be coupled to separose beads and used to isolate the DNA binding protein by passing cell extracts over a column. The eluted protein(s) are then isolated and subjected to partial amino acid sequencing. The sequence obtained can then be used in a database search to identify its identity.

It was previously demonstrated that the expression of the human homologue of *Unp* (*Unph*), was consistently elevated in lung small cell carcinomas and adenocarcinomas (Gray et al. 1995). The murine *Unp* gene is located on chromosome 9 in a region which is syntenic to human 3p21. It has also been shown that this region in humans is frequently rearranged or deleted in human tumours (Kok et. al. 1987. Naylor et al. 1987. Yokota et al. 1987. Daly et al.1993. Yamakawa et al. 1993. Hibi et al. 1992). We can speculate that the over expression in the human tumours results from the loss of repression normally present as a result of the putative element. This loss may be brought about by deletion of the genomic region containing the element. point mutations in the promoter in the tumour cells. or alternatively. deletion of a region on 3p21 which may code a repressor element that binds directly or indirectly to the *Unph* promoter. The latter may be more probable as a result of the identification of three new genes at 3p21 that encode zinc-finger proteins which may act as transcription factors (Calabro et al. 1995).

CONCLUSION

In conclusion it has been demonstrated that the Mpv 20 mutation arose as a consequence of a provirus integrating into an intron of the murine *Npat* gene. This integration event does not cause any observable changes in the methylation status of the

gene, but rather results in the generation of a novel message which we believe contains both *Npat* and proviral sequences.

The intron/exons of the *Unp* gene have been mapped using a combination of Southern blotting and genomic sequencing. The gene was found to contain 22 exons which are spread out over 47.3 kb of DNA. CAT assays using a 1.1 kb fragment positioned at the most 5' end of sequence displayed promoter activity. This activity was augmented when an internal region of the sequence was deleted pointing to the possible presence of a repressor element.

References

Applied Biosystems (1987). Evaluating and isolating synthetic oligonucleotides. User Bulletin 13. Perkin-Elmer Corporation. Norwalk CT. USA.

Arrick, B.A., A.L. Lee, R.L. Grendell, and R. Derynck (1991). Inhibition of translation of transforming growth factor-β3 mRNA by its 5' untranslated region. *Mol. Cell. Biol.* **11:** 4306-4313.

Baker, R. J.W. Tobias, and A. Varshavsky (1992). Ubiquitin-specific proteases of *Saccharomyces cerevisiae* cloning of *UBP2* and *UBP3*, and functional analysis of the *UBP* gene family. *J. of Biol. Chem.* **267:** 23364-23375.

Barker, D.A., H. Wu, S. Hartung, M. Breindl, and R. Jaenisch (1991). Retrovirus-induced insertional mutagenesis: mechanism of collagen mutation in Mov 13 mice. *Mol. Cell. Biol.* **11:** 5154-5163.

Barlow, C., S. Hirotsune, R. Paylor, M. Liyanage, M. Eckhaus, F. Collins, Y. Shiloh, J.N. Crawley, T. Ried, D. Tagle, and A. Wynshaw-Boris (1996). *Atm*-deficient mice: a paradigm of Ataxia-telangiectasia. *Cell* **86:** 159-171.

Bartholomew, C. and J.N. Ihle (1991). Retroviral insertions 90 kilobases proximal to the *Evi-1* myeloid transforming gene activate transcription from the normal promoter. *Mol. Cell. Biol.* **11:** 1820-1828.

Bartholomew, C., K. Morishita, D. Askew, A. Buchberg, N.A. Jenkins, N.G. Copeland, and J.N. Ihle (1989). Retroviral insertions in the Cb-1/fim-3 common site of integration activate expression of the *Evi-1* gene. *Oncogene* **4:** 529-534.

Beck E, Ludwig G, Auerswald EA, Reiss B, Schaller H (1982). Nucleotide sequence and exact localization of the neomycin phosphotransferase gene from transposon Tn5. *Gene* **19**:327-336

Bird, A.P. (1986). CpG-rich islands and the function of DNA methylation. *Nature* **321:** 209-213.

Bird, A.P. (1987). CpG island as gene markers in the vertebrate nucleus. *Trends Genet.* **3:**

342-347.

Breindl,M., K. Harbers, and R. Jaenisch (1984). Retrovirus-induced lethal mutation in collagen I gene of mice is associated with an altered chromatin structure. *Cell* **38:** 9-16.

Brown,A., N.G. Copeland, D.J. Gilbert, N.A. Jenkins, J. Rossant, and R. Kothary (1994). The genomic structure of an insertional mutation in the Dystonia Musculorium locus. *Genomics* **20:** 371-376.

Brown,A., G. Bernier, M. Mathieu, J. Rossant, and R. Kothary (1995a). The mouse *dystonia musculorum* gene is a neural isoform of bullous pemphigoid antigen 1. *Nat. Genet.* **10:** 301-306.

Brown,A., G. Dalpe, M. Mathieu, and R. Kothary (1995b). Cloning and characterization of the neural isoforms of human dystonin. *Genomics* **29:** 777-780.

Buckler, A.J., D.D. Chang, S.L. Graw, J.D. Brook, D.A. Harber, P.A. Sharp, and D.E. Housman (1991). Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. *P.N.A.S. USA* **88:**4005-4009.

Byrd, P.J., P.R. Cooper, T. Stankovic, H.S. Kullar, G.D.J. Watts, P.J. Robinson and M.R. Taylor (1996). A gene transcribed from the bidirectional ATM promoter coding for a serine rich protein: amino acid sequence, structure and expression studies. *Hum. Mol. Gen.* **5:** 1785-1791.

Calabro, V., G. Pengue, P.C. Bartoli, A. Pagliuca, T. Featherstone, and L. Lania (1995). Positional cloning of cDNAs from the human chromosome 3p21-22 region identifies a clustered organization of zinc-finger genes. *Hum. Genet* **95:** 18-21.

Chatot, C.L., C.A. Ziomek, B.D. Bavister, J.L. Lewis, and I. Torres (1989). An improved culture medium supports development of random-bred 1-cell mouse embryos *in vitro*. *J. Reprod. Fert.* **86:** 679-688.

Cheng, S.S., and F. Costantini (1993). Morula decompaction (mdn), a preimplantation recessive lethal defect in a transgenic mouse line. *Dev. Biol.* **156:** 265-276.

Conlon, F.L., K.S. Barth, and E.J. Robertson (1991). A novel retrovirally induced embryonic lethal mutation in the mouse: assessment of the developmental fate of embryonic stem cells homozygous for the 413.d proviral integration. *Development* **111:** 969-981.

Connor. F.. P.D. Cary, C.M. Read. N.S. Preston, P.C. Driscoll. P. Denny. C. Crane-Robinson. and A. Ashworth (1994). DNA binding and bending properties of the post-meiotically expressed Sry-related protein Sox5. *Nuc. Acids Res.* **22:** 3339-3346.

Connor, F.. E. Wright. P. Denny. P. Koopman. and A. Ashworth (1995). The Sry-related HMG box-containing gene *Sox6* is expressed in the adult testis and developing nervous system of the mouse. *Nuc. Acids Res.* **23:** 3365-3372.

Covarrubias, L.. Y. Nishida, and B. Mintz (1986). Early postimplantation embryo lethality due to DNA rearrangements in a transgenic mouse strain. *P.N.A.S. USA* **83:** 6020-6024.

Covarrubias. L., Y. Nishida, M. Terao. P. D'Eustachio, and B. Mintz (1987). Cellular DNA rearrangements and early developmental arrest caused by DNA insertion in transgenic mouse embryos. *Mol. Cell. Biol.* **7:** 2242-2247.

Daly. A.C.. R.-H. Xiang, D. Buchhagen, C.H. Hensel. D.K. Garcia. A.M. Killary. J.D. Minna. and S.L. Naylor (1993). A homozygous deletion on chromosome 3 in a small cell lung cancer cell line correlates with a region of tumor suppressor activity. *Oncogene* **8:** 1721-1729.

Denny, P.. S. Swift. N. Brand. N. Dabhade. P. Barton, and A. Ashworth (1992a). A conserved family of genes related to the testis determining gene. Sry. *Nuc. Acids Res.* **20:** 2887.

Denny, P.. S. Swift. F. Connor. and A. Ashworth (1992b). An *Sry*-related gene expressed during spermatogenesis in the mouse encodes a sequence-specific DNA-binding protein. *EMBO* **11:** 3705-3712.

Di Fruscio M.. H. Weiher, B.C. Vanderhyden. T. Imai. T. Shiomi . T.A. Hori. R. Jaenisch. and D.A. Gray (1997). Proviral inactivation of the *Npat* gene of Mpv 20 mice results in early embryonic arrest. *Mol Cell Biol* **17**:4080-4086.

Dush. M.K.. J.M. Sikela. S.A. Khan. J.A. Tischfield, and P. J. Stambrook (1985). Nucleotide sequence and organization of the mouse adenine phosphoribosyltransferase gene: Presence of a coding region common to animal and bacterial phosphoribosyltransferases that has a variable intron/exon arrangement. *P.N.A.S. USA* **82:** 2731-2735.

el-Hodiri, H.M, and M. Perry (1995). Interaction of the CCAAT displacement protein with shared regulatory elements required for transcription of paired histone genes. *Mol Cell Biol* **15**:3587-3596.

Finley, D., B. Bartel, and A. Varshavsky (1989). The tails of ubiquitin precursors are ribosomal proteins whose fusion to ubiquitin facilitates ribosome biogenesis. *Nature* **338:** 394-401.

Funabiki, H., H. Yamano, K. Kumada, K. Nagao, T. Hunt, and M. Yanagida (1996). Cut2 proteolysis required for sister-chromatid separation in fission yeast. *Nature* **381:** 438-441.

Gray, D.A., J. Inazawa, K. Gupta, A. Wong, R. Ueda, and T. Takahashi (1995). Elevated expression of *Unph*, a proto-oncogene at 3p21.3, in human lung tumors. *Oncogene* **10:** 2179-2183.

Gridley, T., D.A. Gray, T. Orr-Weaver, P. Soriano, D.E. Barton, U. Francke, and R. Jaenisch (1990). Molecular analysis of the Mov 34 mutation: transcript disrupted by proviral integration in mice is conserved in Drosophila. *Development* **109:** 235-242.

Gridley, T., P. Soriano, and R. Jaenisch (1987). Insertional mutagenesis in mice. *TIG* **3:** 162-166.

Gubbay, J., J. Collignon, P. Koopman, B. Capel, A. Economou, A. Münsterberg, N. Vivian, P. Goodfellow, and R. Lovell-Badge (1990). A gene mapping to the sex-determining region of the Y chromosome is a member of a novel family of embryonically expressed genes. *Nature* **346:** 245-250.

Gubbay, J., N. Vivian, A. Economou, D. Jackson, P. Goodfellow, and R. Lovell-Badge (1992). Inverted repeat structure of the *Sry* locus in mice. *P.N.A.S. USA.* **89:** 7953-7957.

Gupta, K., N.G. Copeland, D.J. Gilbert, N.A. Jenkins, and D.A. Gray (1993). *Unp*, a mouse gene related to the *tre* oncogene. *Oncogene* **8:** 2307-2310.

Gupta, K., M. Chevrette, and D.A. Gray (1994). The *Unp* proto-oncogene encodes a nuclear protein. *Oncogene* **9:** 1729-1731.

Hanahan, D. (1983). Studies on Transformation of *Escherichia coli* with plasmids. *Mol. Biol.* **166:** 557-580.

Harbers, K., M. Kuehn, H. Delius, and R. Jaenisch (1984). Insertion of retrovirus into the first intron of α1 (I) collagen gene leads to embryonic lethal mutation in mice. *P.N.A.S. USA* **81:** 1504-1508.

Harbers, K., Muller, U., Grams, A., Li, E., Jaenisch, R., and Franz, T. (1996). Provirus

integration into a gene encoding a ubiquitin-conjugating enzyme results in a placental defect and embryonic lethality. *P.N.A.S. USA* **93**:12412-12417.

Harden. D.G. (1994). The nature of ataxia-telangiectasia: problems and perspectives. *Int. J. Radiat. Biol.* **66:** S13-S19.

Hartung. S., R. Jaenisch, and M. Breindl (1986). Retrovirus insertion inactivates mouse α1(I) collagen gene by blocking initiation of transcription. *Nature* **320:** 365-367.

Heller. D.L.. K.M. Gianola, and L.A. Leinwand (1988). A highly conserved mouse gene with a propensity to form pseudogenes in mammals. *Mol. Cell. Biol.* **8:** 2797-2803.

Hershko. A. (1988).Ubiquitin-mediated protein degradation. *J. of Biol. Chem.* **263:** 15237-15240.

Hibi. K.. T. Takahashi. K. Yamakawa. R. Ueda. Y. Sekido. Y. Ariyoshi. M. Suyama. H. Takagi. Y. Nakamura. and T. Takahashi (1992). Three distinct regions involved in 3p deletion in human lung cancer. *Oncogene* **7:** 445-449.

Holmes. D.S. and Quigley M. (1981). A rapid boiling method for the preparation of bacterial plasmids. *Anal. Biochem.* **114:** 193-197.

Huang, Y.. R.T. Baker. and J.A. Fischer-Vize (1995). Control of cell fate by a deubiquitinating enzyme encoded by the *fat facets* gene. *Science* **270:** 1828-1831.

Huebner. K., L.A. Cannizzaro. T. Nakamura, J. Hillova, R. Mariage-Samson. F. Hecht. M. Hill. and C.M. Croce (1988). A rearranged transforming gene. tre. is made up of human sequences derived from chromosome regions 5q. 17q. and 18q. *Oncogene* **3:** 449-455.

Imai, T., M. Yamauchi. N. Seki. T. Sugawara. T. Saito. Y. Matsuda. H. Ito, T. Nagase. N. Nomura, and T. Hori (1996). Identification and characterization of a new gene physically linked to the ATM gene. *Genome Research* **6:** 439-447.

Jaenisch, R. (1976). Germ line integration and Mendelian transmission of the exogenous Moloney leukemia virus. *P.N.A.S. USA* **73:** 1260-1264.

Jaenisch, R. (1988). Transgenic animals. *Science* **240:** 1468-1473.

Jaenisch. R., D. Jähner, P. Nobis, I. Simon, J. Löhler, K. Harbers, and D. Grotkopp (1981). Chromosomal position and activation of retroviral genomes inserted into the germ line of

mice. *Cell* **24:** 519-529.

Jaenisch, R., K. Harbers, A. Schnieke, J. Löhler, I. Chumakov, D. Jähner, D. Grotkopp, and E. Hoffmann (1983). Germline integration of Moloney murine leukemia virus at Mov 13 locus leads to recessive lethal mutation and early embryonic death. *Cell* **32:** 209-216.

Jähner, D. and R. Jaenisch (1980). Integration of Moloney leukemia virus into the germ line of mice: correlation between genotype and virus activation. *Nature* **287:** 456-458.

Jähner, D. and R. Jaenisch (1985). Retrovirus induced de novo methylation of flanking host sequences correlates with gene inactivity. *Nature* **315:** 594-597.

Jentsch, J. (1992). The ubiquitin-conjugation system. *Annu. Rev. Genet.* **26:** 179-207.

Kaldi, K., P. Diestelkotter, G. Stenbeck, S. Auerbach, U. Jakle, H.J. Magert, F.T. Wieland, and W.W. Just (1993). Membrane topology of the 22 kDa integral peroxisomal membrane protein. *FEBS L.*, **315:** 217-222.

Kalpana, G.V., S. Marmon, W. Wang, G.R. Crabtree, and S.P. Goff (1994). Binding and stimulation of HIV-1 integrase by a human homolog of yeast transcription factor SNF5. *Science* **266:** 2002-2006.

Karls, U., U. Muller, D.J. Gilbert, N.G. Copeland, N.A. Jenkins, and K. Harbers (1992). Structure, expression, and chromosome location of the gene for the β subunit of brain-specific Ca^{+2}/calmodulin-dependent protein kinase II identified by transgene integration in an embryonic lethal mouse mutant. *Mol. Cell Biol.* **12:** 3644-3652.

Kim, J.W., E.I. Closs, L.M. Albritton, and J.M. Cunningham (1991). Transport of cationic amino acids by the mouse ecotropic retrovirus receptor. *Nature* **352:** 725-728.

Kok, K., J. Osinga, B. Carritt, M.B. Davis, A.H. van der Hout, A.Y. van der Veen, R.M. Landsvater, L.F.M.H. de Leij, H.H. Berendsen, P.E. Postmus, S. Poppema, and C.H.C.M. Buys (1987). Deletion of a DNA sequence at the chromosomal region 3p21 in all major types of lung cancer. *Nature* **330:** 578-581.

Kothary, R., S. Clapoff, A. Brown, R. Campbell, A. Peterson, and J. Rossant (1988). A transgene containing lacZ inserted into the dystonia locus is expressed in neural tube. *Nature* **335:** 435-437.

Kothary, R., S. Clapoff, S. Darling, M. D. Perry, L. A. Moran, and J. Rossant (1989).

Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice. *Development* **105:** 707-714.

Kozak, M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messanger RNAs. *Nuc. Acid Res.* **15:** 8125-8148.

Kratochwil, K., K. von der Mark, E. J. Kollar, R. Jaenisch, K. Mooslehner, M. Schwarz, K. Haase, I. Gmachl, and K. Harbers (1989). Retrovirus-induced insertional mutation in Mov 13 mice affects collagen I expression in a tissue-specific manner. *Cell* **57:** 807-816.

Lennard, A.C, and M. Fried (1991). The bidirectional promoter of the divergently transcribed mouse Surf-1 and Surf-2 genes. *Mol. Cell. Biol.* **11:** 1281-1294.

Lewis, J.D., R.R. Meehan, W.J. Henzel, I. Maurer-Fogy, P. Jeppesen, F. Klein, and A. Bird (1992). Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell* **69:** 905-914.

Lewis, S.E. (1978). Developmental analysis of lethal effects of homozygosity for the c25H deletion in the mouse. *Dev. Biol.* **65:** 553-557.

Martin, G.R. (1980). Teratocarcinomas and mammalian embryogenesis. *Science* **209**, 768-776.

Matsuda, Y., T. Imai, T. Shiomi, T. Saito, M. Yamauchi, T. Fukao, Y. Akao, N. Seki, H. Ito, and T. Hori (1996). Comparative genome mapping of the ataxia telangiectasia region in mouse, rat, and syrian hamster. *Genomics* **34**:347-352

Melton, D.A., D.S. Konecki, J. Brennand, and C.T. Caskey (1986). Structure, expression, and mutation of the hypoxanthine phosphoribosyltransterase gene. *P.N.A.S. USA* **81:** 2147-2151.

Mooslehner, K., U. Müller, U. Karls, L. Hamann, and K. Harbers (1991). Structure and expression of a gene encoding a putative GTP-binding protein identified by provirus integration in a transgenic mouse strain. *Mol. Cell. Biol.* **11:** 886-893.

Morishita, K., D.S. Parker, M.L. Mucenski, N.A. Jenkins, N.G. Copeland, and J.N. Ihle (1988). Retroviral activation of a novel gene encoding a zinc finger protein in IL-3-dependent myeloid leukemia cell lines. *Cell* **54:** 831-840.

Morishita, K., E. Parganas, D. Parham, T. Matsugi, and J.N. Ihle (1990). The Evi-1 zinc

finger myeloid transforming gene is normally expressed in the kidney and in developing oocytes. *Oncogene* **5:** 1417-1423.

Mount, S.M. (1982). A catalogue of splice junction sequences. *Nuc. Acids Res.* **10:** 459-472.

Nakamura T, Hillova J, Mariage-Samson R, Onno M, Huebner K, Cannizzaro LA, Boghosian-Sell L, Croce CM, and Hill M. (1992). A novel transcriptional unit of the tre oncogene widely expressed in human cancer cells. *Oncogene* **7**:733-741

Naylor, S.L., B.E. Johnson, J.D. Minna, and A.Y. Sakaguchi (1987). Loss of heterozygosity of chromosome 3p markers in small-cell lung cancer. *Nature* **329:** 451-454.

Panganiban, A.T. (1985). Retroviral DNA integration. *Cell* **42:** 5-6.

Pagano, M., S.W. Tam, A.M. Theodoras, P. Beer-Romero, G. Del Sal, V. Chau, P. R. Yew, G.F. Draetta, and M. Rolfe (1995). Role of the ubiquitin-proteasome pathway in regulating abdundance of the cyclin-dependent kinase inhibitor p27. *Science* **269:** 682-685.

Palombella, V., O.J. Rando, A.L. Goldberg, and T. Maniatis (1994). The ubiquitin-proteasome pathway is required for processing the NF-κB1 precursor protein and the activation of NF-κB. *Cell* **78:** 773-785.

Papa, F.R., and M. Hochstrasser (1993). The yeast DOA4 gene encodes a deubiquitinating enzyme related to a product of the human tre-2 oncogene. *Nature* **366:** 313-319.

Peabody, D.S., and P. Berg (1986). Termination-reinitiation occurs in the translation of mammalian cell mRNAs. *Mol. Cell. Biol.* **6:** 2695-2703.

Perkin-Elemer (1995). DNA Sequencing. Chemistry Guide. Perkin-Elmer Corporation. Norwalk CT. USA.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). MatInd and MatInspector- New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nuc. Acids Res.* **23:** 4878-4884.

Rechsteiner, M. L. Hoffman, and W. Dubiel (1993). The multicatalytic and 26 S proteases. *J. of Biol. Chem.* **268:** 6065-6068.

Reith, A.D., and A. Bernstein (1991). Molecular basis of mouse developmental mutants. *Genes and Devel.* **5:** 1115-1123.

Reynolds, G.A., S.K. Basu, T.F. Osborne, D.J. Chin, G. Gil, M.S. Brown, J.L. Goldstein, and K.L. Luskey (1984). HMG CoA Reductase: A negatively regulated gene with unusual promoter and 5' untranslated regions. *Cell* **38:** 275-285.

Rohdewohld, H., H. Weiher, W. Reik, R. Jaenisch, and M. Breindl (1987). Retrovirus integration and chromatin structure: Moloney murine leukemia proviral integration sites map near Dnase I-hypersensitive sites. *J. Virol.* **61:** 336-343.

Sambrook, J., E.F. Fritsch, and T. Maniatis (1989). Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Lab. Press, Plainview, NY. 2nd Edition.

Sanger, F., S. Nicklen, and A.R. Coulsen (1977). DNA sequencing with chain-terminating inhibitors. *P.N.A.S. USA* **74:**5463-5467.

Schnieke, A., K. Harbers, and R. Jaenisch (1983). Embryonic lethal mutation in mice induced by retrovirus insertion into the α(I) collagen gene. *Nature* **304:** 315-320.

Seliger, B., R. Kollek, C. Stocking, T.Franz, and W. Ostertag (1986). Viral transfer, transcription, and rescue of a selectable myeloproliferative sarcoma virus in embryonal cell lines: expression of the *mos* oncogene. *Mol. Cell. Biol.* **6:** 286-293.

Shiloh, Y. (1995). Ataxia-telangiectasia: Closer to unraveling the mystery. *Eur. J. Hum. Genet.* **3:** 116-138.

Soriano, P., T. Gridley, and R. Jaenisch (1987). Retroviruses and insertional mutagenesis: proviral integration at the Mov 34 locus leads to early embryonic death. *Genes & Dev.* **1:** 366-375.

Stewart CL, Stuhlmann H, Jahner D, and Jaenisch R. (1982). De novo methylation, expression, and infectivity of retroviral genomes introduced into embryonal carcinoma cells. *P.N.A.S. USA* **79:**4098-4102.

Suzuki, K., J. Olvera, and I.G. Wool (1991). Primary structure of rat ribosomal protein S2. *J. of Biol. Chem.* **266:** 20007-20010.

Tobias, J.W., and A. Varshavsky (1991). Cloning and functional analysis of the Ubiquitin-specific protease gene UBP1 of Saccharomyces cerevisiae. *J. of Biol. Chem.* **266:** 12021-12028.

Tsurumi, C., G.N. DeMartino, C.A. Slaughter, and K. Tanaka (1995). cDNA cloning of p40, a regulatory subunit of the human 26S proteasome, and a homolog of the *Mov-34* gene product. *Biochem. and Biophys. Res. Comm.* **210** (2): 600-608.

Turpen, T.H. and O.M. Griffith (1986). Rapid Isolation of RNA by a guanidinium thiocyanate/cesium chloride gradient method. *BioTechniques* **4**:11-15.

Vijaya, S., D.L. Steffen, and H.L. Robinson (1986). Acceptor sites for retroviral integrations map near Dnase I-hypersensitive sites in chromatin. *J. Virol.* **60:** 683-692.

Von Lindern, M., M. Fornerod, S.V. Baal, M. Jaegle, T.D. Wit, A. Buijs, and G. Grosveld (1992). The translocation (6;9) associated with a specific subtype of acute myeloid leukemia, results in the fusion of two genes, *dek* and *can*, and the expression of a chimeric, leukemia-specific dek-can mRNA. *Mol. Cell. Biol.* **12:** 1687-1697.

Wagner, T., J. Wirth, J. Meyer, B. Zabel, M. Held, J. Zimmer, J. Pasantes, F. D. Bricarelli, J. Keutel, E. Hustert, U. Wolf, N. Tommerup, W. Schempp, and G. Scherer (1994). Autosomal sex reversal and Campomelic Dysplasia are caused by mutations in and around the *Sry*-related gene *Sox 9*. *Cell* **79:** 1111-1120.

Wang, H., M.P. Kavanaugh, R.A. North, and D. Kabat (1991). Cell-surface receptor for ecotropic murine retroviruses is a basic amino-acid transporter. *Nature* **352:** 729-731.

Weiher, H., E. Barklis, W. Ostertag, and R. Jaenisch (1987). Two distinct sequence elements mediate retroviral gene expression in embryonal carcinoma cells. *J. Virol.* **61:** 2742-2746.

Weiher, H., T. Noda, D.A. Gray, A.H. Sharp, and R. Jaenisch (1990). Transgenic mouse model of kidney disease: insertional inactivation of ubiquitously expressed gene leads to nephrotic syndrome. *Cell* **62:** 425-434.

Wells J., P. Held, S. Illenye, and N.H. Heintz (1996). Protein-DNA interactions at the major and minor promoters of the divergently transcribed *dhfr* and *rep3* genes during the Chinese hamster ovary cell cycle. *Mol. Cell. Biol.* **16:**634-647.

Williams, T.J., and M. Fried (1986). The MES-1 murine enhancer element is closely associated with the heterogeneous 5' ends of two divergent transcription units. *Mol. Cell. Biol.* **6:** 4558-4569.

Williams, T.J., J. Yon, C. Huxley, and M. Fried (1988). The mouse surfeit locus contains a very tight cluster of four "housekeeping" genes that is conserved through evolution. *P.N.A.S.*

85: 3527-3530.

Xu, Y., T. Ashley, E.E. Brainerd, R.T. Bronson, M.S. Meyn, and D. Baltimore (1996). Targeted disruption of ATM leads to growth retardation, chromosomal fragmentation during meiosis, immune defects, and thymic lymphoma. *Genes and Devel.* **10:** 2411-2422.

Xu, Y., and D. Baltimore (1996). Dual roles of *ATM* in the cellular response to radiation and in cell growth control. *Genes and Devel.* **10:** 2401-2410.

Yamakawa, K., T. Takahashi, Y. Horio, Y. Murata, E. Takahashi, K. Hibi, R. Ueda, T. Takahashi, and Y. Nakamura (1993). Frequent homozygous deletions in lung cancer cell lines detected by a DNA marker located at 3p21.3-p33. *Oncogene* **8:** 327-330.

Yokota, J., M. Wada, Y. Shimosato, M. Terada, and T. Sugimura (1987). Loss of heterozygosity on chromosomes 3, 13, and 17 in small cell carcinoma and on chromosome 3 in adenocarcinoma of the lung. *P.N.A.S USA*. **84:** 9252-9256.

Zachariae, W., and K. Nasmyth (1996). TPR proteins required for anaphase progression mediate ubiquitination of mitotic b-type cyclins in yeast. *Mol. Biol. of the Cell* **7:** 791-801.

Zawacka, R.M., A. Reuter, E. Pfaff, J. Moll, K. Gorgas, M. Karasawa, and H. Weiher (1994). The glomerulosclerosis gene Mpv17 encodes a peroxisomal protein producing reactive oxygen species. *EMBO J* 13: 5129-5134.

Ziomek, C.A. and M.H. Johnson (1980). Cell surface interavtion induces polarization of mouse 8-cell blastomeres at compaction. *Cell* **21:** 935-942.

Appendix 1. Nucleotide sequence of the Unp gene.

GAATTCAGAGATCTGCCTACCTCTGCCTCCTGAGTGCTGGGTTAAAAGCCTGTGCCATCA 60
CCGCCCGGAAGAAAACGGAATCATTGTTTCAGGAAAAAATTAAGATGCCTGGACCTCTGT 120
CGGCTTCGGTGGTTGCTACTCTCTCCTTGACCTGGATGGCTGACAGGATACAAAGTTGTT 180
TAAGACAACTTTGTTTTTTGCGTCTATTTATGACTTTATTTTTGTTTTTTGCAGGTGTGG 240
TGATGTATGCTTTTAATCCCAGCACTCTGGAGAAAAGGGCAAATGGGCCTAAATTCGAGA 300
CTAGCCTAGATTACATAAACAGCTCCAGGACAGCAGCTAGAAGAAACCTGTCCCAAACAA 360
ACAACAAATTTCACTATTGTGTGCCTGTTTGGCGAACGCGTGCCACAGTGTCTGTGTCGA 420
GAACAAAGGACAACTTTAAAGAGTCAGTTTTCTTTTTCCACTTTTACACGGGTTCTGGGG 480
ATCAAACCCAGTTGTCACCCACGTAGCCATCGCGCCGGCCCTTGTTTGAAACAACTATTA 540
ATGGCTGGACCACACCGGACAGTCCAGATCTCAATGCGTCCGTGTGCACTTGGCTACAGG 600
GGCTGCCGAGGCAGTAAATTCAATCTCGCATTTAGGCAGTCTTGCTGAAAGCCGCATGGA 660
ACCGAGTGGAAACAAGAAAAGCAAGCGGCAGGCTCCTCCTGCACGGATTTAGCATCAGGC 720
TAACCTCACAGGCCAGGATGGAGCGAGTGACTGATGGGTCGAGGGCGGGCCAGCAGTCAG 780
ATCGCAGGAGCCGCTCAGACACGGGTGGATGTGGCCGCCGCCTCTGGGCGTCCAGCCTGC 840
TTCTGTCCGCTTCGGCTCCGGAGGGGGCGGGGGGGGTTAACAGCGCGTGGGGGCGGGGA 900
GAAGAGGCGAAAGTCGACGGCGGTGCCGGGCGCGCGCACGTGCGTGGTTACGTGCGTGCG 960
TGGGCGCGGCCCATAGCACGCCGTGCGCTAGGCGGATGTCCGGGCCGGCTGGGCCGGGGC 1020
GGCGGACGAGATGGCGGAAGGCCGGGGCAGCCGTGAGCGACCGGATGTGGAGACCCAGAA 1080
GACGGAGCTCGGAGCCTTGATGGGGACCACGCTCCAACGTGGGGCGCAGTGGTGAGCGTC 1140
CGCCCTGCCCTCCTGCGGCGCAGAGCGTCTAGCCGGCTCGCTGGCCCTCAGGCCGGACCT 1200
CTGTCCTCAGTTTTAGGTGGAGGCGGGCCGGGGGCGGTGCCTGGAGTCTGAAAGAGTGGC 1260
TTGACAGCGGCCGCTGATGAAAGAAACCCCGGGATCAGCTGGGGGACAGCGCTGACAGGC 1320
GTGGGGTTTGTCCCGGTCCATTCGGATTCACAGTAGTCTTGGACCGGCGGGGACAGGGGG 1380
TGTAGAATATAGAAAAGGGGGCGCGCTGGGGTCGGCGGGAGCGGGAAGGTGGGGGAAAGC 1440
ACTTAACCACCGAAGACTTGGGCGCTTGATGGTGACCTTGGTGGGGGAAGTTTCCACCCC 1500
TTACACACCTCCGATCCTTAACAATTAAAACCTCGGTAGCAATTTGTGGCTTAGGACGCT 1560
CCTGGCGTGGAGAGAGGATCCGAAGTGGATACTTTGTAACTTGTAAACTGTGATACTTTG 1620
TAACTTGTAAACCGTGGGACAGGGATTCAAAGCCAAAAACTAGAAGCCAGTTGTCAGATG 1680
GTCCGGGACAATTGTAGCAGTAGTTAGATGTAAGAGTGGGAGGGGAGGGGAAGGGAGGGG 1740
AGCTTGACTCCTGGAGCAAAATAGAAGATCGAAAGTTACAAGACTTTGCCAGGCACGCTG 1800
GCATATGCTTTATTGGTTTGCAGGAGCAGAGGCTGTAGGATTGCTGTAGGTTCAGGGCTG 1860
GTCTCATACAGTGAGCAGCACCCCCCCCCAAAAAAAAAGTTCCAGGTTTACTAGAGCAG 1920
ATTCAAATATGAGGCCTCTTAAGAGATTGTTAACTAGAAGGGATTGTACTCCCCAGTCGG 1980
GGAGAAATGCCTGGGTGGTGCTGCTGTGTAGGAGTCAAATAATAGGGTAGGTTTCTCATT 2040
GATTATGACCAATATGAGGACTGAAAAAGAGAACAGAACTTATGTGTTTGGGGTGTGGTC 2100
TAGGTGTAGTTTATGGTGTTTAAATTGAGCCAGTGAAAGGCCAGACAAATAGAGAGAGGT 2160
ACTTTAGAAATTCTGCAAAGGGATGACACAGAAAGAATATTCAGGTTGGTCTGGATGCTC 2220
ATGCCTGTAGTCTTAGTCCTGGAGGAGTATTGCTGCCAAATTTCAGGGCCAACCTCAGTG 2280
ACATAGTAAAAGCCTTGTCTCAAAAAAAAAAAAAAATGGTGGTGGGTAATGTGATTAATA 2340
ACTTAGTAACATAGTTTTACTTGTAGAAAGAGTCCATTTGAAATGATTGCTCATGGGCTG 2400
GTGAGACTGCTCAGCATCTGAGAACTTAGATTATTCTTGCAGTGCCTGAGTCTGAGGCCC 2460
AGCACTCGGGTTTGGTGCCTCCCTGCCATCTGTAACTCAAGCTCCAGCTCCAGGGATAGC 2520
AACTCCCCTGACCTCCGTGGGCACTGGGAGTCATGCTCAAACTCAGGCACATAACCAAAG 2580
CATAATTAAAAAGAAAAACTCAGCCCAGGAGCGGTGGCACACATCTTTAGTCCCAGCGCT 2640
TGGGAGGCAGAGGCAGGTGGATCTCAGAGTTTGAGGCCAGCTCAGTCTACAGACAGGACA 2700
GACAGAGAAACCATGTTTCGAAAAACTGAAATAAATTAAATCCTAACTAGGCATAGTGTC 2760
ACACAACTTTAATCCCAGCCTCTGGGAGGCATACACAGGTGGATCTCTCTAGTTCAAGGC 2820
TATCCTGGGCTACATACTGAGGGTGCACCTCCCCCCCCCCCCCCCCCCCGCCATCTCCAA 2880
AGTTAATTAGTTTTAAAATTCTTTCAATAATCGTTTATGGTTTTGGTTGTTTGTTTCTGA 2940
GACAAGGTCTCCCTGTGGCCCCTACTGGTTTGGAACTTGAGATGTAGCCAGGCTGGCCAC 3000
AGACTTGCAGACTCTTCTGCATCTGCCTCCTGTGTGCTGGGATTACAGATAAGTGATCGC 3060

AGTGTCTTGTGTGTGAGAGAGAAAGAGACAGGCAGAGAGAGAGGGGGGGAGGGAGAGGA 3120
ATATGAGAATGTATGGCACACGCTATGCATGCAGATGCCTATGCAGAATAGAAGAGGCTG 3180
TTGGATTTCCTGGAGCTAGAGTTGTGCATACCATACATGGGTGCTAGGAACCAAACTTGA 3240
GTCCTTTGGCAGAGCACAAGACTCTCTTAACCTCTGAATTATCTTTCCAGTCCCTTGTTC 3300
ATTTTTTTTCAGACGTGGTCATATTAGGTAACCTTGACTGGTCTGGAATTCCTATGTAGA 3360
CCAAGCTGGACTGGAACTCATAGACATCCAAGTGTCTATGCCTCCTGAGTGCTGGGATGT 3420
AAAGGCCAGCATGTGCTAGCATGTGTAGCTTTGTTTTGTCTTTTACAGATAGGGTCTCTT 3480
GTAGCCCAAGGGAGGGAACCTCAAGCTTACTGTGCAGGGTAGATTGACCTTTTTGCCTGG 3540
ACCTCACAGATGCTGGAATCATAGATGTGGGTCACCACAAACTTCCTCTACTTATTCTCT 3600
TAGCAGTCATTTTTCTTAGTGAATACCTGGCATAGATTAGGCCCTCTCTGCCATGTGGGG 3660
CTAACTTCTGTTCTTGATATTAGGAGTGTGATTTTAGTCGACAGTGGCGGTAATGCACAC 3720
TTCTAGTACCAGCACTCAGGAGGCAGTGGCAGGCAGATCTCTGCATTTGAGGTTAGCCTG 3780
GTCTACAGAGCAAGTTCTAGGACAGCCAGGGCTACACAGAGAAACCCTATCTTGAAAAGC 3840
TCCCCTCCCCCCAAACGTGATTTTGTGAGAATTGTATATTACCTTGGTGAGAAGACCATA 3900
AGGAGCTTAAGGTAGGTGTCTAAAGATGAAAATTAAAGGAGAGCTGCAAAACATAATGCT 3960
TTGGGTTTTTAAAACTGTGTTTCTTTATTCAGCTGTATTTTATATGTAGCTGCTATTTCT 4020
CCCTAGGTATCTTATTGACAGCCGGTGGTTCAAGCAGTGGAAGAAGTATGTGGGCTTTGA 4080
CAGCTGGGACATGTACAATGTCGGGGAGCATAACCTGTTTCCTGGACCTATTGACAACTC 4140
CGGACTCTTCTCAGGTGGAGTATGCTCAGTTTTCTACTGTAGTTCTCATGGGCCAGTCAG 4200
GGGCTGCGTGTATCTCTACCATGGTCCTCATGAGTGTGAAACTTTCATAATGTGAAGACA 4260
TAGACGTGTTCCTTTACATCTGTAGTAGACGTCTTTGCAGATTATTAGGGAAAAGTCAAG 4320
TAGTTATTTTGCCCTAAACAACTACTTCTAGCAGAAGCATGCCTTTTTTGTTACCGTTCT 4380
GTTGCTGTGAAGAGACACTATGAACATGGCAACTTATAAAGAGTCACTTAATTGGGGGCT 4440
GGCTTACAGTTTCAGCAAGTTAGTTCAGGATTATCGTGGTGGAAAGAGTTTGGTACCAGC 4500
AGGTGAGTGTGGCACTAAGAAGTAGCTGAGGGAGGCTGGTGAGCTGGCTCAGTGGTTAAG 4560
AGCACCGACTGCTCTTCCAAAGGTCCTGAATTCAAATTCCAGCAACCACATGGTGGCTCA 4620
CAACCATCCATAACGGAATCTGATGCTCTCTTCTGAAGTGTCTGAAGACACACTAGAGAC 4680
AGACAGAGCTAGGTGAGGGTTTATGAAACCTCAAGACCACCCTTCAGTGACACACCTCTT 4740
TCAGCAAGGCCACACCCCTCCTTATCCTTCCCAAACAGTACTACCATTGGGAACCAAAGA 4800
TTCAAACACGCGAGCTTAAATGGCCCATCCTCATTCGAACGGCCACATTTTATAGCATAC 4860
TAATGTTTTTTTCATATTTTTATTTTTACTGTATATGTGCACAATGTAGGATGTGTATGA 4920
TGTGTGTGCCAGTGAGTCACTGTGTATGTTAAGGCCAAAGTATAGCTTTCTGCAGTTGGT 4980
TCTCAGAAGACATCTTACCTGTCTACCTCGTAACGTTTTGTCTCTACTTTCTGTTTTTTA 5040
ATTGTGTGTATGGATGTGGATCTGAATGAGCTGTGCATGTGAAGTGCAGTTCCTGGGGAT 5100
CCAGAAGGGAATTTAGGCTCCCCTGGAGCAGTAGGTTGTGAGCTGCTCCGCAGAGGTGCC 5160
TAGACCCGCACTTGGTCCTCTGCAGGAGCAACGGTGCTCGTCCTTGCTGAACTGCCTTGC 5220
TGGCCCCTTAGTTTTCTTTTTCTTTTTTAGGTTTTTGGATTTGTATGTAAACAAGAAAGA 5280
AGAATAGTCATTTTAGATTGCGCCTATAGTCTGAAAACTGGGAATCATGCCCTCCCCTTT 5340
TCTTTCATTGCTGTGATCAAATTCAGTATATCCAGTAATTGGTCTGGGCATTGATGCTGT 5400
TTTTCTGTTTTGTTTTGGTTTTTGATTTTGTTTTCAAGAAGAGGGTCTCTCTGCATAGCC 5460
CTGAATGCCCTAGCACTAAGTATATAGACCAGGCTGCCCTTGAACACACAGAGATCTGCC 5520
TGCCTCTGCCTCCCCAGTGCTGGGATTAAAGGTGTGTGCCACTACACTTAGCTGTACCCA 5580
CAAAATGTTGATACCTGTGAAGTAGGACAAATTGCATGCATATTAAACGTCATCAGGCCT 5640
CTCTGTACCGACTGTTTATAGTAGGAATAGCGTAAGGAGAGGTAGAAGGTCAGTAGAGTT 5700
GTGGCTTCTGACCTCGGCGCTGTCCTGTTAAGTGTGTCAGCAGAAGTGATGGATGATGGA 5760
TCCTGTGACTGTGAGGTCAGCTTCCCCAAATGCTCACTCCTGTACCTGTGTCCAGAATCT 5820
GCCAGAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCGTGTATTCATACACTTAAT 5880
ATATTCCTGTGTGTGTGCGAGTGTGCTCATGAGTGTTCTCTGGTGTTTATGTGGTGTGCA 5940
TGTGGAGGTCAGAGTGCTGCTGATGGAAGGTGTTTCTTTCCTCTTACCCTGTGGGATCCA 6000
TCATCTAGGGACCAACTCATGTCCTCAGGCTTGGCAGTAAATGCCGCTACCTGTGGAGCC 6060
GTTTCTCCAGCTCCCTCCATTTCTGTTGGATTTTGCTCAGGCAGTGTCTTCCCCGTGTTG 6120
CCCTGGCTGGCCTTGCAGTTGAGTTAATCCTCCTGCAGTTGCCTCCTTTGTGTTGGAGTT 6180
TCAGGAGTGTGCCACTACAACTATACTCTTGCTTTATTTTACAGATAGAGTTTTGCTGTG 6240
TTCCTTCAAACTTGCTATGTAGCGTAAGCTGACATTGAAATCACTGCCTTCCTGTCTTAG 6300

CCTCCCAAGTTATGAAATTAAAGTCTTGTGCCATTGAGTCTGGCTTTCAGGTTTTTATTA 6360
ATCTTACATTTTTCTAGGTTAAAAATGTATCTGTTTGTGTGTGAGACAGGGTGAGTGCTT 6420
GTGTGCATGGCGTGGTCCAGGGTGTGGATTTAAGGACAACGTCTCTGTCTCTTCTGCAGT 6480
GTGGGTTCTGGGGATCACACTCATGTGGTTCTGCTTGGTAGCAGGTGTCATTACCTGCTG 6540
AGCTGCCTCACCAGCCCTTGTTTTTTTCAACACAGGATCTGAGTTAAGGATCTGGGGTTG 6600
GCCTCAATTTGTGTTGCTCTTTGTTTCAAGTGCTAGGCTCACAGGGACCTGCTGCCAGGC 6660
ATGGCTATTCCATGATGGGTTTAATTTTATTTTAGATATGGTCTCACTATGTAGACCGGG 6720
TTGGCCTTTAACTCACAGAGGTCCTCCTACTTCTGCCTCCTCAGTGCTGAGATTAAAGCC 6780
ATGCCCTATGCTCTACCTACTGGCCTGTTTATTTTTATTTTTACTTGTTTGGAGTTTTTA 6840
TTTGCTTCTTTTGTTTTGAGGTGTGATCTTGTATCATTCAGGCTGGCCTTAAACTACAAT 6900
GTAGCTGAGGTTGACCTTGAACTCATGTACTTCTTGAATTTACTTCCCAGATGCTCAGGA 6960
AGATAATTCTATTTGTTAACGTACTTATTTGAGGCAGGGTCTTGGTATACAACTTGGAT 7020
ACATAGCAAGTTTGTAGCTCAACCTGGTCCAGAACTACATGTATACCCCAGTTAAGTCTT 7080
GAATTTATAGTCTTTCTAACTATAGTGCTGGGATTGCAAATGTCCCCCATAATGTCTGGT 7140
TAAGAGGAACATTTTTTGTTTTCTTATTTGTTGTTTCATGGTGACAGTGGCAGTGTGTTG 7200
GTGGAGGGGAAGGGGTGGTGCTGGTGCTGGTGCTGGTGGTGTTGGTTGAGAGTTTCTCTC 7260
TACTGCCTGTGCTGGCCTGAGAATTGACCAGTCTATTTCAGCCTTCCAGAGCCTGTGGGG 7320
TTCTAGGTCTGAGCATCATGGTGTGGGAATGACTGCTTGACAGATGTATTCAGCAGGAAA 7380
CAGGCCTGTCACCTGCAGAGCAGCCCTTCTTATTTCCACCTGTCTACTGAATTGGACAGT 7440
TTTTTAGGTTTTGCTTTGAAGCCTGACGTAGGCTAGGCCGGGTACACGTGAAGTAGGTCA 7500
GACTGTAGGTCTACATCAAGCATCTCTGTACTGACTTCAGACCAGAGACAGTCAGAGACA 7560
GCGCTTCAGGAGAGTCACCTCCAACCTCCTTGTTCCCAGTCCTCAAACTCCTTCTGTAGC 7620
TGACACCTAGCAGCACCTGGTCCTCCTGTCCCGTCTGCTAGGCACAGGGTTAACAGATGT 7680
ATACTACCACCATGCCTGTCATCTTCTCCTTTTTTTTTCTTTTTTTTTTTTGGTTTTTCG 7740
AGACGGGGTTTCTCTGTGTAGCTCTGGCTGTCCTGGAACTCACTCTGTAGACCAGGCTGG 7800
CCGCCAACTCAGAAATCCGCCTGCCTCTGCCTCCCGAGTTCTGGGATTAAAGGCGTGCGC 7860
CACCATGCCCGGCTTTTTCTTCTTGAGACAAGATGTTACAATATAGACCTGGCTGGTCTG 7920
GAACTCGACAGGGATGCACCAGCCTCTACCTTCTGAGTGCAGGAATCAAGTGTGTGTGCT 7980
ATCACACCTGGTTTCTATTTAAAAAAAAAAAAAAAATCATCTGCCTTGGGTTACAGCTGAG 8040
TTGCACCTTCACGGGTACTTTCTTGTCTACACTGGTCTGAGATCAGTGACTTCTGTCATC 8100
CTCAGTGTGGGTGTTGATAAACAGGCTTAGCAGGGAGTACACTTTGTGGCTGACAGGTGA 8160
TGAATCTTGTTGGATTTTCCGCTGAGTGTTTCATATACCTTTGTGATGGGTGTAAAATGC 8220
ATGTTCAACATGTTACATATTTACTTTCCATTCCACACAGTGTCCATGAGAACTCTTTTT 8280
TTTTTTTAAAGGTTATTTGTTTTATGTATATGAGTACACTGAAGCTGTCTTCAGACACAC 8340
CAGAAGAGGGAATCAAATCCCATTACAGATGGTTGTGAGCCACCATGAGGTTGCTGGGAA 8400
TTGAACTGAGGGCTGTGGGCATGGTTGCTTGGGAAGAAAATACCCTACCCACCCCCACCC 8460
CTCACCCCCCTCCCAGCCCTCAGCACCTCATTCATTTGCCTTCTTAGTAAATGTTTCATG 8520
TGTAGATGGGGTCTGAGTTCAAGCACTATCAATAGGATTGCTATTATAAGTATTTATAAT 8580
TATTAGCCATTGTACTCTCTTCTTCAGGACAGCCAGGGCTACACAGAGAAACCCTGTCTT 8640
GAAAAACAAAGCCAAACAAAAACCTATTCTTATTGAACTCAAGTTCATGGATTTGGCAGC 8700
AAGTACCTTTACCTGCTGAGCCATCTCCCCGGCTCATTTTCCCTCCAGTCTCCTTCCTGT 8760
CATCTCTGCAGTGGAGGCAGGAAGATCAGAAATTAAAGTCATCTTTGGCTGTGTAGAATT 8820
CAAGGTCACGCTGTACTGCATGTGACCCTCAAAACAGACGACAAAAATGAGAAGGGGGCA 8880
AGTAGGGCAGGGGATTTGGGAGGGCTTGGGGAAAGGCAAAAAATATGGCTAAAATATAGC 8940
TCAATAAATTTTATTTTAGAATTACTTTTTACCCCTATATATCCTTGGCTGTCCTGGAGC 9000
TCCTACATAGACCAGGCTGACTCTAAATTCACAAAAATCCACCTACCACTGCCTCCCAAG 9060
TCCTCAGATTAAAGGGTGCACCACCATGTTCACCTTCAATACATTTGACTTCAGAAGAAG 9120
AAAATGAACAGAGAGGCTCCTTTCCTCTCCATCCTTCCTCCTTGTTCTTTTTGACATGGA 9180
GTCTAGCTGTGTTGCCCAAGCTGACCTTGAACTTCGGGGCTTAACCTATTCCTCCTCAGC 9240
TTCCTGCGTAGTGGAGACTATGAGGATGAACCACGGACTCCAGCTCACTTTTCTCTCATT 9300
ATAGAAGAGATACAAGTTCCTCATAGAACAAGACTCAGGATAAAGAGTGGGCAGGTCTGC 9360
TTTGTTTTGTAGCCAGTGACAGTGTTTTGGAATATTGTGCTTGTTCCTGTTGAATCTTTT 9420
CTTATGAAAAATACATAAGAATCTTGTTTCAAAAAACAAGCAACCAAACAAAAAGAAACT 9480
TCTGGGGAAAAGAAAAATGAGATCTGCCCTGATTTCCTTTTTCAGATCCTGAGAGTCAGA 9540

CCTTGAAGGAGCACTTAATCGATGAGCTGGACTATGTGCTGGTCCCAGCCGAAGCCTGGA 9600
ATAAATTGCTGAATTGGTATGGCTGTGTGGAGGGCCAGCAGCCTATTGTCAGAAAAGTGA 9660
GTTGACAAGTTGTTAGACCTGCTAGGTGTGTATGTGAGTTCTCCTTGGGTACACTGACTT 9720
CTCTTTATTTCCCGGGAGGATGAAAACGGTTACATGGTTGAAAATGAAGCAGGTGAGACT 9780
TTGTAATGTTTTTGGATTTTGGTTTGTTTTTGTTTTTGAAACAGGGTTTTTCTGAGTATC 9840
CCTTACTAGTAAGTAGTTCTGGAACTTTCTATGTAGCCCAAACCAGCCTCAAACTCAGAG 9900
ATCCCTAAGTACTGAGATTAAAGATGTACCAGGCTTTAGACCTTGTTTATGTTTCCTGAG 9960
ACAGGATCTCTGTGTATCCCGGGCTGGTCTTGTTCTGGTCATCTCCTGCCTCTTGTATGC 10020
TGGAATTTTGGCCTGAATGACTATTCTACTTCTCTCTGGGATATGAAGTATTTGTGAGTA 10080
ACTCAGGATGTTCCTAAGTGGCCTCCAAATACAAGAGCTGCCTGCTGGTTTGGAATGCCC 10140
ATAGTACTTGAACCCAGTCTCAGTTGGGCTGTGTGGACATAATCGAATGCATTGCACCTT 10200
GCACCGCTTCCTTTTCCCCAGGACTGGGAGTGAACCTAGGGCCGAATAGAATACATGACT 10260
CACTGAGTGTCCTAGCACTGAGCTATATCCAGCCCTCTTTTTTTTTTCTTTGAGATGGTG 10320
TTGTACTGAGTTAACCAGGCCAAATTTGAACTCACTTATAGCTCAAGTAGGCCTAGAACT 10380
TTGACCATTTTGAGCCTTTTCATTGAGGCTCCTGGATGTTGAGATAACTAGTGTGAGCCA 10440
CCATGCTTACCTAACAACTTCTCTTTTTCTTTTTCTTTTTCTTTTCTTTTCTTTTCTTTT 10500
CTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCT 10560
TTTTAGATTTATTTATTTATTAAAGTACACTGTTGTTGTCTTCAGGCACTCCAGAAGAGG 10620
GCGGCAGATCTCATTATGGATGGTTGTGAGCCACCATGTGGTTGCTGGGATCTGAATTCA 10680
GGACCTCTGGAAGAAGTCAGTGCTCTTAACCACTGAGCCATCTCTCCAGCCCTCTAACAA 10740
CTTTTCTAAAGCCATTTCCCCTGTGCTCATGAGTCTGATTGTACTGCTGAGTAGTAAAAA 10800
CGTAGGTGTTTATATGCATAGGAGAACTTCTCAAGTTCTCACACTGAGACACACAGCTTT 10860
GTGCCTTTGGTGGACAGAACACATGAGATCACTGCCTTGTCTCTCAGCTACTAGCTTTTC 10920
TTACTAGAAAGGAATCATAAAAGCCCATTCAGGACTGCTGATACATAGACAATACTACAT 10980
TTCCATGTTGCTTGGGGGGTATGCTCTCCCTCAGTATGACAGAGCTTCATCCTTGTCTGT 11040
CTGTCTGTCTGTCTGTCTGTCTCTCTCTCTCTCTCTCTCTCTGCTTTGCTTATTTTTCAA 11100
GACAGAGTTTCTCTGTATAATCCTGGCTGTCCTGAAATTCAGTGTATAAACCAGGCTAAC 11160
CTCAAATTCAGGATCCACCTGCCTCTGCCTCCTGAGGACTGGGATTAAAGTTGTGGGGCA 11220
CCACTGTCTGGTAAGAGCTGCATTTTTTTTTTTTTTAAGCAGACTTAATTCTTTTATTTT 11280
TCTTGTATAAAAACCCTTATGTGGTAGCCACAGCTGGAGCCTGGGTCCTCTGAACAGAGA 11340
CTCTGGTGTGGGTTTTCACAAGATGATCAGTGAATTCCTGATAAGGAGACTTGGTGAAGA 11400
CAGTCTCTTTCCAGAGGTCGGGGGTCAGGTAGCTGTAAGTCTTGGAGATGGCATCAAAGG 11460
TGGCCTTAGCAAAGTTGCCCAGGGTGGCAGCGCAGCCTCTGGCTGAAGTGTAGCAGTCAT 11520
CTATACCGGCCATCATCAGGAGCTTCTTGGGCACAGGAGCAGAGACAATGCCAGTGCCTG 11580
TGGGGGCAGGGATGAGATGCACCAGCACAGAGCCACAGCAGCCTGTCACATTGCATGGAA 11640
CAGTGTGGGGCTTGCCAATCTTGTTCCCCCAGTAGCCTCTCCTCATAGGGACAATGGAAA 11700
GCTTGGCCAAGATGATGGCCCCTCGGATGGCAGTGGCAACCTCCTGTGGAGCACTTAACA 11760
CCAAGACCAACGTGACCGTTGTGTCCCCAATAGTGACGAAAGCCTTGAACCTGGTCCGCT 11820
GGCCAGCCTGAGTCTGCTTCTGCACTGGCATGATTTTCAGAACCTCATCCTTTAGGGACG 11880
CACCCAGGAAGAAGTCAATGATCTCAGACTCCTTAATAGGCAGGAAGAACAGGTAGATCT 11940
CCTCCAAGGACTTGATCTTCATGTCCTTAACCAGGCGGCCCAGCTTGGTGACGGGGATCC 12000
ACTCCTTGTCTTCAGCTTTACCTCCACGAGCCCCGGGGCCTTGACCACGGCCTCGGCCAT 12060
GACCACGGCCCCTAAGACCGCTGCCGAATCCTCCGCGGAAGCCACCGCGGCCTCCTAATC 12120
CTGGGCCCCGGGTCCTCCGGGCCCTCTCGCTGCACCGGCGTCATCCACCATTTGGTGTTT 12180
TCCCGAAGAAGAAGAAACAAGAGCTGTGTTCTTAAGGAGTCATTATTTCAAATCAGAACC 12240
ACCATGACTACCAACCCACCTGCTCCCTTTTAAAAAAACTGTTTCATGTAGTCTAGGATG 12300
GCTTTAAAGTTACTGTATAGTTCAGGCCGACCTTGAACTCTTTATCCTTCTGCCTCAATA 12360
TACCTTGAGTTGAAATTAGTCTTGTGTCACCATGCAAGGCTTTAGAACTCTTTGTTTTTG 12420
TTTTTGTTTTTTTGACAGTTTGAAACCCTTAAGAGTGCTAAGGAATCAGCATTGTGGGCG 12480
TACAGCCTTCTTCAGAGGTCTGCAAATGCTGTCAGTGTAGAGTGTAAACATTCTGCTGTG 12540
GTTTGACAGACTCTGTAAGGGCCAAAGCAAGAGTTTTCTGTTCAAAGCTGACCTTCTGAG 12600
ACATGTCTTATAGGGCTGGTTTATCCACACAGATCTCTCTGTGGTTCCCTGAAGATATAC 12660
TTGGTTCTTCATTATAGTAGGAAAAATTGTAACTTCTGGGATATATGCTGATAAAATCAG 12720
TCTATACTTTTTTTTTTTTTTGGTTTTTTCGAGACAGGGTTTCTCTGTGTAGCCCTGGCT 12780

GTCCTGGAACTCACTCTGTAGACCAGGCTGTCCTCGAACTCAGAAATCCACCTGCCTCTG 12840
CCTCCCAAGTGCTAGGATCAAAGGTGTGTGCCATCACTGCCTGGCTATTCAGTCTATACT 12900
TTGTAAAATTGAATAAAAATTAGTTTTTCTTAGATCAAATGTAGGGGATACATTTCCTTT 12960
AATTCTTGCACTCAGAAGAGGTAGAGTTCTTTGATTTCAAGACCAGTTTAGTCTACATAG 13020
TGAGTTCCAAGCCAGTCTACATAGTGAGACCTTATCTCAGAAACAAACAATTCCTGACAG 13080
AAGGTTCTTTCTCCTGGGTTCTTTCTCTTTAGCCTTCTGGTGCTGAGCCATGTTGCTTCC 13140
AGGACTGTCCAGATGTCCACCACCGCTCCTTCACCTTTGCCTCTCTTGCTTACAGGTTGT 13200
GGAGCATGGCCTGTTTGTCAAGCACTGCAAAGTGGAAGTGTATTTGCTGGAGCTGAAGCT 13260
CTGTGAGAACAGTGACCCCACCAATGTGCTAAGTTGCCATTTTAGCAAAGCAGACACCAT 13320
TGGTAAGCAGCTGGCAGGGTTCTGCATGTCTGGTGAAAGTATACAGGGTATTGTAGTTGA 13380
TGAGAGGGAACCTGTGTTGTTTTCTTTCACTATTGTTACTCATCTGTGGTCTTCAGTCCT 13440
GACTTGGAAGTCTTCGTGAGGGAATGTTTATGAGGAACCAGACCTGGGAACTGTGTTCTA 13500
TAGAACAGTACTTGGGCAAGCAAAATGGCCAGCAAAGTAGCTCAAGCAGGCCTGAACTCT 13560
GTATAAAGTCATGCATAACCTTGTATTCCTGATCTCCCTGTCTCCACCTCCTAAGTGCCT 13620
GGTTTATATGATCCTAGGGATCGCACTCATGGTCTCAGAACAAAAACCAAACAAAACCAA 13680
TAAAAGCCTTTACTACTTGAATTGGGCATGGTGTCTAACACCTGTAATCCCAGCACTTGG 13740
AGTGCAGAGGCATGAGGATCTGCCAAGAGTTTAAGGCCAGTGTGAGAGCCTATTTCAAAG 13800
AGGAAAAAAAGGAAAGGCTAGGAAGATGGTTTAACAGGTAAGGGCCCGCTGTTCGGGCAT 13860
GAGGAATTGAGTTCAGATCCCTACCATATAGGAAATCCAGGTGTGGTATGTATCTTAAAC 13920
CTTAGTACTGGGGAACGTGAGACAAGCAGGTAGATTCCTGAAGCTCATTGGTCACCACTC 13980
CCAGGTAAATCAATGAACTTTTAGATTCACAGAGAGATCCTGTCTCAGAAAATAAAAGAT 14040
GGGAAGGGATGGAAAGATACCTAAATATGGCCTCTGGTCCCCACAATGACACCCACCCAC 14100
ACACATGCTTATGCATATACCAGCACTAACAAATCACATACCATACACAGAAGACATAAC 14160
AGTGGCTGGAGAGATGGTTCAGTGATTAAGAGCACTGGTTATTCTTCCAGAGGAGTTAGA 14220
CTGAATTTGTAGTATCTACGTGGCAGTTTACAATTGCCTTAGTTACTGTTTAATTCCTGT 14280
GAAGAGACACCAAGACCAAGGCAACACTTATAAAAGAAAACATTTCATTGGGGGCTTGCT 14340
TACAGTTTTAGAGGATTAGTCCCTGAGCGAGCGTCATGGCAGGAAACACACTGGCGTGGT 14400
GCTGCAACAGTATTTTCATACTGATCTGCAGGCTTTCCATACTGACCTCGGAGAGGAGGA 14460
GAGAGCATGAGCAGCACAAGGCCTAGTGAGGGCTTTCGAAGCCTCCCAGGGATCCCCTTC 14520
TCCAACAGGCCACACTTCCTAATCCTTCCAAACGGTTCCTCTACCTGGGAGCCACACATT 14580
CAAAGACATGAACCTATGGGGGCCATTGTCATTCAAACCACCACACTCTCTTTAACTCCA 14640
GTCCCATGGACCCACAGCCTCTGGCCTCTTGAGGGTATCAAACACATTTAGTGTACAAAC 14700
ATACATACAGTCAAAAACGCATTTATACATATAAAAATTAATTAAAAAAAAGAAACCCTA 14760
ACAATCTCAGGGGAGAATGGCTGATTCTGTTTGGGTTGGTTGTTGTTTTGGGGAGCTGGG 14820
GGTCAATCTCAGGACCTTTCACTCATGCTGAGCAAGGGCTCTACACTAAACCACATCCTC 14880
TGTCCACTCTGCCTAGCACCCACTTTCTTTGAGATGAAGCCTTGCTGTGTAGCTCAGATG 14940
ACCTCAAACTCATGATCCTGTGCATCATCAAATCCATTGCTAAACCATGGGCTGGGCTGT 15000
ATGTATCCTGGCTCCGAGGTCACTTTTTCTTCCTTAAAACTTTATTACATCACTCCCCTT 15060
CTCTCTCCTTACTCCACCCCTCCCCTGCTCTCTCAGATCTCTCCTGTGGCCTCGCCCACT 15120
CCATCTAAAATTGGTGAAATAAACTCAGAAGGTTGTGTGTGTGTGTGTGTGTTTGTCTAA 15180
CAGTAATGATCATAGAACAAGCTTCTCTTCTTGTTTGTTTTCTGAGACAGAAAGTTGTGT 15240
GCCACCACGACCTAGCTTCCAGTTTTCTAACTTTTATGTTACTTTTCTTTCTTTATCTTT 15300
TCATCTTTCTTCTTCCCTTCCTCTTTTTTTTATAGTTTGTTTTTTTGAGTCAGGGTTTCT 15360
CTGTGTAGCCCTGGCTGTCCTGGAACTCACTCTGTAGACCAGGTTGGCCTAGAACTCAGA 15420
AATCTGCCTGCCTCTGCCTCCCAAGATGCTGGGATTAAAGGTGAGCACCACCATTACCCA 15480
GCTTGCCTTAACATGAGTTTTTCCTGACCTGATGACAAACAGAGCCAGCTGTAGTGGCGC 15540
ATGCCTGTGTGGCAGTAGAAGGAGGGGGAGTTCAAGGCCACCCTCTGGTATTAGTGAATT 15600
CTAGGGTTTCTTTAACTTTATTTCTGGAGACAATCCTGAGTGATGTTTTTGCCTGTTCTT 15660
AGCAACTATTGAGAAGGAGATGAGGAAGCTCTTCAACATCCCTGCAGAACGTGAAACACG 15720
GCTTTGGAACAAATACATGAGCAACACCTATGAGCAGTTGAGCAAGCTAGACAACACTAT 15780
CCAGGATGCTGGGCTGTACAGGGTCAGGTAAGCGGGCCCAACCCATGCCCAGACAGGCAA 15840
CTCTGACTGGGGTGTGTGTCATGAAGGCTTCTGGCAGTAAGCATGTCAGCAGTAGGGTTT 15900
GTTGCATTGTGCTCATCTGGTTTCCCCTTTCACTCAGGTGCTAGTAATTGAGCCCCAAAA 15960
TGAAGATGGCACATGGCCCCGGCAGAGCCTGCAGTCAAAGTAAGTGAGTCTTCTCCCCCA 16020

TGTACCTGGCTCATAGGTGCAGAGGAAGGGAAGAGGGAAGGCTTGTTCTGAAGTAGGTGG 16080
AAAGCCGGTTTCTCATTCCAAAGCCCCAGTTTTTAAGCAGCCTGCTTGACATTCATGCAC 16140
GAAATAGCTGTCGGGATTCCGTCAGGACAAGAGATGGGGAGGGATCTGATATTGTGGCTC 16200
TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTAAGCTCCTTTATTTTTCTACTT 16260
GTAGTTTAGAGGCTTTGAGAAAGGAAGGAGGATTAGGGTGAGGTCTTGGACGACGTAAGG 16320
GTGATAGCTTGCTCCTACTCGACTTTTCTGTGGCCTCTTACAGATTACAGCTCAGCTACT 16380
AAATTTGGATTCCAAGCACAGTTATGCCAGGAGATATTCATGTGGCTTGTGGAATTGTGC 16440
ACCTTGCAGTAGAAAGTGATGGCTTTTCTCTAGGATGTAGTGCAGGGAAATAACATGAGT 16500
GTTTCCTGACCTGATGACAAACAGAGCCAGCTGTAGTGGCGCATACCTGTGTGGCAGGTA 16560
GAAGGAGGGGAGTTCAAGGCCACCCTCTGGTACTTAGTGAATTCTAGGCTAGCCTGATAT 16620
ATAGGAGACCATATCTACAAAACAAACAAACAAAAGCAAACCTGTGGCTGGGATGATGCC 16680
TCAGCAGGTCAGTTGCTTGCTATGCAAGCATGAAGACCAGAGTTCACATTCCCAGAACTC 16740
ACTGTGTTGGGTGAGTGTGATAGCCTGCCTATAATCCTAGTGCACAGGAGCTGGAGGCAG 16800
GGATCTCTAGAGCAAGCTGTCCAGCCATTTAGCAGAAATGGTCTCACCCTGTATTCAGAG 16860
AGATCTGCCTCAGAAGGTGAAGAGCAGTGAAGGAAGACAACTAACATGAACCTCAGCTTC 16920
CACATGTATGTGTACACACATGCACGTTTCCCACATCTATGCATATGCAAAACAAAACAA 16980
AACAAAACAAAAGTTTAGTAGAAAACACTGTATTGCCACAAAATGTTTGTCCTTATTACA 17040
GAACTTAGACTGCTTTGCAGGTGTGATTAAATTAATTATGTTGAAATCAAGAGACTCTCC 17100
AGGTGGACCCATTGTAATCATGAGACTCCCTGGAGGTGAGAGGGAAGCTGGAGAGTCATA 17160
GTAGGTATGCAATCAGAAGCAGAGGCGAAGCAGTCAAGTGTGAGAAAAGCTTTTACTGAC 17220
TCAGCCAGGACATGGGAACGAACGCCTTTACAAGGATTTAGGGACTGCTGAGAACTGTAA 17280
CTGGGACATCTTGACTTGAACTCAGACCTCCTTTGGGTTTTTTTTTTTTTTTGTATATGT 17340
GAATGGGCATTTTGCTTGTATGTATATCTGTGTATTATGTGTGTGCCTGGTGCTAGAAGA 17400
GGGTATTGGATTCCCTGGCCTGGAGTTGCAGGTGGTTGTAATCGTATGGATGCTGAGTCC 17460
TCTGCAGAAGCAGCAAGTGCTCTTGGTCATGGGAGCCTCTCTGCAGCCCTCCTTAGACTT 17520
TCACATCTATGCTCTTTTATGAGAACTCAGCTGGGTTGTCTGGAAGCCCCTGACTGAGTG 17580
ACACCTGTAACAGAACAGTAGGAAACAGTGGGCAAGTGCAAGCTGGACTGCCCCCTCTTC 17640
ACCAGAGTGCACAAAGCAGACAGCAGCCTCCAGCCTGTGGATCCTCCAGGGCTTGCACAG 17700
GGGTCGCCTGTGACCATGGGAAAACACAAGCAAAATTACAGTTACGAAGTACAATGAAAA 17760
TAATTTTATGGTTGGGGGTCACCCCAGACTGAGGAACTGTGATAAGGGTGTGAACACTGC 17820
CAGGCAGTGGTGGTGCACACTGTTAATCCCAGCACTTGGAAGGCAGAGGCAGAGGGATTT 17880
CTGAGTTCGAAACCAGCCTGGTCTGCAGAGTGAGTTCCAAAACAGCCAGGGCTACACAGA 17940
GAAACCCTGTCAGAGTGTGACCATTAGGAATGGTGAGAACCACTGCTCTAGAGAAAACCC 18000
TCTAGGTGCTCAGCTTGTAGCCTGGGAGGTACAGGTGGTAACCCAGGCAGAGGCAGGCTG 18060
GCTGGCCTGGCCTTTGGACCTCTGTCTTCTTGGTTTTCCTTCTCCCTATCCACCCCACAC 18120
TCTGAAGCCTAAGAATTTTCACTAAGGGAGACAGCTTTTGAAAACGTAGCAGTGCCTCCT 18180
TGAGACTCCATGAAGGGTAGAGTCTCTTGGAGTGGTGTGGTTGATCAAGAGTGAGCATAG 18240
AGTGGAGTGAGGTGCATTCTCTGATCTCTGGTGCAGTACCAGATATTGTTGTGCATGAGG 18300
ATACAGCCATAAAGTGAACAGTTTGTTGGGTCAGGAAAGGAGATCAATAGGAAAAGTACT 18360
TGTGCCCACCGATTCATAAGCAATATAGAAAAAATTGAAGGAGAGAATTATCAAGCTCTA 18420
GGAGTTAGAATGGGGAGGACTTAACCTGCTGAGCCATCACACGAGATGGACACACACACA 18480
CATAGAATATATGTGAATGGATGAATGAATGGATGACTGTGTGTTCGTGTGTGTAAGTGC 18540
AGACACAGGCTTGTGGAGTCGAGGTTGACGTCAGAGGTCAGCCCACCTAGGTTCAGATGG 18600
GAATCTCTCTCAAGTTTGCTGCTGCCCACGCCACCCTAGATGACCGACCAGAGCTGGGCC 18660
CACAGACGTGACAAGTTCTCTCCGCTCTACTTCATAGCTCCTGCGTAGTGTGAAGACAGC 18720
GCTTTTGCAGCTTAACCGAGAGTGGGGAATGTTTTTCCTTCCCTTTTTCCTCTTTCCATG 18780
TAGGCCAGATTGTCCTCAAGCTTTTTGATTCTGGGGATTCAAGTGTGAGCCATGAAGCAC 18840
ATTTAGGAATGTTTGAGACAAGACCTAAAGCAGTGAATTGGGCCTGTACTTTTAAAACAG 18900
TCAGTGTATTGTTTTTTGCTTCCCCTTCCAGATCAAGCACTGCACCTAGCAGAAATTTCA 18960
CTACCTCTTCAAAACCATCCGCAAGTCCCTATTGCTCAGTGTCTGCCTCTCTCATTGCAA 19020
ATGGTGATAGCACTAACAGCTCTGGGATGCACAGCTCCGGTGTCAGCAGGGGGTAAGAAC 19080
AGGCCTTTAACTGTGCTTTCCTTCCACTGCGTCCTTTCTCAGGGGGTACCTGGGCTGCTC 19140
ATCTGTGCCTAAAATAAAGCTGTGGGATTAGAACTGCCTGCAAGAGCATGCTGCATGTGG 19200
CCTCCGCTCCAGCCCATTCCTGCCTCATCCTTCTGACTCTGCTCACTTGTCCTTTTGCTG 19260

TCTGGTGTTTTTTCCTGTCTAAGCCATTGATAATCTTGAAGGCATGGTAGCTGGTGTAGG 19320
GAGATTACAACATTTGTAAAGGGATGGGCTGTTCAGATGGCAGCTGCAAGGGGGCAGTTC 19380
TCAGTTCTGTAGAACCCTGTTTCTCCTTTTAGTAGCTGTGTTCCTTTCTCTATGAATGTT 19440
AGGAGCACCTGCCATGAGATATAAACAGGATGGAAGAGAGTAAGCAAGGAAAGAGTAGAT 19500
GAACACTCAGGCCCAGTCAAGGGTAGTATGGCCTTGTGGATGGTGGCACATGTAATGTAA 19560
GCTCTTGATCTTGGTACCTAAAATATGTTTTCCTGTTATCATGTGATAGCCTTCTAGAAT 19620
GATATAAAATTAGTCGCCAAAGTCTAAAAGAAGAAAATCCAAATAGCATTTCTCCACTAT 19680
ACTTACACCTGCTGACAAGTCAAGTAAAGATGGAGGGGTGGcTCTGGAGGACCAGAGAAG 19740
GTAGGGTACATTATGAGACAACCAGGCACAAAATCTCAATGCCCAAGCAGAGCTCTGAAA 19800
GTCGTCCTGAGGGGATATGTGCAGCACTTGGTTTTATTGCTGATGTGGGTTGTGTTTTTC 19860
AGTGGATCTGGCTTCTCTGCTTCGTATAATTGCCAGGAGCCCCCATCACCTCATATACAG 19920
CCTGGCCTCTGTGGACTTGGAAACCTGGGGAACACTTGCTTCATGAACTCTGCTCTGCAG 19980
GTAGAGGGGAGCATCACCAACTGAATAAAACTGTTGTTGGGGTTTGTTAATTGAAAAATT 20040
AGTATTAGCCCCCTCTCCCCCCACTTCCTTTTAGTGCTTCAGAAGTCTCTGCTAGGATAT 20100
CTCCTACTGGCTGCTTCAGAGTAGCATCATCAGTGTTTATGGTCCTAAAACATCTATTGC 20160
TTTGTTAGTTGGTTGATATTTTTTTGAGACACAGACTCTTACTTTCTCTTAAATAAAAAG 20220
CAAGAAGGCTCGGGGTGGTTGAGGTCTTTCTTGGCACTGTCAAGCATATTAGGCATCTTC 20280
AATGCTGGATTCAGTCCCTAGCACTACTGCTACAGAGCCCAACAGAAACCAGTGTTAGCA 20340
TGGTAGTTTATGACTGTGATCTCAGGGCGTGGGAGTCAGGAAGATCTCAAGTTCAAAGAT 20400
CTTAGGCTGCAGAGGCAGATCTATCTCAGAAATCTCTCTTTATTCATTCATCACATATTT 20460
ATTGGGTAAGCACTATGTAATAACATCTGATAGACATGGGCTGCAAGGCATTAAGAGTCC 20520
AACTCCAAAATTAGGAGGTCATAGTCTCTAAAAATGTCAACTGCAAGTTCAGGGTTTCCT 20580
AAGAAAGCTCTCAATTTTAATACTTTGCTAGAAGGATTTAGGATTTACTTAAAACTATGA 20640
TAGTTAATTTATGATAATTGTTTTTAATGAGGAAAGGATACAGATTAATGGTAGTCAAGA 20700
GAAAATGCAAATAGAGTGTCCAGGATTCCAAATGCAGAACTCTGGTATGCTTCTATATAT 20760
CTCCAATCCCAGGACTTGGAAGGCTGTGGTGGGAGGGTGGCCAAGAGTTTGAAACCAGCT 20820
TGGTCTGTAAGTGTATCAGACAGAAAGGACTTGGTCTGTATTAACTGTTTTCATGCTGTC 20880
TGGAAGGGTTGCAAGCTCTGGCTTTGTTCTGTTTCAGTGCCTGAGCAACACTGCCCCACT 20940
GACTGAGTACTTTCTCAAAGATGAGTATGAGGCCGAGATCAACCGAGACAACCCTCTGGG 21000
GATGAAAGGGGAGATTGCAGAGGCCTATGCAGAGCTCATCAAGCAGATGTGGTCTGGAAG 21060
GGACACTCACGTGGCACCACGGATGTTCAAGGTAGGTTGTGTGCACACTGGAATGCAGTT 21120
GTGCACACATGCATGTAGCATAGCACAAGTGTAGATGTCAGAGGACAACTTGCAGGAATG 21180
GGTCTCTCTAGCGTGTGGATCCTGAGGATCAAACTCAGCCGGTCAGGCCTGGTGGTGGTA 21240
AAGCATCTTTATCTGCTGAGCCTGCGTGTTGGAAGGGAGCGGCTTCATATTGTACTAAGG 21300
CATCCCTTTGCTTGCATTTTGACCTAGGTGAGGGTAGACTCTGAGGGGCCAGCAAGATGG 21360
CTCAGTGGGTAAAGGCAAGCCCAATGACCGGAGCACAGCCCCCAGAACCCATGTTACGGT 21420
AGAATGAAATAATCGGCTCTATGAAGTTTTCTCTGACCTCTGTCATGTTCTGTACTGTG 21480
GCACTCAGTGACACTCAGCACTCATATTGTTTAAAAAAAAAAAAGGTGTAATCTAAGACT 21540
TAAAATAAGTTCATGTTATCAGAACGTCCTTTTCTTTCTTTTAGTGCTGCGGAGTTTTGG 21600
GTGGGGGGCTTTTTATTTTATTTATTTATTTATTTTTATTTTGGTTGTTTGTTTTGTTTT 21660
GTTTTTGGTGGTGTGGGGGTTTTGTTTGAGTCAGGGTCTCATTATGTAGTACTGGCTATC 21720
ATGGAACCCAGTATATAGACCAGGCTGTTCTCAGACTCGGAGAGATGCAGCTGCCTTTGC 21780
CTCTCGAGTACTGGGATTAAAGGCATGTGCTACCAGGCCTGACTCAGTACTACAGTTTTG 21840
TGTGTTTATTTTATTTTTCGAGACAGGGTTTCTCTGTATATCCCTGGCTATCCTGGTCTC 21900
ATTTGTAGACCAGGCTGGCCTCGAATTCAAGAGATCTATGTGCTTCTGCCTCCCAAATA 21960
CTGGGATCAAAGTGTGTGCACTACCACCTGGGCCCTATGATCTAGACTAGACTATAAGCA 22020
TTTGGTGGTGGTGGTTGTTTTGTTTTGTTTTTTGTTTTTTTTTTTGTTTTTTCGAGACGG 22080
TTTCTCTGTGTAGCCCTGGCTATCCGGGAACTCACTCTGTTCTAGGCTGGCCTCGAACTC 22140
AGAAATCCACCTCCTGCCTCTGCCTCCCAAGTGCTGGGATTAAAGGCCTGTGCCACCACG 22200
CCCAGCTTTTTTTTTTTTTTAAGGATTTATTTATTATTATACATAAGTATACTGTAGCTG 22260
TCTTCAGACAACACCAGAAGAGAGTGACAGATCTCATTACAGGTGGTTTTGAGCCACCAT 22320
GTGGTTGCTGGGATTTGATCTCAGAACCTTTTGAAGAACAGTCAGTGCTCTTACCCACCG 22380
AACCATCTTGACAGCCCCTAAACGTTTGGTTTTGACTGCTAAAGTGATGTAATCCGGAAT 22440
CCTGTGTCCCATGTTTCTGTGCTCTAGTCAGGACCTCAGAAGGCAGGGATAGACTTGCAT 22500

TCTCAGAGGGTGATATGGAAGGCAATTCTTTTCATAACAGGGTTTCCTCTGTATAGCCTT 22560
GGCTGTCCTTGCCTCTAGTTCAGAGATCCATTTGCCTCTGTCTTCTTAGTACTGGGATTA 22620
AAGGCACGTGCACCACTACCTAGTTAAAGGCAATTCTTTTTTTTTTTTTTTTAAGATTTA 22680
CTTATTTTATGTAAATGAGTACACCATTGGCTCTCTTCAGACACACCAGAAGAGGGCATC 22740
GGATCCTATTACAGGTGGTTGTGAGCCACTGTGTGGTTGCTGGGGATTGAACTGAGGACC 22800
TCTGGAAGAACAGTCAGTGCTCTTAACCGCTGAGCCATCTCTCCAACCCCGTGAAAGGCA 22860
ATTCTTACTTAGATTTTTTCCCTCAGCGGAAAAAAAAATCTTAAGTGGTCGGTTAATAGT 22920
GAGGACTCTTTCACCTTCTTGGTCACTTGAATCTGTGAGAATTACAAGTCAATTACCATT 22980
TATACACTAAATCCTAAAGGATACTAAGATCGGACGGCAGCACCCTGGACTTGCTCAAGT 23040
GGAAACGACTTACAAATGGCATTTTGGGGCGGCATACTGATGGACACAGTTGATAGCTGC 23100
ACTGTCTGGTGGTAGCCGCTAGCCAGGTGGTTACTGTGACCAAGAATCTGAACAGGGGAA 23160
AAGGGATTGGAGAGATGGCTCTGTAGTTACAAGAACTGGCTGTTCTTCTAGAGGATCTAG 23220
ATTGGATTCCCAGCATCTACATGGTGGTCACAATCATCTGTAAGGCCAGTCCCAGGAAAT 23280
TCAAATTTATCTTCTGGCTTCCATGGGTACCATAGAGGCAGAAGGAAAAAAAAAAAAAAA 23340
AAAAAAAACAACTCCTAGGCGTTAGGGTCTTAAAGGAAGGTTTTGAGCATGCTCAGAAAG 23400
GGGAGGTGAGCTCCATCCCAGCTGCTCTTTTATTTTCAACTAATTTAAATATAAAAAACT 23460
TCATAGTTAGTAGCTAAAATATTACACACCACAGATGTAGAACACTGCTTGACTTATCAC 23520
AAGAGACTATGGGCTCATTCCTGGGGTTTGTGGTTGAGCACTGCTGCTGTGCTGCTGACG 23580
GAAACTACACTCCAAGGGCGGCTGGATACAGCAGCACTGGACACTCAGACCTAGGGATGT 23640
TAGGGCTGTTTCACATATGACTCCTGGATTGATCACTTCTAAGGCTCTATAATTCAACTC 23700
TGCGAACTACAAAAAGACACTGGAGTGGGGTTCACATCACTGCTTGGATCTATCGCTGAG 23760
AGTGCGCCCCTCCCCGTTAGGATAGTGGCCTGACTTCAGGCCTGTAGTTTACTTGGACCT 23820
CTAACCTGAGCCATTCCTCTGGTGTCCCATTGGCCTCAGTGTCGCTTGTAAGGCGTTAGC 23880
GGGACTGTCGCCCAGGTGACTGTAAGGCTGCTGCTGTGGGCCCTGCCCTGACTTTGTTTG 23940
CTTCTTTTCCCCCTTCACTACTTTATTGTGTCTCCTAATGGAGTTGCTTTTGTTGTAGAC 24000
GCAAGTGGGACGTTTTGCCCCTCAGTTTCTGGCTACCAACAGCAAGACTCCCAGGAGCT 24060
GCTAGCCTTTATTCTAGACGGACTGCACGAGGACCTGAACCGCGTAAAGAAGAAGCCTTA 24120
CCTGGAGCCCAAGGACGCCAATGGGCGACCAGATGCGGTATGATACTAAAGGGGTCAGCA 24180
CCCTGTCCCAAAGAGAGTCCCTTGTATTGACAGCATGGGCACAGACCTCTGGGGTTGGAG 24240
TCAGTGCAAGGGCCCCGGTTGCCTAGTCAGAAAGCACGCTAACTATAGTTGCCTTTCCCG 24300
CCAGCACTCCAGATCAAACTCACAGGGCCGTTTGAGTGCTAGACAGGCCTGCAACTTCTG 24360
AGCTACATCCCAGGCCAGCTGTAGAAACTCCTAGCACTTCCTCGGGTTTCTTCAAGTGAG 24420
TTAGTTTGTTTAGTATTGCATTGGCAAGCCAGGACTGAGCGAGCACTGAGAACTTACCAG 24480
CTGACCGTGTGCTCTACCACTATCACTGCTCTTTACAGCTCAGCCAGGGTTGGTTTAGTT 24540
GTGGTACTGGCGATAAAACCTCCAACTTCTTCCTATTGGCCAAGTGCATACCACTGGGCT 24600
GAAGTTATGGCCTACTTGCATTGTTTAGTTTGAAATGCAGTCTTACTACATTGCTCAGAA 24660
TGGTCTTGACCTCAGACTCTAGCCTAGGAAGACCTTGAAATTGGCATCTTTCTGCCTCAA 24720
CTTCCCAAGTAGCCAGGGTTACAGGCCTGTGCCACTAGCCCAGTTTTTAGTGGGACAGGA 24780
TCTCATTAGTACTTTGGGCTGGTGTCTGACTGTATATCTGAGAGTGCCTAGAAATTTTGA 24840
GTCTTTTGCCTCCAGAGTATTAAAGTATCCTCTTCTTTCCTTGATTATTAAAATATCTGG 24900
TTAAAACTAAATTTAAGACATATATACTTGATGCCTTGATTTCTCTGGTACTTGAAAGGG 24960
GAATTTATTTTAGCGATCCTGTTGCATTAAAACAATTGACATGAATGCTTTAAAGTAGCT 25020
TTTAAATCTTAAAAAACAGCTTAACTATATCATGGATTAAAAACTCTCCACTAGTGGCCT 25080
AGAGAGATGGTTTAGAGGTTAAGAGCACTGCCTGCTCTTCCAGAGGATCTGAGTTCAATT 25140
CCCAGCAACCACATGGTGGCTCACAACCGTCTATAATGAAGATCTGGTGCCCTCCTCTAT 25200
CATGCAGGCAGAATGCTGTATACACAATAAATAAATAAATCTTTAAAAGAAAAGGAAAGA 25260
AAATCTCCCTTAAGGCTAGGCAGTGGTGGCCCATGCCCTTAATCCCAGTATTTTGGAGGC 25320
AGAACCAGGTGGATTTCTGAGTTCCAATCCAACCTAGTCTATAGAGTAGCTTCCAGGACA 25380
GCCCGGGCTACACCGAGAAACCTGTCTCAAAAACCCAAAACAAAAATCTTCCCTAAGGAC 25440
CTTTAGATCATTAAAAGCACAAAAGGATTGGTCAATCCCCAGTTGGACTTGTGTGCCTTT 25500
GGTGCTCTCAGAATCCCTTTTTTGCCATAGATTCTTGATGGTTGTGGGTGTAGAGCTAAC 25560
CCGTGGTGAGACTACGGTACACTAAAAGCTCTCGGCTGTTCACTGGGAGAGGACAGATGT 25620
TGCGAATCTCCCTACTTAATCCTTTTCCTTTCGTGTAGGTGGTAGCAAAGGAAGCCTGGG 25680
AAAACCACAGGCTGAGGAATGATTCTGTGATTGTGGATACTTTCCATGGCCTTTTCAAA 25740

CGACTTTGGTTTGCCCAGAATGTGCTAAAGTTTCTGTGACCTTTGACCCATTTTGCTATC 25800
TAACTCTCCCACTGCCTTTGAAGAAGGATCGGATTATGGAGGTCTTCCTGGTTCCTGCTG 25860
ACCCTCAGTGCAGACCTATCCAGGTACCGGAATCCTCACCTTCATAGCACCCCTCACTGA 25920
GTAGACTCTGGTCCCAGATGGGACTACGTTCTATCTGCTGCATGGGAAACACCATTGACT 25980
GTCCCTAGGTCCCCTTGGGCTTAAGTCTCTATGTGCACCTTGGCCTCTCTGAGACCTCTG 26040
TGTAAGCTCATTACTTGTGATGGTCCTTAGGCTAGCTCTGTCACTTCCCCACCGTGTAAA 26100
TTCATGTAGCTGTAGTAACCTGCCTTTCCTGCTAAAGGGCCTGGAATGTTGGGAGGATGC 26160
AGTAGAGTTTCCTTAAGTTTGTCTTCCAGACAGGTGGGAGAACAAGTCAGGTATGGTCAG 26220
TGGGTCTTCATATGCAATTCTGCTGGTAGCAGAGGGCAAGGTGATACCTGCTGGGTGTGG 26280
CTGTGCCCACTTTTTCTCCTGGTCCTCAAGGGAAGCAAAGGCAGGTAGAACAGATCTGAA 26340
GTTGAAGCCAGCTTGTCTTTGTAATGAGGTTCGGGCCATCCAGAGCTACACAGTGAAGCT 26400
GTCTCAAGGGGAAAAAAGGAAAGAGGGGTGATACATAGTATTAAATGGCAATTTATATGC 26460
TTTTAACATTTCTCTAGAGGAAGTGCTGACCCAATCCAGAAGGTAAAGGATACAGCAATA 26520
TTTGGGAGGGAAGGAGGTGGCTGCCTGAGTGTTGCAGCTGTGCTGACAGCAGAGCCTTGG 26580
CTCTGTCAGCTGGCCATATCTGGGCCATGCACTGCCTCTGGAGGAATCCTACGTCTCCAG 26640
GATGTGTCTTCCTCTAGGGTCATTTCTTGTCTCCTGTCACCTTGACTGTTTCTCCTTTGC 26700
AGTACCGTGTGACTGTGCCATTGATGGGGGCCATTTCTGACCTGTGTGAAGCACTCTCCA 26760
AGCTGTCTGGCATTGCTGCAGAAAATGTAAGAAGGAGTCATTAGTGGTCCCTCCCACCAT 26820
CAAATCACTTTATGGTCCTTTGGGTTTACTCTGAAAGTTTTGGGAGATTTTTGTTATTGG 26880
ACAGTGCAAATGTCTTGGATACCTCATAGCTCCTGGTCTTACTCACGGTCCTTTCCAGTG 26940
CAGGGAAACAGTCTTTCTCTCACTGTGGTGGTGCCGCCCATTGTCAGAGGGACAAGAGTA 27000
GGGCTTGCTTTCATCATGGTGCAGGTATAGAGACTGGCTCTGGAAAGCCACTGGGATTTC 27060
TAGGCCATGACCAGTGTCTGCTCAGCACTCCCTTCTCATCGTTCCTCCCATTTGAAGGCA 27120
ATGTCATCAAACAGCCGTGCTGGCAATAGCCTGCTGCTCCCTTTCCGTCTTGGCCATACG 27180
TCGCCTTGGGACCACTCGAATGGTGGTTCTTGGTAGGCAACACTTGAGCAAGGAGTTTCT 27240
GCCGCTGTTTCTCAGTTTTAGTCAGTGAACAGGAGCTCTGTGAGCCATCCTGTCATCTGC 27300
TTGTGCTCTTGATTCTTAGATGGTGGTCACTGATGTATATAATCACCGCTTCCACAAAAT 27360
TTTCCAAATGGATGAAGGTTTAAGCCACATCACGCCTCGAGATGACATTTTTGTGTAAGT 27420
ACTTAGCAGTTTCTGATACATGAGCAGAAGAGCACTGTAGAGACTTTGCTCTGACATAGA 27480
TGTTTGTGTGTATATGTTCTAGTACCTGACCTAGGTGCTTGTGTATATATGTCCTAGTAC 27540
CTGACCTAGGTGCTTGTGTATATATATCCTAGTAGCTGACCCAGGTGCTTGTGTATATAT 27600
GTCCTAGTAGCTGACCCAGTGCTTGTGTATATATGTCCTAGTATCTGGCTTAGGTGTTTC 27660
TGTATATACGTCCTAGTACCTGGCCAGAGAAGCGTGTTCAGGAATCTGGGTATGTGTTAG 27720
TACCAAGGACCCTTTGGCCCCACAGAGCCCTGCACTAACACAGCTTTGGCAACAGATTCT 27780
GTTTGCTAAGAGGACTAATTCAGTGGTTCTGCTGTCAGAGGTTTCTTGCTCAATAAGAAT 27840
CAGCTGGCCAGTTGGTGGTTGCGCACACCTTTGATCCTAGCACTCAGGAGGGAGAAGCAG 27900
GAAGATCTCTGAGTTCTAGGCAAGCCTGGTCTACAGAATGAGTTCCAGGACAAACAGGGC 27960
TACACAGAGAAACCCTGTCTTAGAAAAAAAGAAAAACAAAGAGTCACAGCTGTGTGAAAG 28020
GCCAGGGCAATCAAAGGGCAGTGGTATACCGGGATTTCGGGTTTCTCTGCCCCATTTTCT 28080
GCTTCTCAAGCAGGTACCATGGTGGCACTCTCTCATCTCAGAATGAATGGCTGTTGGCAT 28140
CTCAGGCATTCTATAAAGCTATGAGATTCAGCTTTTCAGTTTTCTTATCACCATGGAAAC 28200
AATACAAAGTGTTGTTTTGTGGCAACTCTTTGGGAAATAGGGTGTGTGTCCTCCTCCTAG 28260
GTACAGAGGTGCAGCTTCCAGCCAGCCATGTCTTACGTATACTCTTTTTTTACTTCCCAT 28320
TCAGAACGGCATCCAGTTGGCTTGCTCAGTTTTGTACTCTCAGCAGAGCCAGGCAGATCG 28380
GATCTCTCTGAATTTGAGGCCAGTTGAGGCTAGAAGTGCTTTTAACTGCCTCTTAACCTC 28440
CCCACACTTGGGAGGCAAAGGCAGGGAGATCTCCGTATTCAAGGCCAGCCTGGTCTACAA 28500
AGCAAGTTCCAGGAGTTTGGTCCCTCCAATCCACATAGTAAAAGGAGAGAACTGACTCTA 28560
CAGAGTTGTCCTCTGACCTCCACACATATTATAACACACACACGCACGCGCACACGCACA 28620
CACACGAGGGCTTGAGTTCAGATTCCCAGCACCTGTGTAAAAAGGTAGACCTAGTGGTGC 28680
ATGTCTGTAATCCCATGCTGTGGAGGCGGAGCTGGGCTTGCTGGCCAGCGAGCTTATCTG 28740
AAATGGTGAGCTCCAACTTCAGTGAGAGACACTGACTCAGAAAACAGAGTTGAGGGGCTG 28800
GTGAGATGGCTCAGTGGGTAAGAGCACCCGACTGCTCTTCCGAAGGTCCAGAGTTCGAAT 28860
CCCAGCAACCACATGGTGGCTCACAACCATCCGTAACAAGATCTGACTCCCTCTTCTGGT 28920
GTGTCTGAAGACAGCTACAGTGTACTTACATATAATAATAAAATAAAAAAAAATTAAAAA 28980

AAAAAAAAAAAAAGAAAACAGAGTTGACTGAGAAAGATGCCAGTGTTGACCTGACCTCTA 29040
CACGAACATGCACCACACACAGAATGAATTGTCTTATGAGAATAGTGGCATAGTGTTTGG 29100
TCTTGGTTGGTTGATACATTGTTTTGAAACAGGGTTGTCTTGTAGTTTACCATGTAGACC 29160
AGGCTGGCCTCAAACTCAGAGTGAACCATGTCCCTCTGCCTCCTGAGTGCTGGGTATAGG 29220
TGTGAAACTTCACTTCTTGGTACAGATAGTATTTTTAGCCATGTTACATGCTATCCTTGT 29280
AATTTAATGCTCACCACAACCCTGGTAGGTGAGTAATCCTACCAAGTACACAGACCTAGA 29340
AGTTAGAGGGTGCCACTGGGTCACTCCAACCTGAAAAGGACACAGTGAACCTGGAACTTG 29400
GAGCAGCCTCACTCGAGGCCCAGCTGCTAACCATTGCTAACTTGAGTCTTTCTTTACAAC 29460
GTGTCTTATGAGTTGTTTTTAAAATCCCTCCTGCACTAGCCAGCAAGGTGGCTCAATAGG 29520
TATGGCTACTTGCTGACAGTCTAACTGCTAGAGTTTAATCTCTCCAACCCACATAGAGGG 29580
AGAGAACCCGCTCCACTCAAATTGTCCTGTGATGGCCACCGCTGTGCTGTGTGACATGTG 29640
CCTGTCAGTTCTCATTTGCGCGCTCGCACACGCGCATACACACACACACAAGCAGTTTTC 29700
ACAAAAGCTGCTTCTGGAGGTTGGGATTTAGGGCAGTAGTTACTGGGGGTAATTATCCTC 29760
CAGAGGACCAGCTTCGGGCTTTAGCACCTGTATCAGGAAGCTCACCGTCACCTTTATCAC 29820
CAGGGTCAGGGGATTTTGTGTTTTCAGGCTTTTGAGGGCTCCATGCTCATGTTTATGGTG 29880
CATAGAAATCATGAAGGCATATACACAAATGCCAATAAAAATAAATTTTTAAAGGATTTT 29940
TGTTTTTCCTTTGTGATTCCTAATGAAAACTGCTGACCAGGCATTGAGGAGGAGTGCTGC 30000
TTGTGCACTCTACACACTGGAATCAGTAAGTGTGTGTGTCTCTATAGGTAGAGAGAGAGA 30060
CTGCAGTCAGTTGCCAGTAATTTTAGTAACATCTCCAGAATAGCAAATGCTACCTCTAAA 30120
CCTCCCTCCTCGTTAGATCAGGTTGGTACTGTTTATTCACTGTTTAAGACACAGATTTCA 30180
AAATAATTTATTTGTACCTAGTAGTTTATTTGGTTTTCTCTGGATTTTTATAAAGTTGGA 30240
TATGGCTGGATGTGGTGGTATGTACCTGTGTCTGGCATTTGGGAGGCAGAGACAGGTGGC 30300
TCTGTGGTGAGTTTGAGACCACCCTGGCCAACATAACAAGTTCCTGACCAGCCCAGACTT 30360
CATAGTAAACTTTTCAAAAAATGAGAAGGTGTGGGGTATAATTATTTTATTTTGCTTGAG 30420
TGATTCTTGAATAACAATGGTAAGTAACTGGTTTTATAAGTTTCCCCACTTAATAAGGCT 30480
CTGGCTAGTAGTCTCAGTGTATATTCCTGTGTTATATTTGGCTTTATTATACAGAAGAGC 30540
ACTAAGAACTCAGTTGGGTTTTTGTTGTTGTTGTTGTTTGTTTTTGTTTTTTTTTGGAGA 30600
CAGGGTTTCTCTGTGTAGCTCTGGCTGTCCTGGAACTCACTCTGTAGACCAGGCTGGCCT 30660
CAAACTCAGAAATCTGCCTGCCTCTGCCTCCCAAGTGCTGGGATTAAAGGAGGTGCCATC 30720
ACTGTCTGGCTAACTCAGTTTCAAAAATTGAGCTCCTAGCTGGGTGATAGTGTTGAACAC 30780
TTTTAATGTGAATGTAGCTTAGTTGGTAGCATGCTTGCCTAGTGTATACAAAGTAGTGAA 30840
TTTGGTCCCTCTCACCAGATAAACCAGGTCTGTAATCCCAGCACTTGGAAGGCAGAGGAG 30900
GAAGGATTCGAGTTTAAGATCCTCCTCAGTTATGTTGTAAGTTTGAGACTTGCCTGGGAT 30960
ATGGAAGGAAGGAGGAAGAGAGACAGAGACTTAAAAAATAACTTTTAAAAGTTGAGCTTT 31020
TTGGTGAGAAAAAGTGAAGTTGGTAGACGAATTCTCCGTGATGGTGCCGATGATGGTCCT 31080
CATTTCCTTTGATGATCAGAGATGGCCCATGGTTTATGTCTTTGGGAGTAAGCAGTATGG 31140
TCCGGTGTCTGAGTTCTGTAGGTTCACATGGATTGACAGCGATCCTTGGTGAGATTTTTT 31200
TTGTTTTGTTTTTGTTTTTTCGAGACAGGGTTTCTCTGTGTAGCCCTGGCTGTCCTAGAA 31260
CTCACTCTGTAGACCAGGCTGGCCTCGAACTCAGAAATCCACCTGCCTCTGCCTCCCAAG 31320
TGCTGGGATCAAAGGCGTGTGCCACCATTGCCCGGCCTTGGTGAGATTATTGACTCAACC 31380
TGTAATCACCTTGGAAATGGCTTATCTCCTTTTGCTTTCTAAAAGGTACTCTGTCTTTTG 31440
TCGTCAAGATAAAAAGCTCAGCACCTTGCAGCAGCTTATCTGTTTTGCTGACGGACCTGC 31500
TGCTAAAGGTTCCCTCTCTGTGACTCATGCAGGTATGAGGTCTGCAACACGTCCATGGAT 31560
GGCTCAGAGTGTATCACTCTTCCAGTCTACTTCAGAGAGAAGAAGTCCAGGCCGTCGAGT 31620
GCTTCCTCGGGGGCTGTGCTCTATGGACAGCCCCTTCTTGTCTCTGTCCCTAAGCATAAG 31680
CTAACCCTAGAGTCTTTGTACCAGGCTGTTTGTGATCGTATCAGGTTGGTACTGGAAATG 31740
TAATCGCATGGTTATCTGAGATGAATGGGTTTTGCACTTGTTTTTCTGTTGATACCATGT 31800
ACTCAGCTTTGAGAAAAGGAACTGTGCCAGACCCTGAGAGGGGCAGAGGGGGGCAGAGGC 31860
TAACAGGACGGAGTAGCTGGCACATGTAATCTTATCAACTACCTTAGGATGGAGGGTGCT 31920
GAATGGCATTCAGCTGGAGGAGAAGAGGGCTTGCCTAGCATGCACAAACCCCTGTGTTGA 31980
CTCTCCACAGCCCCGTAAACCAGGCATGGTCACAGTGCACTCCTGTAACCTTAGCACGTG 32040
AGATGTGGAGGCAGGCGAATTGGGAGCTGAAAGTTTCCCTTAAGTATATAGCAAATTAGA 32100
GGCCAGCCTGGGGACCATAAAACACTGTTGCAATCAAATCAAATAATGCACATGACATTC 32160
CCAGAAGACCTGGAACCAGCCTTTCTCGGTATCTTTATTGACTTAGTAAGTTAACCTGTG 32220

ATCTCTTAGGTGGTTCTTAATTTCTTGAGCCATTACATTATGCTTATGGATATCTGAATA 32280
TGTGAATTACGGTACTGCTTCATAGCTGTATTGTAGAAAAAGAGTTAAATGTTAATACTA 32340
ACAATCTTTTTTTTTTTTTAAGCCGCTACATAAAACAGCCTTTGCCTGATGAGTTTCTCA 32400
GCTCACCCTTAGAGCCTGGGGCCTGCAATGGCTCTAGGAGTAGCTATGAAGGTAAGCACT 32460
CCTGCTGTGGGGCTGTGGGGCTGTGATGACCTCATTGCTCCTGATTTCCCTCCTCTTGGA 32520
AGGTGTTAAACACATGGATTTATTTGAGTCTGGACTGGTAGAGTGTAGGATCTCTTAGTC 32580
ACCATTAACCTTTGGATCTTGTAGAATGGTTTTGGATTCCTGCTGACACACATTGTGTTC 32640
TGTCTCAGACACACTGAGCCTGGAAGTACTGGGTAACGAGGCTAAGGTTATGAAGCAGGG 32700
CCAAGATCTGTTTGATTGTACAGTGCTGGTCATCAAGGCCTTGCATGGTCACTGCCTGCT 32760
GTGCTCCCTTCCCGAGCTCCCGGAGTCCGAGTTTCCTCCTTATCTGTTAGCATTTACTTC 32820
CCTAGCATGGACTCGTACACACCGTTCAGCATGGCTGGGGACCCTTTACTGCCCTAAAAG 32880
GACAGGTTTCCCGGCCAGCTAATCAGGAGGTTGAAGCTGGCTATCTTGGGCATTTATTTC 32940
TGACTTCCTAGGAAACAACTGAGTTTAGGTCATAGGAGAGAAGAAAGTCACAGTTGAAAA 33000
ACAGTGCAGTGGGTTCATTGCAAACCCCAAGGCCTCCGTCCCTTCCTGTGGCCAGTGTGC 33060
CTGGTCTGCAGTGTGTAGGTGATCATGTGCCTTTATTCCTATAAATCCATAAGTCATCTC 33120
TGAATTCTTTGGGTTTTTAAGTTTTTTGTGTATTACTCTTGTATCCGTGTGTGCATTGTC 33180
TATGCATGGGTGTGTGTGGAGGCCGAGGGTGACGTTTTGGAAGTTAGTTCTCTCCATCTG 33240
TTTTTATGCAGGTTCTGGGGAGCAAACTCAGATCCCCACACTCATGCAGCACATGCCTTT 33300
CTCTGTTGACTTTAATGCTCCCTTTGTCTCTCTAACCCGGGCTGGGCTAGAACCCACTGT 33360
GCAGGTTACGCTCGCTTGCTTGCTCTAACCCGGGCTGGGCTAGAGCCCACTGTGCAGATT 33420
ACACTCGCTTGACTTTCCTTCCCAGTCCTTCCAAGCACTGAGATTGCAAGATTTTAAATC 33480
TGACTTTAAGATTTTCTTCCTTGCAAAACTGTTCTTGGGTACTGAGTGTTCCATTTTAGT 33540
GCATCTCTCTGGGAAGCTAATGATGAATCAAGGGGAGAACCATGTGATGTGAGGGGCTGC 33600
TGCCAAAAGACTCACTTAAATTCTCAACTGTGGCGGCAGCGGTAGTGGTGGTAGTGGTGG 33660
TGGTGGTAATGGTGGTAGTGGTGCACACCTTTAATCCCAGCACTGGGAGGCAGAGGCAGG 33720
TGGATCTCTTGAGTTCCAGGCCAGCCTGGTCTATGGAGTAAATTCCAAGACAACCAGGGC 33780
TATTAAAAACAAACAAACAAACAAACAAACAAACACCTCTCTCTTGAGAAACAAAGTAAA 33840
TGCATTCATTCTCTGGGTATCTGGTACCTCATCATGGAATAACATGCTTTGTCTGCCTCT 33900
TGGGGGCAGTCACTTGTTAAATCAAGTATCTAATAACTATGGAATGTATAGGGGTGGATG 33960
CACTTCAATATATACAGCAAAGGGGCTGGAGACATGGCTCAACAGTTAAGAGCACTGACT 34020
GTTCTTCCACAGGTCCTGAGTTCAAATCCCAGCAACCACATGGTGGCTCACAACCATCCA 34080
TAATGAGATCTGATGCCCTCTTCTGGGGTGTCTGAAGACAGCTACAGTGTACTTATGTAT 34140
AATAAATAAATAAATAAATCTTTAAAAATATATATACAACAAAATATACTAGGAAATAAT 34200
AATGCACTTTAGAGTAAATAAATATTTTATGGTTTCTTTGTGGGTTATGTTGTTTTGAGG 34260
TAGGGTCTCCTGTAATCCAGACCAGCTGCAAACTTGTCATGTGGTTGTGGCTGGCCTTGA 34320
ACTCACTGTCCTTTCTCAGCCTCCTGAGTGCTTGGGTTATAGATAGATATCACCATACCT 34380
GGCTAAGTATGGGACTTCACAGATTGTTTTTCATGTGTATCAGGAGATGAAGAGGAAGAA 34440
ATGGATCATCAAGAGGAAGGAAAAGAGCAGCTTTCCGAAGTGGAAGGCAGTGGTGAGGAC 34500
GATCAGGGAGATGACCATAGCGAGAGCGCCCAAAAGGTGAAAGGCCAGCCGAGGCACAAG 34560
AGGCTCTTTACCTTCAGCCTCGTGAACTCCTGTGGAACTGCTGATATCAATTCACTGGCA 34620
ACTGACGGGAAACTCCTCAAACTCAACTGTAAGTGCTTGCTTGCCCCTTCAGGTAGACTT 34680
GGGGCACATAAGAGCTCTTTAGGAATCGTGTTACTTCATATCCCCTACATCAGGTGGTAA 34740
CTCTTTATGTAGCTCTAGGGTCAGACATTTGAGTGCCTTGGAGAACGGTTGAGGACACAC 34800
ATCTCTCAAAGGAGTTTTACCCTGTTCCCTTCGGTAGTAGAGTGGCAAGTTTAAGACAAG 34860
CCCCCCCCCCATTTTTGTTTTTGTTTTTGTTTTTCGAGACAGGGTTTCTCTGTGTAGCCC 34920
TGGCTGTCCTGGCACTCACTTTGTAGACCAGGCTGGCCTCGAACTCAGAAATTCGCCTGC 34980
CTCTGCCTCCCGAGTGCTGGGATTAAAGGCGTGCGCCACCACGCTCGCCCCCCCCCCCCA 35040
TTTTTATGTGTTGACATTTTACCTACTTGTATGTCTGTGCACCATGTTTGTGCCTGGTAT 35100
CAGTGGACACCAGAAGACAGCATTAAATCTCCTACAATGGAAATTACTGCTGTGAACCAC 35160
CAAGTGGATGCTGGGACCTGAACTCTACAAGAGCAAGAAGTATTCCTCACCATCTCTAAC 35220
CCCAGGAGGTTCTAACCCAGACTTCCTATGTAGCTGATGCTGGCCTTGAACTTCTGACCT 35280
TCCTGTCTCTATCTCTTGAGTCTTTAGATTTCAGGCATATGCCACTATGCCCAGCTGTCA 35340
CCTTAATCCTTGATGTTCATGACCTCAGAAGCAAACAGGCTTCCATTGCTCTGCGAGAGC 35400
TTTCCTCTGGCTCAGTGATTCTTAGGGTCATGCTAGGAGCATTTCTGTAGACTTGCTTCA 35460

CTCATAAGAATAGGTGCATTTGTCTCTTTGTTCCGTAAAAATACTTAATTGTATATGATG 35520
AAGCTAGAGGTAGTTGTGGATTGTAATCAGGACAGCGCTTGGGCGCTTGGGGAAAAAGCA 35580
TCTTCTAGGGGTGTGGACATGCATCATGTGCGATGCCTCCATGGGCAGGGCGGACGCTGC 35640
ACCCTTTGTTTTCTACCTTAAATACTTATGTTAGTTTCCCCCCCTTCCCCTTCCCCTCCT 35700
CCCCCCCTTGCCCTTCCTTCCCTTCATTTTTTTTTTTTTTAATTTTTTACCTTCCCTTCA 35760
TTTTAATTAAAAAAAATTTTTTTTTGTTTTGTTTTTCGAGACAGTGTTTCTCTGTATCC 35820
CTGGCCTCAAATTCACAGAGATCCATGCCTCTGCCCTCCAAGTGCTGGGTATTAAAGGTG 35880
CCTGCCACCATTACCCAGCTTTAATTTTAAAAATTAGATTCTGGGGCTGGAGAGATGGCT 35940
CAGCGGTTAAGAGCACTGACTGCTCTTCCGAAGGTCTTGAGTTCAATTACCAGCAACCAC 36000
ATGGTGGCTCACAACCATCTGTAATGGGATCCAATGCCCTCTTCTGGTGTGTCTGAAGGG 36060
AACAACAGTGTACTCACATAAAGTAAATACATAAATAAATCTTTAAAAAAAAAATTAGAT 36120
TCTATGTGTTTGAGTGTTTTCCTGCATATATGTATGTGTACTATGTGTGCCTGGGACCTG 36180
CAGAGGAACTAAAGCTGTAATAAATTATGAGCTCCCGCCTTGGGTGTTAGGGACTGAACC 36240
CTAGTGTTCTACAAGAGTAATAAATGATCTCTTACTACTGAGCCATCTTTCCAGCCCAAC 36300
CCATACCTTTCAAAAAAATTTTATGTGAATGGGTATTTTACATGTATATATATGTGTGCA 36360
CCATATGAGTGCCTGCAGAGGCCAGAAGAGGGCATCCGATCCCCCGGGACTGGAATTTCA 36420
GACAATCAAGAGCCTTTGGGTGCTGGTAATTGAACCTGGGTCTTCAGGAAGAGCAGACAG 36480
GGCACTTAGCTGTCAAGCCATCTTTCCAGCCTCCTCCAGCTTTTTAAATGTTTCATGTAT 36540
TTATCAAGACTAAGGACTGACCTCAGGAACTTGCACATACAAGGCAGAGCTCTAGCACTC 36600
ACTGTAGTTCCTGGTTCCCTTTACACAGTCTTTTATCTGAGTTTTTATTGCTGCCTGTCT 36660
GTAAGCTTACTACATGCATGCCTCGTGCTCACAGAGCCCAGGAGAGGTTGTCAGATCCCC 36720
AGAAACTAGAACTACAGACTACTGTGAGCCACTGTGTGGGTGTAGGGAGCTGAGCTTGGG 36780
TACTCTATAAGAGCAGAAAATGCTCTTAATCTCTGAGCCATATCTACACTCCCATTTACA 36840
TTTTTATTTTGTTGTTTTGTTTTGTTTTTCGAGACATGGTTTCTCTGTGTAGCTCTGGCT 36900
GTCCTGGAACTCACTCTGTAGACCAGGCTGGCCTCAAACTCAGAAATCCGCCTGTCAGAT 36960
TCTGGGTTTTGGGGGTTATCTGTAATATTAGGGGAGACTTTGAAGCTGTGGGAGGTTGAG 37020
ATTCAAAAAGGCTGCCCCAGGTTCTAGCATTGCTCTCCTCAGTGGCTGGTGGTGGGTGAG 37080
CAGGTGTGGTGCTGCTGCCTGAGTGTGGCCCTTCTCACTTGGAGAGCCTCTCCTGGTGTT 37140
GGTCATTCCATAATGAGTGACACTATGGTCCCCTCCACAGCTCGATCCACACTGGCCATT 37200
GACTGGGACAGTGAGACCCGAAGCCTTTACTTTGATGAGCAAGAATCTGAGGTATGTCAT 37260
TTGGGGTTTGTGGTCAGTGCTGTGTTCTGGCTTCTGCTCTGAAAATACATTGATGTTCAT 37320
CTCATTGATTGGACAACAGTTAAATTTTCCTTATGACTTGAAATGAAAAGCACCCAAGAC 37380
TGCCATTCTGTGTGCAGAGACTGCAGAAAGCACGGTATATTCCAGGCAGAGCTTGAGCAT 37440
CCCGTGAGCCGCCAGGAGGCTTCACAGCAGAACTCAGTGCTTCTCTCCGCAGGCCTGTGA 37500
GAAGCACCTGAGCATGTCACAGCCGCAGAAGAAGAAGAAGGCGGCAGTAGCCCTGCGAGA 37560
GTGCATCGAGCTCTTCACCACCATGGAGACCCTTGGGGAGCATGACCCCTGGTATGCAAG 37620
TCCCCGAGCTCTGAGCAGGGACCTATGTACCCTAGGCTCTGAGGTTTGCTGGGACCACCT 37680
GTGTGCACAGAAACCTCCTCCTGTGCCTGGACAGGTCATCTTTTCTTAAGCAGTTCTGTA 37740
AATATTTGTGTGTGAGAGACGACTGTTTGCTGGACTGAGAAAAGCCACTGCTCTGATGAG 37800
AAGGCAGGGGTTTGTTCATAAGTCCTGGGGTTGAGTCCTGCCTGGAAGGTCTCCTTGTGA 37860
CCTCCTGCTGCCCTGGCTAAGTAAGTCTCTTCTGTCTTTAGGTACTGTCCCACCTGTAAG 37920
AAGCACCAGCAGGCAACAAAGAAGTTTGACTTGTGGTCTTTGCCCAAGATCCTGGTGGTT 37980
CACCTCAAGCGTTTCTCCTATAACAGATACTGGCGGGATAAACTTGACACCGTGGTGGAG 38040
TTCCCAGTCAGGTGGGTGTGTGCACAGGTGCTGAGGAGGGAGGTGAGAAGATAGATGCTG 38100
GAGTTAGATTGTAGTACTTGGCTCTGGGAGGTTGACTTGTGACTTGTGGAGCAATCCTTC 38160
CTTCCCCTACTCGTGGGAAAAGTCACTACTAGAAGAGTCCTGCTGATTGTGTCACTGCTG 38220
ATCTGCTTGTATCCTCACACTGGTCTGTGGCCTGTCCTGGCCGTCATTGCTGTCAGGGAA 38280
CTTGGCTTCAGCCAGGACAGCACAGGCTCAGTTCTGTTTCTTATTAACAGGGCGAGGCAA 38340
CCAGGTTAGCCTACCCGCGGGAGCTGATGTTCAGCTGGGTGTGGTGTCGCATGCTTGTAA 38400
TTCCTTGCTCGGGCAGGAGGATCATTGATTGAGTTTAAGGCCACCTTTGTAGTTAAAATT 38460
TTGTTTACTTTTTTTGTTTGTTTGTTTTGGTTTTTTGAGACAGAGTTTCTCTGTGCAGCC 38520
TGGGCAATCCTGGAACTCACTCTGTCATTCAGTGCCAGTCTTAGCCCCTGAGTGCTGGGA 38580
TTGAAGCCTGTGCCTCAGCCTGACTGTTATTTTTTACCTTTTCTTTATTTTTGAAAATTG 38640
TATAATGGGCTGGTGAGACTAGGCAGCTCATTGAGAATCCTTGACCCCCAAGCCTGCTGA 38700

CCTGGATTTGTTCCCAGACTCAGGGTAGAAGCAGAGAACCAACTCCTAAAATTTGTCCCC 38760
AGCTCTTTCTGCAGTGTGGATTTCAGTGCAGAATACAGAGGTAATGCTAACATATCCAGT 38820
TTTTTGGCCCCCAACAACAATTCCCTAATGTCAGCAGAAAGTAATTATGCACGATTGAGT 38880
CACTCCCCTTATTCCTTTGAAGAGCAGGCAGATTTCTGAGTTCGAGGCCAGCCTGGTCCA 38940
CAGAGTGAGTTCCAGGACAGCCAGGGCTACACAGAGAAACCCTGTCTCGAAACACCAAAA 39000
AAAAAAAAAAAAAAAACCAAAAACCAAAAAACAAAATGATGGAATGATAGCCTAGTCAGGT 39060
TAGAGCAGGTTCTTGCTCTCCAGTCTCTGTAGGATCATCTGGGGCTAGGGACATGCAGTT 39120
AGCAGTTGGAGCAATTTGCAGCCTTTAATTGACTAGAAATCTGCTCAAGCAGATACATAT 39180
TAAGGGTAGAAGCTAAAGCTAAGGAGTTTGGGGAAATTTTTTTCTTAGGTGTGCATTAGA 39240
AACTTTTAGGCCCGTGGTCCTAGTAGTTGGGAACGGGAAGTCCCAAGACCAGGGAAAGGG 39300
GGAGAGGAGAGATCCCAGGCTCAGGACCAGGCAGGTGAGGAAACAGGCTGTTGACTGAGT 39360
CCATCTGGAGTCCATTTGACCTGTGAGAGGTCGATTCAAGGCGATTCCCTGAAGCTTCCA 39420
AGAATTGTTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTATTGGT 39480
TTTTTCAAGACAGGGTTTCTCTGTGTAGCCCTGGCTGTCCTGAAGCTCACTTTGTAGACC 39540
AGGCTGGCCTCGAACTCAGAAATCTGCCTGCCTCTGCCTCCCGAGTGCTGGGATTAAAGG 39600
CGTGCGCCACCACGCCTGGCTAGAATTATTTTTTTAAGTTTATAAAAAGTGATGTTTTAG 39660
GTATGTTAGTGTTAACCACTGTTTGTAATCTGTACTGAGAGCCACATGGTAGAGAGAGCA 39720
GCAGACTCATACCTTGCAGTCATAATAAAAGCTTAAGCACCCAGAAATAATTTGTTCTTA 39780
CCCAGTTACTTTCTATGAGACACAGGTGGTTGAGTCCTGACAGTAGTCACTGTAAAGCGG 39840
GCTCTTTACATGAAAGGGCGGAAGAGCTTACTTGAGACCTATTCTGTGAGTTTACCCCGT 39900
TTTAGAGTCTGGCAGCAAGGTAAACTGTCTTTCTCAGCTAGAAGCTTAGGAGATGAGGCC 39960
TGAATTTTGTACACATGAAGAGAACTGAGAACAGCTGAAGCCTTGGCTGCTGCTGTTAAA 40020
CTGACAAATCTGCCCTGAGCCTCGTGGTCAGCACCAGAGACCTGAGAAAGATAAATTTCA 40080
CCTGAGTAAGCAGGAAGTATAGACCAAGCAGCTTCCAAATCTATTTAAAAAGTGACAGAG 40140
ATCCAGGTCACCCTAGGCTCCTCCGTGGTCAGTGGAGGGCGGCATGGCTCTGTAGCCACC 40200
AGGCATAGACGGTCTGACAGACTTTCCTATACAGTAGGAGCTTGGGGGAACTGGCCTACT 40260
TGTCTAGGCAAAGAGGGAGTCGACTCCTCAGTGACCTGCTTGTCCACAGTTTGGACACTG 40320
CTGTGAACAGTCAAGGTGAAGGGCAGTGTTTGGCCCAGTGGCCTGCTTTGCTGCATTTAG 40380
AGTAAGCCTGGGGGGATGGTCTTCTGTGCCCATTATCTTCTTGGGAGGATGCTAGGAGCC 40440
GACTTGTCTCTCTGTCATGAAAAACTTTTCTTATTAAACATTTTAAATACCATATTCTGC 40500
AGATCTCTGAAGTGTTTGAGGACCACCTGTTAACTGTACCTAAACAAACACACTCTGTCT 40560
CTGGTCACTTGTTTTCTGCTTAACTTTGAAAACATGCAGGAACCTTATTCCTATAATTAC 40620
TGTAGTACTTAGCATGATTACAAGTATACTTCTGAACACATTAAAGAATTTGATCATTTA 40680
TAAGGTGACTGACCATGAACTTGCATGTCTTAATTATTCTTAACTGTTAGTTTATGTAAG 40740
GTCACTATTTTGCAGCTTTATCCATTTAGGTTAAGCCTTAAAATTTACAATTTTTAAATT 40800
AAAGCCCTATAAAACTTTTAATCCTAGTTAAGGTCCACAGGGAGGCTGGGAGCTTCCCTG 40860
ACCCTGCCGTGGGCACAGGGAAGAAGCCGCAACAGCCACTCCCTGGGTGAACAGGCAAAA 40920
CATTTATTCTAGACTGCCGTGCAATTTTCAGGGTTAAATGGGCCCTTGAGAAAGGACTAA 40980
AGGGACCTAGAGGCTTGCATTGACCATGAAAAATCAATAGATACCAGTCATGTCAGTGGC 41040
TTGCCTCAGGTTTGAAACTTAACAAAAGAAAATTTCAAAATACAGTAAAGGTTTCCAGGG 41100
CCTGTCTGACATTTAAATTCCTATGGCCTTGTTCGCACTGTGGTTAGTCTCAGCCTTCGC 41160
GTGCAGATGCTGAATTCCTCTTTGGCATCTTTTACCCCGCTTCAACCTCATGGTATGATA 41220
GCACCATGACAGCCTTCTCCCAGTTTCTGCTCCCACAAACCATCACTGTGCCTTGGCCTA 41280
GTCTTCTCGCATAGTTTAGTGTTCTGTGGGGTGAGGGGGGTGCTGGGTTACACTGACCTT 41340
TTGAATTTACCATATTTATGGGCTATTATGTTTTCACTCAGAGCTCTGAACATGTCCGAG 41400
TTTGTCTGTGACCGGTCAGCAAGGCCTTATGTTTATGACCTAATTGCTGTGTCCAATCAC 41460
TATGGAGCCATGGGGGTTGGTCACTGTAAGTATTGGCATCTGCTTCCTTCTGAGGTATGG 41520
AAACGCCCTACCTACTGTGCCTCAGCCTCTTCCTGGATGCCAGCCCTGACGTAGCACCTG 41580
GCTTGTTTTCAGTCTCCGACTTTGGGCTTTCCCAGCCCCTCTCTCCTCCCTACCTTCCTC 41640
AGTGGTGGTAACGATCAAGACAGTCCTGCTGATCGCTTATGCTGGCCTGGGGCTCACCAT 41700
GTAGCTCAGCCAGGGTGGCCTTACTTTCCTGGCACAGTGTGAGCCGCCACTCTGGGTTAC 41760
TCTGGAAACTCTACTAGAGCCAGGACATGACAAAACATTGTTCTCTTTGTTTATTTTGGT 41820
TTTTCAAGTCAGGGTTTCTCTGTGTAACCCCAGCTGTCGTGGGACTCACTCTGTAGACCA 41880
GTGTTTCCGAGCGGGGCACTGTGGTTCAGGTACTCGGGAAAAGAATTCTTAACCGTTTCC 41940

AAAGAAAGTTAATGTTAAAGAGACGTCCCTAATAATTCACCCAACTACACAAAATTAGTG 42000
ACTGTAATCACAGAAGTATCGCAGCGCTGCAAAACCAAAACGTGGAGAAAATGCAAGGGA 42060
AGGAGAGATGCCAAAATGAGTTTACTTTTGGTTTGGTTTTGAGACAGAGCCTCACATCCA 42120
GACCAGGCAGGCTTGTGACTTGCTAGGTAGGCCTCGATGGTCCTGAACCCACGTTCTTTC 42180
CACCCTAGGCACCTAAGTGCTAAGATACAGGTACACATCTCCATATCTGGGAAAATAGTA 42240
CGTTTTTTTCTTTTCTCTATTACCCAGAATTTCTATAACATTTAGGTTCACTAAAACTGT 42300
GTTTTAGTTACACCTGAACCTGCTTTAGGGGCATTGTTTCACTCCCCAGTGCTGGGATGG 42360
GACTTAGAGCCTTGTGCATCCTAGGTAAGTAGTGTGGCCGCCAAGCTTCACCTGCAGCCC 42420
CATCCACTCCAGAGGTTCAGTCACAGCTGTGGAGAGAAGAAAATGTGCACACCCAGCTCT 42480
GCAGACCAGGGAGCACTGCCTTCCTGACCAAAGCGCCCCTGCTGGCTCCACAGGCACCGG 42540
GTGTTCTGTGGAGATGGGGGTGGAGCCTTCTGCTTTGGTTTTGTTTTTTTAATTTTAGC 42600
AGCGTCAGCTCTCTGAGTCCTTCTGTGGCTCTGGATTATACTCGTGGCCATCTATGTTAG 42660
AGATTCAGATGAAAATCTCAGGCGCATGCCCTTTCCTTGTGTTATGTCTCCAGCACTGAT 42720
CACAGTCATTAAATAGACACCAGGGCAGCTGGGGAGAAAGCTCAGTAGTGGATCACTGTT 42780
CTTTACAGGACCTGAGCTAATTCCCAGGACCTAGGTCAGGTGGCTCACAACTGCCTGGAA 42840
CTCTGTGACACCTGTCAGTACCCTCCATGTAATCAAATAATAAATCTTTATTTAAAAATA 42900
GACAAGTCTGAGTGTGGTGGCATCTGCGGTTAATTCCAGCACTCGGGAGGCAGAAGCAGG 42960
TGGATCTGAGTTCAGGGCCAGTGTGGTCCACAGAGTTTCAGAAATACACACTAGCGTGAC 43020
AAGGTGGCTCAGCAGGTAAAGATCCTGGCCACACAAGCCTGGTGACCTGAGTTCAATTCC 43080
AAGAACTCTAAAAAAGTAGAGGGAGAGAACTCACCCCACAGCGTTAACCTCCACACACAA 43140
ATACCATGGCACGTGTACACACACACACACACACACAGCTTTTAAAATTAAATTTAAGGGCT 43200
GGTGAGATGGCTCAGCGGGTAAGAGCACTGACTGCTCTTCCGAAGGTCTCGAGTTCAAAT 43260
CCCAGCAACCATATGGTGGCTCACAACCATCTGTAACAAGATCTGACTCCCTTTTCTGGA 43320
ATGTCTGAAGACAGCTATATTGTACTTATATATAATAAATAAATCTTTTTTAAAAATTAA 43380
ATTTAAAATTATTTTAAATATTAACATTAAGTAGGCTGGTTTTATTGATTAGCATCCCAA 43440
AATTATAGGACACCCTGTGCTCACCCTTTCAAGTTTGTCTGAAAGACCCAGGTTGGGAAA 43500
AAATGGTGAATTATTTCCACTGGGCTGCAGTGCCACACGCCTTTAATCTCTGTGGGAGGC 43560
AGAGGCAGAGGATCTCTGAGTTCAGGACCACCGGGGGACACAGAGAAACTGTGTTTTGAA 43620
AAGCCAAAAACAAACAAAATGTACAAAACCATGAGCCATAGGGGCCGGGGATATTATTTA 43680
GCAGCAGAGTGTGTGCCTAGCAAGTACAAGGCCCTGGGTTTGGATCCCCTAACCTCTGGG 43740
GAAGAAATAAAAAGAACTATAAATTATCGACTATAAATGATCATTACTTATATGTATGTG 43800
CATATTTTAAAAATTATTTCTATAATCTGGTGGTGGTGGTTTTCTGTTTTGAGATGTGGT 43860
TTCACTGTGCAACCTACCTAGACTTGGCCAGGATGGCCTTGAACTTGGGAATAGTACTTC 43920
TGTCTCTGGTTCCTGAGTGGATCCACACTGGGTATACATTATTCTTAGTTTTAGTGTTAA 43980
CATTTTGGGTATGAGTATGTGCGAGTATTTAGTGCTAACCTAATAAGGTTACAAGATTAT 44040
TTTTTTTTTGGTGGCCATTAGAGCATTTTGTTTGTAACTTGGCAGACACTGCATATGCGA 44100
AGAACAGACTGAACGGGAAATGGTATTACTTTGATGATAGCAGTGTGTCCCTGGCCTCTG 44160
AGGACCAGATAGTGGTGAGTACTTTCCCAGGTGCCTGCTCCTCCCTGGGCTCTCAACAAA 44220
TCAACCGAAAAGGTTCCCTCAAGCCAGACGGTTACCTCCTTATCTGCTAGGGCTCTGAGT 44280
GGAGCAGAGGGGACTGCCATTCATCCTAGATCTTTCCTAACGGAGAAGCTCAGTTTTCCT 44340
AGAGGTATCACAGTGCATCTATCTGCATTTGCTAGAGCCATGGCAGGGGCATCAGGAAGC 44400
AGCTAATAAATAAAGGTGGCAGGTTCCCGAGAGTACAGCTGTCCACTCAGTCACTCAGTG 44460
AGTGTGCTTGTTTATTGAGATAGTATCTCTATATCTCTGCCTGCTCTGAAACTCATTTTG 44520
TAAGTTTGTAGCCTCAAACTTAGGGTAATTCTCCTACCTCTGCTTCTTGAGTGCCGGTAT 44580
TACAGATATCCATCCCCACATTCAGTTTAAGATTGCTTTTTTTTTTTTTTAAGATAGGGT 44640
TTCACTCTGAAGCCCAGGCTGGCCTTGACCTTGGAAAAAGAATCCTCCTGTCAGGCCCCC 44700
CTCGCCTTGCTGGGATCACAGGTGCTAGGAAAGGGGTGGGGTGAGTTTATTATCAGTGAA 44760
TGGACCCCATTCCTTGTGTTTTAGACGAAAGCCGCCTACGTGTTGTTCTATCAGCGTCGG 44820
GATGACGAATGCTCCAGCACATCTTCGCTCGGCAGTTTCCCCGGTTCTGATGGAGGGGTG 44880
AAGCTGAGCAGCTCACACCAGGGCATGGGGGATGAGGAGGCTTACAACATGGACACCAAC 44940
TGATGTGGCCTCTGTGACCCTGCCACCCTCCCAACACCAGCATCCTCCCCCAGGAGAATA 45000
TCTGTGGCACTCTTGAAGACTGCTCATATCAATCTCAAACCAGATGAGGACATTACCTCC 45060
TTTTATGAGAAGAGGAAGGAAAGAGAAACCCTGCTTTCTGAGAAGGGCTCTATGTTAAGA 45120
GGCTGGATTATTTTTTCTTTTAAACGGGTGGTGAGAGAAAAACCTGTGGAGCTCAATTGT 45180

GGGCTGGGGTTTGGTGCACAGTGGCATGTGTCTCATGGATTCCAAAGAGTGGAGAGGAAG 45240
ACCCTGCAATGAGCATGGTGCAGCCATGAAGACCCACCTGCAGGGCTGCTCCGAACTGGG 45300
TTTTGGATTCTGGTGCCCACACTGCTCACCCCAGTAGTCTGGCCACAGTAAGACCAAGCC 45360
TCCAGCAAATTTGATGTTGTCCTCTAGTTCTTTGCTGTTCATTGAGATGGGCTTATAGTT 45420
TTCTAAACTGTCTCTGAAACCCGTTTCTATTCTCTTGTTCTCTCAGAATTTTCCAGCCGC 45480
CTTCCCGGTTATGAAATGAAGTTAGTATTGGGCTATTGTTTTCAGTTTCTCACAGACTCC 45540
TAGAAGTTCTGTTTTGTCCCCCACTGCCCTTGATAAAGTGATTCTATGTGGAACTTTCAT 45600
TAGCTCTCCCTTAAAAGAAGTCTATTCATAAAACTGGGTGCAGTGTGTGGCATTGTTTCT 45660
CCACGTTAGCTTCAGAAGAACACAATGGGTATTTTTCAGATCTCCAAACATTCTAGTTAA 45720
TAAGGTTTAAGAAAATACTTTTTCGGTAATAGACCCTCTCCACGACCGGATTCTCTTCCT 45780
TAACCCTTCCCTTACTCCCCTAGTGGCAAAGCCGGCAACCTAACAGGCCAGCTCTAGGGC 45840
CTTTGTTCCCATTGGGGTCATGTGAGTCTTCACCTCGACCTTTAACCTAGTAGGTGGGGG 45900
TTGGGAATCGCTGGTCTTGGAACGTGGACTCAGTGACCGTTGGGCACTCTTGTGCGGCAA 45960
CGGCTGGGTGAGGACTCAAACGGGGCCCCGTCCTGTCTAGGCTGGCCATAGTTTGGGGGC 46020
CCAGACAGCAACTGGGCGCAGTAGTGGGTGGAGGGAAAAGGCTTAGCGGCCTACAGGCTA 46080
AGGAAGCTCTGGGCGGTGAGGCGGGGCGTCTCGGGCCCTCCCCCGTGACCCTGGGTCCTC 46140
TGAGAATCCCGTGACCCTGGGTCCTCTGAGAATGGCGGGTGATCCGCCCCTCCGAGGAAA 46200
AGAACGAAGGATGATGTTGTATAACCGTAGGTCATACTCGAAATCGAGGTTGAATCCCAG 46260
CCTCCTGTATAGAAGCATCATTTTCCCGTTGCTCCGTGCATTTACAATTATCACTTAATA 46320
AAAGAAAGCATGGTCGCTTTTAGGGAAATGTCTGAAAAACTACGAAGATGCAGAAAGGAG 46380
CTGACTGCTGCTATAGACCGAGCCTTTGAAGGAGTCAGGCATTCTCAGGAGTGCACAGCC 46440
CAGCAGAGGCTGGACGCCCCGTCGCTCACCTCCCAGCCGGTACACAGGCTCCTCTGCAGA 46500
AACCCACTGGCTGCCTGCCCCTCTGCTGCCCCATACTCTGGTGCCTCGTGTGCTCCTGAG 46560
AGTGAGAACCCGGCCTTCGGGACACACCATATTCCGGTTAATTCAAAACTTCAGCAGCCT 46620
CTATACCCCAAAAGGAAACCTCTGACCAGCAAGGAAAATGTCTTGATGCAGTCTTCCATT 46680
TTGGCACGTGACAGACAGTTTTGGAGAGCTGCAGGTGATGGGGAAGACTGGAGAAAAGAT 46740
AGTTTAAGGTGATTTTAATTCGATAGCTTTCTAGAAACCGTTGTTTGTTTAAAAAAACAA 46800
ACAAACAAACAGCACTCGGGAGGCAGAGGCAGGCAGATTTCTTGAGTTCGAGGACAGCCA 46860
GGGCTATACAGAGAAACCCTGTCTACAAACAAACAAAAAAGTTGAAGATCTCAGATCTAA 46920
TGACTAAGCTAGCCTCCTCTTTGGAGCGTTGACTGTACACAAGAAATTCATGGTTTTTAG 46980
AATGTGGGCCGGCAAACAGTTTTTATGTGTTTAGGTTGAAAGCCTGGTTATTTTCCAAAC 47040
AGTCATCGAAACTACTGTACCAAAGGGTAATAGGTGTTGAGCTAGTTTTAAATTGTGGGT 47100
AGTGCGAAGGACTGAATCCAAGGCCTCACACATGCTGGGGCAGGCACTCCCGCTGATCTA 47160
CAAGCCTTTGCAATGTTTTGTCTGGTTTGATACTTTAAGGATTTTCCCTAAGTTTTTTTA 47220
ATGTTTGTCTTTTGTACATGTAGATGTTTTGCCTGCAGAGGCCAGAAGGCTGTGGCATCA 47280
CATCCCCTGGAACTAGAGTTACTGACAGTTGTTAGGTACCAGGTGGGATCCGAATTC 47337

Appendix 2: MatInspector analysis of 1.1 kb Eco RI/Sac I fragment containing the *Unp* promoter.

```
MatInspector Release 1.0   September 1995          Mon Mar 18 20:54:08 1996

    Solution parameters:
    --------------------
    sequence file:     /mnt/quandt/matinspector/seq/unp.seq
    selected matrices:
              V$AHR_01              (AhR                          )  re: < 0.01
                  core sim:   0.80      matrix sim: 0.85
              V$AP1FJ_Q2            (AP-1                         )  re:   2.45
                  core sim:   0.80      matrix sim: 0.85
              V$AP1_Q2              (AP-1                         )  re:   1.82
                  core sim:   0.80      matrix sim: 0.85
              V$AP1_Q4              (AP-1                         )  re:   2.48
                  core sim:   0.80      matrix sim: 0.85
              V$AP1_Q6              (AP-1                         )  re:   5.28
                  core sim:   0.80      matrix sim: 0.85
              V$AP2_Q6              (AP-2                         )  re:   4.78
                  core sim:   0.80      matrix sim: 0.85
              V$AP4_01              (AP-4                         )  re:   0.01
                  core sim:   0.80      matrix sim: 0.85
              V$AP4_Q5              (AP-4                         )  re:   0.96
                  core sim:   0.80      matrix sim: 0.85
              V$AP4_Q6              (AP-4                         )  re:   0.50
                  core sim:   0.80      matrix sim: 0.85
              V$ARP1_01             (ARP-1                        )  re:   0.07
                  core sim:   0.80      matrix sim: 0.85
              V$ATF_01              (ATF                          )  re:   0.34
                  core sim:   0.80      matrix sim: 0.85
              V$BRACH_01            (Brachyury                    )  re: < 0.01
                  core sim:   0.80      matrix sim: 0.85
              V$BRN2_01             (Brn-2                        )  re:   0.99
                  core sim:   0.80      matrix sim: 0.85
              V$CDPCR1_01           (CDP CR1                      )  re:   9.83
                  core sim:   0.80      matrix sim: 0.85
              V$CDPCR3HD_01         (CDP CR3+HD                   )  re:   3.06
                  core sim:   0.80      matrix sim: 0.85
              V$CDPCR3_01           (CDP CR3                      )  re:   0.12
                  core sim:   0.80      matrix sim: 0.85
              V$CDP_01              (CDP                          )  re:   0.04
                  core sim:   0.80      matrix sim: 0.85
              V$CDP_02              (CDP                          )  re:   0.12
                  core sim:   0.80      matrix sim: 0.85
              V$CEBPB_01            (C/EBPbeta                    )  re:   2.07
                  core sim:   0.80      matrix sim: 0.85
              V$CEBPB_02            (C/EBPbeta                    )  re:   5.91
                  core sim:   0.80      matrix sim: 0.85
              V$CETS1P54_01         (c-Ets-1(p54)                 )  re:   1.53
                  core sim:   0.80      matrix sim: 0.85
              V$CETS1P54_02         (c-Ets-1(p54)                 )  re:   7.63
                  core sim:   0.80      matrix sim: 0.85
              V$CLOX_01             (Clox                         )  re:   0.12
                  core sim:   0.80      matrix sim: 0.85
              V$CMYB_01             (c-Myb                        )  re:   4.78
                  core sim:   0.80      matrix sim: 0.85
              V$COMP1_01            (COMP1                        )  re:   0.56
                  core sim:   0.80      matrix sim: 0.85
              V$COUP_01             (COUP-TF / HNF-4              )  re:   0.19
                  core sim:   0.80      matrix sim: 0.85
              V$CP2_01              (CP2                          )  re:   0.39
                  core sim:   0.80      matrix sim: 0.85
              V$CREBP1CJUN_01       (CRE-BP1/c-Jun                )  re:   0.22
                  core sim:   0.80      matrix sim: 0.85
              V$CREBP1_01           (CRE-BP1                      )  re:   0.15
                  core sim:   0.80      matrix sim: 0.85
```

```
V$CREBP1_Q2           (CRE-BP1                          )  re:    0.09
    core sim:   0.80      matrix sim: 0.85
V$CREB_01             (CREB                             )  re:    0.40
    core sim:   0.80      matrix sim: 0.85
V$CREB_02             (CREB                             )  re:    1.12
    core sim:   0.80      matrix sim: 0.85
V$CREB_Q2             (CREB                             )  re:    0.34
    core sim:   0.80      matrix sim: 0.85
V$CREB_Q4             (CREB                             )  re:    0.34
    core sim:   0.80      matrix sim: 0.85
V$CREL_01             (c-Rel                            )  re:    2.74
    core sim:   0.80      matrix sim: 0.85
V$DELTAEF1_01         (deltaEF1                         )  re:    2.42
    core sim:   0.80      matrix sim: 0.85
V$E2F_01              (E2F                              )  re:  < 0.01
    core sim:   0.80      matrix sim: 0.85
V$E2F_02              (E2F                              )  re:    0.13
    core sim:   0.80      matrix sim: 0.85
V$E2F_Q6              (E2F                              )  re:    0.06
    core sim:   0.80      matrix sim: 0.85
V$E2_01               (E2                               )  re:    0.29
    core sim:   0.80      matrix sim: 0.85
V$E2_Q6               (E2                               )  re:    0.40
    core sim:   0.80      matrix sim: 0.85
V$E47_01              (E47                              )  re:    0.11
    core sim:   0.80      matrix sim: 0.85
V$E47_02              (E47                              )  re:    0.27
    core sim:   0.80      matrix sim: 0.85
V$E4BP4_01            (E4BP4                            )  re:    0.07
    core sim:   0.80      matrix sim: 0.85
V$ELK1_01             (Elk-1                            )  re:    0.06
    core sim:   0.80      matrix sim: 0.85
V$ELK1_02             (Elk-1                            )  re:    1.93
    core sim:   0.80      matrix sim: 0.85
V$ER_Q6               (ER                               )  re:    1.73
    core sim:   0.80      matrix sim: 0.85
V$EVI1_01             (Evi-1                            )  re:  < 0.01
    core sim:   0.80      matrix sim: 0.85
V$EVI1_02             (Evi-1                            )  re:    0.02
    core sim:   0.80      matrix sim: 0.85
V$EVI1_03             (Evi-1                            )  re:  < 0.01
    core sim:   0.80      matrix sim: 0.85
V$EVI1_04             (Evi-1                            )  re:    0.92
    core sim:   0.80      matrix sim: 0.85
V$EVI1_05             (Evi-1                            )  re:    0.16
    core sim:   0.80      matrix sim: 0.85
V$EVI1_06             (Evi-1                            )  re:    0.02
    core sim:   0.80      matrix sim: 0.85
V$GATA1_02            (GATA-1                           )  re:    2.27
    core sim:   0.80      matrix sim: 0.85
V$GATA1_03            (GATA-1                           )  re:    2.08
    core sim:   0.80      matrix sim: 0.85
V$GATA1_04            (GATA-1                           )  re:    1.82
    core sim:   0.80      matrix sim: 0.85
V$GR_Q6               (GR                               )  re:    6.47
    core sim:   0.80      matrix sim: 0.85
V$HEN1_01             (HEN1                             )  re:    0.12
    core sim:   0.80      matrix sim: 0.85
V$HEN1_02             (HEN1                             )  re:    0.07
    core sim:   0.80      matrix sim: 0.85
V$HFH1_01             (HFH-1                            )  re:    0.12
    core sim:   0.80      matrix sim: 0.85
V$HFH2_01             (HFH-2                            )  re:    2.03
    core sim:   0.80      matrix sim: 0.85
V$HNF1_01             (HNF-1                            )  re:    0.52
    core sim:   0.80      matrix sim: 0.85
V$HNF3B_01            (HNF-3beta                        )  re:    1.98
    core sim:   0.80      matrix sim: 0.85
```

```
V$HNF4_01            (HNF-4                         )  re:    5.75
    core sim:   0.80     matrix sim: 0.85
V$HOX13_01           (Hox-1.3                       )  re:    0.02
    core sim:   0.80     matrix sim: 0.85
V$HSF1_01            (HSF1                          )  re:    9.96
    core sim:   0.80     matrix sim: 0.85
V$IK1_01             (Ik-1                          )  re:    0.86
    core sim:   0.80     matrix sim: 0.85
V$IK2_01             (Ik-2                          )  re:    3.95
    core sim:   0.80     matrix sim: 0.85
V$IK3_01             (Ik-3                          )  re:    0.16
    core sim:   0.80     matrix sim: 0.85
V$IRF1_01            (IRF-1                         )  re:    0.02
    core sim:   0.80     matrix sim: 0.85
V$IRF2_01            (IRF-2                         )  re: <  0.01
    core sim:   0.80     matrix sim: 0.85
V$LYF1_01            (Lyf-1                         )  re:    3.51
    core sim:   0.80     matrix sim: 0.85
V$MAX_01             (Max                           )  re:    0.11
    core sim:   0.80     matrix sim: 0.85
V$MEF2_01            (MEF-2                         )  re: <  0.01
    core sim:   0.80     matrix sim: 0.85
V$MYCMAX_01          (c-Myc/Max                     )  re:    0.05
    core sim:   0.80     matrix sim: 0.85
V$MYCMAX_02          (c-Myc/Max                     )  re:    1.67
    core sim:   0.80     matrix sim: 0.85
V$MYOD_01            (MyoD                          )  re:    0.21
    core sim:   0.80     matrix sim: 0.85
V$MYOD_Q6            (MyoD                          )  re:    0.96
    core sim:   0.80     matrix sim: 0.85
V$MZF1_01            (MZF1                          )  re:    3.84
    core sim:   0.80     matrix sim: 0.85
V$MZF1_02            (MZF1                          )  re:    8.22
    core sim:   0.80     matrix sim: 0.85
V$NF1_01             (NF-1                          )  re: <  0.01
    core sim:   0.80     matrix sim: 0.85
V$NF1_Q6             (NF-1                          )  re:    4.11
    core sim:   0.80     matrix sim: 0.85
V$NFE2_01            (NF-E2                         )  re:    0.12
    core sim:   0.80     matrix sim: 0.85
V$NFKAPPAB50_01      (NF-kappaB (p50)               )  re:    0.05
    core sim:   0.80     matrix sim: 0.85
V$NFKAPPAB65_01      (NF-kappaB (p65)               )  re:    0.08
    core sim:   0.80     matrix sim: 0.85
V$NFKAPPAB_01        (NF-kappaB                     )  re:    0.26
    core sim:   0.80     matrix sim: 0.85
V$NFKB_Q6            (NF-kappaB                     )  re:    0.28
    core sim:   0.80     matrix sim: 0.85
V$NFY_Q6             (NF-Y                          )  re:    0.70
    core sim:   0.80     matrix sim: 0.85
V$NMYC_01            (N-Myc                         )  re:    1.64
    core sim:   0.80     matrix sim: 0.85
V$NRF2_01            (NRF-2                         )  re:    0.16
    core sim:   0.80     matrix sim: 0.85
V$OCT1_01            (Oct-1                         )  re:    0.04
    core sim:   0.80     matrix sim: 0.85
V$OCT1_02            (Oct-1                         )  re:    2.28
    core sim:   0.80     matrix sim: 0.85
V$OCT1_05            (Oct-1                         )  re:    0.03
    core sim:   0.80     matrix sim: 0.85
V$OCT1_06            (Oct-1                         )  re:    0.97
    core sim:   0.80     matrix sim: 0.85
V$OCT1_Q6            (Oct-1                         )  re:    4.82
    core sim:   0.80     matrix sim: 0.85
V$P53_01             (p53                           )  re: <  0.01
    core sim:   0.80     matrix sim: 0.85
V$PAX5_01            (BSAP                          )  re:    0.04
    core sim:   0.80     matrix sim: 0.85
```

```
V$PAX5_02             (BSAP                              )  re: < 0.01
    core sim:    0.80     matrix sim: 0.85
V$PAX6_01             (Pax-6                             )  re:   1.21
    core sim:    0.80     matrix sim: 0.85
V$PBX1_02             (Pbx1b                             )  re:   0.05
    core sim:    0.80     matrix sim: 0.85
V$RORA1_01            (RORalpha1                         )  re:   0.18
    core sim:    0.80     matrix sim: 0.85
V$RORA2_01            (RORalpha2                         )  re:   0.01
    core sim:    0.80     matrix sim: 0.85
V$RSRFC4_01           (RSRFC4                            )  re:   0.02
    core sim:    0.80     matrix sim: 0.85
V$S8_01               (S8                                )  re:   1.73
    core sim:    0.80     matrix sim: 0.85
V$SOX5_01             (Sox-5                             )  re:   1.10
    core sim:    0.80     matrix sim: 0.85
V$SP1_01              (Sp1                               )  re:   7.52
    core sim:    0.80     matrix sim: 0.85
V$SP1_Q6              (Sp1                               )  re:   1.70
    core sim:    0.80     matrix sim: 0.85
V$SREBP1_01           (SREBP-1                           )  re:   0.21
    core sim:    0.80     matrix sim: 0.85
V$SREBP1_02           (SREBP-1                           )  re: < 0.01
    core sim:    0.80     matrix sim: 0.85
V$SRF_01              (SRF                               )  re: < 0.01
    core sim:    0.80     matrix sim: 0.85
V$SRF_Q6              (SRF                               )  re:   0.16
    core sim:    0.80     matrix sim: 0.85
V$SRY_02              (SRY                               )  re:   3.94
    core sim:    0.80     matrix sim: 0.85
V$TAL1ALPHAE47_01     (Tal-1alpha/E47                    )  re:   0.14
    core sim:    0.80     matrix sim: 0.85
V$TAL1BETAE47_01      (Tal-1beta/E47                     )  re:   0.11
    core sim:    0.80     matrix sim: 0.85
V$TAL1BETAITF2_01     (Tal-1beta/ITF-2                   )  re:   0.06
    core sim:    0.80     matrix sim: 0.85
V$TAXCREB_01          (Tax/CREB                          )  re:   0.06
    core sim:    0.80     matrix sim: 0.85
V$TAXCREB_02          (Tax/CREB                          )  re: < 0.01
    core sim:    0.80     matrix sim: 0.85
V$TH1E47_01           (Th1/E47                           )  re:   2.04
    core sim:    0.80     matrix sim: 0.85
V$TST1_01             (Tst-1                             )  re:   5.66
    core sim:    0.80     matrix sim: 0.85
V$USF_01              (USF                               )  re:   1.09
    core sim:    0.80     matrix sim: 0.85
V$USF_Q6              (USF                               )  re:   2.52
    core sim:    0.80     matrix sim: 0.85
V$VJUN_01             (v-Jun                             )  re: < 0.01
    core sim:    0.80     matrix sim: 0.85
V$VMAF_01             (v-Maf                             )  re:   0.99
    core sim:    0.80     matrix sim: 0.85
V$VMYB_01             (v-Myb                             )  re:   2.29
    core sim:    0.80     matrix sim: 0.85
V$YY1_02              (YY1                               )  re:   0.13
    core sim:    0.80     matrix sim: 0.85
V$ZID_01              (ZID                               )  re:   0.33
    core sim:    0.80     matrix sim: 0.85

Explanation for column output:
------------------------------
  ->     Matrix positions correspond to sense strand numbering, but all
         sequences are given in 5'-3' direction.

  ->     n/a in column 'core simil.' indicates, that no core search
         was conducted.

  ->     n/a in column 'matrix simil.' indicates, that no
```

matrix similarity was calculated, because the sequences was too short (only the core was found). In that case the core position is given in brackets in column 'matrix position'.

-> Capital letters within the sequence indicate the core string.

matrix name	matrix position(str)	core simil.	matrix simil.	sequence
Inspecting sequence unp (1 - 982):				
V$AHR_01	(979)(-)	1.000	n/a	GCGT
V$AP1FJ_Q2	146 (+)	1.000	0.868	ctTGACctgga
V$AP1FJ_Q2	159 (+)	1.000	0.908	gcTGACaggat
V$AP1FJ_Q2	209 (+)	1.000	0.897	taTGACtttat
V$AP1FJ_Q2	438 (-)	1.000	0.930	acTGACtcttt
V$AP1FJ_Q2	489 (-)	1.000	0.933	ggTGACaactg
V$AP1FJ_Q2	747 (+)	1.000	0.953	agTGACtgatg
V$AP1FJ_Q2	772 (-)	1.000	0.903	tcTGACtgctg
V$AP1_Q2	159 (+)	1.000	0.894	gcTGACaggat
V$AP1_Q2	209 (+)	1.000	0.876	taTGACtttat
V$AP1_Q2	438 (-)	1.000	0.940	acTGACtcttt
V$AP1_Q2	489 (-)	1.000	0.901	ggTGACaactg
V$AP1_Q2	747 (+)	1.000	0.945	agTGACtgatg
V$AP1_Q2	772 (-)	1.000	0.882	tcTGACtgctg
V$AP1_Q4	209 (+)	1.000	0.860	taTGACtttat
V$AP1_Q4	438 (-)	1.000	0.907	acTGACtcttt
V$AP1_Q4	489 (-)	1.000	0.868	ggTGACaactg
V$AP1_Q4	747 (+)	1.000	0.963	agTGACtgatg
V$AP1_Q4	772 (-)	1.000	0.871	tcTGACtgctg
V$AP1_Q6	(4)(-)	0.806	n/a	TGAA
V$AP1_Q6	75 (-)	0.810	0.857	aaTGATtccgt
V$AP1_Q6	209 (+)	1.000	0.899	taTGACtttat
V$AP1_Q6	438 (-)	1.000	0.930	acTGACtcttt
V$AP1_Q6	489 (-)	1.000	0.859	ggTGACaactg
V$AP1_Q6	747 (+)	1.000	0.956	agTGACtgatg
V$AP1_Q6	772 (-)	1.000	0.901	tcTGACtgctg
V$AP2_Q6	820 (-)	0.905	0.935	cgCCCAgaggcg
V$AP2_Q6	857 (-)	0.857	0.875	ccCCCTccggag
V$AP2_Q6	866 (-)	0.976	0.870	ccCCCCccgccc
V$AP2_Q6	867 (-)	0.976	0.850	ccCCCCcccgcc
V$AP2_Q6	886 (-)	0.976	0.912	cgCCCCcacgcg
V$AP4_Q5	32 (-)	1.000	0.850	ccCAGCactc
V$AP4_Q5	258 (+)	1.000	0.850	ccCAGCactc
V$AP4_Q5	318 (+)	1.000	0.909	aaCAGCtcca
V$AP4_Q5	329 (+)	1.000	0.916	gaCAGCagct
V$AP4_Q5	332 (+)	1.000	0.886	agCAGCtaga
V$AP4_Q5	598 (-)	1.000	0.851	ggCAGCccct
V$AP4_Q5	770 (+)	1.000	0.893	gcCAGCagtc
V$AP4_Q5	881 (+)	1.000	0.876	aaCAGCgcgt
V$AP4_Q6	318 (+)	1.000	0.866	aaCAGCtcca
V$AP4_Q6	332 (+)	1.000	0.859	agCAGCtaga
V$BRN2_01	(3)(-)	0.854	n/a	GAAT
V$BRN2_01	(1)(+)	0.854	n/a	GAAT
V$BRN2_01	703 (-)	1.000	0.914	ttgatgctAAATccgt
V$CDPCR3HD_01	478 (-)	1.000	0.874	gtttGATCcc
V$CDPCR3_01	(972)(-)	1.000	n/a	ATGG
V$CDP_02	(974)(+)	0.806	n/a	ATAG
V$CEBPB_01	85 (+)	0.873	0.860	tgtttcaGGAAaaa
V$CEBPB_01	86 (+)	0.986	0.859	gtttcagGAAAaaa
V$CEBPB_01	265 (+)	0.986	0.891	ctctggaGAAAagg
V$CEBPB_01	308 (-)	0.930	0.940	tgtttatGTAAtct
V$CEBPB_01	366 (-)	0.986	0.908	caatagtGAAAttt
V$CEBPB_01	639 (-)	1.000	0.907	gctttcaGCAAgac
V$CEBPB_01	640 (+)	0.986	0.870	tcttgctGAAAgcc
V$CEBPB_02	308 (+)	0.833	0.879	agaTTACataaaca
V$CEBPB_02	639 (+)	1.000	0.914	gtcTTGCtgaaagc
V$CETS1P54_01	24 (-)	0.926	0.910	tcAGGAggca

V$CETS1P54_01	64	(+)	1.000	0.944	ccCGGAagaa
V$CETS1P54_01	89	(+)	0.926	0.882	tcAGGAaaaa
V$CETS1P54_01	151	(+)	0.852	0.870	ccTGGAtggc
V$CETS1P54_01	163	(+)	0.926	0.907	acAGGAtaca
V$CETS1P54_01	554	(+)	1.000	0.874	acCGGAcagt
V$CETS1P54_01	693	(-)	0.926	0.877	gcAGGAggag
V$CETS1P54_01	734	(+)	0.926	0.897	ccAGGAtgga
V$CETS1P54_01	828	(-)	0.852	0.893	gcTGGAcgcc
V$CETS1P54_01	854	(-)	1.000	0.872	tcCGGAgccg
V$CETS1P54_01	858	(+)	1.000	0.896	tcCGGAgggg
V$CETS1P54_02	23	(-)	0.839	0.866	actcaGGAGgcag
V$CETS1P54_02	62	(+)	1.000	0.967	cgcccGGAAgaaa
V$CETS1P54_02	72	(+)	1.000	0.951	aaaacGGAAtcat
V$CETS1P54_02	87	(+)	1.000	0.953	tttcaGGAAaaaa
V$CETS1P54_02	149	(+)	0.958	0.893	gacctGGATggct
V$CETS1P54_02	161	(+)	0.958	0.936	tgacaGGATacaa
V$CETS1P54_02	252	(-)	0.958	0.863	tgctgGGATtaaa
V$CETS1P54_02	264	(+)	0.839	0.852	actctGGAGaaaa
V$CETS1P54_02	319	(-)	0.839	0.850	gtcctGGAGctgt
V$CETS1P54_02	423	(+)	0.839	0.863	acaaaGGACaact
V$CETS1P54_02	452	(-)	1.000	0.926	aaagtGGAAaaag
V$CETS1P54_02	474	(+)	0.958	0.875	tctggGGATcaaa
V$CETS1P54_02	552	(+)	0.839	0.884	acaccGGACagtc
V$CETS1P54_02	574	(-)	0.839	0.866	cacacGGACgcat
V$CETS1P54_02	653	(+)	1.000	0.903	cgcatGGAAccga
V$CETS1P54_02	663	(+)	1.000	0.924	cgagtGGAAacaa
V$CETS1P54_02	692	(-)	0.839	0.864	gtgcaGGAGgagc
V$CETS1P54_02	700	(+)	0.958	0.933	tgcacGGATttag
V$CETS1P54_02	732	(+)	0.958	0.911	ggccaGGATggag
V$CETS1P54_02	803	(-)	0.958	0.906	cgggtGGATgtgg
V$CETS1P54_02	842	(-)	0.839	0.863	gaagcGGACagaa
V$CETS1P54_02	853	(-)	0.839	0.872	cctccGGAGccga
V$CETS1P54_02	856	(+)	0.839	0.871	gctccGGAGgggg
V$CLOX_01	(974)	(+)	0.807	n/a	ATAG
V$CMYB_01	127	(+)	1.000	0.898	cggtgGTTG
V$CMYB_01	170	(+)	1.000	0.865	acaaaGTTG
V$CMYB_01	186	(-)	1.000	0.865	acaaaGTTG
V$CMYB_01	431	(-)	1.000	0.866	ttaaaGTTG
V$CMYB_01	486	(+)	1.000	0.954	acccaGTTG
V$CMYB_01	532	(-)	1.000	0.871	taataGTTG
V$CMYB_01	880	(-)	0.841	0.875	gcgctGTTA
V$CP2_01	351	(+)	0.909	0.856	tcccaaaCAAA
V$DELTAEF1_01	146	(+)	1.000	0.867	cttgACCTgga
V$DELTAEF1_01	230	(-)	1.000	0.960	ccacACCTgca
V$DELTAEF1_01	720	(+)	1.000	0.857	gctaACCTcac
V$E2F_02	388	(+)	1.000	0.886	TTTGgcga
V$E47_01	228	(+)	1.000	0.931	tttGCAGgtgtggtg
V$E47_02	227	(+)	1.000	0.909	ttttgCAGGtgtggtg
V$E4BP4_01	309	(-)	1.000	0.958	gtttatGTAAtc
V$ELK1_02	61	(+)	1.000	0.955	ccgcccGGAAgaaa
V$ELK1_02	71	(+)	1.000	0.883	gaaaacGGAAtcat
V$ELK1_02	86	(+)	1.000	0.856	gtttcaGGAAaaaa
V$ER_Q6	137	(+)	1.000	0.882	tactctctcctTGACctgg
V$GATA1_02	162	(+)	1.000	0.891	gacagGATAcaaag
V$GATA1_03	167	(+)	1.000	0.902	GATAcaaag
V$GATA1_04	163	(+)	1.000	0.881	acagGATAcaaag
V$GR_Q6	325	(-)	0.922	0.863	cttctagctgcTGTCctgg
V$GR_Q6	369	(+)	0.898	0.862	tttcactattgTGTGcctg
V$GR_Q6	418	(-)	1.000	0.913	aagttgtccttTGTTctcg
V$GR_Q6	425	(-)	0.922	0.856	ctctttaaagtTGTCcttt
V$HEN1_01	(6)	(+)	0.827	n/a	CAGA
V$HFH1_01	313	(-)	1.000	0.913	agctGTTTatgt
V$HFH1_01	355	(-)	1.000	0.863	tgttGTTTgttt
V$HFH2_01	189	(+)	1.000	0.866	cttTGTTttttg
V$HFH2_01	216	(+)	0.823	0.884	ttaTTTTtgttt
V$HFH2_01	217	(+)	0.823	0.860	tatTTTTgtttt
V$HFH2_01	220	(+)	1.000	0.870	tttTGTTttttg
V$HFH2_01	355	(-)	1.000	0.925	tgtTGTTtgttt

V$HNF3B_01	216	(+)	0.848	0.896	ttaTTTTtgttt
V$HNF3B_01	355	(-)	0.987	0.945	tgtTGTTtgttt
V$HNF4_01	266	(+)	1.000	0.876	tctggagaAAAGgcaaat
V$HNF4_01	417	(+)	1.000	0.926	tcgagaacAAAGgacaact
V$HSF1_01	(4)	(-)	0.867	n/a	TGAA
V$HSF1_01	339	(+)	1.000	0.861	AGAAgaaacc
V$HSF1_01	613	(-)	0.867	0.881	TGAAtttact
V$HSF1_01	901	(+)	1.000	0.878	AGAAgaggcg
V$IK2_01	253	(-)	1.000	0.946	tgctGGGAttaa
V$IK2_01	347	(-)	1.000	0.947	gtttGGGAcagg
V$IK2_01	474	(+)	1.000	0.903	tctgGGGAtcaa
V$IK2_01	894	(+)	1.000	0.873	ggcgGGGAgaag
V$IRF1_01	447	(-)	1.000	0.877	ggaaaaaGAAAac
V$IRF2_01	447	(-)	1.000	0.863	ggaaaaaGAAAac
V$LYF1_01	349	(-)	1.000	0.927	tttGGGAca
V$MAX_01	934	(-)	1.000	0.905	cacgCACGtgcgcg
V$MAX_01	934	(+)	1.000	0.905	cgcgCACGtgcgtg
V$MAX_01	(978)	(+)	1.000	n/a	CACG
V$MYCMAX_01	934	(-)	1.000	0.852	cacgCACGtgcgcg
V$MYCMAX_01	934	(+)	1.000	0.852	cgcgCACGtgcgtg
V$MYCMAX_01	(978)	(+)	1.000	n/a	CACG
V$MYCMAX_02	935	(-)	1.000	0.899	acgCACGtgcgc
V$MYCMAX_02	935	(+)	1.000	0.909	gcgCACGtgcgt
V$MYCMAX_02	(978)	(+)	1.000	n/a	CACG
V$MYOD_01	229	(+)	1.000	0.876	ttgCAGGtgtgg
V$MYOD_Q6	230	(-)	1.000	0.910	caCACCtgca
V$MYOD_Q6	487	(-)	0.872	0.886	gaCAACtggg
V$MZF1_01	474	(+)	1.000	0.962	tctGGGGa
V$MZF1_01	894	(+)	1.000	0.982	ggcGGGGa
V$NF1_Q6	387	(+)	1.000	0.901	gttTGGCgaacgcgtgcc
V$NF1_Q6	391	(-)	1.000	0.856	ctgTGGCacgcgttcgcc
V$NF1_Q6	539	(+)	1.000	0.859	taaTGGCtggaccacacc
V$NF1_Q6	588	(+)	1.000	0.920	actTGGCtacaggggctg
V$NF1_Q6	759	(-)	1.000	0.864	tgcTGGCccgccctcgac
V$NFE2_01	(5)	(-)	1.000	n/a	CTGA
V$NMYC_01	394	(+)	1.000	0.893	gaacgCGTGcca
V$NMYC_01	464	(-)	1.000	0.880	gaaccCGTGtaa
V$NMYC_01	798	(-)	1.000	0.890	ccaccCGTGtct
V$NMYC_01	883	(+)	1.000	0.881	cagcgCGTGggg
V$NMYC_01	935	(-)	1.000	0.964	acgcaCGTGcgc
V$NMYC_01	935	(+)	1.000	0.960	gcgcaCGTGcgt
V$NMYC_01	(978)	(-)	1.000	n/a	CGTG
V$NRF2_01	64	(+)	1.000	0.922	cccGGAAgaa
V$OCT1_02	(1)	(-)	0.980	n/a	ATTC
V$OCT1_02	(3)	(+)	0.980	n/a	ATTC
V$OCT1_06	(1)	(-)	0.889	n/a	ATTC
V$OCT1_06	(3)	(+)	0.889	n/a	ATTC
V$OCT1_06	96	(+)	1.000	0.872	aaaattaagATGCc
V$OCT1_06	365	(-)	0.889	0.867	aatagtgaaATTTg
V$OCT1_06	565	(-)	0.833	0.906	cgcattgagATCTg
V$OCT1_06	(972)	(-)	0.833	n/a	ATGG
V$OCT1_Q6	(3)	(-)	0.800	n/a	GAAT
V$OCT1_Q6	(1)	(+)	0.800	n/a	GAAT
V$PAX5_01	(14)	(-)	0.952	n/a	GGCA
V$PAX5_01	(4)	(+)	0.810	n/a	TTCA
V$PAX5_01	(971)	(+)	0.857	n/a	CCCA
V$PAX5_01	(976)	(+)	0.905	n/a	AGCA
V$PAX6_01	(979)	(+)	1.000	n/a	ACGC
V$RORA1_01	147	(-)	1.000	0.964	ccatccaGGTCaa
V$S8_01	93	(+)	1.000	0.962	gaaaaaATTAa
V$S8_01	531	(+)	1.000	0.862	acaactATTAa
V$S8_01	538	(-)	1.000	0.866	ccagccATTAa
V$SOX5_01	81	(-)	1.000	0.985	gaaaCAATga
V$SOX5_01	374	(-)	1.000	0.865	cacaCAATag
V$SP1_01	58	(-)	1.000	0.872	cgGGCGgtga
V$SP1_01	685	(+)	0.917	0.899	gcGGCAggct
V$SP1_01	763	(+)	1.000	0.900	agGGCGggcc
V$SP1_01	865	(+)	1.000	0.948	ggGGCGgggg

V$SP1_01	870 (+)	0.833	0.928	ggGGGGgggt
V$SP1_01	871 (+)	0.833	0.880	ggGGGGggtt
V$SP1_01	892 (+)	1.000	0.974	ggGGCGggga
V$SP1_01	962 (+)	1.000	0.878	tgGGCGcggc
V$SP1_Q6	57 (-)	1.000	0.865	tccgGGCGgtgat
V$SP1_Q6	761 (+)	1.000	0.918	cgagGGCGggcca
V$SP1_Q6	863 (+)	1.000	0.979	agggGGCGggggg
V$SP1_Q6	890 (+)	1.000	0.967	tgggGGCGgggag
V$SREBP1_02	235 (-)	1.000	0.885	caTCACcacac
V$SRF_Q6	(972) (+)	1.000	n/a	CCAT
V$SRY_02	80 (-)	1.000	0.941	tgaaACAAtgat
V$SRY_02	166 (+)	1.000	0.851	ggatACAAagtt
V$SRY_02	172 (-)	1.000	0.881	ttaaACAActtt
V$SRY_02	187 (-)	1.000	0.915	aaaaACAAagtt
V$SRY_02	218 (-)	1.000	0.943	aaaaACAAaaat
V$SRY_02	353 (+)	1.000	0.915	ccaaACAAacaa
V$SRY_02	357 (+)	1.000	0.903	acaaACAAcaaa
V$SRY_02	360 (+)	1.000	0.876	aacaACAAattt
V$SRY_02	373 (-)	1.000	0.910	gcacACAAtagt
V$SRY_02	419 (+)	1.000	0.908	gagaACAAagga
V$SRY_02	518 (-)	1.000	0.881	tcaaACAAgggc
V$SRY_02	527 (+)	1.000	0.872	tgaaACAActat
V$SRY_02	668 (+)	1.000	0.910	ggaaACAAgaaa
V$TH1E47_01	467 (+)	1.000	0.877	cacgggttCTGGgat
V$TH1E47_01	560 (-)	1.000	0.899	attgagatCTGGactg
V$TST1_01	(2) (-)	1.000	n/a	AATT
V$TST1_01	(2) (+)	1.000	n/a	AATT
V$TST1_01	703 (+)	0.895	0.856	acggATTTagcatca
V$USF_01	935 (-)	1.000	0.976	acgCACGtgcgc
V$USF_01	935 (+)	1.000	0.976	gcgCACGtgcgt
V$USF_01	(978) (+)	1.000	n/a	CACG
V$USF_Q6	395 (-)	1.000	0.890	ggCACGcgtt
V$USF_Q6	465 (+)	1.000	0.880	taCACGggtt
V$USF_Q6	551 (+)	0.818	0.855	caCACCggac
V$USF_Q6	585 (+)	0.864	0.877	tgCACTtggc
V$USF_Q6	726 (+)	0.864	0.890	ctCACAggcc
V$USF_Q6	743 (-)	0.864	0.873	gtCACTcgct
V$USF_Q6	799 (+)	1.000	0.908	gaCACGggtg
V$USF_Q6	884 (-)	1.000	0.885	ccCACGcgct
V$USF_Q6	936 (-)	1.000	0.887	cgCACGtgcg
V$USF_Q6	936 (+)	1.000	0.887	cgCACGtgcg
V$USF_Q6	(978) (+)	1.000	n/a	CACG
V$VMYB_01	71 (+)	1.000	0.937	gaaAACGgaa
V$VMYB_01	443 (-)	0.876	0.873	gaaAACTgac
V$VMYB_01	487 (-)	0.876	0.865	gacAACTggg
V$YY1_02	(972) (+)	1.000	n/a	CCAT

In 1 sequences 1 matches to the matrix V$AHR_01 were found.
In 1 sequences 7 matches to the matrix V$AP1FJ_Q2 were found.
In 1 sequences 6 matches to the matrix V$AP1_Q2 were found.
In 1 sequences 5 matches to the matrix V$AP1_Q4 were found.
In 1 sequences 7 matches to the matrix V$AP1_Q6 were found.
In 1 sequences 5 matches to the matrix V$AP2_Q6 were found.
In 0 sequences 0 matches to the matrix V$AP4_01 were found.
In 1 sequences 8 matches to the matrix V$AP4_Q5 were found.
In 1 sequences 2 matches to the matrix V$AP4_Q6 were found.
In 0 sequences 0 matches to the matrix V$ARP1_01 were found.
In 0 sequences 0 matches to the matrix V$ATF_01 were found.
In 0 sequences 0 matches to the matrix V$BRACH_01 were found.
In 1 sequences 3 matches to the matrix V$BRN2_01 were found.
In 0 sequences 0 matches to the matrix V$CDPCR1_01 were found.
In 1 sequences 1 matches to the matrix V$CDPCR3HD_01 were found.
In 1 sequences 1 matches to the matrix V$CDPCR3_01 were found.
In 0 sequences 0 matches to the matrix V$CDP_01 were found.
In 1 sequences 1 matches to the matrix V$CDP_02 were found.
In 1 sequences 7 matches to the matrix V$CEBPB_01 were found.
In 1 sequences 2 matches to the matrix V$CEBPB_02 were found.
In 1 sequences 11 matches to the matrix V$CETS1P54_01 were found.
In 1 sequences 23 matches to the matrix V$CETS1P54_02 were found.

```
In  1 sequences  1 matches to the matrix V$CLOX_01 were found.
In  1 sequences  7 matches to the matrix V$CMYB_01 were found.
In  0 sequences  0 matches to the matrix V$COMP1_01 were found.
In  0 sequences  0 matches to the matrix V$COUP_01 were found.
In  1 sequences  1 matches to the matrix V$CP2_01 were found.
In  0 sequences  0 matches to the matrix V$CREBP1CJUN_01 were found.
In  0 sequences  0 matches to the matrix V$CREBP1_01 were found.
In  0 sequences  0 matches to the matrix V$CREBP1_Q2 were found.
In  0 sequences  0 matches to the matrix V$CREB_01 were found.
In  0 sequences  0 matches to the matrix V$CREB_02 were found.
In  0 sequences  0 matches to the matrix V$CREB_Q2 were found.
In  0 sequences  0 matches to the matrix V$CREB_Q4 were found.
In  0 sequences  0 matches to the matrix V$CREL_01 were found.
In  1 sequences  3 matches to the matrix V$DELTAEF1_01 were found.
In  0 sequences  0 matches to the matrix V$E2F_01 were found.
In  1 sequences  1 matches to the matrix V$E2F_02 were found.
In  0 sequences  0 matches to the matrix V$E2F_Q6 were found.
In  0 sequences  0 matches to the matrix V$E2_01 were found.
In  0 sequences  0 matches to the matrix V$E2_Q6 were found.
In  1 sequences  1 matches to the matrix V$E47_01 were found.
In  1 sequences  1 matches to the matrix V$E47_02 were found.
In  1 sequences  1 matches to the matrix V$E4BP4_01 were found.
In  0 sequences  0 matches to the matrix V$ELK1_01 were found.
In  1 sequences  3 matches to the matrix V$ELK1_02 were found.
In  1 sequences  1 matches to the matrix V$ER_Q6 were found.
In  0 sequences  0 matches to the matrix V$EVI1_01 were found.
In  0 sequences  0 matches to the matrix V$EVI1_02 were found.
In  0 sequences  0 matches to the matrix V$EVI1_03 were found.
In  0 sequences  0 matches to the matrix V$EVI1_04 were found.
In  0 sequences  0 matches to the matrix V$EVI1_05 were found.
In  0 sequences  0 matches to the matrix V$EVI1_06 were found.
In  1 sequences  1 matches to the matrix V$GATA1_02 were found.
In  1 sequences  1 matches to the matrix V$GATA1_03 were found.
In  1 sequences  1 matches to the matrix V$GATA1_04 were found.
In  1 sequences  4 matches to the matrix V$GR_Q6 were found.
In  1 sequences  1 matches to the matrix V$HEN1_01 were found.
In  0 sequences  0 matches to the matrix V$HEN1_02 were found.
In  1 sequences  2 matches to the matrix V$HFH1_01 were found.
In  1 sequences  5 matches to the matrix V$HFH2_01 were found.
In  0 sequences  0 matches to the matrix V$HNF1_01 were found.
In  1 sequences  2 matches to the matrix V$HNF3B_01 were found.
In  1 sequences  2 matches to the matrix V$HNF4_01 were found.
In  0 sequences  0 matches to the matrix V$HOX13_01 were found.
In  1 sequences  4 matches to the matrix V$HSF1_01 were found.
In  0 sequences  0 matches to the matrix V$IK1_01 were found.
In  1 sequences  4 matches to the matrix V$IK2_01 were found.
In  0 sequences  0 matches to the matrix V$IK3_01 were found.
In  1 sequences  1 matches to the matrix V$IRF1_01 were found.
In  1 sequences  1 matches to the matrix V$IRF2_01 were found.
In  1 sequences  1 matches to the matrix V$LYF1_01 were found.
In  1 sequences  3 matches to the matrix V$MAX_01 were found.
In  0 sequences  0 matches to the matrix V$MEF2_01 were found.
In  1 sequences  3 matches to the matrix V$MYCMAX_01 were found.
In  1 sequences  3 matches to the matrix V$MYCMAX_02 were found.
In  1 sequences  1 matches to the matrix V$MYOD_01 were found.
In  1 sequences  2 matches to the matrix V$MYOD_Q6 were found.
In  1 sequences  2 matches to the matrix V$MZF1_01 were found.
In  0 sequences  0 matches to the matrix V$MZF1_02 were found.
In  0 sequences  0 matches to the matrix V$NF1_01 were found.
In  1 sequences  5 matches to the matrix V$NF1_Q6 were found.
In  1 sequences  1 matches to the matrix V$NFE2_01 were found.
In  0 sequences  0 matches to the matrix V$NFKAPPAB50_01 were found.
In  0 sequences  0 matches to the matrix V$NFKAPPAB65_01 were found.
In  0 sequences  0 matches to the matrix V$NFKAPPAB_01 were found.
In  0 sequences  0 matches to the matrix V$NFKB_Q6 were found.
In  0 sequences  0 matches to the matrix V$NFY_Q6 were found.
In  1 sequences  7 matches to the matrix V$NMYC_01 were found.
In  1 sequences  1 matches to the matrix V$NRF2_01 were found.
```

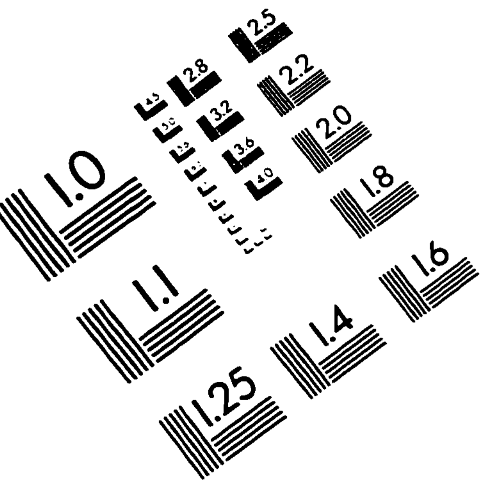
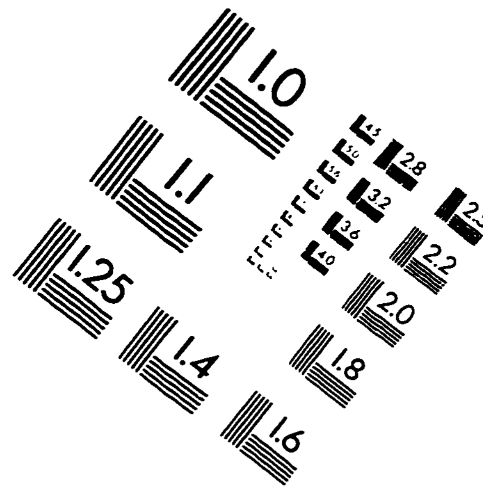
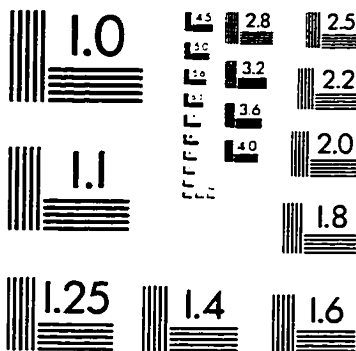
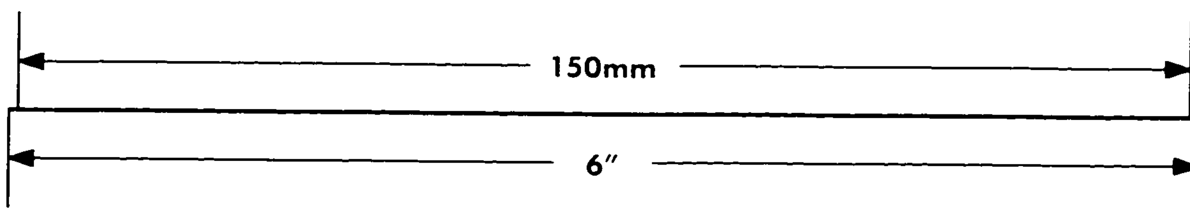
```
In  0 sequences   0 matches to the matrix V$OCT1_01 were found.
In  1 sequences   2 matches to the matrix V$OCT1_02 were found.
In  0 sequences   0 matches to the matrix V$OCT1_05 were found.
In  1 sequences   6 matches to the matrix V$OCT1_06 were found.
In  1 sequences   2 matches to the matrix V$OCT1_Q6 were found.
In  0 sequences   0 matches to the matrix V$P53_01 were found.
In  1 sequences   4 matches to the matrix V$PAX5_01 were found.
In  0 sequences   0 matches to the matrix V$PAX5_02 were found.
In  1 sequences   1 matches to the matrix V$PAX6_01 were found.
In  0 sequences   0 matches to the matrix V$PBX1_02 were found.
In  1 sequences   1 matches to the matrix V$RORA1_01 were found.
In  0 sequences   0 matches to the matrix V$RORA2_01 were found.
In  0 sequences   0 matches to the matrix V$RSRFC4_01 were found.
In  1 sequences   3 matches to the matrix V$S8_01 were found.
In  1 sequences   2 matches to the matrix V$SOX5_01 were found.
In  1 sequences   8 matches to the matrix V$SP1_01 were found.
In  1 sequences   4 matches to the matrix V$SP1_Q6 were found.
In  0 sequences   0 matches to the matrix V$SREBP1_01 were found.
In  1 sequences   1 matches to the matrix V$SREBP1_02 were found.
In  0 sequences   0 matches to the matrix V$SRF_01 were found.
In  1 sequences   1 matches to the matrix V$SRF_Q6 were found.
In  1 sequences  13 matches to the matrix V$SRY_02 were found.
In  0 sequences   0 matches to the matrix V$TAL1ALPHAE47_01 were found.
In  0 sequences   0 matches to the matrix V$TAL1BETAE47_01 were found.
In  0 sequences   0 matches to the matrix V$TAL1BETAITF2_01 were found.
In  0 sequences   0 matches to the matrix V$TAXCREB_01 were found.
In  0 sequences   0 matches to the matrix V$TAXCREB_02 were found.
In  1 sequences   2 matches to the matrix V$TH1E47_01 were found.
In  1 sequences   3 matches to the matrix V$TST1_01 were found.
In  1 sequences   3 matches to the matrix V$USF_01 were found.
In  1 sequences  11 matches to the matrix V$USF_Q6 were found.
In  0 sequences   0 matches to the matrix V$VJUN_01 were found.
In  0 sequences   0 matches to the matrix V$VMAF_01 were found.
In  1 sequences   3 matches to the matrix V$VMYB_01 were found.
In  1 sequences   1 matches to the matrix V$YY1_02 were found.
In  0 sequences   0 matches to the matrix V$ZID_01 were found.
```

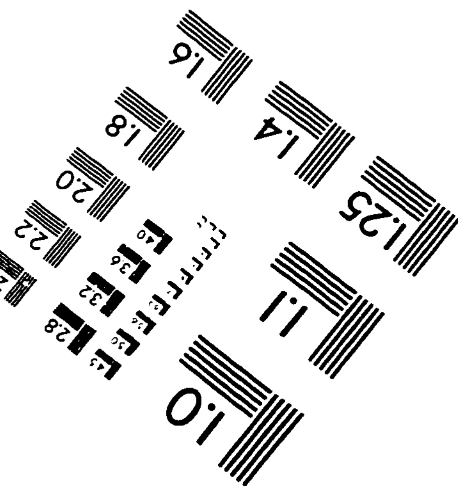
Don't forget to cite:

Quandt, K. Frech, K. Karas, H. Wingender, E. and Werner, T.
MatInd and MatInspector - New fast and versatile tools for detection of consensus matches in nucleotide sequence data
Nucleic Acids Research 23, 4878-4884 (1995)

Back to the MatInspector input form

IMAGE EVALUATION TEST TARGET (QA-3)



APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved