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.

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.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600



Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/466

UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

SEQUENCE AND ANALYSIS OF TWO REGIONS OF MOUSE CHROMOSOME 1 INVOLVED IN INTERFERON RESPONSE AND BONE MINERAL DENSITY REGULATION

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

In partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

Stéphane D. Deschamps

Norman, Oklahoma

2002

UMI Number: 3045845



UMI Microform 3045845
Copyright 2002 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

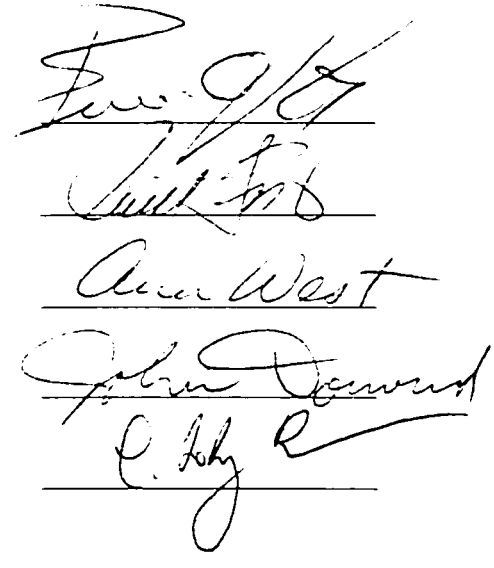
ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

©Copyright by Stéphane D. Deschamps 2002
All Rights Reserved.

SEQUENCE AND ANALYSIS OF TWO REGIONS OF MOUSE CHROMOSOME 1 INVOLVED IN INTERFERON RESPONSE AND BONE MINERAL DENSITY REGULATION

A Dissertation APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY



Acknowledgements

I wish to thank my advisory committee, Drs Paul F. Cook, Ann H. West, Leroy Blank, John Downard and Bruce A. Roe for their continued support and efforts. The work for my dissertation was done entirely in the laboratory of Dr. B. A. Roe, my major professor. I wish to express him my sincere gratitude for accepting me into his laboratory and providing me with the guidance, knowledge, and support, which allowed me to complete my dissertation research.

All the members of Dr. Roe's laboratory, past and present, deserve many thanks for their help and support. My sincere thanks goes to past members, Guozhong Zhang, Feng Chen, Huaquin Pan, Steve Toth, YingPing Wang, Judy Crabtree, Min Zhan, Victoria Lao and Ping Hu, for their help, support and guidance, to the informatic team, Steve Kenton, HongShing Lai, and Jim White, for their invaluable help and patience with my very limited computer skills, and to present members, graduate and undergraduate students, technicians, secretaries, for their invaluable help, support, and friendship.

I would like to thank our collaborators, Dr. Weikuan Gu from the JLP VA Medical Center at Loma Linda, California, and Dr. Peter Lengyel, from Yale University, as well as members of their research groups, for initiating the two main projects mentionned in this dissertation research.

My deepest gratitude goes to my parents, my brother, and my family. They deserve immense thanks for their love, help and support. They always helped me in being what I wanted to be. Finally, my deepest love and gratitude goes to my wife, Marigel. Her continuous love and support was invaluable to me and made me who I am now.

Table of contents

List of tables

List of Figures

List of Abbreviations

Abstract

Chapter I: Introduction 1

I-1 Human and mouse genomes 1

I-1-1 DNA structure 1

I-1-1-1 DNA is the carrier of hereditary information 1

I-1-1-2 DNA structure 3

I-1-2 From structure towards function 5

I-1-3 DNA sequencing technology 12

I-1-4 Genome organization 17

I-2 Interferon-inducible proteins 23

I-2-1 Interferons and their mode of action 23

I-2-2 The "200 family" of proteins 26

I-2-2-1 The murine Ifi200 family 26

I-2-2-1-1 The Ifi202a gene 28

I-2-2-1-2 The Ifi203 gene 33

I-2-2-1-3 The Ifi204 gene 33

I-2-2-1-4 The D3 gene 35

I-2-2-2 The human "200 family" of proteins 35

I-2-2-2-1 The MNDA gene 36

I-2-2-2-2 The IFI-16 gene 38

I-2-2-2-3 The AIM2 gene 39

I-3 Genetic aspects of bone density regulation 40

Chapter II: Materials and Methods 44

II-1 Random shotgun sequencing 44
II-1-1 Large scale DNA isolation 45
II-1-2 Physical shearing 47
II-1-3 End repair 47
II-1-4 Size selection and DNA extraction 48
II-1-5 Shotgun library construction 49
II-1-6 Miniprep isolation of DNA subclones 50
II-1-7 Cycle sequencing reaction and purification 51
II-1-8 Data assembly 52
II-2 Gap closure 52
II-2-1 Primer walking 53
II-2-2 PCR-based methods 54
II-2-3 LI-COR-based DNA sequencing 56
II-2-4 M.J. basestation-based DNA sequencing 57
II-3 Proofreading and correcting the assembled data 57
II-4 Sequence analysis 60
II-4-1 Detecting regions with repeated sequences 60
II-4-2 Gene prediction 60
II-4-3 Database similarity searches 61
II-4-4 Promoter and transcription factor binding sites searches 61
II-4-5 Sequence comparison 62
II-4-6 Motif searches 63
Chapter III: Results and Discussion 64
III-1 Murine Ifi200 family of genes 65
III-1-1 Ifi202a gene 67
III-1-2 Ifi202b gene 72
III-1-3 Ifi203 genes 78

III-1-3-1 Ifi203b gene 79
III-1-3-2 Ifi203a gene 84
III-1-3-3 Ifi203c gene 90
III-1-4 Ifi204 gene 91
III-1-5 Ifi201 gene 96
III-1-6 Ifi202c pseudogene and others possible genes 98
III-1-7 Repeat elements in the mouse Ifi200 cluster 113
III-1-8 Mouse Ifi200 gene cluster: a summary 116
III-1-9 Comparative analysis of the mouse and human Ifi200 gene clusters 118
III-1-9-1 Human/mouse cluster genomic sequence comparison 121
III-1-9-2 Human/mouse cluster 5'flanking sequence comparison 129
III-1-9-3 Human/mouse cluster repeat sequence comparison 135
III-1-10 Evolutionary aspects of the mouse Ifi200 gene cluster 137
III-1-11 Closing the gap between the two contigs 138
III-1-12 Two-dimensional dot-plot comparison of the human and mouse Ifi200 clusters 140
III-1-13 Postulated evolution of the Ifi200 gene cluster 141
III-2 Genetic aspects of bone mineral density 143
III-2-1 Mouse BAC clone rp23-145f9 145
III-2-1-1 Mouse CD150/SLAM gene 145
III-2-1-2 Mouse CD48/BCM1 gene 150
III-2-1-3 Mouse BC011154 gene 154
III-2-1-4 Mouse AK002863 gene 160

III-2-1-5 Repeat elements in the mouse clone rp23-145f9 160
III-2-2 Mouse BAC clone rp23-77a8 162
III-2-2-1 Mouse Ly108 gene 163
III-2-2-2 Mouse CD84 gene 168
III-2-2-3 Repeat elements in the mouse clone rp23-77a8 173
III-2-3 Mouse BAC clone rp23-21118 174
III-2-4 Mouse BAC clone rp23-395h6 177
III-2-4-1 Mouse BAC clone rp23-395h6: a preview 177
III-2-4-2 Mouse CAR gene 179
III-2-4-3 Mouse myelin protein zero (P0) gene 185
III-2-4-4 Mouse apoA-II gene 189
III-2-4-5 Mouse NDUFS2 gene 192
III-2-4-6 Mouse beta4Gal-T3 gene 196
III-2-4-7 Mouse Ppox gene 201
III-2-4-8 Mouse USP23 gene 206
III-2-4-9 Mouse DEDD gene 210
III-2-4-10 Mouse Nit1 gene 214
III-2-4-11 Mouse Adamts4 gene 218
III-2-4-12 Mouse Fc epsilon RI gamma subunit gene 224
III-2-4-13 Mouse L27´processed pseudogene 227
III-2-4-14 Mouse RIKEN1 gene 228
III-2-4-15 Mouse RIKEN2 gene 231
III-2-4-16 Mouse RIKEN3 gene 234
III-2-4-17 Repeat elements in the mouse BAC clone rp23-395h6 236
III-2-5 Mouse BAC clone rp23-9p22 238

III-2-6 Fine mapping of the QTL of bone mineral density 240

Chapter IV: Conclusion 242

Bibliography 256

List of tables

1.1- Summarized functions and interactions of the 200 family of proteins 28

1.2- Summarized functions and interactions of the human 200 family of proteins 35

3.1- The four mouse BAC clones used in this study 65

3.2- Exon sizes and contig positions of the Ifi202a gene 69

3.3- Neural network splice sites predicted sequences (Ifi202a) 70

3.4- Potential binding sites for transcription factors 71

3.5- Exon sizes and contig positions of the Ifi202b gene 74

3.6- Neural network splice sites predicted sequences (Ifi202b) 75

3.7- Comparison of percentage nucleotide identity of Ifi202a and Ifi202b 77

3.8- Exon sizes and contig positions of the Ifi203b gene 79

3.9- Neural network splice sites predicted sequences (Ifi203b) 81

3.10- The six poly(A) sites located at the 3' end of the Ifi203 cDNA 83

3.11- Exon sizes and contig positions of the Ifi203a gene 86

3.12- Neural network splice sites predicted sequences (Ifi203a) 86

3.13- Comparison of percentage nucleotide identity of Ifi203a and Ifi203b 89

3.14- Exon sizes and contig positions of the last three exons of the Ifi203c gene 90

3.15- Neural network splice sites predicted sequences (Ifi203c) 91

3.16- Exon sizes and contig positions of the Ifi204 gene 92

3.17- Neural network splice sites predicted sequences (Ifi204) 93

3.18- Identification of the repeat elements on contig 1 114

3.19- Identification of the repeat elements on contig 2 115

3.20- Contig position, biological effect and tissue distribution of the Ifi200 genes 116

3.21- Identification of the repeat elements in the human Ifi200 gene cluster 137

3.22- Synonymous and non-synonymous substitution rates 138

3.23- Summary of the genes present in the mouse BAC clones 144

3.24- Exon sizes and contig positions of the SLAM gene 148

3.25- Neural network splice sites predicted sequences (SLAM) 149

3.26- Exon sizes and contig positions of the CD48 gene 152

3.27- Neural network splice sites predicted sequences (CD48) 152

3.28- Exon sizes and contig positions of the 19A24 gene 158

3.29- Neural network splice sites predicted sequences (19A24) 159

3.30- Identification of the repeat elements on the mouse clone rp23-145f9 162

3.31- Exon sizes and contig positions of the Ly108 gene 166

3.32- Neural network splice sites predicted sequences (Ly108) 167

3.33- Exon sizes and contig positions of the CD84 gene 170

3.34- Neural network splice sites predicted sequences (CD84) 171

3.35- Identification of the repeat elements on the mouse clone rp23-77a8 174

3.36- Identification of the repeat elements on the mouse clone rp23-21118 176

3.37- Clone position and biological activities of the genes on clone rp23-395h6 179

3.38- Exon sizes and contig positions of the CAR gene 182

3.39- Neural network splice sites predicted sequences (CAR) 184

3.40- Exon sizes and contig positions of the P0 gene 187

3.41- Neural network splice sites predicted sequences (P0) 187

3.42- Exon sizes and contig positions of the apo A-II gene 190

3.43- Neural network splice sites predicted sequences (apo A-II) 190

3.44- Exon sizes and contig positions of the NDUFS2 gene 194

3.45- Neural network splice sites predicted sequences (NDUFS2) 195

3.46- Exon sizes and contig positions of the beta4Gal-T3 gene 198

3.47- Neural network splice sites predicted sequences (beta4Gal-T3) 199

3.48- Exon sizes and contig positions of the Ppox gene 204

3.49- Neural network splice sites predicted sequences (Ppox) 205

3.50- Exon sizes and contig positions of the USP23 gene 209

3.51- Neural network splice sites predicted sequences (USP23) 209

3.52- Exon sizes and contig positions of the DEDD gene 213

3.53- Neural network splice sites predicted sequences (DEDD) 213

3.54- Exon sizes and contig positions of the Nit1 gene 217

3.55- Neural network splice sites predicted sequences (Nit1) 218

3.56- Exon sizes and contig positions of the Adamts4 gene 220

3.57- Neural network splice sites predicted sequences (Adamts4) 221

3.58- Exon sizes and contig positions of the Fcer1g gene 226

3.59- Neural network splice sites predicted sequences (Fcer1g) 226

3.60- Exon sizes and contig positions of the RIKEN1 gene 229

3.61- Neural network splice sites predicted sequences (RIKEN1) 230

3.62- Exon sizes and contig positions of the RIKEN2 gene 232

3.63- Neural network splice sites predicted sequences (RIKEN2) 233

3.64- Exon sizes and contig positions of the RIKEN3 gene 235

3.65- Neural network splice sites predicted sequences (RIKEN3) 235

3.66- Identification of the repeat elements on the mouse clone rp23-395h6 238

3.67- Identification of the repeat elements on the mouse clone rp23-9p22 240

3.68- Phylogenic analysis of the proteins coded by the genes present on the two 0.37- and 1.4-Mb mouse contigs 252

List of Figures

1.1- Two-dimensional structure of the Watson and Crick's double helix model 4

1.2- Three-dimensional structure of the Watson and Crick's double helix model 6

1.3- Mechanism of gene activation by interferons 24

1.4- Summarized functions of the proteins encoded by Ifi202a and Ifi204 27

1.5- Summarized functions of the proteins encoded by the human 200 family 36

2.1- Summarized steps of the random shotgun sequencing strategy 44

3.1- Chromosomal location of the mouse BAC clones 64

3.2- Location of the genes of the mouse Ifi200 cluster on the two contigs 66

3.3- GENSCAN analysis of contig 1 67

3.4- Structural organization of the Ifi202a gene and its encoded protein 68

3.5- Alignment of the amino acid sequences of p202a and p202b 73

3.6- Dot-plot analysis of the Ifi202a and Ifi202b genomic regions 76

3.7- Dot-plot analysis of the 5' flanking regions of the Ifi202a and Ifi202b genes 78

3.8- Structural organization of the Ifi203b gene and its encoded protein 80

3.9- Alignment of the amino acid sequences of p203a and p203b 85

3.10- Dot-plot analysis of the Ifi203a and Ifi203b genomic regions 88

3.11- Dot-plot analysis of the 5' flanking region of the Ifi203a and Ifi203b genes 89

3.12- Structural organization of the Ifi204 gene and its encoded protein 92

3.13- Dot-plot analysis of the 5' flanking region of the Ifi202a and Ifi204 genes 95

3.14- Clustal alignment of the amino acid sequences of the p202a, p202b proteins and the putative p202c protein 99

3.15- GENSCAN analysis of contig 2 100

3.16- Dot-plot analysis of the Ifi203b gene and the putative "203-like" pseudogene 102

3.17- Dot-plot analysis of the "203-like" pseudogene and the Ifi203 cDNA 102

3.18- Dot-plot analysis of the Ifi204 gene and the putative "204-like" gene 103

3.19- Dot-plot analysis of the putative "204-like" gene and the Ifi204 cDNA 104

3.20- Alignment of the cDNA sequences of the Ifi204 gene and the "204-like" putative gene 109

3.21- Alignment of the amino acid sequences of p204 and the putative "p204-like" 110

3.22- Alignment of the amino acid sequences of the D3 protein and the virtual "D3-like" protein 112

3.23- Structural organization of the putative "204-like" and its encoded protein 112

3.24- Dot-plot analysis of contig 1 and itself 117

3.25- Dot-plot analysis of contig 1 and contig 2 118

3.26- Percent identity plot for the mouse Ifi200 family gene cluster 123

3.27- Dotplot comparative scale of mouse and human cluster sequences 125

3.28- Percent identity plot for the human Ifi200 family gene cluster 127

3.29- CLUSTALW alignment comparison between the 5' flanking region of IFI-16, MNDA, Ifi202a, Ifi204 and Ifi203a genes 132

3.30- CLUSTALW alignment comparison between the 5' flanking region of Ifi202a, Ifi204, Ifi203a and MNDA 134

3.31- CLUSTALW alignment comparison between the 5' flanking regions of MNDA and IFI-16 135

3.32a- Dot-plot comparison of the mouse and human Ifi200 clusters 140

3.33a- Postulated evolution of the Ifi200 gene cluster 142

3.32- ALIGN comparison between the human and mouse SLAM proteins 147

3.33- GENSCAN analysis of the mouse BAC clone rp23-145f9 147

3.34- Structural organization of the SLAM gene and its encoded protein 148

3.35- Structural organization of the CD48 gene and its encoded protein 151

3.36- ALIGN comparison between the human and mouse CD48 proteins 153

3.37- CLUSTALW comparison between the human and mouse 19A24 cDNAs 156

3.38- ALIGN comparison between the human and mouse 19A24 proteins 157

3.39- Structural organization of the 19A24 gene and its encoded protein 158

3.40-TMHMM search for transmembrane domain of the 19A24 protein 159

3.41- Position of the genes on the contig generated by merging BAC clones rp23-145f9 and rp23-77a8 162

3.42- GENSCAN analysis of the mouse BAC clone rp23-77a8 164

3.43- Structural organization of the Ly108 gene and its encoded protein 165

3.44- ALIGN comparison between the human and mouse Ly108 proteins 168

3.45- Structural organization of the CD84 gene and its encoded protein 169

3.46- ALIGN comparison between the human and mouse CD84 proteins 172

3.47- GENSCAN analysis of the mouse BAC clone rp23-21118 175

3.48- GENSCAN analysis of the mouse BAC clone rp23-395h6 177

3.49- Position of each genes and pseudogenes on the clone rp23-395h6 178

3.50- ALIGN comparison between the human and mouse CAR proteins 181

3.51- Structural organization of the CAR gene and its encoded protein 183

3.52- Structural organization of the myelin P0 gene 186

3.53- ALIGN comparison between the human and mouse myelin P0 188

3.54- Structural organization of the Apo A-II gene 190

3.55- ALIGN comparison between the human and mouse apo A-II proteins 191

3.56- Structural organization of the NDUFS2 gene 193

3.57- ALIGN comparison between the human and mouse NDUFS2 proteins 196

3.58- Structural organization of the beta4Gal-T3 gene and its encoded protein 198

3.59- ALIGN comparison between the human and mouse beta4Gal-T3 proteins 200

3.60- TMHMM search for transmembrane domain of the beta4Gal-T3 protein 200

3.61- ALIGN comparison between the human and mouse Ppox protein 203

3.62- Structural organization of the Ppox gene 204

3.63- ALIGN comparison between the human and mouse Usp23 protein 208

3.64- Structural organization of the Usp23 gene 208

3.65- ALIGN comparison between the human and mouse DEDD protein 212

3.66- Structural organization of the DEDD gene and its encoded protein 213

3.67- ALIGN comparison between the human and mouse nit1 protein 215

3.68- Structural organization of the last six exons of the nit1 gene 217

3.69- Structural organization of the ORF of the putative Adamts4 gene and its encoded protein 220

3.70- ALIGN comparison between the human ADAMTS4 protein and its putative mouse counterpart 222

3.71- ALIGN comparison between the human and mouse Fcer1g protein 225

3.72- Structural organization of the Fcer1g gene and its encoded protein 225

3.73- Dot-plot analysis between the L27 pseudogene and the L27 cDNA 227

3.74- Structural organization of the hypothetical RIKEN1 gene 229

3.75- ALIGN comparison between the mouse Tom40 and the RIKEN1 proteins 231

3.76- Structural organization of the hypothetical RIKEN2 gene 232

3.77- ALIGN comparison between the CGI-126 and the RIKEN1 proteins 234

3.78- Structural organization of the first five exons of the RIKEN3 gene 234

3.79- ALIGN comparison between the human cybL and the RIKEN1 proteins 236

3.80- GENSCAN analysis of the mouse BAC clone rp23-9p22 238

3.81- Corrected map of the QTL region of bone density 240

3.82- Phylogenic tree for the concatenated set of pxf, Usp23, prefoldin subunit 2, NDUFS2 and beta4Gal-T3 proteins 253

List of Abbreviations

ABI - Applied Biosystems Inc.

Adamts4 - A Disintegrin And Metalloprotease with Thrombospondin motifs 4

AH - Absorptive Hypercalciuria

AIM2 - Absent In Melanoma 2

ApoA-II - Apolipoprotein A-II

ATP - Adenosine Tri-Phosphate

ATPase - ATP hydrolase

BAC - Bacterial Artificial Chromosome

bp - base pairs

BCM1 - CD48 antigen

Beta4Gal-T3 - UDP-Gal:beta-Glc NAc beta-1,4-galactosyltransferase 3

BLAST - Basic Local Alignment Search Tool

BMD - Bone Mineral Density

cAMP - cyclic adenosine mono-phosphate

CAP II - Computer Assembly Program II

CAR - Constitutive Androstanol Receptor *or* Constitutive Activator of Retinoic acid response elements

cDNA - complementary DNA

CFTR - Cystic Fibrosis Transmembrane Regulator

Consed - Consensus editor

CREB - Cyclic AMP Regulatory Element Binding protein

CYP - Cytochrome P450

dbEST - EST database

dCTP - deoxycytosine tri-phosphate

DD - Death Domain

ddNTP - dideoxynucleotide tri-phosphate

DED - Death Effector Domain

DEDD - DED-containing DNA-binding protein

dGMP - deoxyguanosine mono-phosphate

dGTP - deoxyguanosine tri-phosphate

dITP - deoxyinosine tri-phosphate

DMF - Dimethyl Formamide

DMSO - Dimethyl Sulfoxide

DNA - Deoxyribonucleic Acid

DNAse - deoxyribonuclease

dNTP - deoxynucleotide tri-phosphate

dsRNA - double-stranded RNA

EBV - Epstein-Barr Virus

EDTA - Ethylene Diamine Tetraacetic Acid

eIF2 - eukaryotic Initiation Factor 2

EMBL - European Molecular Biology Laboratory

EST - Expressed Sequence Tag

FADD - Fas-Associated Death Domain

Fcer1g/Fc epsilon RI - Fc epsilon RI gamma subunit

FGENESH - FGENES Human (Human gene prediction program)

FIGE - Field Inversion Gel Electrophoresis

FISH - Fluorescence *In Situ* Hybridization

GAS - interferon Gamma Activation Site

GET - Glucose EDTA Tris

GRAIL 2 - coding region recognition program

GTPase - GTP hydrolase

HDL - High Density Lipoprotein

HTGS - High Throughput Genome Sequencing

IFI-16 - human interferon-activatable 200 family gene

IFN - Interferon

IgE - Immunoglobulin E

IGF-1 - Insulin Growth Factor-1

Ik - Ikaros

IL-6 - Interleukin-6

IPTG - isopropyl-beta-D-thiogalactopyranoside

IRF - Interferon Regulatory Factor

ISRE - Interferon-Stimulated Response Element

JAK - Janus Kinase

kb - kilobase pairs

KOAc - potassium acetate

LB - Luria-Bertani broth

LINE - Long Interspersed repeat Element

LTR - Long Terminal Repeat

mAb - monoclonal Antibody

Mb - megabase pairs

MHC - Major Histocompatibility Complex

MMCV - Mouse Mammary Cytomegalovirus

MNDA - Myeloid Nuclear Differentiation Antigen

mRNA - messenger RNA

MZEF - Michael Zhang Exon Finding program

NADH - Nicotinaminide Adenine Dinucleotide

NaOAc - sodium acetate

NDUFS2 - Nicotinaminide adenine Dinucleotide (NADH): Ubiquinone oxidoreductase complex I Fe-S subunit 2

NF - Nuclear Factor

Nit1 - Nitrilase 1

NMR - Nuclear Magnetic Resonance

NNPP - Neural Network Promoter Prediction program

NNSSP - Neural Network Splice Site Prediction program

ORF - Open Reading Frame

P0 - myelin Protein zero

PAAD/DAPIN - Pyrin, AIM (absent in melanoma), ASC (apoptosis-associated Speck-like protein containing a caspase recruitment domain (CARD), and Death-domain (DD)-like/Domain in APoptosis and INterferon response

PBRE - Phenobarbital Response Element

PCR - Polymerase Chain Reaction

PDGF - Platelet-Derived Growth Factor

Phred/Phrap - Phil's (Phil Green) read editor/Phil's read assembly program

PIP - Percent Identity Plot

PMCA - Plasma Membrane Calcium-ATPase

PPi - inorganic pyrophosphate

Ppox - Protoporphyrinogen Oxidase

pRb - Retinoblastoma protein

QTL - Quantitative Trait Locus

RACE - Rapid Amplification of cDNA Ends

rATP - riboadenosine tri-phosphate

rDNA - ribosomal DNA

RE - Regulatory Element

RIKEN - genomic sequencing center

RNA - Ribonucleic Acid

RT-PCR - Reverse Transcription -PCR

SAP - SLAM-Associated Protein

SDS - Sodium Dodecyl Sulfate

SERCA - Sarco(endo)plasmic Reticulum Calcium-ATPase

SINE - Short Interspersed repeat Element

SLAM - Signaling Lymphocytic Activation Molecule

SNP - Single Nucleotide Polymorphism

SREBP-1 - Sterol Regulatory Element Binding Protein-1

STAT - Signal Transducer and Activator of Transcription

STS - Sequence Tagged Site

Taq - *Thermus aquaticus*

TE - Tris EDTA

TGF - Transforming Growth Factor

TM - Tris Magnesium

TMHMM - Transmembrane topology prediction method based on a Hidden Markov Model

tRNA - transfer RNA

UBF-1 - Upstream Binding Factor-1

USP23 - Ubiquitin-Specific Protease 23

UTR - UnTranslated Region

VNTR - Repeat polymorphism of variable length

X-Gal - 5-bromo-4-chloro-3-indolyl-beta-D-galactoside

XGAP - X window-based Genome Assembly Program

XLP - X-linked Lymphoproliferative disease

YENB - Yeast Extract Natural Broth

YY1 - Ying and Yang 1

Abstract

The random shotgun sequencing approach, followed by a new strategy to sequence numerous and strong repeats in the BAC clones containing genomic DNA, were used to sequence two regions totaling ~2 Mb of mouse chromosome 1q during this dissertation research.

One region of mouse chromosome 1q contains the murine interferon-activatable Ifi200 gene cluster which contains at least ten 200 family genes and pseudogenes, including Ifi201, Ifi202a, Ifi202b, Ifi202c, Ifi203a, Ifi203b, Ifi203c, "203-like", Ifi204 and "204-like", coding for proteins which are involved in cell growth regulatory, immunomodulatory, antimicrobial and cell differentiation activities. This genomic region likely arose by a duplication or triplication of a segment including two adjacent genes. Three new 200 family genes, "204-like", Ifi203a and Ifi203c, were localized in this region in addition to two putative pseudogenes ("203-like" and Ifi202c), and 5 previously identified genes (Ifi201, Ifi202a, Ifi202b, Ifi203, and Ifi204). A putative spliced variant of the "204-like" gene product exhibited over 86% sequence identity with an mRNA encoded for another member of the 200 family of proteins (D3). The sequences of the Ifi202a and b genes, as well as of the Ifi203a and b genes, were over 99% identical. The translation product of Ifi203b differed from that of Ifi203a by only 1 amino acid out of 408. Identification of an additional transcribed coding segment corresponding to a region in introns 4 of Ifi203a and b suggests the possibility of a previously unreported alternatively splice product. Comparative analysis of the mouse and the human 200 clusters indicated a relative absence of sequence conservation in non-coding regions.

The second region of mouse chromosome 1q corresponds to a Quantitative Trait Locus (QTL) of Bone Mineral Density (BMD), and is located near the Ifi200 gene cluster. Bone mineral density is the strongest determinant of osteoporotic fracture, and 60 to 70% of the normal variability in human bone density is genetically determined. To accelerate the

identification of genes present in QTL of bone density. 6 BACs, totaling about 1.4 Mb, were sequenced. Their sequences were merged with the sequences of 7 other BAC clones mapped to the same region, and sequenced by other laboratory members, to create two contigs of approximately 1.4 Mb and 0.37 Mb, containing at least 40 genes. Analysis of this QTL region indicates the presence of immune- and inflammation-associated genes. At least 5 putative calcium-ATPase genes, whose function may be involved in BMD regulation, also were present in the QTL. A metabolic disorder caused, at least partially, by an increase bone resorption, and known as absorptive hypercalciuria (AH), was mapped to the QTL of BMD (Reed *et al.*, 1999). The putative calcium-ATPases whose genes are located on the QTL may be involved in maintaining calcium homeostasis inside the bone cell, and a calcium-ATPase-encoding gene defect therefore could be responsible for AH.

Chapter I

Introduction

The major goals of the work completed and reported in this dissertation were to investigate two closely related regions on the q arm of mouse chromosome 1, one encoding several interferon-inducible proteins, and another encoding region implicated in modulating and/or regulating bone mineral density.

I-1 Human and Mouse Genomes

I-1-1 DNA structure

I-1-1-1 DNA is the carrier of hereditary information

In the 1860s, linear bodies in the eukaryotic cell nuclei, named "chromosomes", were observed by staining with chemical dyes. Then, in 1866, G. Mendel, who was working on a series of genetic crosses between true-breeding strains of garden peas (*Pisum sativum*), differing in certain traits such as shape and color, discovered the basic laws of heredity when he reported that the pairs of contrasting traits of the progeny all resulted from a "factor" (now known as a gene) that had alternative "forms" (allelic forms), and that different traits were independantly inherited. His work, largely ignored by his contemporaries, was rediscovered at the turn of the 20th century and sparked a renewed interest in the scientific community. In 1903, W. Sutton, realizing that genes and chromosomes behave in a parallel fashion, formulated the chromosomal theory of heredity, in which he hypothesized that genes were part of chromosomes. The pace of discovery greatly improved when T.H. Morgan began working with *Drosophila*

melanogaster as an experimental model. This small insect was easily maintained in laboratory conditions where it was able to produce a new generation every 14 days. In succeeding years, the chromosomal location of many *Drosophila* genes was determined by crossing experiments and the notion of linked genes lying on the same chromosome was introduced. In 1902, A. Garrod was the first to suggest a specific connection between gene and enzyme by working on individuals with alkaptonuria. He showed that this metabolic disorder linked to the absence of an enzyme metabolizing the homogentisate that alkaptonurics excrete actually was inherited in a Mendelian fashion. In 1940, G. Beadle and E. Tatum worked on irradiated mutants of *Neurospora crassa* and showed that there was a one-to-one correlation between a mutation and the lack of a specific enzyme, linking genes and proteins together. However, none of the above studies, while delimiting the cellular aspect of heredity and its linkage to chromosome and genes, were able to determine the molecular factors involved in it.

Nucleic acids were first discovered in 1869 by F. Miescher from the nuclei of leukocytes from discarded surgical bandages. However, it was not until about 75 years after their discovery that their correlation to genetic studies was made. Indeed, it was widely believed in the 1930s and in the 1940s that nucleic acids were actually made of a monotonously repeating pattern of the four bases adenine, cytosine, guanine and thymine, so that no genetic function could be derived from them. The idea that genetic material was nucleic acid came in fact from the discovery of transformation by Griffith in 1928. Griffith injected mice with a mixture of live non pathogenic and heat-killed virulent pneumococci, resulting in the death of most of the mice. The surprising fact was that the blood of the dead mice contained live pneumococci that somehow had been transformed from non-pathogenic to virulent forms. Furthermore, the progeny of the transformed pneumococci were also virulent, suggesting that the transformation was permanent. In 1944, O. Avery discovered that the "transforming principle" that transformed the non-pathogenic pneumococci to virulent ones in Griffith's experiment

was DNA, based on the fact that it was unaffected by trypsin, chymotrypsin and ribonuclease but totally inactivated by DNAse (Avery *et al.*, 1944). DNA was therefore the carrier of genetic information. The next step was to demonstrate that DNA was in fact the hereditary molecule of many organisms. In 1952, Hershey and Chase infected bacteria with T_2 phages specifically radiolabeled in their DNA (with ^{32}P) or in their proteins (with ^{35}S). The infected bacteria were agitated in a kitchen blender and two fractions were obtained by centrifugation. One, containing the empty phage capsides, was shown to contain most of the ^{35}S, whereas the other one, containing the infected bacteria, contained most of the ^{32}P. Furthermore, 30% of the ^{32}P and only 1% of the ^{35}S appeared in the progeny phages. Hershey and Chase therefore concluded that only the phage DNA was essential for the production of the progeny and that DNA was the carrier of the hereditary information (Hershey and Chase, 1952).

I-1-1-2 DNA structure

The determination of the structure of DNA by Watson and Crick (1953) is said to have marked the birth of what is now known as molecular biology. Not only did Watson and Crick correctly predict the 3D structure of DNA (see Fig. 1.1), but they also suggested the molecular mechanism of heredity. The Watson-Crick structure of DNA, known as the double helix structure, was based on several earlier studies. First, E. Chargaff had observed in 1949 that there was a one-to-one ratio of adenine to thymine and cytosine to guanine in DNA, thus determining an accurate base ratio of DNA (Chargaff *et al.*, 1949). Second, Jerry Donohue, a laboratory mate of Watson and Crick, had established by X-ray, NMR and spectroscopic studies that the correct tautomeric form of the bases was actually the keto form. This was of great importance as, in 1953, guanine and thymine were widely believed to be in their enol forms in order to maximize their resonance stability and the knowledge of the dominant keto form was a

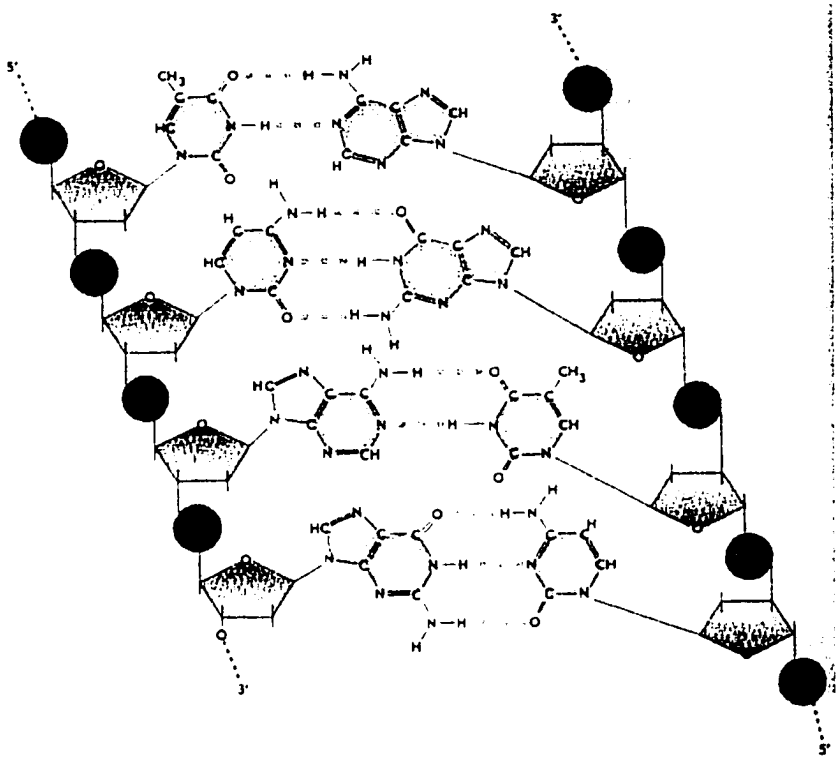


Fig. 1.1: Two dimensional structure of the Watson and Crick's double helix model

prerequisite for the prediction of the correct hydrogen bonding associations between the bases. Third, an X-ray diffraction photograph of a DNA fiber taken by R. Franklin (Wilkins *et al.*, 1951) had allowed Crick, an X-ray crystallographer by training, to deduce that DNA was a helical molecule whose aromatic bases formed a stack of parallel rings that was parallel to the fiber axis. This information allowed Watson and Crick to intellectually model the structure of the DNA double helix (Watson and Crick, 1953). In Watson and Crick's double helix model, the double stranded DNA molecule consists of two polynucleotide strands that wind about a common axis with a right-handed twist to form a ~20Å diameter double helix. The two strands are antiparallel and wrap around each other (Fig. 1.1). The bases occupy the center of the helix and the sugar-phosphate backbone is coiled about its periphery (Fig. 1.2). Each base is hydrogen bonded to a base on the opposite strand to form a planar base pair. This hydrogen bonding is the phenomenon that results in the specific association of the two strands of the double helix. The Watson-Crick secondary structure of DNA has two deep exterior grooves (minor groove and major groove) that wind between its sugar-phosphate backbone. These two exterior grooves are of unequal size because each edge (top and bottom) of each base pair is structurally distinct and also because the deoxyribose residues are asymmetric. A remarkable feature of this structure is that it can only accomodate two types of base pairing. Each adenine residue must pair with a thymine residue and each guanine residue must pair with a cytosine residue. The Watson-Crick structure can accomodate any sequences of bases on one strand, as long as the opposite strand has the complementary sequence, thus accounting for the rules described by Chargaff (1951), and suggesting that the sequence of bases on either strand is the carrier of hereditary information.

I-1-2 From structure toward function

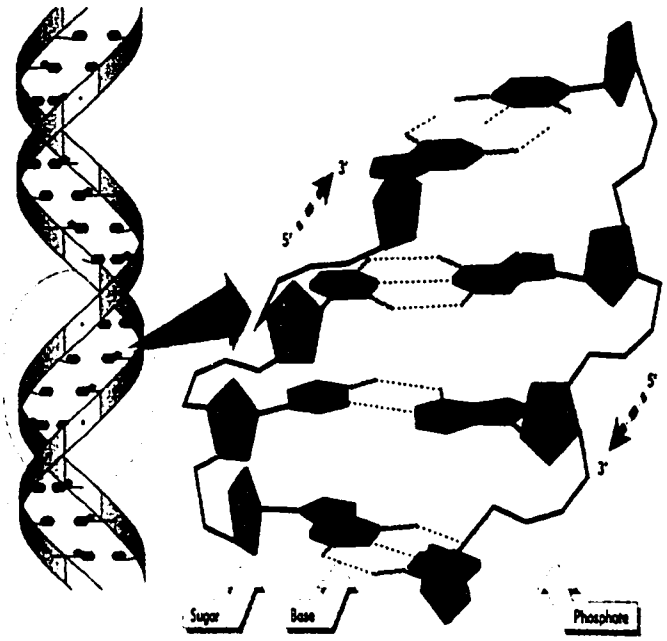


Fig. 1.2: Three-dimensional structure of the Watson and Crick's double helix model.

In 1958, Crick proposed what eventually became the "Central Dogma" of molecular biology: DNA directs its own replication and its transcription to RNA which, in turn directs its translation to proteins. This hypothesis was confirmed by a series of experiments. In 1959, A. Pardee discovered that a protein, known as the *lac* repressor, inhibited the synthesis of *lac* operon proteins (Pardee *et al.*, 1959). The nature of the *lac* repressor's target molecule was deduced in 1961 by Jacob and Monod (Jacob and Monod, 1961), and their observations led them to the conclusion that proteins are synthesized in a two stage process, the first involving the synthesis of a messenger RNA (mRNA) whose sequence was complementary to the sequence on the coding strand of DNA and the second involving translation of the information in the mRNA into protein. This model developped by Jacob and Monod later was confirmed by experiments designed by Brenner *et al.* (1961) in which they showed that mRNAs transiently associates with ribosomes, which are essential for polypeptide synthesis. In addition, the development of the RNA-DNA hybridization technique by S. Spiegelman demonstrated that all cellular RNAs were transcribed from a DNA template and that these mRNA sequences were complementary to the sequence of the DNA template (Hall and Spiegelman, 1964). RNA polymerase, the enzyme responsible for the DNA-directed synthesis of RNA, was discovered independently in 1960 by S. Weiss and J. Hurwitz. The binding of RNA polymerase to DNA was demonstrated in prokaryotes through electron micrograph studies of bacteriophage-infected *E. coli* cells (Williams, 1977) and the existence of such binding sites, known as promoters, was recognized through mutation studies enhancing or diminishing the transcription rates of certain genes, including those of the *lac* operon. The "Central Dogma" of molecular biology was further illustrated by a series of experiments made in 1961 by F. Crick and S. Brenner (Crick *et al.*, 1961). In these experiments, they collected a series of proflavin-induced mutations, designated FC0, FC1, FC2, FC3... in the rIIB cistron of bacteriophage T_4. They discovered that the growth of these mutant phages on a permissive *E. coli* B host

resulted in the spontaneous appearance of some phenotypically wild type phages, as demonstrated by their ability to grow on a restrictive *E. coli* K12 host. These phages were not genotypically wild type but rather double mutants containing two successive mutations, FC0 and FC1 or FC1 and FC2... In addition, recombination studies showed that recombinants containing three odd-numbered or three even-numbered mutations were phenotypically wild type. They concluded from these data that each mutation was a suppressor of the following mutation, i.e., each mutation was a single nucleotide insertion or a single nucleotide deletion, followed by another mutation that was respectively a single nucleotide deletion or a single nucleotide insertion. They also concluded that the genetic code was a triplet code, in which each triplet codes for a particular amino acid. The next step in deciphering the genetic code came in 1961 when M. Nirenberg and J. Matthaei established that UUU was the codon specifying phenylalanine (Phe) (Nirenberg and Matthaei, 1961), by demonstrating that the addition of poly(U) to a cell-free protein translation system stimulated the synthesis only of poly(Phe). Nirenberg then deciphered the entire genetic code by using other ribonucleotide polymers in cell-free protein translation system and by triplet binding assays. Nirenberg's triplet binding assays were based on the discovery that trinucleotides, as well as mRNAs, are able to promote the ribosomal binding of specific tRNAs (Nirenberg and Leder, 1964), thus allowing the various triplets to be identified. The existence of tRNAs was first hypothesized by F. Crick who, in 1955, proposed that translation was occuring through the mediation of "adaptor molecules" in order to somehow translate base sequences into amino acid sequences. In 1965, R. Holley reported the base sequence of yeast alanine tRNA (Holley *et al.*, 1965) and predicted a "cloverleaf" secondary structure. Ribosomes were first seen in cell extracts by dark field microscopy in the late 1930s by A. Claude, who referred to them as "microsomes". However, it was not until 1955, and the work of P. Zamecnik who demonstrated that ^{14}C-labeled amino acids are transiently associated with ribosomes before they appear in

proteins, that ribosomes were considered the site of protein synthesis. Finally, the "Central Dogma" of molecular biology was amended by the discovery of a RNA-directed DNA polymerase (reverse transcriptase) in retroviruses, which are RNA-containing eukaryotic viruses . This enzyme, which was independently discovered in 1970 by H. Temin and D. Baltimore, synthesizes DNA in the 5' to 3' direction from a primed template, in the same way as DNA polymerases. However, in this case, the template is RNA. Reverse transcriptases are a particulary useful tool in molecular biology because of their ability to transcribr mRNAs to complementary strands of DNA (cDNA).

There are three classes of eukaryotic RNA polymerases based on the RNA they synthetize. Class I, located in the nucleolus, synthesizes precursors of most ribosomal RNAs. Class II, located in the nucleoplasm, synthesizes messenger RNA precursors. Class III, also located in the nucleoplasm, synthesizes the precursor of 5S ribosomal RNA, tRNAs and other small nuclear and cytosolic RNAs. Since class II synthesizes RNAs that are necessary for a proper metabolism of the cell, sequences recognized by class II RNA polymerases are the most well characterized promoters. These promoters are longer and more diverse than those of prokaryotic genes. The genes expressed in all tissue, which are thought to be constitutively transcribed, exhibit one or more copies of the sequence GGGCGG or its complement (the GC box) upstream from their transcription start sites. Genes that are selectively expressed in certain tissues or certain types of cells often lack these GC-rich sequences but rather contain an AT-rich sequence (the TATA box) located 25 to 30 bp upstream from their transcription start sites. The region extending between 50 and 110 bp upstream from the transcription start site also contains promoter elements. Among them are a conserved consensus sequence CCAAT (the CCAAT box) whose alteration greatly reduces the transcription rate. It appears that the consensus sequences located upstream of the TATA box form the initial DNA-binding sites for RNA polymerase II and other proteins involved in the initiation

of transcription known as transcription factors. Such sequences, known as enhancers, apparently mediate much of the selective gene expression in eukaryotes. Three different classes of transcription factors have been implicated in regulating the initiation of transcription by RNA polymerase II. The first class contains general transcription factors that are required for the synthesis of all mRNAs. Many of these, such as TFIIA, TFIIB, TATA-Binding Protein (TBP), are multiprotein complexes that interact in an ordered pathway with RNA polymerase II and promoter sequences near the transcriptional start site. The second class contains upstream transcription factors that bind DNA sequences 5' to the transcriptional start site and thereby either stimulate or repress transcriptional initiation by RNA polymerase II when it is complexed with general transcription factors. Many of these upstream transcription factors have been shown to be important in regulating cellular development. For example, upstream transcription factors such as GATA-1, NF-E2, NF-E3, NF-E4, Ik-1, Ik-2 or Sp1 are essential elements in controlling the differential expression of the various globin genes in the embryo, fœtus and adult. The third class contains inducible transcription factors that function similarly to upstream transcription factors but must be activated, or repressed, by phosphorylation or specific ligand binding in order to bind to their target DNA sequence. They are synthesized and activated in specific tissues at particular times and therefore mediate gene expression in a timely fashion. Steroid receptors, an example of inducible transcription factors, are activated by steroid hormone binding as the newly formed complex translocates into the nucleus where it binds to specific DNA segments known as response elements, and activates the initiation of transcription of specific genes.

The early products of eukaryotic transcription, the primary transcripts, are not immediate functional entities. The experiments by P. Sharp (Berget *et al.*, 1977) and R. Roberts (Chow *et al.*, 1977) demonstrated that primary transcripts were processed by the excision of non coding internal sequences. This process, known as RNA splicing,

was correlated to the structure of eukaryotic genes in which coding sequences (exons) are interspersed with unexpressed regions (introns). The size of exons varies among eukaryotic genes, from several to thousands of base pairs (bp). However, the average exon size is one or two hundred bp. The number of exons also varies among eukaryotic genes, from single exon genes to genes containing more than 70 exons, such as the ~2500 kb human dystrophin gene. In higher eukaryotes, introns generally are much larger than exons, usually spanning several thousand bp and typically constitute over ~80% of a vertebrate primary gene transcript. Most introns do not appear to carry any particular function and are often considered only as a result of millions of years of evolution and chromosomal rearrangements. However, this idea is difficult to rationalize and there are several studies that support the concept that introns play a positive role in regulating eukaryote gene transcription. Although it is difficult to explain why the evolution of such a fine tuned machinery as the splicing machinery offers any selective advantages over the simple elimination of the split genes, it has been proposed that introns were actually genetic "fossils", DNA that had an essential function at an earlier stage during evolution (Gilbert *et al.*, 1986). Many protein-coding genes arose during evolution as collections of exons assembled by recombination between intron sequences. For instance, pyruvate kinase is encoded by ten exons, each coding for a discrete element of the protein's secondary structure (Lonberg and Gilbert, 1985). Introns therefore might be remnants of a process that facilitated protein evolution by combining a series of small protein-coding units and expressing them as a single protein. This process, known as exon shuffling, might also explain why different genes in the same organism, or in different organisms, show sequence similarity between some of their exons. However, if introns are remnants of some primitive exon shuffling that could have arisen before the speciation of intron-containing species, such as plants or vertebrates, it is difficult to explain why they are absent from lower organisms, i.e., prokaryotes. One explanation is that introns might have been selectively eliminated in

lower organisms because of a stronger selective pressure and a quicker turn-over rate (intron-early theory). Another hypothesis is that introns actually appeared in eukaryotes, thus disrupting the structure of early eukaryotic genes containing no introns (intron-late theory). The origin and function of introns remain enigmatic but their existence may demonstrate that the need for structural diversity in eukaryotic organisms may have been answered by a series of phenomenons involving coding, as well as non- coding, regions of the genome.

I-1-3 DNA sequencing technology

The understanding that DNA is the carrier of hereditary information in a cell and that this information is encoded by the order of bases in DNA led to the emergence of new technologies to determine the primary structure of DNA, i.e., its sequence. Determining the base order of any DNA has become a powerful approach for understanding the primary structure, and the expression, of the proteins performing the biological activities of a cell.

Two basic methods to sequence DNA were developed. In the method developed by Maxam and Gilbert (1977), the DNA to be sequenced was radiolabeled at its 5' end. Four different chemical cleavages of the DNA then were performed separately, addressing the base immediately adjacent to the site of cleavage. One chemical cleavage was made by hydrazine which, alone, cleaves after cytosine and guanine, and with salts, cleaves only after cytosine. Dimethylsulfate with heat cleaves after guanine and, with acids, cleaves after adenine, but also weakly guanine. The DNA fragments generated by each cleavage reaction then were loaded separately on an acrylamide gel. Electrophoresis allowed migration according to their size. Because each fragment was radiolabeled at its 5' end, the DNA sequence could be read by autoradiography on the resulting gel, according to the position of the fragments and the identity of the chemical used to cleave

after a specific base. It must be noted that this method is very sensitive to the presence of secondary structure and compression. Consequently, the probability for each base to be cleaved greatly depends on its structural environment. In the method developed by Sanger *et al.* (1977), the DNA fragments are not generated by chemical cleavage but rather synthesized by enzymatic activity. In the original method, the enzyme used, a DNA polymerase, extended a primer radiolabeled at its 5' end using the DNA to be sequenced as a template. Again, four different reactions were performed in which the DNA template, the DNA polymerase and the radiolabeled primers were mixed with the four dNTPs and one of four ddNTP terminators. The DNA polymerase can not extend a polynucleotide chain if the 3' hydroxyl group usually present on the deoxyribose of the last base is missing. Thus the DNA polymerase-catalyzed chain elongation stopped every time a ddNTP was randomly introduced to the elongating DNA chain, this resulted in four base-specific reactions containing a series of fragments with different lengths that terminated with the same base. After loading the DNA fragments from the four reactions separately on a gel and then electrophoresing them, they migrated according to their size. Since each fragment was radiolabeled at its 5' end, an autoradiography was performed on the gel that has been dried and the exposed X-ray film was examined or read to determine the succession of bases according to the position of the fragment on the gel by noting the order of the fragments based on the identity of the terminating ddNTP. Again, this method is sensitive to the presence of secondary structure and compression that influence the progression of the DNA polymerase along the DNA template. However, improvements have been made to reduce this problem (see below).

The Sanger method has been widely used and is currently the method of choice in large scale genomic sequencing projects. However, several improvements to the method have been made over the years in order to increase both its efficiency and its accuracy. In 1986, Smith *et al.* reported a new method using fluorescent 5'-labeled

primers as an alternative to the use of radiolabeled primers for signal detection. Then, the use of fluorescent labeled ddNTP terminators for sequencing reaction was reported by Prober *et al.* in 1987. These improvements not only eliminated the need of manipulating radioactive materials, and the possible dangers linked to it, but also greatly increased the efficiency of any sequencing reactions by making it possible to perform the four base-specific reactions simultaneously in the same tube instead of being divided into four different tubes. The use of four base-specific fluorescent-labeled ddNTP terminators further allowed the four sequencing reactions to be loaded simultaneously on one lane of the electrophoresis gel, instead of four lanes as was the case with early sequencing experiments using the original Sanger method. Furthermore, fluorescent signal automated detection and computer assisted data collection and analysis increased the efficiency and accuracy of the sequencing data while decreasing the time to collect them. Finally, fluorescent-labeled ddNTPs reduced compressions because of the presence of a rather large dye moiety bound to the ddNTP that prevents base stacking because of its size. Background also is reduced, as abortive elongations remain undetected if they are not terminated by a fluorescent-labeled ddNTP. These fluorescent labeling methods have been commercialized by several companies and regularly updated to improve signal detection and quality. Two sequencing chemistry kits (BigDye Prism kit and dRhodamine kit) are currently available from Applied Biosystems Inc. and one from Amersham (ET kit), which mainly differ by the fluorescent dyes bound to each ddNTPs. Each dye emits light at a specific wavelength when excited by laser light and it has been shown experimentally that the quality of signal in regions difficult to sequence depends on the type of dyes used during the sequencing reaction.

Another improvement in DNA sequencing was developed by McBride *et al.* in 1989 and later optimized by Craxton in 1991. Here the sequencing reactions are cycled at three temperatures instead of incubated at a constant temperature. Cycle sequencing reactions are thermally cycled, with each cycle including three different denaturation,

annealing and elongation temperatures. The cycle sequencing method yields larger amounts of fluorescent-labeled DNA fragments, and consequently increased levels of signal. Furthermore, because a thermostable DNA polymerase (see below) is used during the reaction, the elevated temperature reduces DNA template secondary structure and compression, thus increasing the accuracy of the DNA sequence.

Other studies resulted in improved enzymes for DNA polymerization during the sequencing reaction. The original DNA polymerase used by Sanger was a fragment of the *E.coli* DNA polymerase I, called the Klenow fragment, from which any 5'-3' exonuclease activity had been removed by limited proteolysis (Klenow *et al.*, 1971). Since then, other DNA polymerases have been introduced that include the modified bacteriophage T_7 DNA polymerase (Sequenase) (Tabor and Richardson, 1989), the *Bacillus stearothermophilus* (Bst) DNA polymerase (Stenesh and Roe, 1972) and the *Thermus aquaticus* (Taq) DNA polymerase (Innis *et al.*, 1988). These enzymes belong to the same family of DNA polymerases but differ sharply in their enzymatic activities, their thermostabilities and their efficiency in incorporating ddNTPs. The modified T_7 DNA polymerase has a very high tendency to incorporate ddNTPs in the elongating DNA chain. Tabor and Richardson (1995) discovered that this ability to efficiently incorporate ddNTP was due to the Tyr residue at position 526. In contrast, the wild type Taq DNA polymerase has a Phe residue at the corresponding position (position 667). Because of its thermostability over a broad range of temperature, Taq DNA polymerase is the enzyme of choice for cycle sequencing reactions. Not surprisingly however, wild type Taq DNA polymerase has a very high rate of discrimination against incorporating ddNTPs, thus requiring a very high concentration of ddNTP terminators during the sequencing reaction. Fortunately, Tabor and Richardson (1995) discovered that a single amino acid substitution at position 667 (F667Y) on the wild type Taq DNA polymerase allows a much more efficient incorporation of ddNTPs, dramatically decreasing discrimination against it by 250 to 8000 fold. Based on this discovery, a new set of Taq

DNA polymerases was commercialized. For example, the AmpliTaq-FS (Applied Biosystems Inc.) and the Thermosequenase (Amersham) both carry the F667Y amino acid substitution. To further improve the "Klenow-like" properties, N-terminal modifications also were introduced to remove the 5'-3' exonuclease activity of these enzymes.

Finally, the use of automatic sequencers has been fundamental in obtaining rapid and accurate sequencing data. Several automated DNA sequencing instruments have been developped and commercialized. Until recently, polyacrylamide gel electrophoresis was the method of choice to run a sequencing reaction. This method, although accurate, was time-consuming (10 hours per run) and initially only sixteen reactions could be run together on a gel. Much progress has been made in this area to increase the number of reactions loaded on each gel but, before analysis, each sequence needed to be manually tracked and extracted, adding time to the already time-consuming process and introducing a certain margin of error in the tracking process. The introduction of capillaries filled with a synthetic polymer instead of polyacrylamide slab gels to resolve the sequencing reaction nested fragment set not only reduced the time needed to run a reaction (from 10 to 2 hours per run) but also eliminated the time consuming, error prone tracking of the sequencing data. Initially, several automated sequencers using slab gel electrophoresis, among them the Perkin Elmer ABI Model 377 (the last model of a series of sequencers including Model 370 and Model 373) and the LI-COR Long Read IR 4200 sequencer, were used in this dissertation research (see Materials and Methods). However, capillary sequencers that were used for the bulk of the data in this dissertation to collect shotgun sequencing data (see Materials and Methods) were manufactured by Perkin Elmer, the ABI Model 3700, and Molecular Dynamics, the MegaBACE 1000 sequencer. Although these sequencers differ in the media for migration of the DNA sample, they all consist of three major components: an electrophoresis instrument, a fluorescent signal detection unit and a computer including software tools for data

collection and analysis. An argon ion laser excites the ddNTP terminator-bound fluorescent dyes and the photons emitted from the excited dyes are detected by a photomultiplier tube (PMT) in the slab gel instruments or a charge coupled device (CCD) camera in the capillary instruments. Using the light intensities collected from the CCD camera or PMT, a computer software records and transforms the emission signals in a series of peaks of specific intensity and wavelength. Each peak corresponds to a specific ddNTP terminator-bound fluor dye and thus to a specific base. Typically, 64 reactions now can be loaded on the ABI 377 slab gels and 96 reactions can be loaded simultaneously on the capillary sequencers and each reaction generates a sequence with an average read length of 500 to 600 bp on either instrument.

I-I-4 Genome organization

The recent completion of a working draft of the human genome by a public sequencing consortium (Human Genome Project) (Lander *et al.*, 2001) and a private effort led by Celera (Venter *et al.*, 2001) shed a new light on whole genome analysis and set a new pace for the field of genomics. The genomic era started over 20 years ago with the pioneering work in Sanger's laboratory (Anderson *et al.*, 1981; Fiers *et al.*, 1978; Sanger *et al.*, 1982). Through this work, that now has resulted in a massive growth in genomic sequencing data, the ability of having a global view of genomes can greatly enhance biomedical research by giving the opportunity to consider genetic diseases in a global and comprehensive fashion. Sequencing the human genome was made possible by the emergence of a new strategy for large-scale sequencing (random shotgun sequencing strategy; see Anderson *et al.*, 1981; Bodenteich *et al.*, 1993; Chissoe *et al.*, 1995; Putney *et al.*, 1983) and a program launched by Botstein *et al.* in 1980 to create a physical map of the human genome by systematically screening PAC and BAC clones from a human library for STS markers, fingerprinting the positive clones, integrating

them into the map and selecting the largest clones with minimal overlap for sequencing (McPherson *et al.*, 2001). Although some characteristics of a vertebrate genome were described by numerous previous studies, none of them were fully integrated in a comprehensive fashion, allowing new questions to emerge. The Human Genome Project has allowed the research community free access to the working draft of the human genome, thereby providing a fairly complete view of the genomic sequence. In addition, work in progress on other vertebrates, such as mouse, provides an additional opportunity to study comparative genomics. Conserved DNA segments between human and the mouse model organism can be used to identify likely orthologous genes and to speculate on their mode of transcriptional regulation. Evolutionary studies are greatly enhanced by the study of conserved segments among genomes from diverse species. Also, determining the sequence similarity among genes of known functions in model organisms accelerates the study of human gene function, although the actual function of these genes in human needs to be firmly established by additional studies.

One of the main characteristics of a vertebrate genome is the presence of DNA repeat sequences. The recognition that genomes can contain a large amount of DNA repeat elements originates from what is known as the C-value paradox (Gall, 1981). One might expect the morphological complexity of an organism to be roughly correlated to the amount of DNA in its haploid genome (C-value). However, many simple organisms have an unexpectedly large C-value. For instance, *Amoeba dubia* has a genome size that is 200 times larger than that of human (Gregory and Hebert, 1999). Such large C-values actually are related to the presence of an enormous amount of DNA repeats. In human, DNA repeats, although in lower amount than in *Amoeba dubia*, still account for at least 50% of the genome content. Most of the DNA repeats in human, as well as other vertebrates, are transposon-derived repeats. In mammals, transposon-derived repeats fall into four classes: 1) Long Interspersed Elements (LINEs), 2) Short Interspersed Elements (SINEs), 3) LTR retrotransposons and 4) DNA transposons. LINEs, SINEs

and LTR retrotransposons transpose through RNA intermediates, whereas DNA transposons transpose directly as DNA. The two predominant forms of transposon-derived repeats in the human genome, accounting for 60% of all transposon-derived repeats, are the L1 family and the Alu family. The L1 repeats (a LINE family member) are about 6 kb long, harbour an internal polymerase II promoter and encode two open reading frames (ORFs) (Adams *et al.*, 1980). The Alu repeats (a SINE family member) are about 300 bp long, harbour an internal polymerase III promoter, encode no proteins and are composed of two head-to-tail repeats of a sequence sharing significant homology with 5' and 3' portions of 7SL RNA (Quentin, 1992). The two predominant forms of SINEs in mouse are the B1 and B2 families. B1 repeats, considered as the rodent equivalent of the human Alu repeats, are composed of about 140 bp of a consensus sequence followed by a short A-rich region (Maraia, 1991). B2 repeats are specific to the rodent lineage, are about 180 bp long and share some sequence homology with serine tRNA genes (Daniels and Deininger, 1985). Considering their enormous number in most higher eukaryotic genomes and the absence of experimental evidence for any possible functions, DNA repeats are often seen as "junk" DNA, molecular parasites that have disseminated themselves, over many generations, throughout the entire genome. Because they are present in both human and mouse, as well as many other vertebrate genomes, the appearance of LINEs and SINEs predated human and mouse speciations, indicating the extremely slow rate with which these DNA sequences are removed from veretebrate genomes. In terms of natural selection, the rate of elimination of repeat DNA in slowly growing eukaryotes would be effectively counterbalanced by its rate of propagation, allowing its maintenance over the years in vertebrate genomes. Data from the Human Genome Project (Lander *et al.*, 2001) indicate that the overall activity of all transposon-derived repeats in human has dramatically decreased over the past 35-50 million years, with the possible exception of L1 repeats. However, such decline has not been observed in the mouse genome,

indicating a higher rate of recombination in rodent lineages. Interestingly, although LINEs occur mostly in gene-poor AT-rich regions of the human genome, SINEs show an opposite trend, occuring mostly in gene-rich GC-rich regions. The same pattern is observed for Alu-like B1 and tRNA-derived SINEs (such as B2) in mouse. In addition, young Alu repeats actually have a preference for AT-rich regions, whereas older Alu repeats are progressively found in GC-rich regions. Alu repeats might be selectively eliminated from AT-rich regions over several generations, thus concentrating the older Alu repeats in GC-rich regions. Also, the observation that positive selection favors Alu repeats in GC-rich regions might explain their bias towards GC-rich regions. Such a mechanism has been proposed by Schmid (1998). In many species, SINEs are transcribed under condition of stress and the resulting RNAs bind protein kinase R, blocking its ability to inhibit translation (Chu *et al.*, 1998; Li *et al.*, 1999; Liu *et al.*, 1995). Therefore, SINEs would be positively selected in gene-rich GC-rich regions because of their ability to activate protein translation under stress conditions.

Besides DNA repeats, many genes are present in multiple copies in vertebrate genomes. Multiple copies of the gene for ribosomal and transfer RNAs are present in these genomes, because of the large amounts of these RNAs that are required by the organism. Several proteins are encoded by genes organized in clusters. For example, histone proteins are encoded by a series of identical, repeated genes (Maxson *et al.*, 1983). Histones must be synthesized in large amount during the S phase of the cell cycle and this is made possible by the multiple reiteration of histone genes. Mammalian globin genes also occur in clusters. The genes specifying globin alpha and beta-like subunits are arranged in two developmentally ordered clusters on different chromosomes (Karlsson and Nienhuis, 1985). Here, the genes in each globin cluster are arranged 5' to 3' on the coding strands, in the order of their developmental expression and this order is conserved in humans, mice and many other vertebrates.

One question that remained unanswered until the completion of the working draft of the human genome concerns the total number of genes present in the human genome. Surprisingly, the number of genes present in the human genome appears to be relatively low, according to the level of complexity of the human organism. Estimation of the number of genes is based on several parameters, including cross-species sequence comparison based on sequence similarity to previously identified genes in other species, EST and mRNA alignments with genomic sequences and recognition of groups of exons using various computer programs (see Materials and Methods). So far, the most accurate analysis estimates the number of protein-coding genes in human to be about 30000–40000. This is about only twice as many as much simpler organisms such as the worm or fly. However, genes in human are more complex, with many alternatively spliced mRNAs generated from one gene, yielding a larger number of proteins. Certain exons in one type of cells may actually be introns in other types. About 38% of human genes have their expression modulated by alternative splicing sites, as determined by EST alignments (Brett *et al.*, 2000). This figure is much greater than the worm *C. elegans*, in which alternative splicing has been predicted for 22% of genes for which ESTs have been identified. Preliminary studies on human chromosomes 19 and 22 show that about 70% of alternative splice forms affect the coding sequence, rather than changing the 5' or 3' UTR (Lander *et al.*, 2001).

The ultimate goal in sequencing the human genome is to compile a complete list of all human genes and their encoded proteins (the "proteome"). To facilititate gene identification, a comparison with a set of mouse cDNAs sequenced by the Genome Exploration Group of the RIKEN Genomic Sciences Center shows that about 81% of the 15294 cDNAs selected from a collection of nearly one million mouse ESTs from different tissues and developmental stages exhibit sequence similarity with portions of the human genome and 69% exhibit sequence similarity to the human genes predicted so far (Lander *et al.*, 2001). In addition, many segments of the human genome are

conserved in the mouse genome. A map of linkages and synteny between mouse and human has been described by Nadeau (1989). This map indicates that in many cases, not only the chromosomal assignment, but also the gene order, are conserved in syntenic segments. One hundred and eighty three segments of the human genome containing at least two genes whose order is conserved in the mouse genome without interruption have been identified (Lander *et al.*, 2001). The largest conserved segment is on human chromosome 4, where about 90.5 megabases of DNA are conserved with mouse chromosome 5. The likely presence of orthologous genes (genes conserved between species) in these conserved segments will allow an in-depth analysis of the overall regulation of these genes in human as well as a better study of the proteins they encode.

Finally, sequence similarity with proteins of known function in model organisms offers further opportunities to study the function of the proteins encoded by the human genes. The complete genomic sequences of the yeast, *S. cerevisiae*, fly, *Drosophila melanogaster*, and worm, *C. elegans* allow comparative studies of their proteomes. It appears that only 7% of the families of proteins reported by Lander *et al.* (2001) (each family reflects a specific physiological function) are specific for vertebrates. Not surprisingly, these families of proteins are mostly involved in specific vertebrate functions such as the nervous system or the immune response, indicating the recent emergence of these functions in vertebrates. Also, the number of protein domains is much higher in human (Lander *et al.*, 2001), indicating a higher rate of rearrangements between pre-existing sub-domains during vertebrate evolution, through various mechanisms including exon shuffling. This higher rate of rearrangement mostly involves the recent evolution of extracellular and transmembrane domains. Finally, 60% of the protein families are more numerous in human than in the other species mentioned above, indicating that gene duplication has had a major impact during vertebrate evolution. Many of the families that have expanded in human are mostly involved in

specific vertebrate physiological functions, such as the immune response or the olfactory response (Fuchs *et al.*, 2001).

I-2 Interferon-inducible proteins

I-2-1 Interferons and their mode of action

Interferons (IFNs) were discovered in 1957 as agents interfering with virus replication (Isaacs and Lindenmann, 1957). They are soluble mediators involved in cell-to-cell communication (cytokine) that regulate vertebrate resistance against viral and parasitic infections through their antimicrobial, immunomodulatory and cell growth regulatory functions. They also are components of the host defense against certain tumors (for review, see Lengyel, 1982; Petska *et al.*, 1987; Samuel, 1988; Samuel, 1991). IFNs are classified as IFN alpha, beta and gamma. In human, there are at least 18 IFN alpha genes, four of which are pseudogenes. They constitute, together with the single IFN beta gene a cluster of genes on chromosome 9. One interesting characteristic of IFN alpha and beta genes is that they all lack introns. The reason for this lack of introns remains obscure but may be related to the postulate that all the IFNs alpha and beta genes are thought to have evolved from a single ancestral gene (Weissmann and Weber, 1986). However, in contrast to the IFNs alpha and beta, the single IFN gamma gene, located on chromosome 12, has three introns. Interestingly, most cell types are capable of producing IFNs alpha and beta whereas only lymphocytes T and Natural Killer (NK) cells produce IFN gamma. The nature of the IFN produced depends on both the producing cells and the inducer. Biological inducers enhancing IFN biosynthesis include viruses, bacteria, protozoa, and exposure to other cytokines and growth factors. As an example, IFN gamma synthesis by NK cells is stimulated by Interleukin-2 as well as double-stranded RNA (dsRNA) which is thought to be a side

product or an intermediate produced during virus replication (Ciccarone *et al.*, 1990). The control of IFN production is primarily at the level of IFN transcription and the induction is transient, even in the continued presence of inducer. IFN gene 5'-flanking regulatory regions contain several regulatory elements that are bound by specific proteins such as the transcriptional activator IRF-1 or the transcriptional repressor IRF-2. Interestingly, IRF-1 synthesis is induced by both IFN alpha and gamma in many cell lines (Harada *et al.*, 1998).

The treatment of cells with IFNs alters the rate of synthesis of many cellular mRNAs and of their corresponding proteins (see Fig. 1.3).

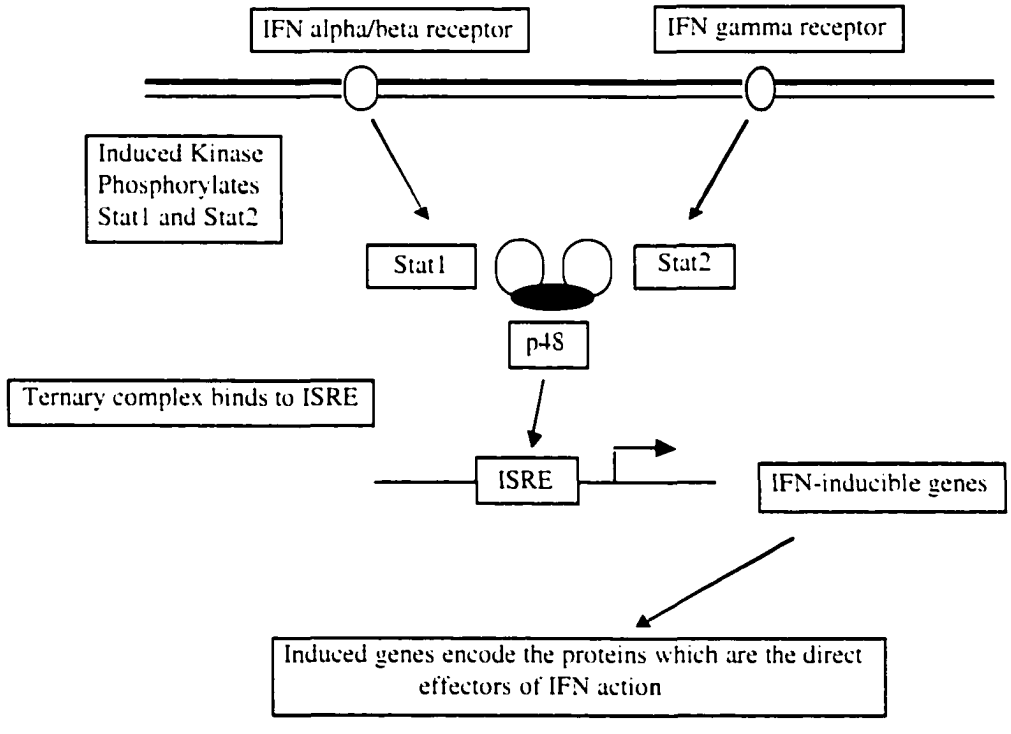


Fig. 1.3: Mechanism of gene activation by interferons.

The first step of gene activation by IFNs is their binding to specific cell surface receptors. IFNs alpha and beta compete for binding to the same receptor whereas IFN gamma binds to a specific receptor. One or more member of the JAK family of tyrosine kinases associated with the cytoplasmic domain of the IFN receptor is then activated

after IFN-receptor interaction, leading to the tyrosine phosphorylation of one or more of a family of latent cytoplasmic transcription factors called STATs. Then, after phosphorylation, heterooligomerization with the DNA-binding protein p48, and nuclear translocation, these proteins then act as transcriptional activators by binding to a consensus cis-acting DNA sequence called the IFN-Stimulated Response Element (ISRE), which confers inducibility by IFN and is present in the 5'-flanking region of all known IFN-inducible genes (Benech *et al.*, 1987).

The mRNAs encoded by the IFN-activatable genes are translated into the IFN-inducible proteins. One IFN-induced protein is the dsRNA-dependent protein kinase (Meurs *et al.*, 1990). Binding of dsRNA elicits autophosphorylation and activation of this latent enzyme which then phosphorylates a serine residue to inactivate the alpha subunit of the translation initiation factor eIF2. Another enzyme activated by dsRNA, called (2'-5')$(A)_n$ synthetase, catalyzes the conversion of ATP into (2'-5')$(A)_n$ and PPi (Hovanessian, 1991). Binding of (2'-5')$(A)_n$ to RNAse L activates this enzyme which primarily cleaves ssRNA after UU, UA, UC and UG residues. All the above three enzymes are mediators of the antiviral action of IFNs. Other IFN-inducible proteins are thought to mediate, through IFN-induced alterations in the level of expression of their corresponding genes, the immunomodulatory and cell growth regulatory functions of IFNs (Staeheli, 1990). As an example, T cells-derived IFN gamma induces the enhanced expression of MHC class II antigens on accessory cells, such as monocytes and macrophages, which either do not express class II antigens constitutively or at a very low level (Liu and Janeway, 1990).

The mode of action of the IFNs is best known in the case of picorna and influenza viruses. Various data demonstrate that the IFN-induced (2'-5')$(A)_n$ synthetase/RNAse L pathway can inhibit the replication of picorna virus (Chebath *et al.*, 1987; Rice *et al.*, 1985). The IFN-inducible murine Mx1 protein, located in the nucleus,

blocks the replication of influenza virus in cultured cells as well as in mice (Staheli, 1990), possibly by impairing viral transcription.

I-2-2 The "200 family" of proteins

The binding of interferons to cell surface receptors activates the transcription of many genes. The proteins encoded by these genes are the direct effectors of interferon action. One related family of IFN-induced proteins is the "200 family" of proteins. This family is encoded by several structurally related genes located on murine chromosome 1 (Ifi200 cluster) and by three homologous genes (MNDA, IFI 16 and AIM2) located on human chromosome 1, within a large linkage group conserved between mouse and human.

I-2-2-1 The murine Ifi200 family

The interferon-activatable Ifi200 gene family in mouse consists of at least five members designated 202a, 202b, 203, 204 and D3. Southern-blot analysis of cosmids containing the 202, 203 and 204 genes have further revealed 1) the cross-hybridization of the 202 and 204 genes with a third gene designated as the 201 gene, 2) the hybridization of a partial 203 cDNA to three genes of the 203 gene family and 3) the close linkage of the 203 family gene to the 201, 202 and 204 genes, indicating the existence of a cluster of at least eight closely linked, IFN-activatable genes on murine chromosome 1 (Kingsmore *et al.*, 1989; Opdenakker *et al.*, 1989).

The Ifi200 gene family encodes the "200 family" of proteins: p202a, p202b, p203, p204 and D3. Comparison of their predicted amino acid sequences has revealed a peculiar structural organization with one or two partially conserved 200 amino acid long segments (*a* type and *b* type), supporting the hypothesis that these genes were

generated, at least partially, by repeated gene duplication of an ancestral gene encoding one copy of the conserved 200 amino acid segment. The p202a, p202b and p204 proteins each have one *a* and one *b* type segments, whereas p203 and D3 have only one such segment of either the *a* type or the *b* type. Comparison of 959 nucleotides from the 5'-flanking regions of the 201, 202a and 204 genes also reveals strong sequence similarity (96% sequence similarity in this region between the 202a and 204 genes), with two short IFN-responsive elements (GA box and Friedman-Stark sequence) identified in these segments (Choubey *et al.*, 1989).

The best characterized 200 family proteins are p202a and p204. Their functions are summarized in Fig. 1.4. The functions and interactions of the 200 family of proteins are summarized in Table 1.1.

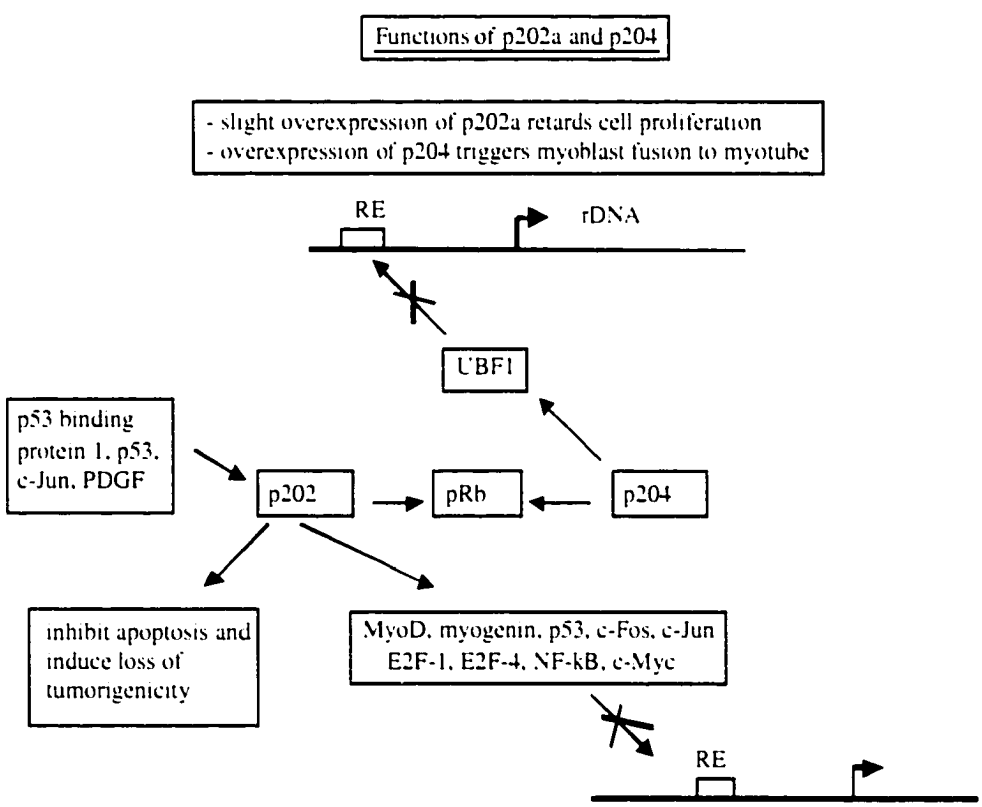


Fig. 1.4 summarized functions of the proteins encoded by Ifi202a (p202a) and Ifi204 (p204). See text for details.

Proteins	Genes	Functions	Interactions
p202a	Ifi202a	-retards cell proliferation	MyoD, myogenin,

		-inhibit apoptosis -induce loss of tumorigenicity	p53, c-Fos, c-Jun, E2F-1, E2F-4,NF-kB c-Myc, p53 binding protein 1, pRb
p203	Ifi203	-unknown	unknown
p204	Ifi204	-retards cell proliferation -inhibit rDNA transcription -triggers myoblast fusion	pRb, UBF1
D3	Ifi205	-myeloid differentiation?	unknown

Table 1.1: Summarized functions and interactions of the 200 family of proteins. See text for details.

I-2-2-1-1 The Ifi202a gene

The 202a cDNA, isolated from a cDNA library prepared from mRNA of IFN beta-treated Ehrlich ascite tumor cells (Samanta *et al.*, 1984), encodes p202a, a 52 kD phosphoprotein that contains several potential sites for phosphorylation by several protein kinases (e.g., cyclin-dependent kinase 2, casein kinase II, protein kinase C, cAMP-dependent protein kinase) (Choubey and Lengyel, 1993). Its basal level of expression in growing mouse cells is relatively low and increases up to 16-fold after a 48 hour IFN exposure. Interestingly, the IFN-inducibility of p202 is poor in confluent mouse cells, indicating that the extent of IFN-inducibility of p202 is affected by the growth conditions. p202 is primarily located in the cytoplasm. However, after a prolonged exposure of the cell to IFN (36 hours exposure), most of the p202 is located in the nucleus (Choubey and Lengyel, 1993). p202 binds non specifically to double-stranded DNA and to single-stranded DNA *in vitro*, raising the possibility that it could bind DNA *in vivo* (Choubey and Gutterman, 1996).

Recent studies have demonstrated that the overexpression of p202 in yeast and murine transfected cells retards proliferation. IFNs can inhibit the growth of various cultured cells by prolonging phases of the cell cycle or by arresting the growth at the G1

phase (Lembo *et al.*, 1995; Min *et al.*, 1996). Interestingly, co-immunoprecipitation experiments of cell extracts prepared from IFN-treated murine AKR-2B cells transformed with an expression construct encoding pRb have shown that p202 can bind the hypophosphorylated active form of the retinoblastoma protein pRb (Choubey and Lengyel, 1995). The 105 kD pRb is a negative growth regulator that retains the cells in the G1 phase of the cell cycle. The effect of p202 binding on pRb function is unknown. However, it is possible that the binding of p202 may impair the phosphorylation of pRb (IFN treatment impairs this process) or the phosphorylation-dependent release of pRb from its association with the nucleus (Choubey and Lengyel, 1995). At least two segments of pRb bind p202, the second segment is the region to which different DNA tumor virus oncoproteins bind. Interestingly, these oncoproteins preferentially bind the hypophosphorylated form of pRb (Dyson *et al.*, 1989; Horowitz *et al.*, 1989).

One of the ways in which pRb regulates cell proliferation is by binding to the E2F-1/DP-1 transcription factor (TF), as shown by co-immunoprecipitation experiments on IFN-treated murine AKR-2B cells, and inhibiting its activity (Choubey *et al.*, 1996). E2F-1/DP-1 promotes the transcription of various growth-promoting genes whose expression is required for the G1-S transition (DeGregori *et al.*, 1995). The hypophosphorylated form of pRb present in the G1 phase of the cell cycle binds E2F-1/DP-1 and thereby inhibits its activity (Chellappan *et al.*, 1991). At the end of the G1 phase, pRb becomes phosphorylated and loses its ability to bind E2F-1/DP-1. Interestingly, p202 can bind E2F-1/DP-1 and inhibit its activity in a transient transfection assay. Since p202 and pRb bind different regions of E2F-1/DP-1, complexes between all three proteins may occur (Choubey *et al.*, 1996).

Besides E2F-1, p202a can inhibit the activity of numerous transcription factors, including c-Fos (Min *et al.*, 1996), c-Jun (Min *et al.*, 1996), NF-kB (Min *et al.*, 1996), c-Myc (Wang *et al.*, 2000), MyoD (Datta *et al.*, 1998), myogenin (Datta *et al.*, 1998) and p53 (Datta *et al.*, 1996). The protein p53 is a major tumor supressor protein that

controls cell growth in response to distinct events. While pRb responds to the presence or absence of mitogenic signals, p53 responds mostly to DNA damage. The p53 protein acts as a checkpoint to regulate cell cycle arrest in the G1, G2/M and G0 phase of the cell cycle. p53 activates, through its specific DNA-binding activity, the transcription of genes with p53-binding sites (Farmer *et al.*, 1992). Among the genes activated by p53 is the mdm2 protooncogene whose gene product, Mdm2, binds to and down-modulates p53 function (Fakharzadeh *et al.*, 1991). The p21 gene is a second p53-activated gene whose gene product, p21, binds to and inhibits cyclin/cyclin-dependent kinase (cdk) complexes, inducing cell cycle arrest in the G1 phase (Harper *et al.*, 1993). Luciferase reporter assays have shown that the overexpression of p202 in murine transfected cells inhibits the p53-dependent expression of reporter genes driven by elements from the mdm2 or the p21 promoters (Datta *et al.*, 1996). Gutterman and Choubey (1999) demonstrated that the induced expression of p202a in murine AKR-2B fibroblasts was accompanied by an increase in p21 mRNA and protein levels. Transient transfection of a p202a-encoding plasmid in Saos-2 cells, which lack a wild type p53 protein, also results in an increase in p21 protein levels, suggesting the possibility that p202a modulates p21 protein levels at a post-transcriptional level (Gutterman and Choubey, 1999).

NF-kB controls the expression of various genes of the immune and inflammatory systems, as well as of various viral genes (Siebenlist *et al.*, 1994). The NF-kB protein is present in most eukaryotic cells and serves as an inducible transcription factor under the effect of various factors, including cytokines (interleukin 1 and tumor necrosis factor alpha), bacterial lipopolysaccharides and double-stranded RNA (Baeuerle and Henkel, 1994). The AP-1 transcription factors (c-Fos and c-Jun) also are induced in response to various inducers, including growth factors and UV irradiation. Their function is modulated by phosphorylation/dephosphorylation at several sites (Angel and Karin, 1991). MyoD and myogenin are members of a family of

transcription factors known as muscle regulatory factors (Buckingham, 1994). The muscle regulatory factors form heterodimers serving as skeletal muscle-specific transcription factors that bind DNA sites in transcriptional enhancers of muscle differentiation genes (Buskin and Hauschka, 1989). MyoD usually is expressed in proliferating myoblasts whereas myogenin is not expressed until the myoblasts exit the cell cycle (Olson and Klein, 1994). The c-Myc transcription factor is a member of the Myc oncoprotein family whose functions have been correlated with proliferation, apoptosis and tumorigenesis in many cell types (Henrikson and Luscher, 1996). The expression of c-Myc is induced by various factors, including mitogens, cytokines and growth factors. Ectopic overexpression of c-Myc can inhibit or delay differentiation (Henrikson and Luscher, 1996) and even induce apoptosis in conjunction with p53 (Packham and Cleveland, 1995). There is evidence that c-Myc forms a transcription factor by forming a heterodimer with the protein Max (Blackwood and Eisenman, 1991) and both subunits are required for the heterodimer binding to DNA (Ferre-D'amare *et al.*, 1993). Both c-Myc and Max are required to modulate transcription, apoptosis and cell proliferation (Amati *et al.*, 1993). In the case of c-Fos, c-Jun, NF-kB, MyoD and myogenin, p202a binds to the transcription factor *in vitro* and *in vivo*, inhibiting its sequence-specific binding to DNA. Besides modulating its activity, p202a also inhibits the expression of MyoD in murine transfected cells. In addition, the overexpression of p202a in murine transfected cells increases during the fusion of myoblasts to myotubes (Datta *et al.*, 1998). p202a binds c-Myc *in vitro* and *in vivo* and therefore inhibits its binding to the Max protein, as well as the transcriptional activity of c-Myc (Wang *et al.*, 2000).

The activity of p202a in turn is inhibited *in vivo* by the binding of a segment from the p53 binding protein 1 (Datta *et al.*, 1996). In addition, D'Souza *et al.* (2001) demonstrated that p53 binds to a p53 DNA-binding consensus sequence present in the 5'-regulatory region of the 202a gene to repress the activity of a reporter gene whose

transcription is driven by the 5'-regulatory region of the 202a gene. Therefore, p202a which regulates p53 functions and inhibits its transcriptional activity, is in turn transcriptionally regulated by p53. Geng *et al.* (2000) recently showed that transfection with a JunD-encoding plasmid along with a reporter plasmid in which transcription of a reporter gene was driven by the 5'-regulatory region of the 202a gene, into murine AKR-2B fibroblasts, resulted in an increased activity of the reporter gene, suggesting that one of the AP-1 like DNA binding sites present in the 5'-regulatory region of 202a might be selectively bound by the Jun transcription factor. These data suggest that the Jun transcription factor might positively regulate the transcriptional activity of the 202a gene. Flati et al. (2001) showed that the Ifi202 gene was transcriptionally activated by the mitogenic platelet-derived growth factor (PDGF), independantly from IFN signaling, suggesting that Ifi202 can be directly activated by mitogenic stimuli.

Finally, the p202a protein also can inhibit apoptosis (Koul *et al.*, 1998) and is associated with an antiproliferative effect on human prostate cancer cells (Yan *et al.*, 1999). Interestingly, human prostate tumorigenic cells expressing p202a show reduced ability to grow in soft agar, indicating a loss of transformation phenotype (Yan *et al.*, 1999). Also, Wen *et al.* (2000) revealed that p202a was able to induce a loss of tumorigenicity of breast cancer cells *in vivo* and that p202a expression actually sensitized breast cancer cells to apoptosis induced by tumor necrosis factor alpha treatment. One suggested mechanism is that p202a, because of its interaction with NF-kB, might actually contribute to its inactivation and the resulting sensitization of breast cancer cells to apoptosis in the presence of tumor necrosis factor alpha. Wen *et al.* (2001) demonstrated recently that p202 expression was associated with multiple antitumor activities in human pancreatic cancer xenograft models, including inhibition of tumor growth, reduced tumorigenicity, prolonged survival and suppression of metastasis and angiogenesis, providing a scientific basis for a p202-based gene therapy in pancreatic cancer treatment. Also, Ifi202 was positioned as a candidate gene for lupus

by Rozzo *et al.* (2001), using microarray studies in mice congenic for the locus contributing to systemic lupus susceptibility.

I-2-2-1-2 The Ifi203 gene

Another member of the Ifi200 family, 203, is expressed in mouse strain C57BL/6-derived cell lines stimulated with IFNs. Several techniques, including reverse transcriptase PCR and rapid amplification of cDNA ends (RACE), have been employed to clone the 203 cDNA (Gribaudo *et al.*, 1997). The 203 gene previously was identified as the upstream neighbor of the 202a gene (Engel *et al.*, 1988), about 5 kb upstream of the 5'-flanking region of the 202a gene. The constitutive expression of the 203 gene is restricted to some myeloid and lymphoid tissues (thymus, bone marrow and spleen) (Gribaudo *et al.*, 1999). In addition, the 203 and 202a genes exhibit a reciprocal pattern of expression in various mouse strains, indicating that p203 and p202a might have a similar function in a mouse strain in which only one of them is expressed (Gribaudo *et al.*, 1999).

I-2-2-1-3 The Ifi204 gene

The 204 gene encodes a 72 kD protein whose levels in cells increases up to 75-fold after IFN alpha/beta exposure. Studies involving fractionation of cell lysates and immunofluorescence analysis of control and IFN-treated AKR-2B fibroblasts indicates that p204 is distributed between the nucleolus and the nucleoplasm (Choubey and Lengyel, 1992). Like p202a, p204 contains several potential sites for phosphorylation by several protein kinases (casein kinase II, protein kinase C, cAMP-dependent protein kinase). Analysis of cell extracts of ^{32}P-labeled IFN-treated AKR-2B cells shows that p204 is phosphorylated. Its overexpression in murine transfected cells inhibits cell

proliferation (Lembo *et al.*, 1998). p204 delays cell progression from G1 into the S phase and cells accumulate with a DNA content equivalent to cells arrested in late G1 phase. Hertel *et al.* (2000) demonstrated that p204 can directly bind the pRb tumor suppressor protein *in vivo*, suggesting that p204, like p202a, contributes to the IFN antiproliferative activities by targetting the pRb regulatory system. Further studies showed that p204 can inhibit the transcription of ribosomal RNA *in vivo* and *in vitro* (Liu *et al.*, 1999). This inhibition is the consequence of the binding of p204 to the ribosomal DNA-specific UBF transcription factor. UBF (Upstream Binding Factor) transcription factor promotes the binding of the murine protein complex TF-1B (consisting of the TATA Binding Protein and three associated factors) to the rDNA promoter, and the resulting complex promotes the binding of RNA polymerase I (Beckmann *et al.*, 1995; Bell *et al.*, 1988). The binding of p204 to UBF results in the inhibition of the binding of UBF to its specific binding site in the rDNA 5'-regulatory region. In addition, like p202a, the level of p204 increases during the fusion of myoblasts to myotubes. Also, p204 is predominantly cytoplasmic during myoblast fusion and its translocation from the nucleoplasm to the cytoplasm is linked to its phosphorylation (Liu *et al.*, 2000). The increased level of p204 during myoblast fusion is related to the binding of MyoD transcription factor to several MyoD-specific sequences in the 204 gene 5'-regulatory region, and the subsequent MyoD-dependent activation of transcription. Interestingly, the increase of p204 level during myoblast fusion does not depend on interferons, suggesting that p204 is involved in muscle differentiation independently from its interferon action. It has been demonstrated that the expression of the 204 gene is restricted to cells of the myelomonocytic lineage, suggesting its potential involvement in the differentiation and maturation of this cell lineage (Gariglio *et al.*, 1998). Finally, p204 is required by mouse cytomegalovirus (MMCV) for its replication, as MMCV replication is strongly inhibited in cells lacking a functional p204 (Hertel *et al.*, 1999).

I-2-2-1-4 The D3 gene

The D3 cDNA clone initially was isolated from a murine macrophage cDNA library. The D3 gene is inducible in macrophages by lipopolysaccharides and IFNs and is not expressed in IFN or plateled-derived growth factor-stimulated fibroblasts (Tannenbaum *et al.*, 1993). Its expression is restricted to myeloid cell lineages at or beyond the promyelocyte stage of differentiation, indicating a potential role for the D3 protein in myeloid differentiation (Weiler *et al.*, 1999).

I-2-2-2 The human "200 family" of proteins

Much less is known about the human counterpart of the mouse 200 family of proteins. To date, only three members of the human 200 family of proteins have been characterized (IFI-16, MNDA and AIM2), and their respective roles remain to be clearly established. Their functions and interactions are summarized in Table 1.2 and Fig. 1.5.

Proteins	Genes	Functions	Interactions
MNDA	MNDA	-unknown	nucleolin, YY1 transcription factor, nucleophosmin
IFI-16	IFI-16	-transcriptional repressor	p53
AIM2	AIM2	-retards cell proliferation -increases susceptibility to apoptosis -tumorigenic suppression?	unknown

Table 1.2: Summarized functions and interactions of the human 200 family of proteins. See text for details.

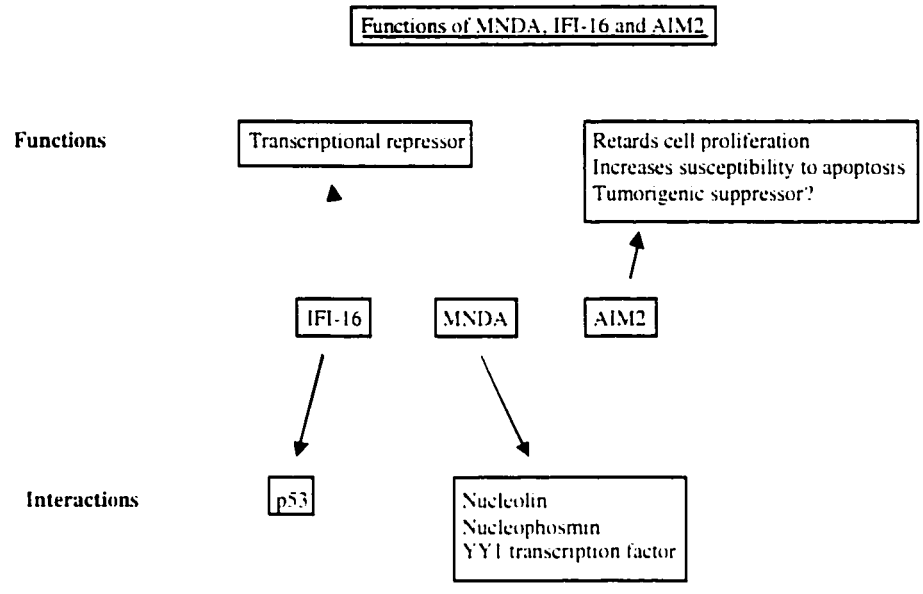


Fig. 1.5 summarized functions of the proteins encoded by the human 200 family. See text for details.

I-2-2-2-1 The MNDA gene

During immunoblot screening of nuclear antigens specifically expressed in myeloid cells, a cDNA was discovered and the corresponding protein was named Myeloid Nuclear Differentiation Antigen (MNDA) (Burrus *et al.*, 1992). This cDNA encodes a 403 amino acid protein containing a single copy of the 200 amino acid conserved segment common to all the proteins of the "200 family" and showing extensive regions of sequence similarity with the protein products of the interferon inducible 204 gene (Briggs *et al.*, 1992). The MNDA gene is specifically expressed in granulocytes, monocytes and earlier stage cells of these lineages, but is upregulated by IFN alpha in monocytes only. Inducibility of MNDA expression by IFN alpha also has been reported in the human myeloid leukemia cell lines HL-60, U937 and THP-1 (Briggs *et al.*, 1994b). The expression of MNDA mRNA is modulated by the myelomonocytic differentiation process, since induction of monocytic differentiation

with phorbol esters is accompanied by a decrease in mRNA level of MNDA and this effect is not due to the inhibition of cell proliferation (Briggs *et al.*, 1994c). In contrast, induction of granulocyte differentiation by retinoic acid does not alter the mRNA level of MNDA.

Characterization of the MNDA gene shows that it is a single copy gene located on human chromosome 1, within a large linkage group conserved between human and mouse containing the mouse Ifi200 family (Briggs *et al.*, 1994a). The 5'-flanking region of the gene contains two Interferon-Stimulated Response Elements (ISRE) flanking a multiple transcription start site region identified as a TATA-less promoter with other consensus elements for binding of cellular transcription factors Myb, Ets, PU-1 and Sp1 (Kao *et al.*, 1996). Analysis of deletion mutants showed that Sp1, but not Myb or PU-1, is one of the factors contributing to the restricted pattern of expression of the myelomonocytic lineage specific MNDA gene (Kao *et al.*, 1997). The MNDA protein was shown to bind the ubiquitous 100 kDa nucleolin protein (Xie *et al.*, 1995). Nucleolin is a protein that binds other macromolecules and exhibits features consistent with a role in signal transduction, production of ribosomes, nuclear matrix structure and regulation of transcription. This result indicates that the function of the protein MNDA most likely is related to interactions with other proteins contributing to cell/lineage-specific processes by ubiquitous proteins such as nucleolin. Xie *et al.* (1997) further demonstrated that the protein MNDA binds the nucleolar protein nucleophosmin, suggesting that additional interactions define overall MNDA functions. Both nucleolin and nucleophosmin binds the ubiquitous zinc finger transcription factor YY1 and MNDA binds YY1 directly under *in vivo* and *in vitro* conditions (Xie *et al.*, 1998). The MNDA-YY1 interaction enhances the affinity of YY1 for DNA and decreases its rate of dissociation. In addition, MNDA forms a ternary complex with YY1 and the YY1 target DNA site, suggesting a role for MNDA as a cofactor to impair lineage-specific features to YY1 functions. The ability of MNDA to act as a co-regulator altering gene expression

through its binding to specific transcription factors or other proteins is similar to the ability of mouse protein p202a and p204 to bind specific transcription factors and alter their DNA-binding activities, although tissue/cell-specific expression of the mouse Ifi200 cluster genes has only been reported for the D3 gene.

I-2-2-2-2 The IFI-16 gene

The IFI 16 cDNA was isolated by screening a human CTL-derived cDNA library with a rabbit antiserum raised against the cytolytic granules obtained from the same cell source (Trapani *et al.*, 1992). The IFI 16 mRNA encodes a 729 amino acid protein containing the two 200 aminoacid conserved segments (*a* type and *b* type) common to all members of the "200 family" of proteins, and shows significant sequence similarities to the mouse genes 202a and 204. The IFI 16 gene expression is strongly induced in myeloid cells by IFN gamma exposure (Trapani *et al.*, 1992) and is associated with the differentiation of myeloid precursor cell lines (Dawson and Trapani, 1995). IFI 16 is present throughout monocytic development but its expression rapidly declines in erythroid and polymorphonuclear precursor cells, suggesting a possible role in myelopoiesis (Dawson *et al.*, 1998). This pattern of IFI 16 expression throughout lymphoid development and within monocyte precursors and peripheral blood monocytes contrasts with MNDA, whose expression is limited to myeloid cells from the promyelocytic stage onwards (Dawson *et al.*, 1995).

The gene encoding IFI 16 has been mapped to human chromosome 1 by DNA hybridization to a panel of human-mouse hybrid cell lines, within the same linkage group conserved between human and mouse and containing the mouse Ifi200 gene cluster and the MNDA gene (Trapani *et al.*, 1992). Its exon-intron organization has been determined (Trapani *et al.*, 1994). Exon 2 of the IFI 16 gene encodes a lysine-rich amino terminal region possessing DNA-binding activities. The 5'-untranslated region

contains an IFN alpha/beta-stimulated response consensus element (ISRE), as well as an IFN gamma-activation site consensus element. Immunofluorescence studies showed that the IFI 16 protein is located in the nuclei of monocytes, neutrophils and lymphocytes in normal peripheral blood but not in granulocytes (Dawson and Trapani, 1995). Alternative mRNA splicing accounts for the existence of three isoforms of the IFI 16 protein (Johnstone *et al.*, 1998). IFI 16 fused to the GAL4 DNA-binding domain can act as a repressor of transcription when positioned in proximity to a promoter containing consensus GAL4 DNA elements (Johnstone *et al.*, 1998). The transcriptional repression domains have been mapped to the 200 amino acid segments conserved in all human and mouse "200 family" protein members. Functional interactions between p53 and IFI 16 have been reported (Johnstone *et al.*, 2000). Whether this interaction of IFI 16 with p53 and the role of IFI 16 as a transcriptional repressor are involved in regulating myeloid differentiation remains to be established.

I-2-2-2-3 The AIM2 gene

Introduction of a normal copy of human chromosome 6 suppresses the tumorigenicity of melanoma cell lines (Trent *et al.*, 1990). The AIM2 cDNA was isolated by substractive cDNA selection in a chromosome-suppressed melanoma cell line (DeYoung *et al.*, 1997). It encodes a 344 amino acid protein containing part of the 200 amino acid *a* type segment common to all members of the "200 family" of proteins. The AIM2 consensus sequence shares strong sequence similarity with IFI 16, MNDA and 204 genes. The AIM2 gene, located on human chromosome 1, is within a linkage group conserved between the syntenic human and mouse regions of chromosome 1. The 39 kD AIM2 protein, unlike IFI 16 and MNDA, is present primarily in the cytoplasm (Choubey *et al.*, 2000). The overexpression of AIM2 in murine transfected fibroblasts retards cell proliferation and, under reduced serum conditions, increases the

susceptibility of the cells to apoptosis, suggesting that AIM2 may be a direct effector of IFNs and relay their growth-inhibitory effects in the human (Choubey *et al.*, 2000).

The AIM2 mRNA has been detected in the spleen, small intestine, peripheral blood leukocytes and can be induced by IFN gamma in HL60 cells. The expression pattern of AIM2 is consistent with data obtained with IFI 16 and MNDA, suggesting that genes in this family may be involved in the blood cell-specific response to IFNs. However, the AIM2 gene also is expressed in non-lymphoid tissues and interestingly, up-regulated in a tumor cell line whose tumorigenicity has been suppressed by an extra copy of human chromosome 6, raising the possibility of its involvement in tumorigenic suppression (DeYoung *et al.*, 1997).

I-3 Genetic aspects of bone density regulation

Osteoporosis is a common metabolic bone disease characterized by a reduction in bone mineral density (BMD), microarchitectural deterioration of bone tissue and an increased risk of fracture. Susceptibility to osteoporosis is multifaceted with genetic factors, endogenous factors depending mostly on nutrition and hormones and environmental factors, such as physical exercise, influencing it (Audi *et al.*, 1999). Peak bone mass occurs at the end of the second decade of life and is an important determinant of osteoporosis susceptibility later in life (Teegarden *et al.*, 1995; Theintz *et al.*, 1992). Low peak bone density is considered to be a major risk factor for osteoporotic fracture in elderly adults. It has been estimated that approximately 40% of postmenopausal women will suffer at least one osteoporotic fracture (Cummings *et al.*, 1985). Major regulatory factors support the adult skeletal system, including environmental factors, such as nutrition or physical exercise, and endogenous factors, such as reproductive status, aging, disease states (cancer, diabete, hyperparathyroidism), hormones and mineral levels. Numerous studies have demonstrated significant

heritability of peak bone mass at various skeletal sites (Seeman and Hopper, 1997), suggesting that the effects of all such factors to support the adult skeletal system are genetically regulated. These studies have demonstrated that 60-70% of the normal variability in bone density is genetically determined. Although many regions of the genome have been involved with bone density regulation, identifying genes participating in regulation of the adult skeleton is still at an early stage. The main strategies for identification of such genes are based on looking for evidence of an association between a phenotypic trait, such as bone density, and a series of polymorphic genetic markers (Stewart and Ralster, 2000). The genetic markers are regions of the genome analyzed by PCR-based methods on DNA extracted from peripheral blood. There are two main types of markers: repeat polymorphisms of variable length (VNTR) and single nucleotide polymorphisms (SNPs). Genetic studies involve analyzing a large number of markers throughout the whole genome, or specific markers concentrated in regions of interest (candidate loci) or specific genes of interest (candidate genes). Regions of chromosomes containing alleles that influence phenotypic traits such as BMD are termed Quantitative Trait Loci (QTL). The search for QTL in population-based studies often gives contradictory results, because of the natural genetic heterogeneity of the human population coupled with important environmental differences. Comparison within twin pairs or among siblings is one method to keep environmental factors relatively constant. However, this approach requires large numbers of twins, which are not frequent in the general population. Studies in animal models provides an additional method of identifying QTL. In particular, inbred mice offer a large number of genetically identical twins whose environment can be strictly controlled. Furthermore, each inbred strain is genetically different from every other inbred strain and cross mating experiments of mouse strains with low BMD and high BMD make it possible to study the segregation of genes essential to bone density regulation. Genetic studies on inbred strains of mice have been widely successful in identifying QTL and, in several

cases, candidate gene association studies have isolated the main regulators of bone metabolism, such as calciotrophic hormones, bone matrix proteins, steroid hormones and local regulators of bone metabolism, whose genetic polymorphism likely contributes to BMD variability. Among them are vitamin D receptor (Morrison *et al.*, 1994), collagen 1A1 (Uitterlinden *et al.*, 1998), IGF-1 (Rosen *et al.*, 1998), apolipoprotein E (Shiraki *et al.*, 1997), calcitonin receptor (Masi *et al.*, 1998), bone-specific alkaline phosphatase (Harris *et al.*, 1998), TGF beta-1 (Langdahl *et al.*, 1997), alpha 2HS-glycoprotein (Zmuda *et al.*, 1998), estrogen receptor (Deng *et al.*, 1998), IL-6 (Ota *et al.*, 1999) and osteocalcin (Dohi *et al.*, 1998). In spite of the controversial status of the evidence for some of these genes, it is clear from the wide range of factors influencing bone density that it is a quantitative trait dependent upon the action of numerous genes that individually have modest effects on bone density.

The search for possible associations between polymorphic variations in any candidate genes located on chromosome 1q and BMD regulation has led to the description of several associations. First, osteocalcin is an abundant non-collagenous component of the bone matrix and this highly conserved bone-specific protein is synthesized by the bone-forming osteoblast cells under transcriptional regulation through a vitamin D-response element (Dohi *et al.*, 1998). A role for osteocalcin in bone resorption has been previously suggested because of its ability to influence recruitment and differentiation of the bone-resorbing osteoclast cells. Dohi *et al.* (1998) described a polymorphism in the osteocalcin gene promoter that is related to BMD and prevalence of osteoporosis in postmenopausal Japanese women. Raymond *et al.* (1999) mapped the human osteocalcin gene to 1q25-1q31 by somatic cell hybridization and described a new microsatellite (CA) polymorphism linked to the gene associated with BMD in a white postmenopausal American population. These data suggest that genetic polymorphism at the osteocalcin gene could impact BMD status in postmenopausal women and predispose some women to osteoporosis. Second, absorptive hypercalciura

(AH) is a common cause of kidney stones and is due to intestinal hyperabsorption of calcium. In some severe cases, this condition may be accompanied by low bone density (Pietchmann *et al.*, 1992). A candidate locus for AH has been mapped by linkage analysis studies to human chromosome 1q23.3-1q24 (Reed *et al.*, 1999). Interestingly, no candidate genes of interest are present in this region, raising the possibility that a still unknown gene in this region could be involved in the regulation of intestinal calcium absorption and possibly BMD regulation. Finally, pycnodisostosis is a rare autosomal recessive skeletal disorder characterized by short stature, wide cranial sutures and increased bone density and fragility. A candidate locus for this condition has been mapped by linkage analysis studies to human chromosome 1q21 (Polymeropoulos *et al.*, 1995). Homology searches of ESTs mapped within this region to DNA sequences of genes of interest identified the cathepsin K gene as a primary candidate for this condition (Johnson *et al.*, 1996). Cathepsin K is a cysteine protease highly expressed in bone-resorbing osteoclast cells and is thought to be involved in the degradation of the bone matrix because it digests collagen (which composes up to 90% of the bone matrix) (Inaoka *et al.*, 1995). Interestingly, a cytosine to thymine transition at nucleotide 862 of the cathepsin K coding sequence has been found in the DNA of an individual affected with pycnodysostosis, resulting in an arginine to stop codon alteration at amino acid 241 (Johnson *et al.*, 1996). These data suggest that cathepsin K could have an important role in BMD regulation and osteoporosis that could result, at least partially, from excess cathepsin K activity or increased susceptibility of the bone matrix to cathepsin K activity.

Chapter II

Materials and Methods

The large scale genomic DNA sequencing approach used in this dissertation is based on a random shotgun sequencing method to collect the majority of the sequencing data, followed by custom synthetic primer and PCR-based methods to close the remaining gaps, assembly proofreading and sequence analysis. The following protocols are taken from our laboratory protocol manual which was published (Chissoe *et al.*, 1991) and is updated on our laboratory website at URL: http://www.genome.ou.edu/proto.html.

II-1 Random shotgun sequencing

Random shotgun sequencing involves eight major steps (see Fig. 2.1).

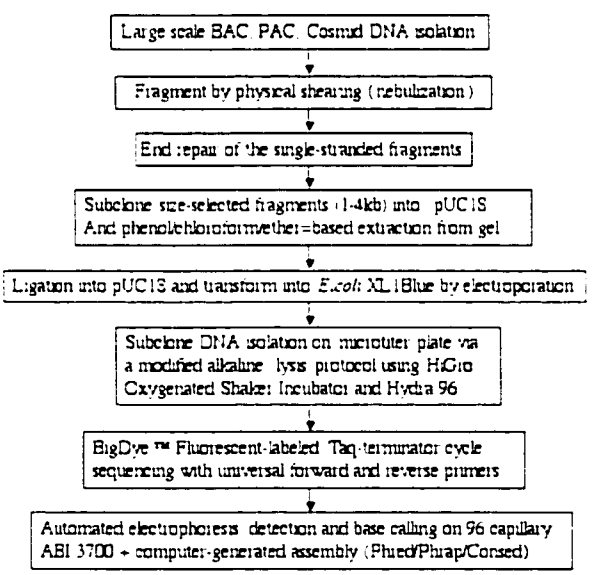


Fig. 2.1 Summarized steps of the random shotgun sequencing strategy. See text for details.

(1) A large scale DNA isolation for each clone harboring the DNA insert of interest; (2) a physical shearing via nebulization of 20 to 50 μg of the isolated DNA clone, generating fragments between approximately 0.8 and 6 kb; (3) end-repair of the single-stranded ends of the nebulized DNA fragments; (4) a size-selection of the now blunt-ended fragments between 1 and 4 kb using a low melting point agarose gel and a phenol/chloroform/ether-based extraction method; (5) the size-selected blunt-ended fragments are ligated into SmaI-treated pUC18 vector using T_4 DNA ligase and the ligation mix is transformed into *E. coli* XLI Blue MRF' electrocompetent cells; (6) miniprep isolations are performed to isolate DNA subclones; (7) the DNA subclones are sequenced using the Taq DNA polymerase-catalyzed reaction with ABI BigDye fluorescent-labeled ddNTP terminators. After purification, the reactions are loaded onto an automated Perkin Elmer ABI 3700 capillary sequencer which performs electrophoresis, signal detection and base calling; (8) the data are assembled into large contigs with computer programs Phred, Phrap and Consed on local Sun work stations.

II-1-1 Large scale DNA isolation

A modified alkaline lysis procedure, involving a second acetate precipitation step to remove remaining sheared *E. coli* chromosomal DNA contamination, was used for the large scale isolation of DNA clone.

First, a smear of colonies harboring the DNA clone of interest is transferred into a 12X75 mm Falcon tube containing 3 ml of LB medium supplemented with the appropriate antibiotic and incubated at 37°C for 8 to 10 hours with shaking at 250 rpm. The culture is transferred to an Erlenmeyer flask containing 50 ml of a similar medium and incubated for another 8 to 10 hours under similar conditions. The culture then is transferred to a flask containing 1 liter of a similar medium and incubated for an additional 8 to 10 hours under similar conditions. Cells are harvested into 500 ml

centrifuge bottles by centrifugation at 7000 rpm for 20 minutes in a RC5-B centrifuge. Cell pellets can be stored at -70°C at this point.

The DNA clone is isolated according to the following protocol. The cell pellet is resuspended in 35 ml of GET/Lysozyme solution by gently teasing the pellet with a spatula and incubated at room temperature for 10 minutes. 70 ml of alkaline lysis solution are added to the resuspended cells. The solution is mixed very gently to avoid the shearing of *E. coli* chromosomal DNA and incubated in an ice-water bath for no ore than 5 minutes. *E. coli* chromosomal DNA, membranes and proteins are precipitated by adding 52.5 ml of 3M NaOAc, followed by a very gentle mixing and incubation in an ice-water bath for 15 minutes. The lysate is cleared of precipitated SDS, proteins, memebrane and chromosomal DNA by pouring through a double layer of cheesecloth. Further clearing is performed by transferring the supernatant into a 250 ml centrifuge bottle and centrifuging at 10000 rpm for 30 min in a RC5-B centrifuge. If neccessary, an additional centrifugation is performed to make sure that every insoluble material is removed. An equal volume of isopropanol is added to the supernatant and, after incubation for 5 minutes at room temperature, the solution is centrifuged at 9000 rpm for 20 minutes in the RC5-B centrifuge. The supernatant is decanted and the pellet resuspended in 18 ml of 10:1 TE buffer. The solution is divided equally into two 50 ml Corning centrifuge tubes and 4.5 ml of 7.5M KOAc are added to each tube. After incubation for 30 minutes at -70°C, the solution is thawed and centrifuged in Beckman GS-6R centrifuge at 2000 rpm for 10 minutes. The resulting supernatant is incubated with 100 μg/ml (final concentration) of DNAse-free RNAse A and 40 μl/100 ml from a 100U/μl stock of RNAse T_1 in a 37°C water bath for 1 hour. After incubation, 30 ml of cold 100% ethanol are added to the solution and the solution is incubated in an ice-water bath for 15 minutes. The solution then is centrifuged in Beckman GS-6R centrifuge at 3000 rpm for 25 minutes. The resulting DNA pellet is washed with 30 ml of 70% ethanol. The DNA pellet then is dried in a vacuum oven and resuspended in a total of 2

ml of sterile ddH_2O. The DNA concentration and purity are estimated by agarose gel electrophoresis.

II-1-2 Physical Shearing

The isolated DNA clone is sheared into fragments of approximately 0.8 to 6 kb by nebulization. First, an aqueous solution containing 20 to 50 μg of DNA is mixed in a nebulizer with 0.5 ml of sterile glycerol and sterile ddH_2O to a final volume of 2 ml. Nebulization is performed in an ice-water-salt bath at -5°C for 2.5 minutes with a nitrogen pressure of 8 psi. When the pressurized nitrogen gas is applied, the DNA solution is forced to enter a small chamber in the nebulizer top and, then, after exiting through an orifice, the solution comes in contact with a spherical surface to form small droplets. These droplets contain DNA sheared by surface tension over a size range which is controlled by the gas pressure, the temperature and the viscosity of the solution. Glycerol is added to the solution to increase its viscosity, and also to protect the DNA solution from freezing. The nebulized sample is collected by briefly centrifuging the nebulizer in a Beckman GS-6R centrifuge at 1500 rpm. The solution then is equally divided into four 1.5 ml microcentrifuge tubes, ethanol precipitated, vacuum dried and resuspended in 24 μl of sterile ddH_2O per microcentrifuge tube prior to proceeding with fragment end-repair.

II-1-3 End Repair

The nebulization generates for the most part DNA fragments with single-stranded ends. Since the ligation step involves ligating DNA fragments to linearized blunt-ended pUC18 vector with dephosphorylated 5' ends, the DNA fragments need to be end-repaired to make them blunt-ended and phosphorylated at their 3' ends. 24 μl of

the DNA solution are mixed with 6 μl of 10X TM buffer, 6 μl of 10X kinase buffer, 6 μl of 10 mM rATP, 10 μl of 0.25 mM dNTPs, 2 μl of T_4 polynucleotide kinase, 3 μl of T_4 DNA polymerase and 3 μl of Klenow DNA polymerase. The use of two polymerase activities is justified by the very active single-stranded 3'-5' exonuclease activity of T_4 DNA polymerase to degrade 3' overhangs. The solution then is incubated in a 37°C water bath for 30 minutes. The reaction is stopped by cooling the sample on ice.

II-1-4 Size selection and DNA extraction

Following step II-1-3, the reaction mixture is loaded into a 0.8% low melting point agarose gel. After an electrophoretical migration for 30 to 60 minutes at 100-120 mA, DNA fragments are visualized in a dark room on a long wave UV light box with appropriate molecular weight markers (lambda DNA digested by HindIII and phage X174 digested by HaeIII). The fragments in the desired range (1-4 kb) are excised from the gel with a razor blade, put into a microcentrifuge tube and frozen at -70°C for at least 15 minutes.

The gel slice containing the DNA fragments of interest is melted by incubation in a 70°C water bath for 30 minutes. An equal volume of TE-saturated phenol is added. After vortexing for 30 seconds, the solution is centrifuged at room temperature for 5 minutes. The upper phase is recovered and mixed with an equal volume of a solution containing half a volume of TE-saturated phenol and half a volume of chloroform. The solution is vortexed for 30 seconds and centrifuged again. The upper phase is recovered and further extracted by adding an equal volume of chloroform. A final round of extraction is finally performed by adding an equal volume of water-saturated ether after what the lower aqueous phase is recovered, ethanol precipitated and vacuum dried. The DNA pellet then is resuspended in 10 to 20 μl of sterile ddH_2O.

II-1-5 Shotgun Library Construction

Library construction requires two steps: ligation and transformation. The ligation is performed by mixing 100 to 1000 ng of insert DNA fragments, 20 ng of pUC18 vector previously linearized and blunt-ended with SmaI and dephosphorylated with calf intestinal alkaline phosphatase, 1 μl of T_4 DNA ligase and 1 μl of 10X ligation buffer, with sterile ddH_2O up to a final volume of 10 μl. The reaction mixture is incubated overnight at 4°C. A negative control is performed in the absence of insert DNA fragments to test for any possible vector self-ligation. The transformation is performed by electroporation. Electrocompetent cells are prepared by inoculating 3 ml of YENB medium supplemented with 12.5 μg/ml tetracycline with a colony picked from a plate previously inoculated with an aliquot of glycerol stock of XL1 Blue MRF' *E. coli* cells and containing 12.5 μg/ml tetracycline. After incubation overnight, 5 to 10 μl of the culture are inoculated into 1 liter of YENB medium supplemented with 12.5 μg/ml tetracycline. The cells are incubated at a shaking of 250 rpm at 37°C until they reach an A_{600} of about 0.5. After being harvested in four 500 ml centrifuge bottles by centrifugation at 5000 rpm for 10 minutes in the RC5-B centrifuge, the cells are washed twice with 100 ml of cold sterile water and once with 40 ml of 10% cold sterile glycerol solution. Cells then are resuspended in 3 ml of 10% cold sterile glycerol solution, divided into small aliquots of 40 μl and frozen at -70°C for storage. For the transformation step, 40 μl of electrocompetent cells are mixed with 3 μl of ligation mix and incubated at 4°C for 1 minute. the mixture then is electroporated in a Bio Rad *E. coli* pulser. One ml of fresh YENB medium is added and the sample is incubated at 37°C at a shaking of 350 rpm for no more than 60 minutes. Cells are harvested by centrifugation in a Beckman GS-6R centrifuge at 2500 rpm for 5 minutes and resuspended in 600 μl of YENB medium, to which are added 90 μl of 20 mg/ml IPTG (in ddH_2O) and 90 μl of 24 mg/ml X-Gal (in DMF). Cells then are plated out on LB

plates supplemented with 100 μg/ml ampicillin (LB/Amp plates). The plates are incubated 18 to 20 hours at 37°C.

II-1-6 Miniprep isolation of DNA subclones

The resulting white colonies are picked with a Flexys automated colony picker (Genomic Solutions) and inoculated into each well of a 96-well flat bottom microtiter plate (Dynayech) containing 150 μl of TB medium supplemented with 100 μg/ml ampicillin. The plates are incubated for 18 hours at 37°C in an HiGro oxygenated shaker incubator (Gene Machines) with shaking at 520 rpm. Cells are harvested by centrifugation in a Beckman GS-6R centrifuge at 2500 rpm for 10 minutes and the cell pellets are stored at -70°C for at least 2 hours. Cells then are resuspended by adding 60 μl of TE-RNAse A solution into each well of the 96-well flat bottom microtiter plates (Dynatech) by using an Hydra 96 (Robbins). After shaking on a bench top shaker for 20 minutes, cells are lysed by adding 60 μl of alkaline lysis solution into each well of the microtiter plates using the Hydra 96 (Robbins) and incubating the plates on a bench top shaker for at least 30 minutes. 60 μl of 3M NaOAc are added to precipitate chromosomal DNA, membrane and proteins and the plates are incubated for 30 minutes in the HiGro oxygenated shaker incubator (Gene Machines) with a shaking at 520 rpm. The plates then are stored at -20°C overnight. After thawing, the plates are centrifuged in a Beckman GS-6R centrifuge at 3200 rpm for 30 minutes. Using the Hydra 96 (Robbins), 60 μl of the supernatant are removed from each well and transferred to a new V-bottom microtiter plate (Dynex). 130 μl of ethanol 95% then are manually added using a 12-channel manually pipeter and the plates are centrifuged in a Beckman GS-6R centrifuge at 3200 rpm and at 4°C for 30 minutes. The resulting DNA pellet is washed with 150 μl of ethanol 70%. The DNA pellets then are dried in a vacuum oven, resuspended in 50 μl of sterile ddH_2O and stored at -20°C for later use.

II-1-7 Cycle sequencing reaction and purification

In the cycle sequencing reaction with ABI BigDye fluorescent-labeled ddNTP terminators, 2 μl of subcloned DNA are mixed with 1 μl of forward or reverse universal primers (6.5 μM), 1 μl of 25% DMSO (in ddH_2O) and 2 μl of ABI BigDye terminator mix previously diluted to one-third by adding two volumes of 5X TM buffer. The ABI BigDye terminator mix (provided by the supplier) contains fluorescent-labeled ddNTP terminators, dNTPs, Taq DNA polymerase and $MgCl_2$. Typically, 384 reactions are prepared into each well of a 384-well thermocycle plate (Robbins) (or each well of a 384-well Viper plate) by using the Hydra 96 (Robbins) and performed simultaneously. 60 cycles are performed for each reaction in a Perkin Elmer GeneAmp PCR system 9600 thermal cycler according to the following cycle conditions: 95°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes. After the reactions, the unincorporated dye-terminators are removed and the reaction products purified by ethanol precipitation. Briefly, 15 μl of sterile ddH_2O are added to each reaction using the Hydra 96 (Robbins). After a brief centrifugation at 1500 rpm in a Beckman GS-6R centrifuge, the reactions are transferred from the 384-well thermocycle plate (or from the 384-well Viper plate) to four 96-well thermocycle plate using the Hydra 96 (Robbins). 60 μl of 95% ethanol/0.12N NaOAc are added to each well by using a 12-channel manually pipeter and the plates are centrifuged in a Beckman GS-6R centrifuge at 4°C and at 3200 rpm for 30 minutes. The supernatant then is removed by briefly centrifuging the plates in a Beckman GS-6R centrifuge at 400 rpm. The DNA pellet is washed with 70 μl of 70% ethanol, dried in a vacuum oven and stored at -20°C.

Before loading onto the Perkin Elmer ABI 3700 capillary sequencer, each DNA sample is resuspended in 20 to 50 μl of sterile ddH_2O. Each sample then is electrophoretically loaded onto a capillary containing ABI 3700 POP-5™ polymer.

Electrophoresis is performed for 3 hours at 6.5 kV and, after data collection, the raw data are transferred to a local SUN work station for further analysis.

II-1-8 Data assembly

The sequence read length generated on the Perkin Elmer ABI 3700 sequencer is typically between 500 and 600 bp. Because the position of the reads generated from a random shotgun library are unknown, computer programs are used on local SUN work stations to assemble these reads according to the potential sequence overlap between them. A program called PhredPhrap (P. Green; University of Washington) performs this task. PhredPhrap is a highly accurate program that assemble short reads into large contigs, first, by estimating the quality of each base and, second, by using this quality evaluation to assemble the reads together. Another program, called Consed (P. Green; University of Washington), creates a consensus sequence from the assembly and performs several other tasks, such as sequence editing.

Such a random assembly usually results in an average 6 to 8 fold coverage of shotgun sequence data and the generation of 10 to 20 contiguous contigs covering typically 95 to 99% of the sequence of the clone of interest.

II-2 Gap closure

After 6 to 8 fold coverage of shotgun sequence data is collected and assembled into a database, a more directed strategy is applied, in which the remaining sequence of the clone is obtained to close the small gaps between contiguous contigs. The presence of gaps in the sequence at this high depth of coverage is not intuitive because, if the subclones picked for shotgun sequencing were truly picked randomly, then a typical clone could be completely sequenced by random shotgun sequencing. However,

because of a natural statistical bias in the random sequencing approach, some regions of the clone have a lower extent of coverage (a 6 to 8 fold coverage is actually an average coverage). Furthermore, ends of contigs are usually made of low quality reads or reads shorter than the average read length. Then uncovered sequences may remain because they are not covered by shotgun reads. In addition, some regions are difficult to sequence because they have a high percentage of G/C which results in secondary structure that prevents read through the region. Other regions that contain long stretches of di- or tri- nucleotide repeats also are difficult to sequence through.

To overcome the above, several gap closing strategies were used during this dissertation research.

II-2-1 Primer walking

From the previous discussion, the approach taken entailed sequencing the shotgun subclones in both the forward and the reverse directions, using respectively forward and reverse universal primers, to generate forward and reverse sequence reads respectively. Since the average sequence read length was 500 to 600 bp, but the length of a subclone can be up to ~4 kb, the middle portion of a shotgun subclone typically is not sequenced using the forward and reverse universal primers. However, if the forward and reverse shotgun reads of a same subclone are located at the end of two different contigs and pointing towards the end of these contigs, the gap between these two contigs is covered by the unsequenced region of this subclone. Therefore, a custom-synthetic primer walking approach can be used to obtain the sequence covering the gap and thereby connect the original two adjacent contigs. To accomplish this approach, custom synthesized primers can be synthesized using a Mermade oligonucleotide synthesizer after their sequences are determined using a primer picking program such as PrimOU. If the gap has a high G/C content or contains long stretches of polynucleotide repeats,

alternative sequencing reagents can be used. For example, it has been shown that G/C-rich gaps are more efficiently closed by using a dGTP-containing dye reaction mix instead of the dITP-containing BigDye reaction mix while A/T-rich gaps are closed more efficiently using the dRhodamine dye reaction mix instead of BigDye reaction mix. Additionally, adding 5% to 10% DMSO or 1M Betaine to the cycle sequencing reaction may more easily denature the DNA regions with a high G/C content during the reaction. After cycle sequencing reaction and excess dye-terminator removal by purification through G-50 column, the mixture is loaded onto a 6% polyacrylamide gel in a Perkin Elmer ABI 377 automatic sequencer and then electrophoresed for 8 hours. After initial processing on the ABI 377's associated computer, the data is transferred to a local SUN work station for further analysis.

II-2-2 PCR-based methods

If no shotgun subclone was found covering the gap, then this region was amplified by a polymerase chain reaction (PCR) using custom synthesized primers to generate a DNA sequencing template covering the gap. The reaction mixture for a PCR typically contained 10 to 20 ng of BAC or cosmid clone DNA, 100 pmol of each primer, 5U of AmpliTaq DNA polymerase (Perkin Elmer), 20 mM of each dNTP, 10 μl of 10X PCR buffer (500 mM KCl; 100 mM Tris-HCl, pH 7.6; 10 mM $MgCl_2$ in sterile ddH_2O) and sterile ddH_2O up to a total volume of 100 μl. 25 cycles of 95°C for 1 minute, 55°C for 1 minute and 72°C for 2 minutes are performed in a Perkin Elmer GeneAmp PCR system 9600 thermal cycler. The concentration and size of the PCR product is estimated by agarose gel electrophoresis. The PCR product is purified by mixing 10 μl of PCR reaction mixture with 10U of exonuclease I (USB) and 1U of shrimp alkaline phosphatase (USB) and incubating the reaction mixture in a thermocycler at 37°C for 30 minutes followed by enzyme deactivation at 80°C for 15

minutes. 2 μl of this reaction mixture then is used for cycle sequencing reaction with custom synthesized primers. An alternative PCR approach, in which dGTP is replaced by 7-deaza dGTP in the PCR reaction mixture appears to be very effective in closing gaps with a high G/C content. Here, the PCR template containing 7-deaza dGMP is more easily denatured during the cycle sequencing reaction, as 7-deaza dGTP does not base pair with dCTP.

If the contig order, and consequently the combination of the two custom synthesized primers used for the PCR reaction, is not known, the Multiplex PCR (MPCR) can be used. In this method, more than one PCR reaction are performed simultaneously in the same tube, because of the simultaneous presence of multiple custom synthesized primers included in the reaction. Typically, 20 pmol of up to 20 primers are mixed in the same tube, dried in a vacuum oven and resuspended in 10 μl of sterile ddH_2O. Then the MPCR is performed by mixing 250 ng of the BAC or cosmid clone DNA, 10 μl of concentrated primers, 25 mM of each dNTPs, 2U of Taq polymerase XL (Perkin Elmer), 2.5 μl of DMSO, 10 μl of 10X PCR buffer (83 mM $(NH_4)_2SO_4$; 335 mM Tris-HCl, pH 9.0; 33.5 mM $MgCl_2$; 50 mM ß-mercaptoethanol and 850 μg/ml bovine serum albumin) and sterile ddH_2O up to a total volume of 50 μl. The reaction mixture then is incubated in a thermocycler for 6 minutes at 94°C, followed by 30 to 40 cycles of 94°C for 30 seconds, 55°C for 30 seconds and 65°C for 4 minutes each. After the concentrations and sizes of the MPCR products are estimated by agarose gel electrophoresis, the MPCR products are purified by adding 10U of shrimp alkaline phosphatase (USB) and 100U of exonuclease I (USB) to the reaction mixture. This mixture then is incubated at 37°C for 30 minutes and the enzymes are deactivated at 80°C for 10 minutes. The MPCR products then are extracted once with one volume of phenol/chloroform {1:1}, precipitated with 2.5 volumes of 95% ethanol/0.12N NaOAc, washed with ethanol 70%, dried in a vacuum oven and resuspended in 50 μl of sterile ddH_2O. Typically 2 μl of the MPCR products are used individually for cycle

sequencing reaction, one reaction being performed with each of the custom synthesized primers used for the MPCR.

II-2-3 LI-COR-based DNA sequencing

The average sequence read length generated by the Perkin Elmer ABI 3700 and 377 sequencers is about 500 to 600 bp. However, the LI-COR automatic sequencer often can generate sequence reads up to ~1000 bp. The LI-COR dye-terminator sequencing reaction differs from the Perkin Elmer ABI BigDye ddNTP terminator sequencing reaction by using the same fluorescent-labeled terminator-dye for each ddNTP instead of four different fluorescent-labeled terminator dyes. Consequently, four reactions, one specific for each ddNTP terminators, are prepared and incubated separately. Because the four reactions are loaded separately, no "spectral deconvolution" is needed and therefore longer read lengths are possible.

If the gap between two contigs is less than a few hundreds bp long, and if the sequence of a shotgun read is less than 500 bp away from and pointing toward the end of a contig, then the corresponding shotgun subclone can be resequenced on the LI-COR according to the following conditions. Four cycle sequencing reactions are performed separately. First, a "master mix" is prepared by mixing 1 to 2 μl of subclone DNA (10 to 20 ng), 1 μl of forward or reverse universal primers (6.5 μM), 1 μl of reaction buffer, 2 μl of thermosequenase (USB) and sterile ddH_2O up to a final volume of 26 μl. 6.5 μl of the "master mix" are aliquoted into four wells of a 96-well thermocycle plate (Robbins) to which are added 2 μl of "terminator-mix" specific for each ddNTP. 35 cycles are performed in a Perkin Elmer GeneAmp PCR system 9600 thermal cycler according to the following conditions: 95°C for 30 seconds, 50°C for 30 seconds and 72°C for 45 seconds. After the unincorporated dye-terminators are removed by ethanol precipitation as described above, the reaction mixture is dried in a

vacuum oven and stored at -20°C. Before loading onto the LI-COR automatic sequencer, the samples are resuspended in 3 μl of loading buffer (10 mg/ml blue dextran and 5 mM EDTA in deionized formamide) and heated at 95°C for 2 minutes. 1 μl of each sample is loaded onto a polyacrylamide gel (0.2 mm) and electrophoresis is performed for 10 hours. After preliminary data processing, the data are transferred to a local SUN workstation and analyzed as described above.

II-2-4 M.J. Basestation-based DNA sequencing

The Basestation, M.J. Research Inc., is a horizontal slab-gel DNA sequencer which is capable of sequence read lengths in excess of 1000 bases from the priming site as it uses thinner gels (50 μm) instead of the 200 μm thick gels used in the ABI 377. Here, samples are applied after incubating in the thermal cycler under the identical conditions as used for the shotgun sequencing reactions described above. The electrophoresis run conditions of 2200 volts for 6.7 hours, followed by base calling on the associated computer, often can produce data which is in the 800-1000 base read length range and thereby close many gaps without excessive additional procedures being needed.

II-3 Proofreading and correcting the assembled data

Although PhredPhrap gives a highly accurate assembly, some errors may occasionally occur, especially in complex repetitive regions that are longer than the average read length of a shotgun read. In this case, false joins will occur between shotgun reads that otherwise belong to different regions. Several strategies are available to correct these misassembly problems.

During random shotgun sequencing, shotgun subclones are sequenced both in the forward and in the reverse direction by using the forward and reverse universal primers (generating respectively a forward read and a reverse read). Because shotgun subclones are first size-selected during random shotgun sequencing, and their size usually cannot exceed ~4 kb, the distance between forward and reverse reads cannot be longer than ~4 kb. After random shotgun sequencing, several contigs are generated whose length usually spans several kb. If the forward and reverse reads from a same shotgun subclone are located in the middle part of two different contigs and therefore more than ~4 kb appart, it is very likely that one, or both, of these two reads is misplaced in the PhredPhrap assembly. Thus, the corresponding shotgun subclone is a primary candidate for assembly proofreading. A list of these shotgun subclones with forward and reverse reads located on two different contigs can be obtained using a locally modified program ("printrev_phrap") and the corresponding forward and reverse reads can be directly located in the assembly by using another locally designed program ("exgap"). The combination of these two programs is used to select several subclones that can be totally sequenced in both the forward and reverse directions by primer walking. Here, to accomplish this, custom primers are designed from the forward and reverse reads of a same subclone and synthesized. Cycle sequencing reactions are performed in both the forward and reverse directions by using these custom synthesized primers. The reads generated are used to design and synthesize more custom primers that will be used for more cycle sequencing reactions. It is of upmost importance that the reads generated in both the forward and reverse directions by this series of primer walking reactions initially be assembled in a separate assembly, and only then the consensus read generated from this assembly is transferred to the main PhredPhrap assembly. Since the consensus reads generated from the selected subclones have a read length of up to ~4 kb and, it can span a unique region of the clone immediately adjacent to a repetitive region. Any repetitive region which likely caused the original misassembly

then will be positionned correctly in the new assembly. However, it must be noted that, even though the consensus assembly is assembled correctly, because of their small size, the underlying shotgun reads corresponding to this consensus sequence may not be positionned at the correct repetitive region in the new assembly.

The Phrap-based sequence data assembly accuracy also can be validated by showing an agreement between the assembly generated by PhredPhrap and the assembly generated by other programs, such as CAPII and XGAP. Assembly accuracy can be further verified by alignments performed with available genomic or cDNA sequences present in databases, such as the GenBank database. In addition, it also may be useful, in the case of very large repeated sequences, to experimentally generate a series of nested deleted clones by minitransposon random insertion (Chatterjee and Coren, 1997) and end-sequencing these deleted clones with specifically designed custom primers. Here, 1 μg of clone DNA, previously dried up in a vacuum oven and resuspended in 2 μl of sterile ddH_2O, is mixed with 16 pmol of primer (16 μM), 1 μl of 25% DMSO (in ddH_2O) and 2 μl of ABI BigDye terminator mix and 99 cycles are performed in a Perkin Elmer GeneAmp PCR system 9600 thermal cycler according to the following conditions: 95°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes. Then, the reaction products are concentrated by ethanol precipitation, dried in a vacuum oven and stored at -20°C before loading onto a Perkin Elmer ABI 377 sequencer. The position of the end-sequencing reads in the assembly is used to validate the assembly by comparing their distance to the end of the assembly with the size of the deleted clone previously assessed by field inversion gel electrophoresis (FIGE).

Finally, to ensure that each base in the assembly is called correctly, the so-called "rule of three" is applied, to ensure that each base in the assembly is read unambiguously at least once in the forward or reverse direction and at least twice in the opposite direction, usually resulting in less than one ambiguous base every 10 kb as determined by Consed. If any base is not covered according to the "rule of three",

further sequencing reads must be generated using custom synthesized primers on a template(either subclone or the original BAC or cosmid, or PCR products) spanning the weak region of interest. These reads then are added to the assembly until the Consed error rate is less than one every 10 kb.

II-4 Sequence analysis

The final stage of a sequencing project consists of analyzing the consensus sequence generated from the assembly to extract biological information. Programs are available to perform different tasks, depending on the nature of the research conducted on the newly obtained sequence. The different tasks performed, and the corresponding programs, are as follows.

II-4-1 Detecting regions with repeated sequences

A program called RepeatMasker detects repeat sequences in the consensus sequence by comparing the query genomic sequence with a database of common repeat sequences usually present in vertebrate DNA, such as SINEs, LINEs and other known repeated sequences. The query sequence file is in FASTA format and RepeatMasker generates an output sequence file in FASTA format containing masked portion corresponding to detected repeat sequences in the genomic sequence of interest. In addition, RepeatMasker also calculates the GC% of the consensus sequence.

II-4-2 Gene prediction

Several programs, such as MZEF (Zhang, 1997), FGENESH (Salamov and Solovyev, 2000), GRAIL2 (Uberbacher and Mural, 1991) or GENSCAN (Burge and

Karlin, 1997), are currently available to predict potential coding regions in the genomic sequence of interest. Potential coding regions prediction is based on codon usage in open reading frame regions, hexanucleotide distribution and the occurence of a consensus exon-intron splice site, and although each of the above programs have specific features, GENSCAN often is the program of choice because of its somewhat superior performance. However, further studies such as database similarity searches usually are needed to validate any predicted genes.

II-4-3 Database similarity searches

Gene prediction validation, as well as the identification of known genes in the genomic sequence of interest, usually initially are performed by BLAST (Basic Local Alignment Search Tool) (Altschul *et al.*, 1990). The BLAST program can search for homology between a query sequence and the known sequences present in databases such as GenBank or EMBL (European Molecular Biology Laboratory) databases. These database sequences are from various species and include cDNA sequences, high troughput genomic sequences (working draft) or finished genomic sequences. Other databases, such as the dbEST database, also can be compared to a query sequence. When homologies are found, alignments are performed automatically between the homologous regions of a query and database sequences and given in the BLAST output along with the similarity probability score.

II-4-4 Promoter and transcription factor binding sites searches

Potential transcription factor binding sites usually are searched by the MOTIF program which compares a query nucleotide sequence with a database containing a listing of vertebrate transcription factor binding sites (see http://motif.genome.ad.jp).

The query sequence file is in FASTA format and MOTIF generates a table output file in which the potential transcription factor binding sites are positioned along the query sequence.

Several programs are available to search for promoters, although their performance in terms of sensitivity and specificity is generally less than 50% (Fickett and Hatzigeorgiou, 1997). One such program, Neural Network for Promoter Prediction (NNPP) (see http://www.fruitfly.org/seq_tools/promoter.html), predicts promoters using a neural network and gives relatively accurate results. However, because of their relatively low accuracy and the high number of false positive predictions, promoter prediction typically only is performed on relatively short sequences near the possible transcription start site, where the probability of having a promoter region is strong.

II-4-5 Sequence Comparison

Several specialized programs are available to perform direct sequence comparisons, using short cDNA or genomic sequence files in FASTA format. Two examples are the DOTTER (Sonnhamer and Durbin, 1995) and CROSSMATCH (P.Green) programs which can compare reasonably large sequences (up to a few hundreds kb). However, only relatively homologous sequences will give significant results using these two programs. The CROSSMATCH program generates a list of the homologous regions of two sequences while the DOTTER program generates a dot plot, in which a dot is plotted every time a base from the first query sequence is similar to a base from the second query sequence. PIPmaker (Schwartz *et al.*, 2000) is a program that aligns sequences using a high stringency implementation of BLAST, called BLASTZ and is available at URL: http://bio.cse.psu.edu. This program is especially useful for comparative genomics and segments of these alignments are then drawn according to their position in the first query sequence, on a percent identity plot

(PIP) along the length of the first sequence. Only segments with an identity of 50% or more are plotted so regions that match poorly appear blank. Finally, sequence alignments also can be performed by using ALIGN (Myers and Miller, 1989), which performs alignments between two sequences, and the more widely used CLUSTALW (Eddy, 1995), which can perform multiple alignments between two or more sequences.

II-4-6 Motif searches

Prediction of functional protein motifs can be performed with ProfileScan (see http://www.isrec.isb-sib.ch/software/PFSCAN_form.html), and with MOTIF (see http://motif.genome.ad.jp), which compare an amino acid query sequence with the PROSITE and the Pfam protein profile databases. Protein targetting sites can be predicted by PSortII (Nakai and Horton, 1999), a program which evaluates different parameters including signal sequence and protein secondary structure based on the amino acid sequence and estimates the subcellular localization of the protein.

Chapter III

Results and Discussion

In this dissertation research, nine mouse BAC clones have been sequenced to completion, generating 1,928,332 base pairs of DNA sequences. Four BAC clones (mgs1-68e20, mgs1-166n14, rp22-225p5 and mgs1-423c02) which map to chromosome 1q21-23 were provided by Dr. P. Lengyel at Yale University. Three BAC clones (rp23-21118, rp23-9p22 and rp23-395h6) which map to chromosome 1q were provided by Dr. W. Gu at the JLP VA Medical Center, Loma Linda, CA. Two BAC clones (rp23-145f9 and rp23-77a8) which also map to chromosome 1q were provided by Dr. D. Chambliss at the University of Texas Southwestern Medical Center at Dallas, TX. The chromosomal locations of BAC clones mgs1-68e20, mgs1-166n14, rp23-21118, rp23-9p22, rp23-395h6, rp23-145f9 and rp23-77a8 are shown on figure 3.1.

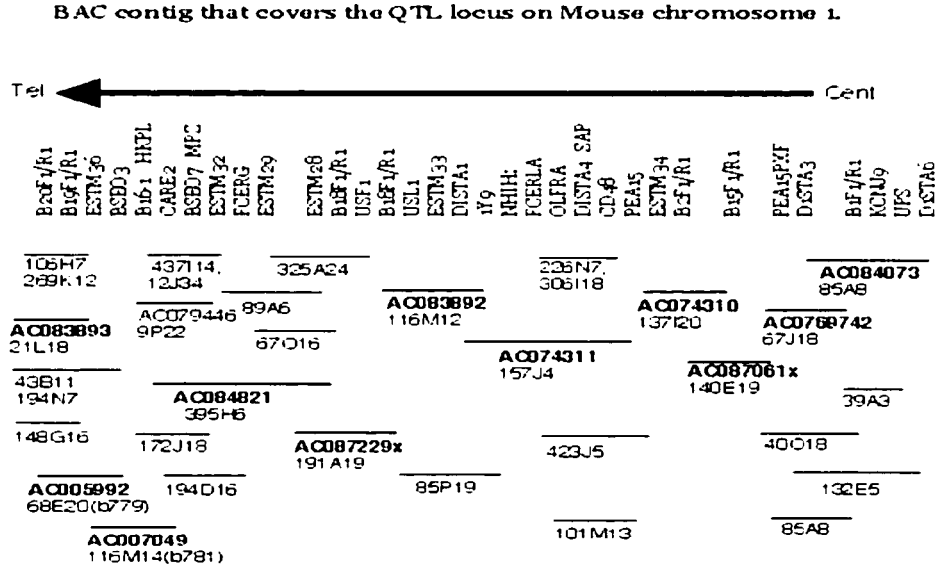


Fig. 3.1: chromosomal locations of the mouse BAC clones, as provided by Dr. W. Gu. The clones marked with a thick line are being sequenced at the Advanced Center for Genome Technology. The BAC clones rp23-145f9 and rp23-77a8 map between the rp23-157j4 and rp23-137i20 clones. The BAC clones rp22-225p5 and mgs1-423c02 have been mapped to mouse chromosome 1q21-23 and likely are located in a neighboring region.

During this work, all contigs larger than 2 kb were progressively deposited in the high-throughput genome sequencing (HTGS) division of Genbank with no restriction on public access and the sequences for each BAC clone were given the Genbank accession numbers AC005992 for mouse BAC mgs1-68e20, AC007049 for mouse BAC mgs1-166n14, AC006944 for mouse BAC rp22-225p5, AC008100 for mouse BAC mgs1-423c02, AC083893 for mouse BAC rp23-21l18, AC079446 for mouse BAC rp23-9p22, AC084821 for mouse BAC rp23-395h6, AC091521 for mouse BAC rp23-145f9 and AC091523 for mouse BAC rp23-77a8, respectively.

III-1 Murine Ifi200 family of genes

Mouse genomic sequences corresponding to the Ifi200 family gene cluster located on mouse chromosome 1q21-23 was obtained through random shotgun sequencing of four mouse BAC clones (table 3.1) which formed two contigs (contig 1 and contig 2) totaling 390 372 bp that are separated by a gap of unknown size (see Fig. 3.2). Sequencing the numerous repetitive regions on mouse BAC mgs1-423c02 was accomplished by introducing a series of consensus reads generated by assembling in a separate assembly "walking" reads from the repeat spanning subclones into the Phrap assembly (see Materials and Methods).

BAC	Size (bps)	Position	GC%
mgs1-166n14 (b781)	85931	1q21-23	37.62
mgs1-68e20 (b779)	154284	1q21-23	38.03
rp22-225p5 (b225)	120856	1q21-23	38.00
mgs1-423c02 (b637)	220020	1q21-23	38.24

Table 3.1: The four mouse BAC clones (official name and given name), their positions, GC% and sizes, used in this study.

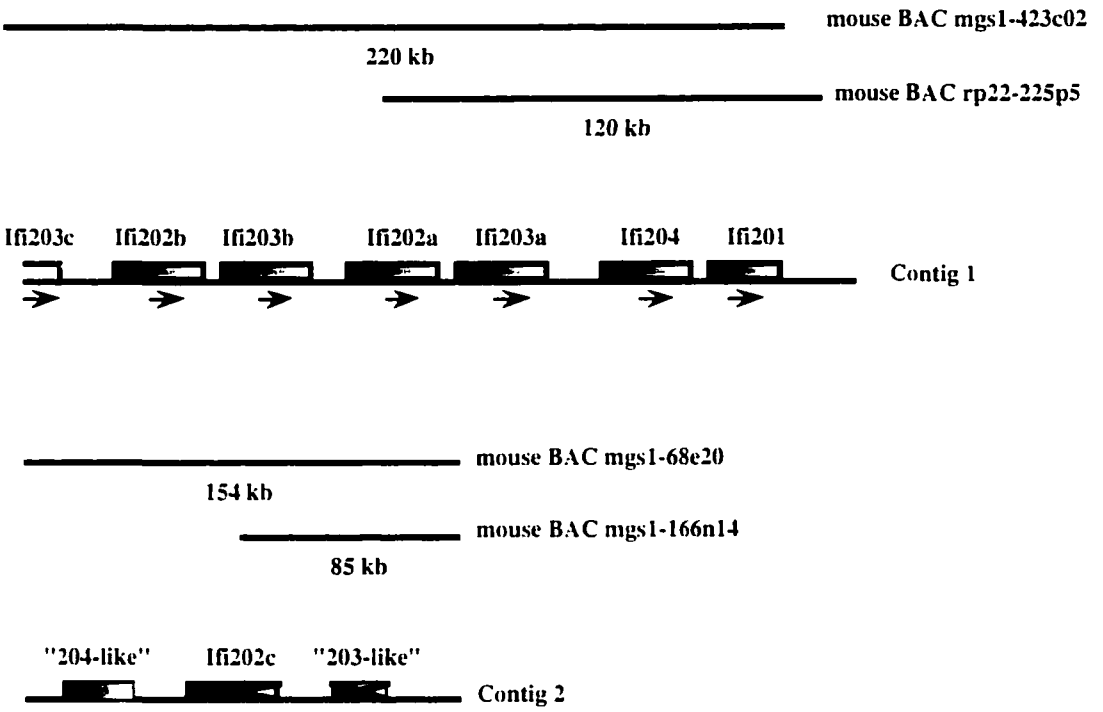


Fig. 3.2: Location of the genes of the mouse Ifi200 cluster on the two contigs sequenced during this dissertation research. ? indicates that the expression of "204-like" remains to be established.

Initial analysis of the mouse sequence for the above four BAC clones with BLAST and GENSCAN revealed the presence of seven genes from the mouse 200 family gene cluster, which all are in the same orientation in the single contiguous sequence obtained by merging BAC clones mgs1-423c02 and rp22-225p5 (contig 1) (Fig. 3.2). Two of the seven genes showed a strong sequence conservation with Ifi203, and therefore have been named Ifi203a and Ifi203c, whereas the original Ifi203 gene was renamed Ifi203b. Although all the exons of Ifi203a and Ifi203b were present in this sequenced region, only the last three exons of Ifi203c were present in BAC mgs1-423c02. Ifi205, the gene encoding the interferon-activatable protein D3 (Tannenbaum *et al.*, 1993) was not observed in any of the sequenced BACs and therefore its exact genomic location remains unresolved. The exon-intron organization of the sequenced mouse Ifi202b, Ifi202a, Ifi203b and Ifi204 genes were deduced by comparison to their published cDNA sequences. The exon-intron organization of Ifi203a and Ifi203c genes

were determined based on their sequence identity to Ifi203b cDNA sequence, because their cDNA sequences have not been determined. The exon-intron organization of Ifi201 gene was determined by its BLAST homology to other Ifi genes after the protein coding region was predicted by GENSCAN. The 5' flanking region of Ifi201 followed by the beginning of the first exon (GenBank accession number M31421) appeared to be located ~5 kb upstream of the first coding exon predicted by GENSCAN. The pseudogene Ifi202c (Wang *et al.*, 1999) was present in the sequence obtained by merging BAC clones mgs1-68e20 and mgs1-166n14 (contig 2) (Fig.3.2). This region also encodes a series of other previously unknown putative coding regions. After BLAST analysis revealed that they are directly related to the mouse 200 family gene cluster, these additional putative coding regions appear to form, along with Ifi202c, a cluster of 200 family-related genes and pseudogenes.

III-1-1 Ifi202a gene

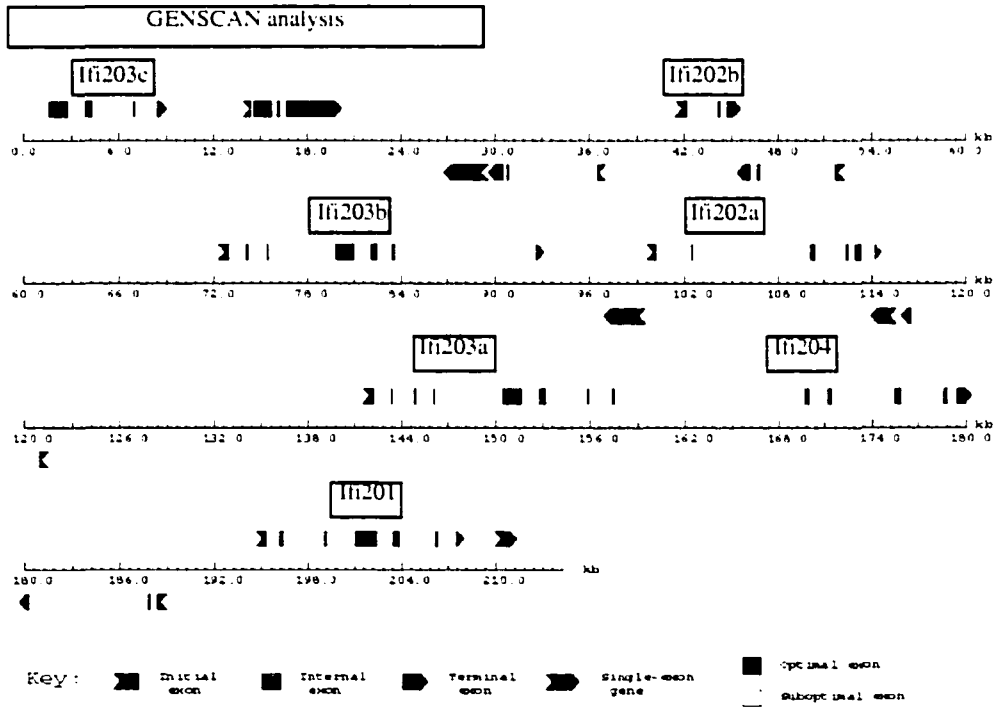


Fig. 3.3: GENSCAN analysis of contig 1

As shown on Fig. 3.3, a GENSCAN analysis of contig 1 revealed that a putative coding region was present in a segment located between base pairs 98646 and 125808. A BLAST search of the GenBank non-redundant nucleic acid database indicated that this segment of contig 1 shared sequence homology with the two GenBank entries corresponding to Ifi202a and Ifi202b cDNAs. An ALIGN comparison made between this segment of contig 1 and the Ifi202a and Ifi202b cDNAs revealed over 99% sequence similarity between the exonic regions of this putative gene and the Ifi202a cDNA. Therefore, this genomic region corresponds to the eight exons of the Ifi202a gene which spans 27 kb (table 3.2 and Fig. 3.4).

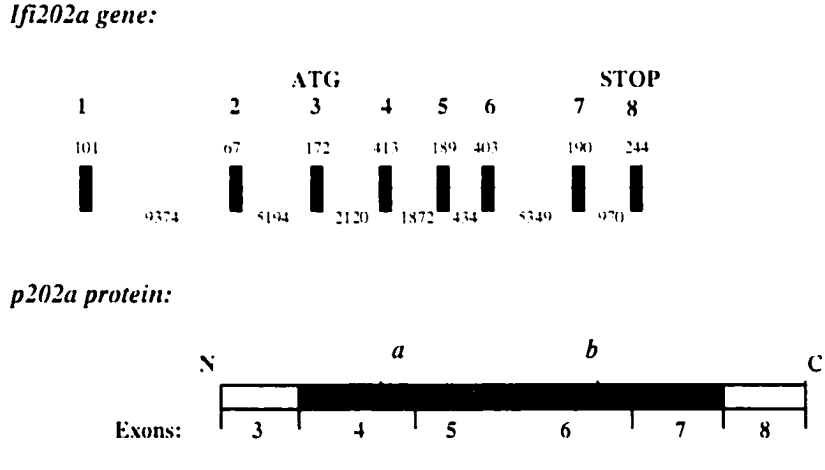


Fig. 3.4: Structural organization of the Ifi202a gene and its encoded protein, p202a. *(Top)*. Size in base pairs are given below the introns and above the exons. The exons are indicated by black boxes, and the connecting introns by horizontal lines. (Bottom) *a* and *b* represents the two 200-amino acid domain conserved among the Ifi200 family of proteins

The Neural Network Splice Site Prediction program (NNSSP) confirmed the presence of all potential splicing sites at the exon-intron borders of the Ifi202a gene that were predicted by a CROSSMATCH comparison between this genomic DNA segment and the Ifi202a cDNA (table 3.3). Although the exon-intron organization of Ifi202a has been previously determined (Wang *et al.*, 1999), the length of intron 1 was unknown

and these data now suggest that intron 1 is 9.374 bp long. Interestingly, exon 3 of the Ifi202a gene is the first coding exon, and exons 1 and 2 correspond to the 5' untranslated region of the mRNA.

Exon number	Size (bp)	Contig position (bp)
1	101	98716-98817
2	67	108191-108258
3	172	113452-113624
4	413	115744-116157
5	189	118029-118218
6	403	118652-119055
7	190	124404-124594
8	244	125564-125808

Table 3.2: Exon sizes and contig positions of the Ifi202a gene.

Donor Site Sequences		Acceptor Site Sequences	
Exon/Intron	Sequence	Intron/Exon	Sequence
1/1	98819 caggaagGTgagtt	1/2	108190 catttatcttccttctcaaAGctgacacactctgccttgt
2/2	108260 cggcttgGTgagtc	2/3	113451 ttttcttttcttttcaaacAGaagaactcaatcaaaaagt
3/3	113626 tggacctGTaagtc	3/4	115743 ccctgtatttttttcaatgcAGaacaaaaagtcacacac
4/4	116159 agagaagGTaatgt	4/5	118028 atgttaattttacttttgcAGaaaaaagtgaatgataaa
5/5	118220 catcaagGTaggaa	5/6	118651 ttttgtatatattctttgtAGggagaaaagctactaaaa
6/6	119057 atacaggGTaagct	6/7	124403 gaatgttaatttctttgcAGaaaagttccggaaataat
7/7	124596 catggagGTatgtg	7/8	125563 ttcatcattttcttatcctAGgtcatcatgcctgaaaaat

Table 3.3: Neural Network splice site predicted sequences: exons and introns are numbered according to their order in the gene and the numbers on the top of the donor and acceptor sites correspond to the contig positions. The numbered bases are underlined.

There are two contiguous 200 amino acid domains (domains *a* and *b*; see Introduction) with unknown function in the protein encoded by Ifi202a and these two domains are encoded by exons 4 and 5 and exons 6 and 7 respectively. Initial analysis of the region 1000 bp immediately upstream of the first exon and the first 200 bp of the first exon of Ifi202a with MOTIF using the TFMATRIX matrix table (see Table 3.4) showed that the 5' flanking region of Ifi202a contained two shorts interferon stimulated response elements (GA Box and Friedman-Stark Sequence). A Friedman-Stark Sequence is located within exon 1 of the Ifi202a gene, 94-122 bp downstream from the 3' end of exon 1, which is downstream of sequence of the 5' termini of the mRNAs. A GA Box is located immediately upstream from the 5' end of exon 1 of the Ifi202a gene. The 5' flanking region of Ifi202a also exhibit several potential transcription factor binding sites for transcription factors that have been shown previously to be expressed in, or to be important for, the development of cells of the hematopoietic lineages. These regions include the CCAAT/enhancer-binding protein b(C/EBPb), c-Myb, c-Ets, GATA1, GATA2, GATA3, MZF1, NF-AT, NF-E2, Ik-1 and Ik-2 transcription factor binding sites (for review, see Kerhl, 1995). Interestingly, the 5' flanking region of the Ifi202a gene also exhibits potential binding sites for transcription factors whose regulatory activity is inhibited by the binding of the protein encoded by Ifi202a (p202a), that include c-Fos, c-Jun, c-myc, E2F and MyoD. This region also contains potential binding sites for transcription factors that mediate the response to a large number of cytokines and growth factors, termed Signal Transducers and Activators of Transcription (STAT) transcription factors. In addition, two potential binding sites for p53 (which recently have been shown to modulate the transcriptional activity of p202a; see D'Souza *et al.*, 2001) were found in the 5' flanking region of Ifi202a, 357-379 bp

and 683-705 bp upstream from the 5' end of exon 1. As stated above, the putative translation start site (ATG) present inside the third exon of the Ifi202a gene is located at contig position 113505.

Gene	ISRE	Potential binding sites	Potential binding sites for transcription Factors bound by 200 family of proteins
Ifi202a/b	GA box and F.S. sequence	c-Myb, c-Ets, GATA, MZF1, NF-AT, NF-E2, Ik-1, Ik-2, STAT	c-Fos, c-Jun, c-myc, E2F-1, MyoD, p53
Ifi203a/b	GA box	c-Myb, c-Ets, GATA, MZF1, Ik-1, Ik-2	c-Fos, c-Jun, MyoD, NF-kB
Ifi204	GA box and F.S. sequence	c-Myb, c-Ets, NF-AT, GATA, Ik-1, Ik-2, SOX5, STAT	c-Fos, c-Jun, c-myc, MyoD, p53
Ifi201	F.S. sequence	MZF1, NF-AT, Lyf-1, Ik-1, Ik-2, GATA, c-Myb, STAT	c-Fos, c-Jun, NF-kB, MyoD
"204-like"	GA box	c-Ets, c-Myb, GATA, Ik-2, STAT	c-Fos, c-Jun, MyoD, E2F

Table 3.4: Potential binding sites for transcription factors, determined using MOTIF, in the first 1000 bp located immediately upstream from the genes of the mouse 200 family cluster. (ISRE) Interferon-Stimulated Response Element. (F.S.) Friedman-Stark.

The promoter prediction by NNPP did not predict any TATA box-containing promoter region within one kb upstream of the first exon. This observation is not surprising since the MNDA protein (see Introduction) also has been shown to lack a TATA box. However, a MOTIF search on the 5' flanking region of Ifi202a revealed the presence of potential TATA box elements located 769-783 bp and 290-304 bp, respectively, upstream from the 5' end of the Ifi202a gene. Although a profile database

(PROSITE) search did not indicate the presence of any particular motif on the protein encoded by Ifi202a (p202a), the p202a protein has been shown to carry several short sequences rich in basic amino acids in its amino terminus that serve, in other proteins, as nuclear targetting signals. The PSortII program predicts that the p202a protein would be localized in the nucleus, confirming the observation that p202a is localized in the nucleus after a prolonged exposure of the cells to interferons. The lack of PROSITE-predicted DNA-binding motifs and the observation that p202a is not released from isolated nuclei by DNAse I digestion suggest that p202a interacts with proteins, rather than DNA, in the nucleus.

III-1-2 Ifi202b gene

The Ifi202b gene was isolated by screening a mouse 129sv genomic DNA library with a Ifi202a cDNA probe. Besides the BAC clones positive for Ifi202a, four other BAC clones, whose sequences were determined to differ from Ifi202a cDNA by several base pairs were obtained. Because the 129sv mouse strain is an inbred strain that should be homozygous at all genetic loci, these two groups of BAC clones clearly contain two different genes, Ifi202b and Ifi202c (Wang *et al.*, 1999). The sequence of these genomic BAC clones confirmed the earlier observation that the Ifi202b cDNA sequence differs from that of Ifi202a by one dinucleotide substitution and two single nucleotide substitutions in exon 3, one single nucleotide substitution in exons 4 and 5, two single nucleotide substitutions in exon 6 and one single nucleotide substitution in exon 8 (Wang *et al.*, 1999). The protein encoded by the Ifi202b gene (p202b) would differ from that encoded by the Ifi202a gene (p202a) by only 7 out of 445 amino acids (Fig. 3.5). This is consistent with the observation that the structural homology between p202a and p202b is so high that mice knocked out for the Ifi202a gene have a wild type

phenotype. A mechanism involving dosage compensation at the posttranscriptional level by Ifi202b in the knockout mice has been suggested (Wang *et al.*, 1999).

```
                10        20        30        40        50        60
p202a.  MSNRNLRSSTNSEFSEGQHQTPSSDSSGHGEDQPQASPGPNKKSHTPKKNISKGAVLHEK
        ::::::::::::::..:::::::::::::::::::::::::::::::::::::::::::
p202b   MSNRNLRSSTNSDLAEGQHQTPSSDSSGHGEDQPQASPGPNKKSHTPKKNISKGAVLHEK
                10        20        30        40        50        60

                70        80        90       100       110       120
p202a.  PMTVMVLTATEPFNYKEGKENMFHATVATESQYYRVKVFNMDLKEKFTENKFITISKYFN
        ::::::::::::::::::::::::::::::::::.:::::::::::::::::::::::::
p202b   PMTVMVLTATEPFNYKEGKENMFHATVATESKYYRVKVFNMDLKEKFTENKFITISKYFN
                70        80        90       100       110       120

               130       140       150       160       170       180
p202a.  SSGILEINETATVSEAAPNQMFEVPKNIIRSAKETLKISKIKELDSGTLIYGVFAVEKKK
        ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
p202b   SSGILEINETATVSEAAPNQMFEVPKNIIRSAKETLKISKIKELDSGTLIYGVFAVEKKK
               130       140       150       160       170       180

               190       200       210       220       230       240
p202a.  VNDKSITFKIKDNEDNIKVVWDKEQHNINYEKGDKLQLFSFHLRKGNGKPILHSGNHSFI
        ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::.
p202b   VNDKSITFKIKDNEDNIKVVWDKEQHNINYEKGDKLQLFSFHLRKGNGKPILHSGNHSFV
               190       200       210       220       230       240

               250       260       270       280       290       300
p202a.  KGEKLLKESFEGDGYHKGPKQVVALKATKLFTYDSIKSKKMFHATVATDTEFFRVMVFEE
        ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
p202b   KGEKLLKESFEGDGYHKGPKQVVALKATKLFTYDSIKSKKMFHATVATDTEFFRVMVFEE
               250       260       270       280       290       300

               310       320       330       340       350       360
p202a.  NLEKKFIPGNTIALSDYFGMYGSLAIHEYSSVSEVKSQNKEDSSSSDERPIEHLKICDLH
        ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
p202b   NLEKKFIPGNTIALSDYFGMYGSLAIHEYSSVSEVKSQNKEDSSSSDERPIEHLKICDLH
               310       320       330       340       350       360

               370       380       390       400       410       420
p202a.  LQTEERLFDGEFKVYRKSSGNNCICYGIWDDTGAMKVVVSGQLTSVNCEIGNTIRLVCFE
        :::.::: ::::::::::::::::::::::::::::::::::::::::::::::::::::
p202b   LQTKERLVDGEFKVYRKSSGNNCICYGIWDDTGAMKVVVSGQLTSVNCEIGNTIRLVCFE
                          380       390       400       410       420

               430       440
p202a.  LTSNADEWFLRATRYSYMEVIMPEK
        :::::::::::::::::::::::::
p202b   LTSNADEWFLRATRYSYMEVIMPEK
               430       440
```

Fig. 3.5: Alignment of the amino acid sequence of p202a and p202b. The substituted amino acids are in bold.

A GENSCAN analysis (Fig. 3.3) of contig 1 reveals that a putative coding region was present in a segment located between base pairs 24697 and 51811. A BLAST search of the Genbank non-redundant nucleic acid database followed by ALIGN comparison between this segment of contig 1 and the IFI202b cDNA revealed that this segment corresponded to the Ifi202b gene. Although the sequence of exon 1 is not present in the Ifi202b cDNA sequence, its position was determined by a CROSSMATCH comparison between this genomic segment and the Ifi202a cDNA, based on the assumption that there is a strong homology between the Ifi202a and the Ifi202a sequences. This analysis revealed that the Ifi202b gene, similarly to the Ifi202a gene, contains eight exons spanning 27 kb of genomic DNA (table 3.5; see also Fig. 3.4) and has the exon-intron borders confirmed by the neural network splice site prediction program as shown in table 3.6.

Exon number	Size (bp)	Contig position (bp)
1	101	24762-24863
2	67	34236-34303
3	172	39485-39657
4	413	41777-42190
5	189	44062-44253
6	403	44687-45090
7	190	50447-50637
8	244	51607-51851

Table 3.5: Exon sizes and contig positions of the Ifi202b gene.

Donor site sequences		Acceptor site sequences	
Exon/Intron	Sequence	Intron/Exon	Sequence
	24865		34235
1/1	caggaagGTgagttc	1/2	catttatcttctactcaaAGctgacacactctgccttg
	34305		39484
2/2	cggcttgGTgagtcc	2/3	ttttcttttcttttaaacAGcagaactcaatcaaaaagt
	39659		41776
3/3	tggacctGTaagtca	3/4	ccctgtatttttcaatgcAGaacaaaaagtcacacac
	42192		44061
4/4	agagaagGTaatgtg	4/5	atgttaattttacttttgcAGaaaaaagtgaatgataaa
	44255		44686
5/5	cgtcaagGTaggaag	5/6	ttttgtatatattctttgtAGggagaaaagctactaaaa
	45092		50446
6/6	atacaggGTaagcta	6/7	gaatgttaatttcttttgcAGaaaagttccggaaataat
	50639		51606
7/7	catggagGTatgtgc	7/8	ttcatcattttcttatcctAGgtcatcatgcctgaaaaat

Table 3.6: Neural Network splice site predicted sequences: exons and introns are numbered according to their positions and the numbers on the top of the donor and acceptor sites correspond to the contig positions. The numbered bases are underlined.

As with the p202a protein, the Ifi202b gene product p202b also contains two contiguous 200 amino acid domains (domains *a* and *b*) that are encoded by exons 4 and 5 and exons 6 and 7 respectively (see Fig. 3.4). A search for potential transcription factor binding sites using MOTIF on a segment spanning 1000 bp immediately upstream from the 5' end of the first exon of Ifi202b followed by the first 200 bp of the first exon revealed results similar to those obtained for the 5' flanking region of Ifi202a (see Table 3.4), suggesting a similar mode of transcriptional regulation for both the Ifi202a and Ifi202b genes. A putative translation start site (ATG) is located at contig position 39538, within the third exon of the Ifi202b gene. Promoter prediction by NNPP did not reveal any TATA-containing promoters within one kb upstream from the 5' end of the first exon. However, a MOTIF search of the 5' flanking region of Ifi202b indicated the presence of two potential TATA box elements at positions similar to the ones observed in the 5' flanking region of Ifi202a. A PROSITE database search did not

indicate the presence of any specific motifs on the p202b protein and the PSortII program predicted that p202b would be localized in the nucleus.

Because of the extensive sequence similarity between the Ifi202a and Ifi202b gene exonic regions (over 99%), the Ifi202a genomic region was compared to the Ifi202b genomic regions using DOTTER (Fig. 3.6).

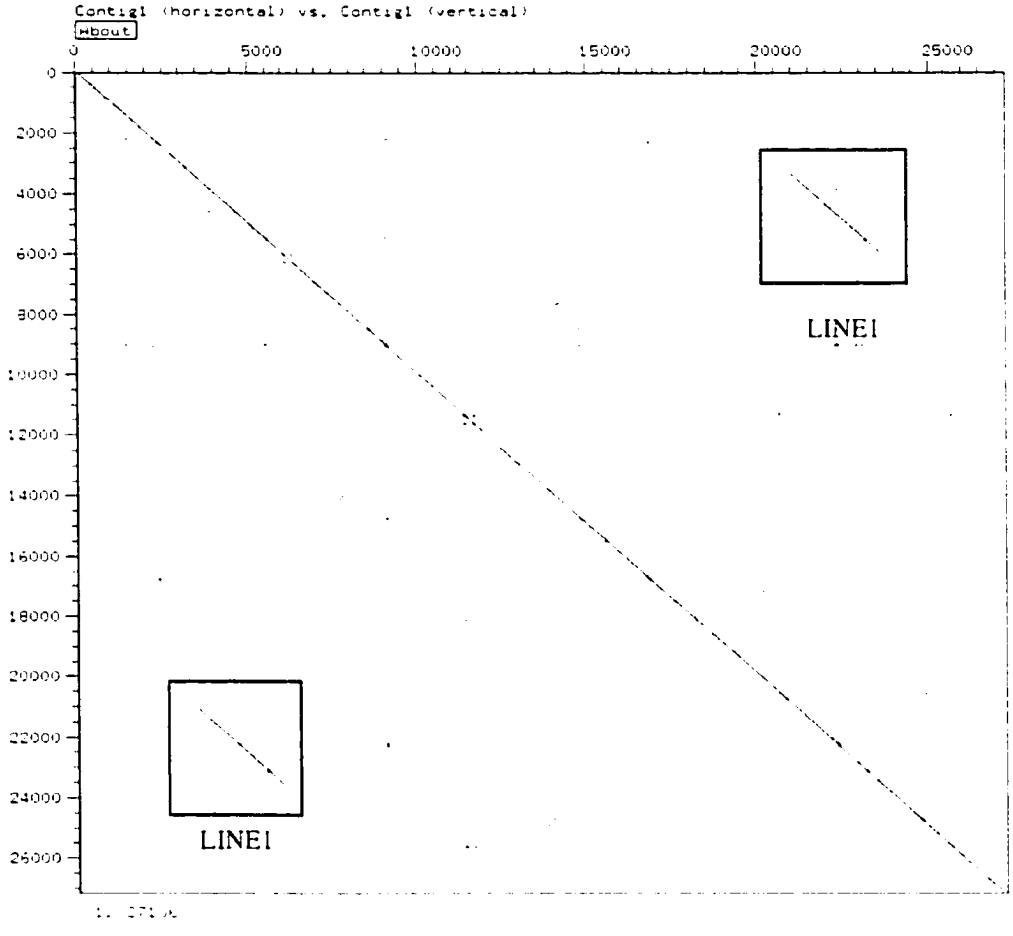


Fig. 3.6: Dot-plot analysis, using DOTTER, of the Ifi202a (horizontal) and Ifi202b (vertical) genomic regions. LINE1 are repeat elements disseminated in the mouse genome. They are indicated by a box in the dot-plot.

A further genomic DNA sequence comparison using CROSSMATCH revealed that the nucleotide sequence for the genomic regions located between the first exon (exon 1) and the last exon (exon 8) of the Ifi202a and Ifi202b genes were 99.65%

similar. Although the coding sequences are extremely well conserved between Ifi202a and Ifi202b, much of the intronic sequences flanking individual exonic sequences also are highly conserved (table 3.7).

	Ifi202a/Ifi202b
	% sequence identity
exon 1	100
intron 1	99.65
exon 2	100
intron 2	99.63
exon 3	98.26
intron 3	99.76
exon 4	99.76
intron 4	99.89
exon 5	99.48
intron 5	100
exon 6	99.51
intron 6	99.49
exon 7	100
intron 7	100
exon 8	99.52

Table 3.7: Comparison of percentage nucleotide identity of exons and introns of murine Ifi202a and Ifi202b genes using CROSSMATCH.

A dot plot analysis using DOTTER made between the first 2000 bps immediately upstream of the 5' ends of the first exons of the Ifi202a and Ifi202b genes also revealed extensive sequence similarity (Fig. 3.7) suggesting a similar mode of transcriptional regulation. Interestingly, in mice knocked out for p202a, the expression of Ifi202b mRNA persists at the same level as in wild type mice (Wang *et al.*, 1999), suggesting that the Ifi202b gene is transcriptionally active in wild type mice and is not subject to dosage compensation at the transcriptional level in the knock-out mice.

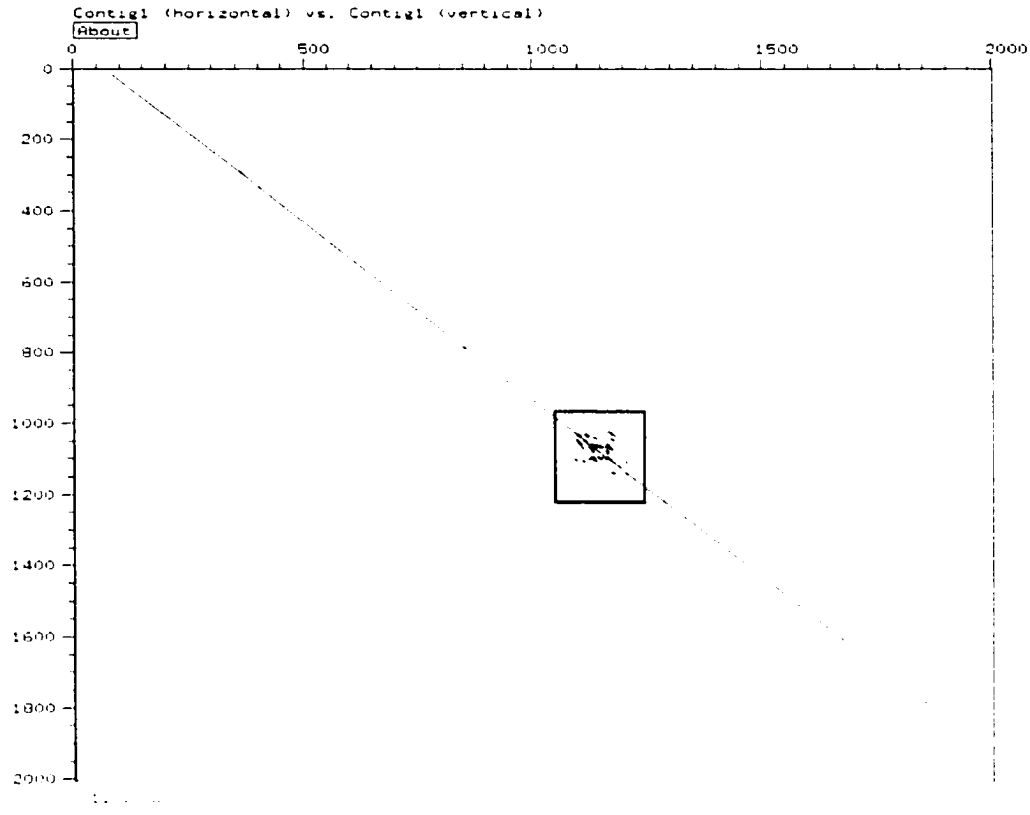


Fig. 3.7: Dot-plot analysis, using DOTTER, of the 5' flanking region of the Ifi202a (horizontal) and Ifi202b (vertical) genes. The first 2000 bp upstream from the initiation of transcription are analyzed. The boxed area corresponds to dinucleotide repeats.

III-1-3 Ifi203 genes

A GENSCAN analysis (Fig. 3.3) of contig 1 revealed the presence of three putative coding regions located at contig positions 3843-10558, 69310-91294, 143300-165288, respectively. A BLAST search of the Genbank non-redundant nucleic acid database revealed that these three regions share sequence similarity with the Ifi203 cDNA. An ALIGN comparison between each of the three genomic segments of contig 1 and the Ifi203 cDNA confirmed that the three genomic segments shared sequence similarity with the Ifi203 cDNA, further revealing over 99% sequence similarity between the exonic regions of the second segment (at contig positions 69310-91294) and the

Ifi203 cDNA, indicating that this region contains the Ifi203 gene. Analysis of the second genomic segment revealed seven single nucleotide substitution and two dinucleotide substitution, when compared to the Ifi203 cDNA, located at the 3' end of the last exon of the Ifi203 gene. The first genomic region (at contig position 3843-10558) shared over 90% sequence similarity only with the last three exons of the Ifi203 gene region (see below). The Ifi203 gene with highest homology to the cDNA has been renamed Ifi203b, the "203-like" gene located at contig position 143300-165288 has been named Ifi203a, and the "203-like" gene whose last three exons are located at contig positions 3843-10558 has been named Ifi203c, in the order of their contig positions.

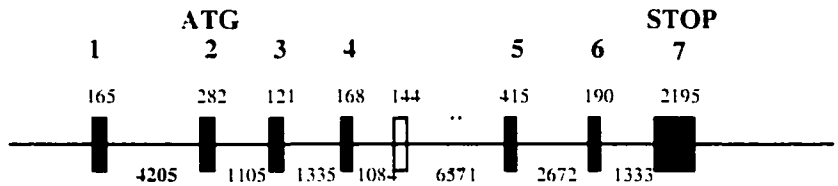
III-1-3-1 Ifi203b gene

The Ifi203b gene contains seven exons, spanning 22 kb of genomic DNA (table 3.8 and Fig. 3.8).

Exon number	Size (bp)	Contig position (bp)
1	165	69310-69475
2	282	73680-73962
3	121	75067-75188
4	168	76523-76691
5	415	84489-84904
6	190	87576-87766
7	2195	89099-91294

Table 3.8: Exon sizes and contig positions of the Ifi203b gene.

Ifi203b gene:



p203b protein:

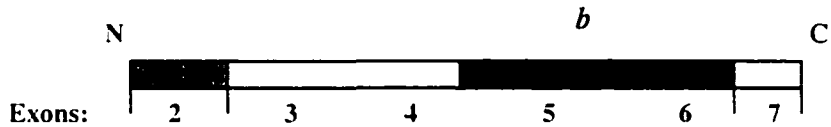


Fig. 3.8: Structural organization of the Ifi203b gene and its encoded protein, p203. (Top) Sizes in base pairs are given below the introns and above the exons. The exons are indicated by black boxes, and the connecting introns by horizontal lines. The white box indicates the presence of a potential 144-bp-long alternatively spliced exon. The shaded white box indicates the presence of a potential "fossilized" exon. (Bottom) ***b*** represents the b 200-amino acid domain conserved among the Ifi200 family of proteins. The gray box indicates the presence of a PAAD/DAPIN domain at the N-terminus of p203b.

Further analysis of the Ifi203b gene region by the neural network splice site prediction program confirmed the presence of all the potential splicing sites that were observed by a CROSSMATCH comparison between this genomic segment and the Ifi203 cDNA (table 3.9).

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
1/1	69477 cacggagGTaagtct	1/2	73679 tgtgtctttaatttttttAGgttaatctataatttgagag
2/2	73964 gcaaaagGTaataag	2/3	75066 gtgttttctttaccectacAGttcaagagaaaaagaaagga
3/3	75190 ttcaaagGTaaagtg	3/4	76522 aaagacaatgttgcttttcAGaaaaagagaaaacagatcac
4/4	76693 aaacaagGTaacact	4/5	84488 tctcatgtgtgttttatgtAGgcaccaaggagaggaactgt

	84906		87575
5/5	atttaagGTaagcca	5/6	aatgttaatttgctttttcAGaaaagtgagaggcatgagtg
	87768		89098
6/6	catgcagGTatgtgc	6/7	ttcattgttttcttctttaAGgtcatcaatgaaggaaagcc

Table 3.9: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

In earlier studies of the Ifi203b gene (Gribaudo *et al.*, 1997), it was reported that the p203b protein was hydrophilic and contained 408 amino acids. Since the p203b protein contains only one copy of the 200 amino acid domains (domain *b*) encoded by exons 5 and 6, it has a unique molecular structure among the Ifi200 family proteins (see Fig. 3.8). The region including the first 1000 bp located immediately upstream from the first exon followed by the first 200 bp of the first exon of the Ifi203b gene was analyzed for the presence of potential transcription factor binding sites using MOTIF (Table 3.4). Interestingly, the 5' flanking region of the Ifi203b gene does not contain any Friedman-Stark sequences which typically are transcribed in the 5' flanking region of many interferon-activatable genes, among which are Ifi202a and Ifi202b. The 5' flanking region of the Ifi203b gene does contain a GA box that is located only a few bp upstream from the 5' end of the first exon of the Ifi203b gene. Besides the presence of many potential binding sites for transcription factors usually expressed in cells from the hematopoietic lineage (for review, see Kehrl, 1995), the 5' flanking region of the Ifi203b gene also exhibits potential c-Fos, c-Jun, MyoD and NF-kB binding sites whose regulatory activity is inhibited by the protein encoded by Ifi202a (p202a). However, the potential binding sites for p53 found in the 5' flanking region of the Ifi202a and b genes are absent in the 5' flanking region of the Ifi203b gene. Taken together, these data suggest a differential mode of transcriptional regulation for the Ifi203b gene. Since the putative translation start site (ATG) is located at contig position 73701, inside the second exon of the Ifi203b gene, it appears that the first exon of the Ifi203b gene is

non-coding. Gribaudo *et al.* (1997) demonstrated previously that a stop codon (TGA) was located at position 1412 in the Ifi203 cDNA and that a large portion of the large last exon (exon 7) of the Ifi203b gene was non-coding, therefore corresponding to a 3' untranslated region. The promoter prediction program NNPP did predict a potential TATA-containing promoter located 881-931 base pairs upstream from the 5' end of the first exon that is located upstream from the GA box which confers the ability of the Ifi203b gene to be induced by interferons. This promoter region was confirmed by a MOTIF search made on the 5' flanking region of Ifi203b. However, further experiments, including reporter genes, are needed to determine whether this putative promoter play any role in the initiation of transcription of the Ifi203b gene *in vivo*. Consistent with the results of an immunoblot analysis of subcellular fractions of interferon-treated Balb/c 3T3 cells which indicated that the p203b protein is present in the nucleus (Gribaudo *et al.*, 1997), a PROSITE search revealed the location of two potential nuclear localization motifs. A Pfam protein profile database search revealed the existence of a PAAD/DAPIN protein domain at the amino terminus of p203b, which is entirely encoded by exon 2 of the Ifi203b gene (Fig. 3.8). The PAAD/DAPIN [*p*yrin, AIM (*a*bsent in melanoma), ASC (*a*poptosis-associated speck-like protein containing a caspase recruitment domain (CARD), and *d*eath-*d*omain (DD)-like/*d*omain in *ap*optosis and *in*terferon response] (Pawlowski *et al.*, 2001; Staub *et al.*, 2001) are unique to vertebrates or vertebrate-specific viruses and are present at the amino terminus of several different proteins involved in apoptosis and interferon response. One of these proteins is the ASC protein (Masumoto *et al.*, 1999). The ASC protein sequence contains a carboxy terminal caspase recruitment domain (CARD) that binds to other CARDs during the apoptotic response, suggesting that ASC is directly involved in apoptosis. The PAAD/DAPIN domain also is present at the amino terminus of the human 200-family of proteins. Based on secondary structure prediction and sequence alignment, Martinon *et al.* (2001) suggested that the PAAD/DAPIN domain is an additional

member of the death-domain-fold superfamily, similar in structure to the DED (*d*eath-*e*ffector *d*omain) and the DD (*d*eath *d*omain) domains and involved in some protein-protein interactions between apoptotic effectors during the apoptotic response. Taken together, these data suggests that p203b may be involved in apoptosis, via some protein-protein interactions. In addition, PSortII predicted that p203b would be localized in the nucleus. These data suggest that, in interferon-treated cells, p203b translocates to the nucleus. However, because no DNA-binding motif has been detected in p203b, it appears that p203b does not bind DNA but rather may be involved in protein-protein interactions in the nucleus. Alternative polyadenylation sites have been reported in the 3' untranslated region of the Ifi203b gene (Gribaudo *et al.*, 1997) which does not share homology with the 3' UTR of the Ifi202a gene. Here, six potential poly(A) sites were identified within the 3' UTR of the Ifi203b gene (table 3.10) and northern-blot analysis, combined with sequencing analysis, revealed that two of these sites, corresponding to the second and the sixth poly(A) sites, predominantly are used in BALB/c 3T3 cells (Gribaudo *et al.*, 1997).

poly(A) sites	Ifi203 cDNA position (bp)
1	1466-1471
2	1636-1641
3	1710-1715
4	2045-2050
5	3361-3366
6	3513-3518

Table 3.10: the six poly(A) sites located at the 3' end of the Ifi203 cDNA and their sequence position.

The 5' end of the Ifi203b 3.8 kb cDNA hybridizes to a 3.8 kb mRNA and a less abundant 1.8 kb mRNA in northern-blot analysis, suggesting two different transcription start sites or different mRNA processing (splicing and polyadenylation site selection) (Gribaudo *et al.*, 1997). Interestingly, a BLAST search of the Ifi203b gene sequence revealed the presence of a short sequence located inside intron 4 that shares 95%

sequence similarity with the sequence of exon 4 (Fig. 3.8). Whether this sequence corresponds to a "fossilized" exon or an important functional sequence possibly subject to alternative splicing has not been established. Potential splice sites GT/AG consensus sequences are present at both sides of this putative exon. However, this sequence is not present in any Ifi203 mRNAs present in the Genbank database. In addition to this putative "fossilized" exon, a sequence comparison between the Ifi203 cDNA and a cDNA corresponding to the 3' end of the Ifi203 cDNA (Genbank accession number L14559) reveals the presence of an additional 144-bp long sequence in the shorter cDNA. This additional sequence, absent from the Ifi203 cDNA sequence, also is located inside intron 4 of Ifi203b, at position 77775-77918 on the BAC clone mgs1-423c02 and between consensus GT/AG exon-intron border sequences (Fig. 3.8). It encodes a putative 48-amino acid polypeptide that is absent from the amino acid sequence of the protein encoded by Ifi203b. Although the origin of this DNA segment in the cDNA corresponding to the 3' end of the Ifi203 cDNA remains to be established, it is likely that this segment corresponds to an additional exon and that the cDNA corresponds to the 3' end of a splice variant of the Ifi203 gene.

III-1-3-2 Ifi203a gene

The Ifi203a gene exhibits strong sequence similarity with the Ifi203 cDNA. An ALIGN comparison made between the Ifi203a gene (at contig position 143300-165288) and the Ifi203b gene sequence reveals that the exonic sequences of the Ifi203a gene differ from those of Ifi203b by a single nucleotide substitution in exon 3, several nucleotides differences in the 3' untranslated region (3' UTR), and in exon 7, where in Ifi203a, it is shorter than exon 7 of Ifi203b. Thus, the proteins encoded by Ifi203a and Ifi203b differ by only one out of 408 amino acid (Fig. 3.9).

```
               10         20         30         40         50         60
p203a. MAEYKNIVLLKGLENMEDYQFRTVKSLLRKELKLTKKMQEDYDRIQLADWMEDKFPKDAG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
p203b  MAEYKNIVLLKGLENMEDYQFRTVKSLLRKELKLTKKMQEDYDRIQLADWMEDKFPKDAG
               10         20         30         40         50         60

               70         80         90        100        110        120
p203a. LDKLIKVCEHIKDLKDLAKKLKTEKAKVQEKKKGKCKMAGKKKGQDELSSSESLFINKES
       :::::::::::::::::::::::::::::::::::: :::::::::::::::::::::::
p203b  LDKLIKVCEHIKDLKDLAKKLKTEKAKVQEKKKGKCKTAGKKKGQDELSSSESLFINKES
               70         80         90        100        110        120

              130        140        150        160        170        180
p203a. YKSVPSSKKKRKQITKTEGGKKKKLTQEQAQLPETSGTNIKKEEDCLQNPHKSPPTPSSS
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
p203b  YKSVPSSKKKRKQITKTEGGKKKKLTQEQAQLPETSGTNIKKEEDCLQNPHKSPPTPSSS
              130        140        150        160        170        180

              190        200        210        220        230        240
p203a. SSNKAPRRGTVPKEPSREEGHHQGPKQVMVLKVTEPFTYDFEETKRMFHATVATETEFFR
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
p203b  SSNKAPRRGTVPKEPSREEGHHQGPKQVMVLKVTEPFTYDFEETKRMFHATVATETEFFR
              190        200        210        220        230        240

              250        260        270        280        290        300
p203a. VKVFDTALMSKFIPGKIIAISHYIGCNGFLEIYRASCVSDVNINPTMIISNTLSESAIAT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
p203b  VKVFDTALMSKFIPGKIIAISHYIGCNGFLEIYRASCVSDVNINPTMIISNTLSESAIAT
              250        260        270        280        290        300

              310        320        330        340        350        360
p203a. PKISYLLSQAKGTFVNGEFVVFKKSERHECICYGIGDDTGKMAVVVYGRLTNVRCEPGSK
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
p203b  PKISYLLSQAKGTFVNGEFVVFKKSERHECICYGIGDDTGKMAVVVYGRLTNVRCEPGSK
              310        320        330        340        350        360

              370        380        390        400
p203a. LRLVCFELTSTKDVCLLRSVRHSYMQVINEGKPLNPDSVRRNSLEPYF
       ::::::::::::::::::::::::::::::::::::::::::::::::
p203b  LRLVCFELTSTKDVCLLRSVRHSYMQVINEGKPLNPDSVRRNSLEPYF
              370        380        390        400
```

Fig. 3.9: Alignment of the amino acid sequences of p203a and p203b.The substituted amino acid is in bold.

Based on its strong sequence similarity with Ifi203b, the exon-intron structure of the Ifi203a gene was deducted from a CROSSMATCH comparison between this genomic segment and the Ifi203 cDNA, revealing that the Ifi203a gene, similarly to the

Ifi203b gene, contains seven exons, spanning 22 kb of genomic DNA (table 3.11 and see Fig. 3.8).

Exon number	Size (bp)	Contig position (bp)
1	165	143300-143465
2	282	147682-147964
3	121	149069-149190
4	168	150525-150693
5	415	158495-158910
6	190	161583-161773
7	2186	163102-165288

Table 3.11: Exon sizes and contig positions of the Ifi203a gene.

The presence of these splice sites was confirmed by the Neural Network splice site prediction program (table 3.12).

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
1/1	143467 cacggagGTaagtct	1/2	147681 tgtgtctttaatttttttAGgttaatctataatttgagag
2/2	147966 gcaaaagGTaataag	2/3	149068 gtgttttctttacccctacAGttcaagagaaaaagaaagga
3/3	149192 ttcaaagGTaaagtg	3/4	150524 aaagacaatgttgcttttcAGaaaaagagaaaacagatcac
4/4	150695 aaacaagGTaacact	4/5	158494 tctcatgtgtgttttatgtAGgcaccaaggagaggaactgt
5/5	158912 atttaagGTaagcca	5/6	161582 aatgttaatttgcttttcAGaaaagtgagaggcatgagtg
6/6	161775 catgcagGTatgtgc	6/7	163101 ttcattgttttcttctttaAGgtcatcaatgaaggaaagcc

Table 3.12: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A MOTIF search on the segment spanning 1000 bp upstream of the first exon followed by the first 200 bp of the first exon of the Ifi203a gene gave results similar to those obtained for the 5' flanking region of Ifi203b (see Table 3.4). Because of the high level of sequence similarity between the 5' flanking regions of the two genes, it is likely that the Ifi203a and Ifi203b genes have a similar mechanism for transcriptional regulation. A putative translation start site is located at contig position 147703, inside the Ifi203a gene second exon and five poly(A) sites were identified within the 3' UTR of the Ifi203a gene. Interestingly, one of the two poly(A) sites (the sixth poly(A) site; see table 3.10 and Gribaudo *et al.*, 1997) predominantly used in BALB/c 3T3 cells for the Ifi203b gene is absent in the 3' UTR of the Ifi203a gene. A search of the Pfam protein profile database revealed the existence of a PAAD/DAPIN domain at the amino terminus of p203a, similar to the one found at the amino terminus of p203b. In addition, a CROSSMATCH comparison between the Ifi203a gene sequence and the sequence of the putative 144-bp long putative alternatively spliced exon found in intron 4 of Ifi203b reveals a sequence similar to the 144-bp long putative exon in intron 4 of Ifi203a, at position 151777-151920. This suggests the existence of a splice variant form of the Ifi203a gene, containing an additional segment encoded by the putative exon located between exon 4 and 5 of the Ifi203a gene.

Because the sequence similarity among the Ifi203a and Ifi203b exonic regions is extensive (over 99%), the Ifi203a genomic region (exonic and intronic regions) was compared to the Ifi203b genomic region by dot plot analysis using DOTTER (Fig. 3.10). This result and a CROSSMATCH comparison reveal that the sequence for the genomic regions located between the first exon (exon 1) and the last exon (exon 7) of the Ifi203a and Ifi203b genes have over 99% similarity. Therefore, the Ifi203a gene shares strong sequence similarity with Ifi203b, as observed for Ifi202a compared with Ifi202b. Interestingly, the putative protein encoded by Ifi203a differs from p203 by only

one out of 408 amino acid, whereas p202a differs from p202b by seven out of 445 amino acids.

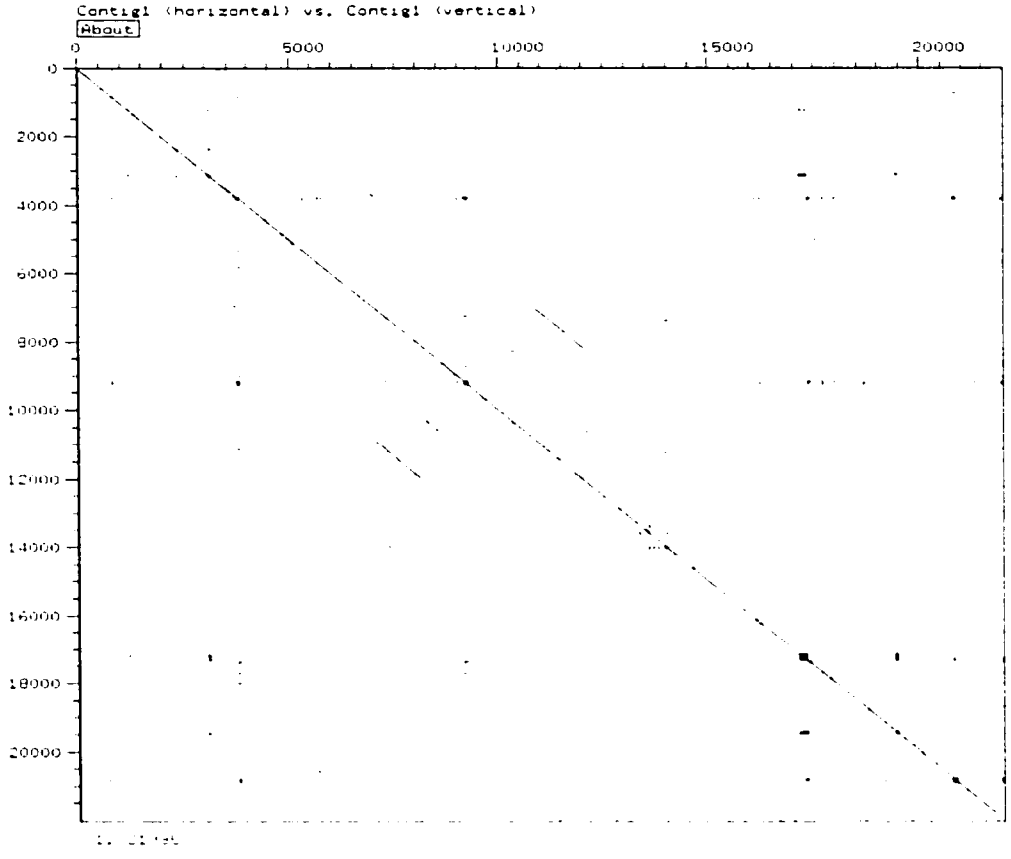


Fig. 3.10: Dot-plot analysis, using DOTTER, of the Ifi203a (horizontal) and Ifi203b (vertical) genomic regions.

As was the case for Ifi202a and b, the intronic sequences flanking individual exonic sequences are well conserved between the Ifi203a and Ifi203b genes (table 3.13).

	Ifi202a/Ifi202b	**Ifi203a/Ifi203b**
	% sequence identity	% sequence identity
exon 1	100	100
intron 1	99.65	99.41
exon 2	100	100
intron 2	99.63	99.63
exon 3	98.26	99.19

intron 3	99.76	99.48
exon 4	99.76	100
intron 4	99.89	99.62
exon 5	99.48	100
intron 5	100	99.81
exon 6	99.51	100
intron 6	99.49	99.4
exon 7	100	99.81
intron 7	100	N/A
exon 8	99.52	N/A

Table 3.13: Comparison of percentage nucleotide identity of exons and introns of murine Ifi202a and Ifi202b genes and murine Ifi203a and Ifi203b genes using CROSSMATCH.

A dot plot comparison of the first 1000 bp located immediately upstream of the 5' ends of the first exons of the Ifi203a and b genes reveals extensive sequence similarity, suggesting a similar mode of transcriptional regulation (Fig. 3.11), as was the case for the Ifi202a and Ifi202b genes.

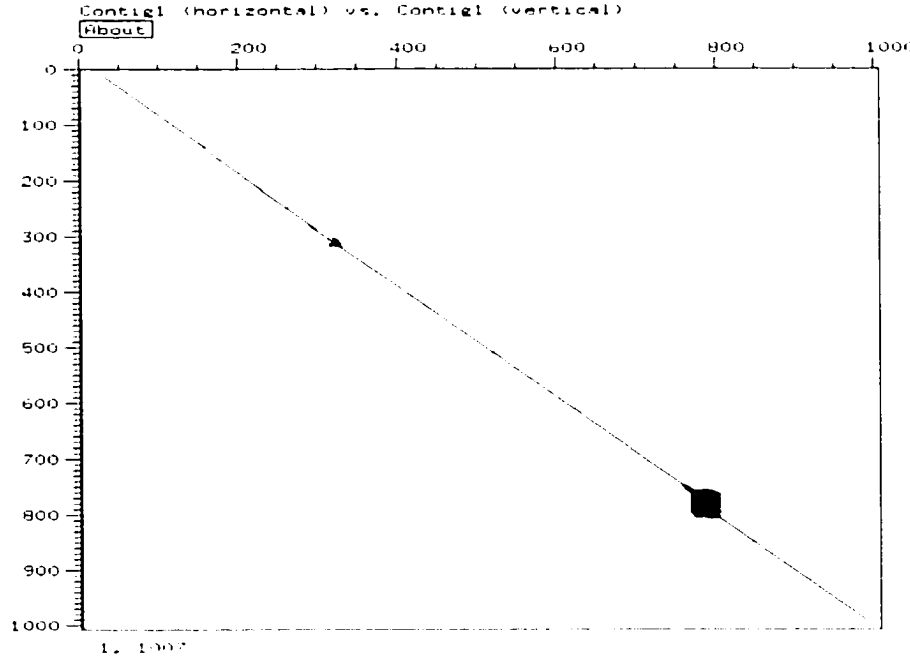


Fig. 3.11: Dot-plot analysis, using DOTTER, of the 5' flanking region of the Ifi203a (horizontal) and Ifi203b (vertical) genes. The first 1000 bp located upstream from the initiation of transcription are compared.

III-1-3-3 Ifi203c gene

An ALIGN comparison between the first genomic segment (at contig position 3843-10558) and the genomic region encompassing the last three exons of the Ifi203b gene, reveals that the coding regions corresponding to the last three exons of the Ifi203c gene (not including the 3' UTR located at the end of exon 7) differ from those of Ifi203b by only seven single nucleotide substitutions in exon 5. The exon-intron borders of the last three exons of the Ifi203c gene, as determined by a CROSSMATCH comparison between this genomic segment and the Ifi203 cDNA, revealed a strong sequence similarity between the Ifi203b and Ifi203c genomic sequences (table 3.14).

Exon number	Size (bp)	Contig position (bp)
5	415	3843-4258
6	190	6926-7116
7	2129	8429-10558

Table 3.14: Exon sizes and contig positions of the last three exons of the Ifi203c gene.

The results from the Neural Network splice site prediction program confirmed the presence of potential splice sites at the exon-intron borders corresponding to those delimited by a CROSSMATCH comparison between this genomic segment and the Ifi203 cDNA (table 3.15).

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
5/5	4260 atttaagGTaagcca	5/6	6925 aatgttaatttgctttttcAGaaaagtgagaggcatgagtg
6/6	7118 catgcagGTatgtgc	6/7	8428 ttcattgttttcttctttaAGgtcatcaatgaaggaaagcc

Table 3.15: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

Analysis of the 3' UTR of the Ifi203c gene also showed the presence of six polyadenylation sites at positions that were similar to the ones found in the 3' UTR of the Ifi203b gene (see table 3.10 and Gribaudo *et al.*, 1997). Whether the Ifi203c gene is transcriptionally active is still unknown because the 5' end of the Ifi203c gene remains to be sequenced.

III-1-4 Ifi204 gene

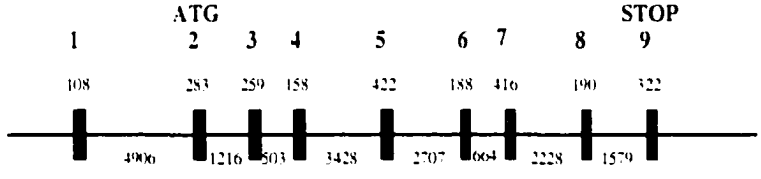
A GENSCAN analysis (Fig. 3.3) of contig 1 revealed the presence of a putative coding region at position 170378-189955 which had sequence similarity to the Ifi204 cDNA. An ALIGN comparison of this region and the Ifi204 cDNA confirmed the presence of the Ifi204 gene in contig 1. Only three nucleotides differed (one in exon 7 and two in exon 9) between the Ifi204 cDNA and the Ifi204 gene exonic sequences. A CROSSMATCH comparison between this genomic region of contig 1 and the Ifi204 cDNA revealed that the Ifi204 gene consists of nine exons spanning a genomic region of ~19 kb (table 3.16 and Fig. 3.12). Interestingly, the exon-intron structure of the Ifi204 gene is similar to that of the Ifi202a and b genes. This Ifi204 gene also contains an additional putative exon encoding a seven amino acid repeat (TSTAGAR) domain that would be located at the amino terminus of p204, but was not observed in p202a and p202b. This seven amino acid repeat domain also is present in the amino terminus of the D3 protein (encoded by the Ifi205 gene) where it consists of four perfect and three imperfect repeats, which could form a largely alpha helical structure. In addition, similar to p202a, the amino terminus of p204 also has several short sequence regions that are

rich in basic amino acids and are similar to sequences that have been shown to serve as nuclear targetting signals in other proteins.

Exon number	Size (bp)	Contig position (bp)
1	108	170378-170486
2	283	175392-175675
3	259	176891-177150
4	158	177653-177811
5	422	181239-181661
6	188	184368-184556
7	416	185220-185636
8	190	187864-188054
9	322	189633-189955

Table 3.16: Exon sizes and contig positions of the Ifi204 gene.

Ifi204 gene:



p204 protein:

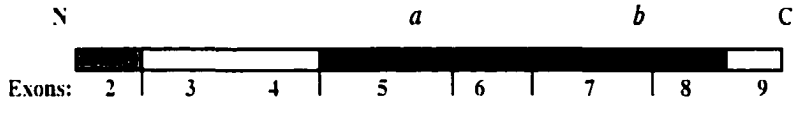


Fig. 3.12: Structural organization of the Ifi204 gene and its encoded protein, p204. (Top) Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines. (Bottom) *a* and *b* represent the two 200-amino acid domains conserved among the Ifi200 family of proteins. The gray box indicates the presence of a PAAD/DAPIN domain at the N-terminus of p204.

The results of the Neural Network splice site prediction program analysis confirmed the presence of potential splice sites at the exon-intron borders indicated by the CROSSMATCH comparison of this genomic region and the Ifi204 cDNA (table 3.17).

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
	170488		175391
1/1	caggaagGTgagttc	1/2	tgccaatttttttcttccAGgctcagcaacaacttcaaag
	175677		176890
2/2	tcagaagGTaacaca	2/3	aaaaagtctccacatcctAGtaacaggagaaacatcac
	177152		177652
3/3	agcccagGTgagcct	3/4	gggcagtggatatatttcAGaaaagaaaaagtatgaga
	177813		181238
4/4	ggctaagGTacaagc	4/5	ccctgtatttgttcaacacAGaaccaaaaatcacaaccca
	181663		184367
5/5	acacaagGTaaggtg	5/6	atgttaatcttacttttgcAGaaaacagtgaaccgaaaga
	184558		185219
6/6	catcaagGTtagtgtg	6/7	tcatatgtgtgttttatgtAGatatcaaagagaggaaatgt
	185638		187863
7/7	aaataagGTaagccg	7/8	aatgttaatttgcttttttcAGaaaacggacaggaataaatt
	188056		189632
8/8	catgcagGTacgtgc	8/9	ttcattgttttcttctttaAGgtcatcaatgctagaaagtg

Table 3.17: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A region including the first 1000 bp located immediately upstream from exon 1 followed by the first 200 bp of exon 1 of the Ifi204 gene was analyzed for potential transcription factor binding sites using MOTIF (table 3.4). Interestingly, the results were highly similar to those obtained by comparing the 5' flanking regions of the Ifi202a and b genes. In addition, two Interferon Stimulated Response Elements (ISRE),

a Friedman-Stark sequence and a GA box, are located a few base pairs upstream from the transcription initiation region of the Ifi204 gene. The Friedman-Stark sequence is located at the border between exon 1 and the 5' flanking region of the Ifi204 gene. The 5' flanking region of the Ifi204 gene also contains several potential transcription factor binding sites that have been implicated in the development of hematopoietic lineage cells, such as Ik-1, Ik-2, c-Myb, c-Ets, NF-AT or SOX5. These regions may be important in this context as Gariglio *et al.* (1998) demonstrated that the Ifi204 gene had an expression pattern that was restricted to cells of the monomyelocytic lineage, suggesting its involvement in the differentiation and maturation of this cell lineage. In addition, the 5' flanking region of the Ifi204 gene has potential binding sites for transcription factors whose activity is modulated by members of the Ifi200 family, including c-Myc, MyoD, c-Fos and c-Jun, as well as for STAT transcription factor that mediates the cell response to a variety of cytokines and growth factors. The 5' flanking region of the Ifi204 gene also shows two potential binding sites for p53 at 104-126 and 699-721 bases upstream from the 5' end of exon 1. Taken together, these data suggest that Ifi202a, Ifi202b and Ifi204 have a similar mode of transcriptional regulation. A DOTTER comparison between the 5' flanking regions of Ifi202a and Ifi204 (the first 1000 bp immediately upstream from the first exon followed by the first 200 bp of the first exon) revealed their extensive sequence similarity (Fig. 3.13), consistent with the thought that Ifi202a and Ifi204 have a similar mode of transcriptional regulation. Since the translation start site (ATG), as defined by Choubey *et al.* (1989), is located at contig position 175419, inside exon 2, exon 1 of the Ifi204 gene contains the 5' untranslated region. Promoter prediction by NNPP did not predict any TATA box-containing promoter in the 1000 bp located upstream from the 5' end of exon 1, suggesting that the Ifi204 gene is another example of a gene lacking TATA box sequences in its 5' flanking region. However, a MOTIF search on the 5' flanking region of Ifi204 predicted the presence of TATA box elements, 763-777 bp upstream from the 5' end of the Ifi204 gene. The poly(A) site

described by Choubey *et al.* (1989) is located at contig positions 189933-189938, 611 bp downstream from the 3' end of the last exon (exon 9). A search of the Pfam protein profile database indicates the presence of a PAAD/DAPIN domain at the amino terminus of p204 (Fig. 3.12), similar in terms of structure and sequence to the ones found at the amino termini of p203a and p203b. Finally, a PROSITE search of the protein profile database predicted the presence of potential bipartite nuclear localization signal, consistent with earlier indirect immunofluorescence microscopy experiments on interferon-treated cells reporting that the p204 protein is located in the nucleolar and nucleoplasmic fraction (Choubey and Lengyel, 1992).

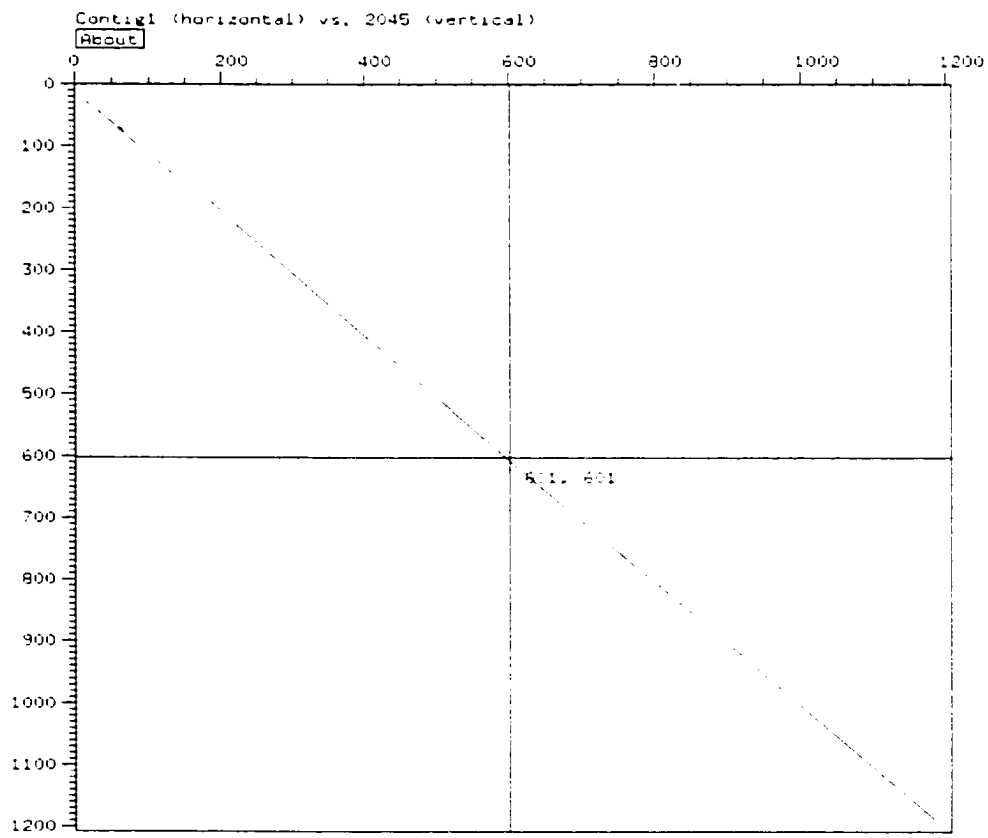


Fig. 3.13: Dot-plot analysis, using DOTTER, of the 5' flanking regions of the Ifi202a (horizontal) and Ifi204 (vertical) genes. The first 1000 bp located upstream from the initiation of transcription, followed by the first 200 bp of the first exon of each gene, are compared.

Ifi204 expression apparently is restricted to cells of myeloid origin (Gariglio *et al.*, 1998), in contrast with the Ifi202a gene that is widely expressed and can be induced in various organs, such as spleen, heart and brain, all of which are negative for p204 induction. The protein encoded by Ifi204 (p204) has two contiguous amino acid domains (domains *a* and *b*), similar in structure to the *a* and *b* domains of p202a and p202b. Even though these two domains are encoded for p202a and p202b by exons 4 and 5 and exons 6 and 7, respectively, but by exons 5 and 6 and exons 7 and 8, respectively, for p204 (Fig. 3.12), they likely arose from duplication of an ancestral domain. However, p202a, p202b and p204 also possesses unique sequences that may reflect their individual functions. As mentioned above, p204 (as well as D3) has a large insertion containing a series of short seven amino acid repeat sequences of unknown function, as well as a PAAD/DAPIN domain absent from the amino terminus of p202a and p202b. The p204 and the D3 proteins have a narrow range of tissue distribution, as they are selectively induced *in vivo* by interferons in macrophages only (Lembo *et al.*, 1998). This expression pattern, that differs widely from the one observed for p202a, could indicate that each member of the Ifi200 family is endowed with specific functions in specific cell lineages. It also appears that both p202a and p204 are involved in protein-protein interactions inside the nucleus, but their structural variations may indicate different interactions with various nuclear proteins are required to exert their antiproliferative activities.

III-1-5 Ifi201 gene

The Ifi201 gene is the fifth member of the murine Ifi200 gene family cluster sequenced during this dissertation research. Since its cDNA has not been cloned or sequenced previously, it was difficult to determine the exact position of the Ifi201 gene in the Ifi200 gene cluster. However, the 5' flanking region followed by exon 1 of a

putative Ifi201 gene was located at contig position 194862-195810 by a CROSSMATCH comparison between this segment of the Ifi201 gene (Genbank accession number M31421) and contig 1. In addition, a GENSCAN analysis of contig 1 locates a putative gene at contig position 199986-213291 (Fig. 3.3) that could correspond to the coding region of the Ifi201 gene. A search of the 5' flanking region followed by the first exon for potential transcription factor binding sites using MOTIF, reveals several interesting features (Table 3.4). First, the 5' flanking region of the Ifi201 gene contains potential binding sites for transcription factors that are expressed in cells of the hematopoietic lineage, such as MZF1, NF-AT, Lyf-1, Ik-1, Ik-2, GATA1, GATA2, GATA3, CCAAT/Enhancer Binding Protein (C/EBP), c-Myb, for STAT transcription factors and for transcription factors whose activity is modulated by other members of the Ifi200 family of proteins, such as c-Fos, c-Jun, NF-kB and MyoD. Furthermore, although an Interferon-Stimulated Response Element (Friedman-Stark sequence) is located inside exon 1, similar to the Friedman-Stark sequence located inside exon 1 of the Ifi202a and b genes, no GA box was detected using MOTIF. Finally, no p53 binding site similar in sequence to those found in the 5' flanking regions of the Ifi202a, Ifi202b and Ifi204 genes, was detected in the 5' flanking region of the Ifi201 gene (see Table 3.4). Taken together, these data suggest a mode of transcriptional regulation for the Ifi201 gene that may be similar to that determined for the Ifi203a and b genes. However, since the Ifi203a and Ifi203b genes lacked a Friedman-Stark sequence while containing a GA box in their 5' flanking regions, an observation opposite to that found for the Ifi201 gene, the Ifi201 gene likely has a different mode of interferon transcriptional activation. Finally, a Pfam search on the putative protein encoded by Ifi201, p201 (whose sequence was deduced from GENSCAN) revealed the existence of a PAAD/DAPIN domain at the amino terminus of p201. These results indicate that p201, along with p203a, p203b and p204, but not p202a or p202b, share a

common structural feature (the PAAD/DAPIN domain) indicative of their putative role in apoptosis.

III-1-6 Ifi202c pseudogene and others possible genes

An Ifi202c pseudogene, along with the Ifi202b gene discussed above, initially was discovered by screening a mouse 129sv genomic DNA library with a 202a cDNA probe (Wang *et al.*, 1998). When compared to the Ifi202a exonic sequences, Ifi202c had a single nucleotide substitution in exon 4, two single nucleotide substitutions and a nucleotide deletion in exon 6, one single nucleotide substitution in exon 7 and one single nucleotide substitution in exon 8. The single nucleotide deletion in exon 6 results in a translational frameshift mutation, and any resulting translated protein encoded by Ifi202c would be truncated because of an in-frame stop codon. (Fig. 3.14).

```
p202a       MSNRNLRSSTNSEFSEGQHQTPSSDSSGHGEDQPQASPGPNKKSHTPKKN
p202b       MSNRNLRSSTNSDLAEGQHQTPSSDSSGHGEDQPQASPGPNKKSHTPKKN
p202c       MSNRNLRSSTNSEFSEGQHQTPSSDSSGHGEDQPQASPGPNKKSHTPKKN
            ***********:::**********************************

p202a       ISKGAVLHEKPMTVMVLTATEPFNYKEGKENMFHATVATESQYYRVKVFN
p202b       ISKGAVLHEKPMTVMVLTATEPFNYKEGKENMFHATVATESKYYRVKVFN
p202c       ISKGAVLHEKPMTVMVLTVTEPFNYKEGKENMFHATVATESQYYRVKVFN
            ******************.**********************:********

p202a       MDLKEKFTENKFITISKYFNSSGILEINETATVSEAAPNQMFEVPKNIIR
p202b       MDLKEKFTENKFITISKYFNSSGILEINETATVSEAAPNQMFEVPKNIIR
p202c       MDLKEKFTENKFITISKYFNSSGILEINETATVSEAAPNQMFEVPKNIIR
            **************************************************

p202a       SAKETLKISKIKELDSGTLIYGVFAVEKKKVNDKSITFKIKDNEDNIKVV
p202b       SAKETLKISKIKELDSGTLIYGVFAVEKKKVNDKSITFKIKDNEDNIKVV
p202c       SAKETLKISKIKELDSGTLIYGVFAVEKKKVNDKSITFKIKDNEDNIKVV
            **************************************************

p202a       WDKEQHNINYEKGDKLQLFSFHLRKGNGKPILHSGNHSFIKGEKLLKESF
p202b       WDKEQHNINYEKGDKLQLFSFHLRKGNGKPILHSGNHSFVKGEKLLKESF
p202c       WDKEQHNINYEKGDKLQLFSFHLRKGNGKPILHSGNHSFIKGEKLLKESF
            ******************************************:***********

p202a       EGDGYHKGPKQVVALKATKLFTYDSIKSKKMFHATVATDTEFFRVMVFEE
p202b       EGDGYHKGPKQVVALKATKLFTYDSIKSKKMFHATVATDTEFFRVMVFEE
p202c       EGDGYHKGPKQVVALKATKLFTYDSIKSKKMFHATVATDTEFFRVMVFEE
```

```
                ************************************************

p202a            NLEKKFIPGNTIALSDYFGMYGSLAIHEYSSVSEVKSQNKEDSSSSDERP
p202b            NLEKKFIPGNTIALSDYFGMYGSLAIHEYSSVSEVKSQNKEDSSSSDERP
p202c            NLEKKFIPGNTIALSDYFGMYGSLANMNIPACLR----------------
                 ************************  : .:  .

p202a            IEHLKICDLHLQTEERLFDGEFKVYRKSSGNNCICYGIWDDTGAMKVVVS
p202b            IEHLKICDLHLQTKERLVDGEFKVYRKSSGNNCICYGIWDDTGAMKVVVS
p202c            --------------------------------------------------

p202a            GQLTSVNCEIGNTIRLVCFELTSNADEWFLRATRYSYMEVIMPEK
p202b            GQLTSVNCEIGNTIRLVCFELTSNADEWFLRATRYSYMEVIMPEK
p202c            ---------------------------------------------
```

Fig. 3.14: Clustal alignment of the amino acid sequences of the p202a, p202b proteins and the putative p202c protein.The mouse protein sequences are indicated on the left. The comparison was performed using the CLUSTALW program. (*) Identical or conserved residues in all sequences of the alignment; (:) conserved substitutions; (.) semi-conserved substitutions.

Wang *et al.* (1998) reported that no 202c mRNA was detected by RT-PCR or by sequencing the ESTs from several murine tissues and cell lines before and after interferon treatment. Thus, they proposed that this putative Ifi202c gene might be a pseudogene. Additional fluorescence *in situ* hybridization (FISH) colocalized the Ifi202a and b genes and the Ifi202c pseudogene on the same distal region of mouse chromosome 1, indicating that this pseudogene is present in the same genomic region as the Ifi202a and b genes (Wang *et al.*, 1999).

A GENSCAN analysis of contig 2 revealed that a putative seven exon-coding region was present in a genomic segment located between base pairs 61679-79950 that represents exon 2 to exon 8 of the putative Ifi202c pseudogene described by Wang *et al.* (1999) (Fig. 3.15). The resulting exon 2, 3, 4, 5, 6, 7 and 8 sequences (Genbank accession numbers AF141664, AF141665, AF141666, AF141667, AF141668, AF141669 and AF141670, respectively) were merged and an ALIGN comparison made between the genomic segment and the sequence generated by merging the Genbank

entries corresponding to exons 2 to 8 of the Ifi202c cDNA confirmed that this segment corresponds to the Ifi202c pseudogene.

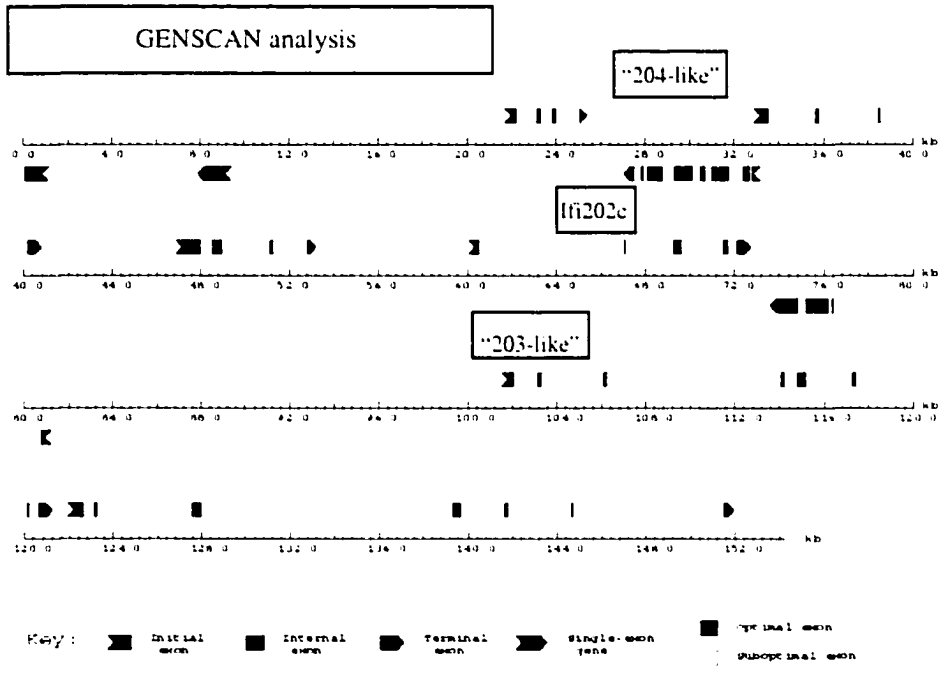


Fig. 3.15: GENSCAN analysis of contig 2. The genes and pseudogenes present on this clone are indicated.

The predicted translation start site (ATG) of the truncated protein encoded by the Ifi202c pseudogene is located inside exon 3, in a position similar to the translation start sites in the Ifi202a and b genes. Based on the observation of a strong sequence similarity between Ifi202a and Ifi202c exonic sequences, a CROSSMATCH comparison was made between exon 1 of Ifi202a and a genomic segment corresponding to the first 10000 base pairs located immediately upstream from the 5' end of exon 2 of Ifi202c. Interestingly, no sequence similar to Ifi202a exon 1 sequence was found, indicating that there is no exonic sequences homologous to the first exon of the Ifi202a gene in the region upstream from the second exon of the Ifi202c pseudogene. This suggests that the duplication event leading to the appearance of the Ifi202c pseudogene

resulted in the loss of exon 1 and, likely, of any transcriptional regulatory region located upstream from the first exon. A CROSSMATCH comparison between the first 1000 bp located upstream from the first exon of the Ifi202a gene and the first 10000 bp located upstream from the second exon of the Ifi202c pseudogene and a MOTIF analysis made on the first 1000 bp located upstream from the second exon of the Ifi202c pseudogene did not indicate the presence of any significant transcription factor binding sites, including interferon-stimulated response elements. Thus, it is likely that the Ifi202c pseudogene is not expressed since it lacks any known transcriptional regulatory regions.

A GENSCAN analysis of contig 2 revealed several other putative coding regions, besides Ifi202c (Fig. 3.15). A BLAST search made on a predicted coding region at contig position 97371-104776 revealed sequence similarity with the first five exons of the Ifi203b gene. A dot plot comparison, using DOTTER, between the contig region at position 97371-104776 and the genomic region between the first and the last exons of the Ifi203b gene (Fig. 3.16) revealed sufficient sequence similarity with part of the Ifi203b gene to speculate that a partially duplicated region of the Ifi203b gene might be present. Another dot plot comparison between the genomic region at contig position 97371-130000 and the Ifi203 cDNA (Fig. 3.17) further confirmed the presence of sequences similar to the first five exons of Ifi203b within this partially duplicated region. A BLAST search made on the genomic region at contig position 97371-130000 against a mouse-specific dbEST database did not reveal any significantly homologous EST sequences. Although the origin of this genomic region is unknown, it is likely that this region contains a truncated Ifi203 pseudogene that was generated by duplication of an ancestral Ifi203 gene or by progressive mutations in a duplicated Ifi203 gene that rendered it non-functional, during the evolution of the mouse Ifi200 gene family cluster.

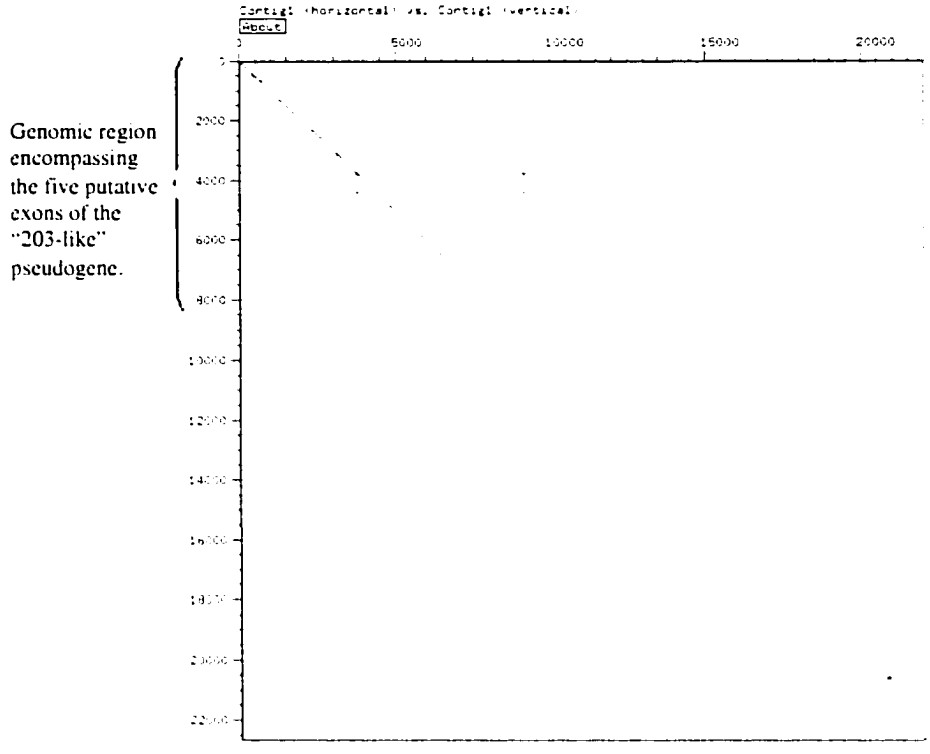


Fig. 3.16: Dot-plot analysis, using DOTTER, of the Ifi203b gene (horizontal) and the putative "203-like" pseudogene (vertical). The vertical axis corresponds to the "203-like" pseudogene followed by ~15 kb of genomic sequence, to demonstrate the absence of homology after the fifth exon of the "203-like" pseudogene.

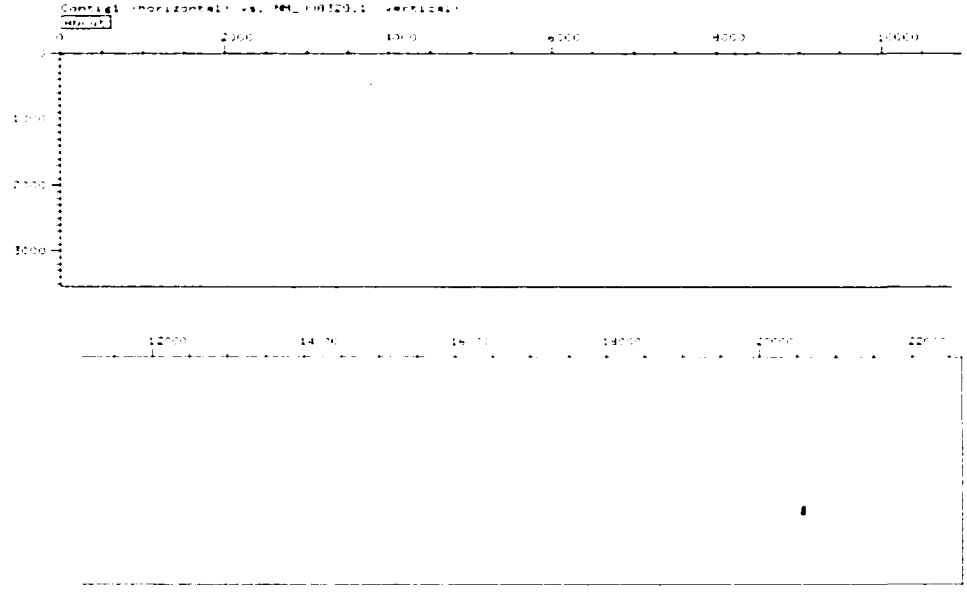


Fig. 3.17: Dot-plot analysis, using DOTTER, of the "203-like" pseudogene followed by ~15 kb of genomic sequence (horizontal), and the Ifi203 cDNA(vertical).

A BLAST search on a predicted coding region at contig position 18705-53100 revealed sequence similarity with the Ifi204 cDNA sequence. A dot plot comparison, using DOTTER, between the genomic segment located between the first and the last exons of the Ifi204 gene and the contig region at position 18705-53100 indicated regions of sequence similarity (Fig. 3.18).

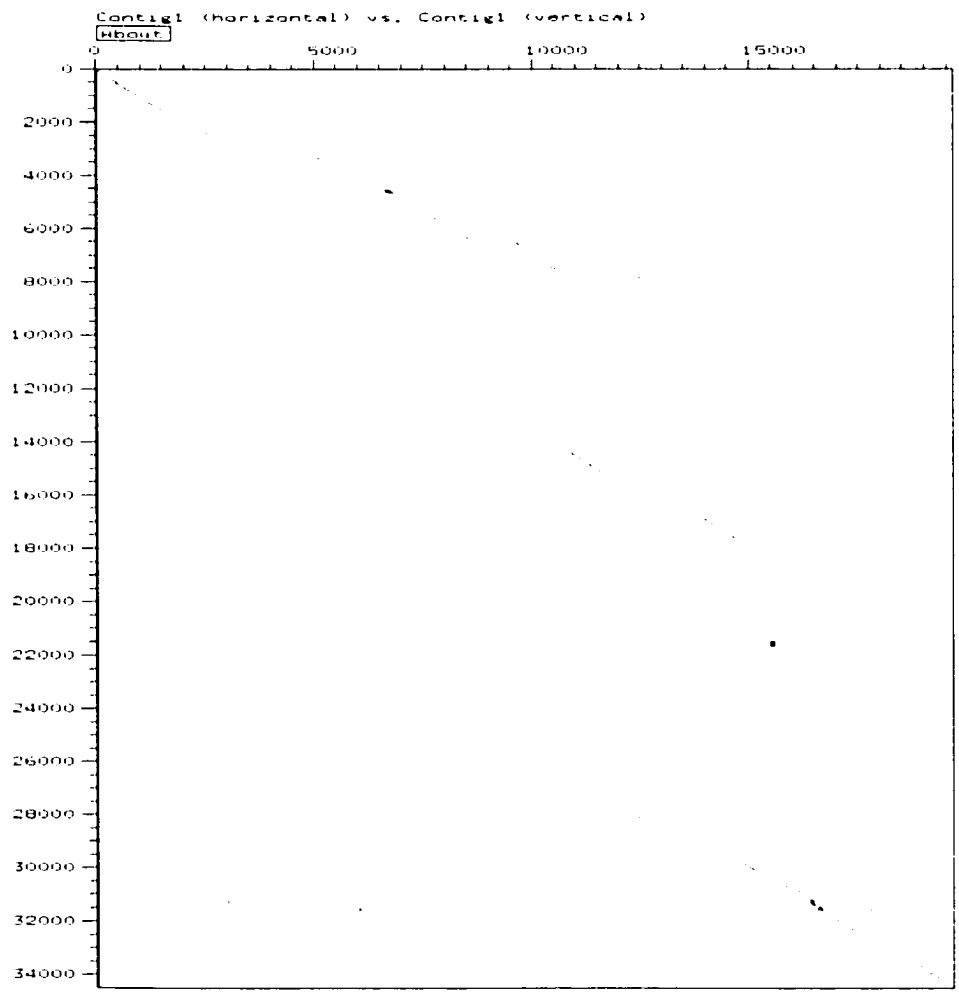


Fig. 3.18: Dot-plot analysis, using DOTTER, of the Ifi204 gene (horizontal) and the putative "204-like" gene (vertical).

Another dot plot comparison made between the contig region at position 18705-53100 and the Ifi204 cDNA revealed the presence of nine putative exons similar in

length to the nine exons of the Ifi204 gene (Fig. 3.19), suggesting the existence of a gene encoding a protein similar to p204.

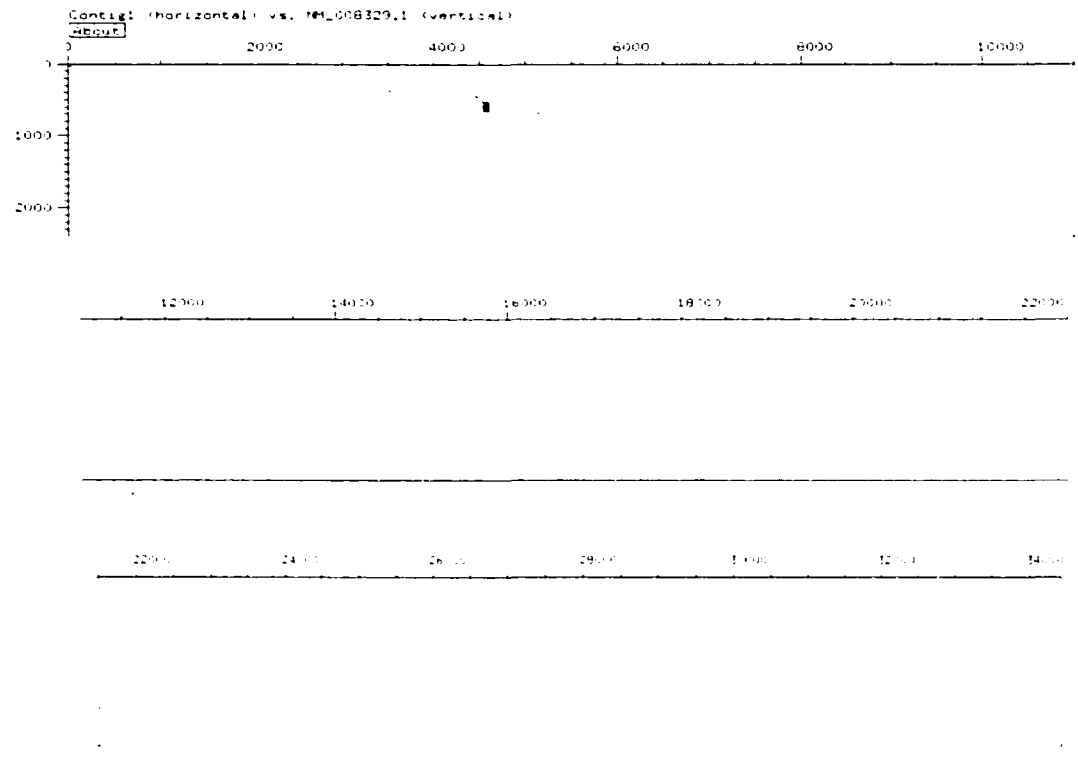


Fig. 3.19: Dot-plot analysis, using DOTTER, of the putative "204-like" gene (horizontal) and the Ifi204 cDNA (vertical).

This putative gene would span a genomic region of ~34 kb, which is longer than the region spanned by the Ifi204 gene. A putative translation start site (ATG) is located at contig position 21986, but because of a seven nucleotide insertion, a poly(A) site similar to that in Ifi204, as defined by Choubey *et al.* (1989), is absent. However, other possible candidate poly(A) sites are present at contig position 52897-52902 and 53133-53138. The search for potential transcription factor binding sites, using a MOTIF search of the first 1000 bp located immediately upstream from the 5' end of this putative gene revealed that, while the Ifi204 5' flanking region contains two ISREs (Friedman-

Stark sequence and GA box), only one ISRE (GA box) could be found in the 5' flanking region of this putative gene. Also, besides the presence of binding sites for transcription factors important for the development of hematopoietic cells, such as c-Ets, c-Myb, Ik-2, GATA1, GATA2, GATA3, and for STAT transcription factors, several potential binding sites for transcription factors whose activity is regulated by other members of the Ifi200 gene family, such as c-Fos, c-Jun, MyoD and E2F, were found (see Table 3.4). The sequences of each exon of this putative "204-like" gene were deduced after a GENSCAN and a CROSSMATCH analysis, based on their strong sequence similarity with the Ifi204 cDNA. The virtual cDNA sequence of the "204-like" gene was deduced by merging the sequence of each exons, and comparing it, using ALIGN, to the sequence of the Ifi204 cDNA. At the nucleotide level, these two sequences shows a 85.4% similarity (Fig. 3.20).

```
               10         20         30         40         50         60
Ifi204 TTTCTCATTTACTGACTTATCTGCCTACCTACTCAAGCCAAGCAGGCCACTTCTTGACCC
       ::::::: ::::::: :::::::::::::::::::: ::::::::::::::::::
Contig TTTCTCAGTTACTGAATTATCTGCCTACCTACTCAAACCAAGCAGGCCACTTCTG---TT
               10         20         30         40         50

               70         80         90        100        110        120
Ifi204 GGTGAAGGTCTCAGGATCTGTACATCACTGCAGAAATATCCAGGAAGGCTCAGCAACAAC
       : ::::: :::::: : ::::::::  :::: ::::: :::::: ::  : : :::::::
Contig GTTGAAGATCTCAGCACCTGTACATTGCTGCCGAAAT-TCCAGGGAGTATAACCAACAAC
        60         70         80         90        100        110

              130        140        150        160        170        180
Ifi204 TTCAAAGATGGTGAATGAATACAAGAGAATTGTTCTGCTGAGAGGACTTGAATGTATCAA
       :: :::::::: ::::::::: :::::: ::::::::::  ::::::::::::::::::
Contig TTGAAAGATGGAGAATGAATATAAGAGACTTGTTCTGCTGGAAGGACTTGAATGTATCAA
        120        130        140        150        160        170

              190        200        210        220        230        240
Ifi204 TAAGCATTATTTTAGCTTATTTAAGTCATTGCTGGCCAGAGATTTAAATCTGGAAAGAGA
       ::::::: : :: :  :::::::::::::::: ::: :: ::::::::::::::::  :::
Contig TAAGCATCAATTCAATTTATTTAAGTCATTGATGGTCAAAGATTTAAATCTGGAAGAAGA
        180        190        200        210        220        230

              250        260        270        280        290        300
Ifi204 CAACCAAGAGCAATACACCACGATTCAGATTGCTAACATGATGGAAGAGAAATTTCCAGC
       :::::::::: :::: :::::: :::::::::::::::::::: : ::::::::::::
Contig CAACCAAGAGAAATATACCACGTTTCAGATTGCTAACATGATGGTAAAGAAATTTCCAGC
        240        250        260        270        280        290

              310        320        330        340        350        360
Ifi204 TGATTCTGGATTGGGCAAACTGATTGAGTTTTGTGAAGAAGTACCAGCTCTTAGAAAACG
       :::: ::::::::: :::::::::  : :::::::::   :::::: :::::: ::::::
Contig TGATGCTGGATTGGACAAACTGATCAACTTTTGTGAACGTGTACCAACTCTTAAAAAACG
```

```
        300       310       320       330       340       350

              370       380       390           400       410
Ifi204 AGCTGAAATTCTTAAAAAAGAGAGATCAGAAGTA---ACAGGAGAAACATCACTGGAAAA
        :::::::::::::::::::::::::::::::::   ::::::::::::::::::::::
Contig TGCTGAAATTCTTAAAAAAGAGAGATCAGAAGTAGTAACAGGAGAAACATCACTGGAAAT
        360       370       380       390       400       410

       420       430       440       450       460       470
Ifi204 AAATGGTCAAGAAGCAGGTCCTGCAACACCTACATCAACTACAAGCCACATGTTAGCATC
       :::: : ::::::::::::::::::::::::::::::::::::::::::::::::::::::
Contig AAATAGGCAAGAAGCAGGTCCTGCAACACCTACATCAACTACAAGCCACATGTTAGCATC
        420       430       440       450       460       470

       480       490       500       510       520       530
Ifi204 TGAAAGAGGCGAGACTTCTGCAACCCAGGAAGAGACTTCCACAGCTCAGGCGGGGACTTC
       :::::::::::::::::: :::::::::::::::::::::::::: ::
Contig TGAAAGAGGCGAGACTTCCGCAACCCAGGAAGAGACTTCCACAGCCCA------------
        480       490       500       510       520

       540       550       560       570       580       590
Ifi204 CACAGCTCAGGCGAGGACTTCCACAGCTCAGGCGAGGACTTCCACAGCTCAGGCGAGGAC
                                                          :: ::::::
Contig ----------------------------------------------------GGTGAGGAC
                                                             530

       600       610       620       630       640       650
Ifi204 TTCCACAGCTCAGGCGAGGACTTCCACAGCTCAGGCGGGGACTTCCACAGCCCAGAAAAG
       ::::::::::::::::: :::::::::::::::::::::::::::: ::::::::::::::
Contig TTCCACAGCTCAGGCGGGGACTTCCACAGCTCAGGCGGGGACTTCTACAGCCCAGAAAAG
           540       550       560       570       580       590

       660       670       680       690       700       710
Ifi204 AAAAAGTATGAGAGAAGAAGAGACTGGAGTGAAAAAGAGCAAGGCGGCTAAGGAACCAGA
        ::: ::::::: :::::: :::::: ::::: :::: :::::: :: : :: ::::
Contig GAAAGGTATGAGTGAAGAAAAGACTGACGTGAAGAAGATCAAGGCATCTGGGAAAGCAGA
           600       610       620       630       640       650

       720       730       740       750       760       770
Ifi204 TCAGCCTCCCTGTTGTGAAGAACCCACAGCCATGTGCCAGTCACCAATACTCCACAGCTC
       :::::::::::::::::::: :::::::::::  ::::::::::::::  :::   :::
Contig TCAGCCTCCCTGTTGTGAAGGACCCACAGCCACATGCCAGTCACCAATATCCCAGGTCTC
           660       670       680       690       700       710

       780       790       800       810       820       830
Ifi204 ATCTTCGGCTTCATCTAACATTCCTTCGGCTAAGAACCAAAAATCACAACCCCAGAATCA
       :::::::::::::::::::::::::::::::::::::::::::::::::::::::::: ::
Contig ATCTTCGGCTTCATCTAACATTCCTTCGGCTAAGAACCAAAAATCACAACCCCAGAACCA
           720       730       740       750       760       770

       840       850       860       870       880       890
Ifi204 GAATATTCCCAGAGGTGCTGTTCTCCACTCAGAGCCCCTGACAGTGATGGTGCTCACTGC
       ::: ::::::::::::::::::::::::::::::::::::::::::::::::::::::::
Contig GAACATTCCCAGAGGTGCTGTTCTCCACTCAGAGCCCCTGACAGTGATGGTGCTCACTGC
           780       790       800       810       820       830

       900       910       920       930       940       950
Ifi204 AACAGACCCATTTGAATATGAATCACCAGAACATGAAGTAAAGAACATGCTTCATGCTAC
       :::::::::::::::::::::::::::::::::::::::::::::::::: ::::::::::
Contig AACAGACCCATTTGAATATGAATCACCAGAACATGAAGTAAAGAACATGTTTCATGCTAC
           840       850       860       870       880       890

       960       970       980       990      1000      1010
Ifi204 AGTGGCTACAGTGAGCCAGTATTTCCATGTGAAAGTTTTCAACATCAACTTGAAAGAAAA
```

```
      :::::::::::::::::::::::::::::::::::::::::::::::::::: : :::::::: ::
Contig AGTGGCTACAGTGAGCCAGTATTTCCATGTGAAAGTTTTCAACATCGATTTGAAAGAGAA
           900        910        920        930        940        950

      1020       1030       1040       1050       1060       1070
Ifi204 GTTCACAAAAAAGAATTTTATCATCATATCCAATTACTTTGAGAGCAAAGGCATCCTGGA
       :::::::::::: :::::::::: :::::::::::::::::::::::::::::::::::
Contig GTTCACAAAAAATAATTTTATCACCATATCCAATTACTTTGAGAGCAAAGGCATCCTGGA
           960        970        980        990       1000       1010

      1080       1090       1100       1110       1120       1130
Ifi204 GATCAATGAGACTTCCTCTGTGTTAGAGGCTGCTCCTGACCAAATGATTGAAGTGCCCAA
       :::::::::::::::::::::::::::::::::::: : ::::::::::::::::::::
Contig GATCAATGAGACTTCCTCTGTGTTAGAGGCTGCTCCTAAACAAATGATTGAAGTGCCCAA
           1020       1030       1040       1050       1060       1070

      1140       1150       1160       1170       1180       1190
Ifi204 CAGTATTATCAGAAATGCAAATGCCAGTCCTAAGATCTGTGATATTCAAAAGGGTACTTC
       : :::::: ::::::::::::::::::::::::::::::::::::::::::::::::::::
Contig CTGTATTACCAGAAATGCAAATGCCAGTCCTAAGATCTGTGATATTCAAAAGGGTACTTC
           1080       1090       1100       1110       1120       1130

      1200       1210       1220       1230       1240       1250
Ifi204 TGGAGCAGTGTTCTATGGAGTGTTTACATTACACAAGAAAACAGTGAACCGAAAGAACAC
       :::: ::::::::::::::::::::::::::::::::::: :::::::   : :::::::
Contig TGGAACAGTGTTCTATGGAGTGTTTACATTACACAAGAAAAAAGTGAAAACACAGAACAC
           1140       1150       1160       1170       1180       1190

      1260       1270       1280       1290       1300       1310
Ifi204 AATCTATGAAATAAAAGATGGTTCAGGAAGCATAGAAGTGGTGGGGAGTGGAAAATGGCA
       :: :::::::::::::::::::::::::::: :::::::::::::::::::: :::::::
Contig AAGCTATGAAATAAAAGATGGTTCAGGAAGTATAGAAGTGGTGGGGAGTGGACAATGGCA
           1200       1210       1220       1230       1240       1250

      1320       1330       1340       1350       1360       1370
Ifi204 CAACATCAACTGCAAGGAAGGAGATAAACTCCACCTCTTCTGCTTTCACCTGAAAACAAT
       :::::::::::: :::::::::::::: :::::::::::::::::::::::::::::: :
Contig CAACATCAACTGTAAGGAAGGAGATAAGCTCCACCTCTTCTGCTTTCACCTGAAAAGAGA
           1260       1270       1280       1290       1300       1310

      1380       1390       1400       1410       1420       1430
Ifi204 TGACAGGCAACCAAAGTTAGTGTGTGGAGAACACAGTTTCATCAAGATATCAAAGAGAGG
           : :::::::::::::::::::::::: ::::::::: :::::: : ::: ::::::
Contig AAGAGGACAACCAAAGTTAGTGTGTGGAGACCACAGTTTCGTCAAGACACCAAGGAGAGG
           1320       1330       1340       1350       1360       1370

      1440       1450       1460       1470       1480       1490
Ifi204 AAATGTACCAAAGGAGCCTGCTAAGGAAGAAGATCACCATCATGGTCCCAAACAAGTGAT
       :: ::::::::::::::::: ::: :::::::: ::::::::: :::::::::::::::::
Contig AACTGTACCAAAGGAGCCTTCTAGGGAAGAAGGTCACCATCAAGGTCCCAAACAAGTGAT
           1380       1390       1400       1410       1420       1430

      1500       1510       1520       1530       1540       1550
Ifi204 GGTGCTGAAAGTAACAGAACCATTTACATATGACCTGAAAGAGGATAAAAGAATGTTTCA
       :::: ::::::::::::::::::::::::::::::::: ::::  ::::::  ::::::::
Contig GGTGTTGAAAGTAACAGAACCATTTACATATGACCTGGAAGAAAATAAAATGATGTTTCA
           1440       1450       1460       1470       1480       1490

      1560       1570       1580       1590       1600       1610
Ifi204 TGCTACCGTGGCTACTGAAACTGAGTTCTTCAGAGTGAAGGTTTTTGACACAGCTCTAAA
       :::::: ::::::::::::::::::::::::::::::::::::::::::::::::::::::
Contig TGCTACAGTGGCTACTGAAACTGAGTTCTTCAGAGTGAAGGTTTTTGACACAGCTCTAAA
           1500       1510       1520       1530       1540       1550
```

```
      1620       1630       1640       1650       1660       1670
Ifi204 GAGCAAGTTCATCCCAAGAAATATCATTGCCATATCAGATTATTTTGGGTGCAATGGGTT
       ::::::::::::::::  :::::::::::::::::::::::::::::::  :::::::::
Contig GAGCAAGTTCATCCCAGAAAATATCATTGCCATATCAGATTATTTTGGGCTCAATGGGTT
          1560       1570       1580       1590       1600       1610

      1680       1690       1700       1710       1720       1730
Ifi204 TCTGGAGATATACAGAGCTTCCTGTGTCTCTGATGTGAACGTTAATCCAAGAATGGTTAT
       :::::::::: :::: : ::::: ::: :::::::::::::  :: :::  ::::::: :
Contig TCTGGAGATACACAGCGATTCCTCTGTGTCTGATGTGAACGAAAACCCACCAATGGTTGT
          1620       1630       1640       1650       1660       1670

      1740       1750       1760       1770       1780       1790
Ifi204 CTCAAATACACTGAGACAAAGAGCTAATGCAACTCCTAAAATTTCTTATCTTTTCTCACA
       :::: :::::::::  :::: :::::: :::::::: ::::::::::::::::::::::
Contig CTCAGATACACTGAAGCAAAAAGCTAAAGCAACTCCCAAAATTTCTTATCTTTTCTCACA
          1680       1690       1700       1710       1720       1730

      1800       1810       1820       1830       1840       1850
Ifi204 AGCAAGGGGGACATTTGTGAGTGGAGAGTACTTAGTAAATAAGAAAACGGAGAGGAATAA
       : ::: ::::: :::::::: ::::   :  ::::::: : ::::::: ::::::::: :
Contig AACAATGGGGAGATTTGTGAATGGAATATTTTTAGTAATTGAGAAAACTGAGAGGAATGA
          1740       1750       1760       1770       1780       1790

      1860       1870       1880       1890       1900       1910
Ifi204 ATTCATTTACTATGGAATTGGAGATGATACAGGGAAAATGGAAGTGGTGGTTTATGGAAG
       : :::::::::::::::: :::::::::::::::::::::::::::::::::::::: :
Contig AGTCATTTACTATGGAATTGAAGATGATACAGGGAAAATGGAAGTGGTGGTTTATGGACG
          1800       1810       1820       1830       1840       1850

      1920       1930       1940       1950       1960       1970
Ifi204 ACTCACCAATGTCAGGTGTGAACCAGGCAGTAAACTAAGACTTGTCTGCTTTGAATTGAC
       :::::::::: ::: :::::::::::::::::::: :::::::::::: ::: :::::::
Contig ACTCACCAATATCAAGTGTGAACCAGGCAGTAAACTTAGACTTGTCTGTTTTAAATTGAC
          1860       1870       1880       1890       1900       1910

      1980       1990       2000       2010       2020       2030
Ifi204 TTCCACTGAAGATGGGTGGCAGCTGAGGTCTGTAAGGCACAGTTACATGCAGGTCATCAA
       :::::::::::::::::::::::::::::::::::::::::::::::::: ::::::::::
Contig TTCCACTGAAGATGGGTGGCAGCTGAGGTCTGTAAGGCACAGTTACATGGAGGTCATCAA
          1920       1930       1940       1950       1960       1970

      2040       2050       2060       2070       2080       2090
Ifi204 TGCTAGAAAGTGAAGGAAAGCCACTCAACCCAGACTCAGTCGGGAGAACCTCTCTGGAAC
       : :::::::::::::::::::::::: ::::::::: : ::::: ::::::::: ::
Contig TCCTAGAAAGTGAAGGAAAGCCACTCAATCCAGACTCAGCCAGGAGATCCTCTCTGGTAC
          1980       1990       2000       2010       2020       2030

         2100       2110       2120       2130       2140       2150
Ifi204 C------ATACTTCTGAAAACCTGAATGCCAATGATATTTTTTTGTGGAGATAAGATTCA
       :       :::::::::::::::: :::::::::::: : :::   : ::: ::::::::
Contig CTGGGTGATACTTCTGAAAACCTGGATGCCAATGATACTGTTTCA-GAAGACAAGATTCA
          2040       2050       2060       2070       2080       2090

         2160       2170       2180       2190       2200       2210
Ifi204 ATTACAGAAAATAAATGTGTAGAAG-CCTATTGAAATATCAGTCCTATAAAGATTATCTC
       : :: ::::::::::::::: : : : ::: :::::::: ::  ::::: :: :  : :::
Contig AGTAGAGAAAATAAATGTGCATATGGCCTGTTGAAATACCATCCCTATTAAAACCAGCTC
          2100       2110       2120       2130       2140       2150

         2220       2230       2240       2250       2260
Ifi204 TTAATTCTAGGAAATGGTATTTTCTTATATCTTTAC-ACATTTT-CTAT-ATCTAAATT-
       ::::::::::::::::::  :::::::: : ::::  :::::::: :::: :::::::::
Contig TTAATTCTAGGAAATGGTGCTTTCTTATGTTTTTAATACATTTTTCTATTATCTAAATTT
```

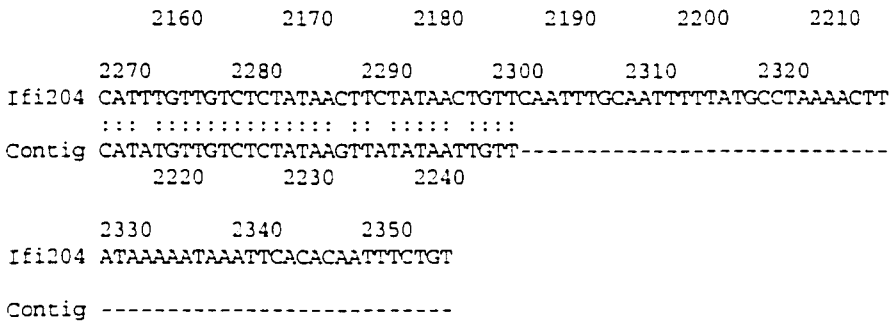
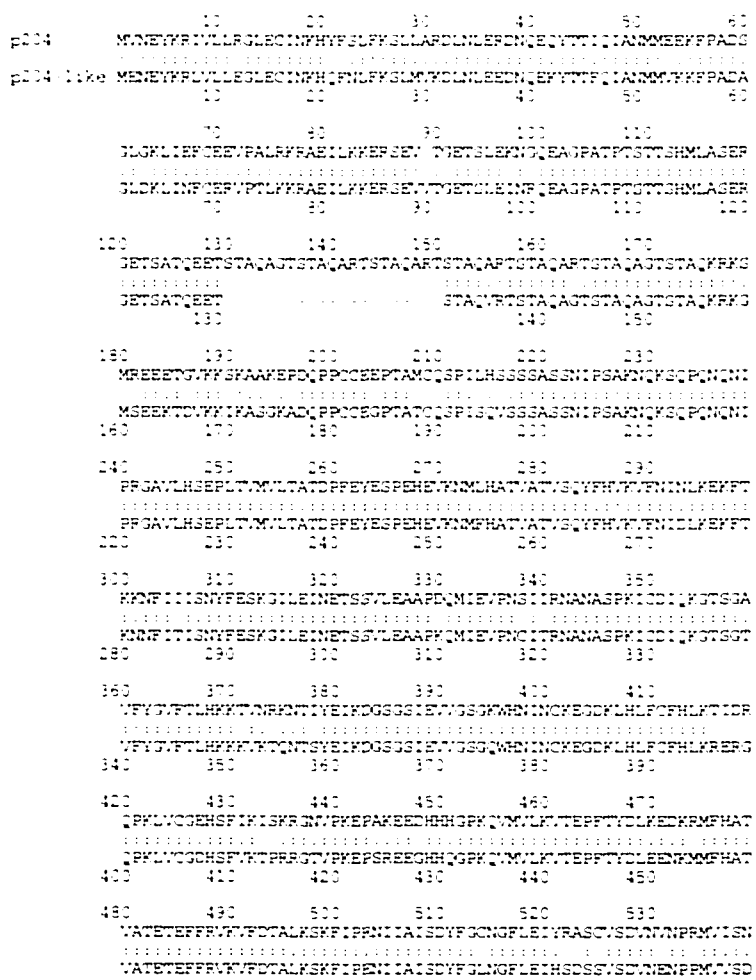


Fig. 3.20: Alignment of the cDNA sequences of the Ifi204 gene (top) and the "204-like" putative gene (bottom). The two sequences exhibit 85.4% sequence homology.

The corresponding amino acid sequence of the putative "p204-like" protein was deduced from the sequence of each putative exon and compared, using ALIGN, to the sequence of p204 (Fig. 3.21).



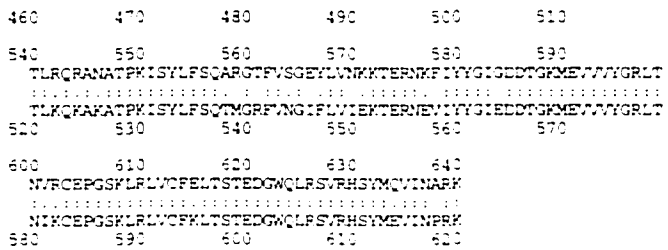


Fig. 3.21: Alignment of the amino acid sequences of p204 (top) and the putative "p204-like" protein (bottom)

This comparison reveals that p204 and the putative "p204-like" protein shares 81.4% sequence similarity at the amino acid level, and shows that three of the seven amino acid repeats present at the amino terminus of p204 are absent at the amino terminus of the putative "p204-like" protein. Taken together, these data suggest that the putative "204-like" gene encodes for a new member of the 200-family of proteins, similar, in terms of sequence, to the protein encoded by Ifi204.

The p204 protein contains two *a* and *b* domains that are conserved among the members of the 200 family of proteins and these two domains are encoded by exons 5 and 6, and exons 7 and 8, respectively, of Ifi204. The *a* and *b* domains of p204 shows over 90% sequence similarity with two segments of the putative "p204-like" protein, encoded by exons 5 and 6, and exons 7 and 8, respectively, indicating that "p204-like" contains two *a* and *b* domains. The D3 protein does not contain a *b* domain and its carboxy-terminus is made of a short non-conserved domain following the *a* domain. Because of the strong sequence similarity between the protein encoded by the D3 gene and the amino terminus of p204 (Tannenbaum *et al.*, 1993), a CROSSMATCH analysis was made between the genomic sequence of the "204-like" gene and the D3 cDNA. The CROSSMATCH comparison between the genomic sequence of the "204-like" gene and the D3 cDNA revealed the first six exons of the putative "204-like" gene, as well as the presence of an additional exon containing a STOP codon located inside the intron 6 of

the putative "204-like" gene. The sequence of this putative additional exon was deduced, using GENSCAN and CROSSMATCH, based on its strong sequence similarity with the D3 cDNA. When merged with exons 1-6, this putative "204-like" gene created a virtual open reading frame from the seven exons. The resulting cDNA was translated, and the deduced amino acid sequence compared, using ALIGN, to the sequence of the D3 protein. Interestingly, the two sequences shares 86.4% sequence identity (Fig. 3.22).

```
               10        20        30        40        50        60
pD3.fa MVNEYKRIVLLRGLECINKHYFSLFKSLLARDLNLERDNQEQYTTIQIANMMEEKFPADS
       : :::::.:::.:::::::: :.:::::::..:::::.:::.::::::: .:::::.
virt|V MENEYKRLVLLEGLECINKHQFNLFKSLMVKDLNLEEDNQEKYTTFQIANMMVKKFPADA
               10        20        30        40        50        60

               70        80        90       100       110       120
pD3.fa GLGKLIEFCEEVPALRKRAEILKKERSEVTGETSLEKNGQEAGPATPTSTTSHMLASERG
       :: :::.:::.::.:.:::::::::::::::::::: : ::::::::::::::::::::
virt|V GLDKLINFCERVPTLKKRAEILKKERSEVTGETSLEINRQEAGPATPTSTTSHMLASERG
               70        80        90       100       110       120

              130       140       150       160       170       180
pD3.fa ETSATQEETSTAQAGTSTAQAGTSTAQAGTSTAQKRKSMREEETGVKKSKAAKEPDQPPC
       ::::::::::::::. ::::::::::::::::::::::::.: :::: ::: ::. . :::::
virt|V ETSATQEETSTAQVRTSTAQAGTSTAQAGTSTAQKRKGMSEEKTDVKKIKASGKADQPPC
              130       140       150       160       170       180

              190       200       210       220       230       240
pD3.fa CEEPTAMCQSPILHSSSSASSNILSAKNQKSQPQNQNIPRGAVLHSEPLTVMVLTATDPF
       :: ::: ::::: . :::::::: :::::::::::::::::::::::::::::::::::
virt|V CEGPTATCQSPISQVSSSASSNIPSAKNQKSQPQNQNIPRGAVLHSEPLTVMVLTATDPF
              190       200       210       220       230       240

              250       260       270       280       290       300
pD3.fa EYESPEHEVKNMFHATVATVSQYFHVKVFNIDLKEKFTKNNFITISNTFESKGILEINET
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
virt|V EYESPEHEVKNMFHATVATVSQYFHVKVFNIDLKEKFTKNNFITISNTFESKGILEINET
              250       260       270       280       290       300

              310       320       330       340       350       360
pD3.fa SSVLEAAPKQMIEVPNCITRNANASPKICDIQKGTSGTVFYGVFTLHKKKVKTQNTSYEI
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
virt|V SSVLEAAPKQMIEVPNCITRNANASPKICDIQKGTSGTVFYGVFTLHKKKVKTQNTSYEI
              310       320       330       340       350       360

              370       380       390       400       410       420
pD3.fa KDGSGRIEVVGSGQWHNINCKEGDKLHLFCFHLKRERGQPKLVCGDHSFVKVTKAGKKKE
       ::::: ::::::::::::::::::::::::::::::::::::::::::::::::::::::::
virt|V KDGSGSIEVVGSGQWHNINCKEGDKLHLFCFHLKRERGQPKLVCGDHSFVKVTKAGKKKE
              370       380       390       400       410       420


pD3.fa ASTVQ----------------------
       :::::
virt|V ASTVQSSTKNEEENDYPKVGIKVEMPK
              430       440
```

Fig. 3.22: Alignment of the amino acid sequences of the D3 protein (top) and the virtual amino acid sequence deduced from the alternatively spliced form of the putative "204-like" gene (bottom).

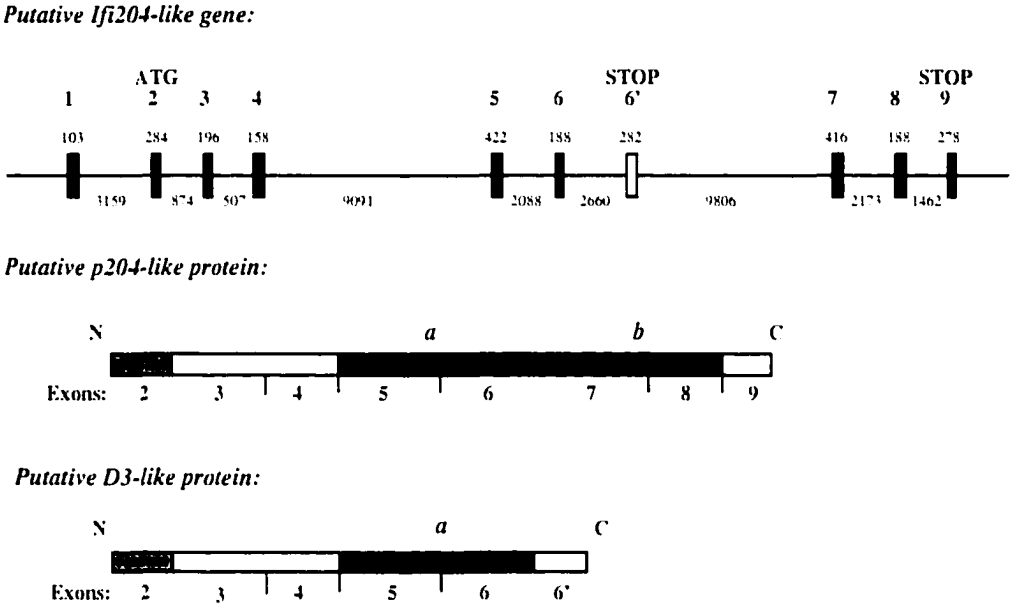


Fig. 3.23 Structural organization of the putative "204-like" gene and its encoded proteins. (Top) Sizes, in base pairs, are given below the introns and above the exons. The exons are indicated by black boxes and the connecting introns by horizontal lines. (Bottom) *a* and *b* represent the two 200-amino acid domains conserved among the Ifi200 family of proteins. The gray box indicates the presence of a putative PAAD/DAPIN domain.

Taken together, these data raise the possibility of a new "204-like" gene whose translation product shares 81.4% sequence identity with p204. In addition, the presence of the additional exon inside intron 6 of this "204-like" gene suggests the possibility of an alternatively spliced form of "204-like" that could contain the first 6 exons of "204-like" and the additional exon, which would result in a translation product that shares 86.4% sequence identity with the D3 protein. A similar approach used with Ifi204 did not reveal that the D3 gene was an alternatively spliced form of Ifi204. It remains to be established whether D3 corresponds to a splice variant of a still unknown gene sharing some sequence similarity with Ifi204 , or whether the duplication of the "204-like" putative gene led to the appearance of the Ifi204 and D3 genes. The organization of the

putative "204-like" gene, along with its corresponding putative p204-like and D3-like translation products, are summarized on Fig. 3.23.

III-1-7 Repeat elements in the mouse Ifi200 cluster

Repeat elements in the mouse Ifi200 gene cluster were identified using Repeatmasker and localized using PIPmaker (see Materials and Methods).

Analysis of the repeat elements in the 236088 base pairs of contig 1 showed that they make up to 39.63% of the sequence of this contig (table 3.18). Interestingly, most of the repeat elements (~75%) were members of the LINE1 family. Other abundant repeat elements in the mouse lineage, such as the SINE B1 and B2 repeat elements, represent only 4 to 5% of the repeats. Inspection of the distribution of the repeat elements in the contig also shows that the LINEs are equally distributed within large intronic and intergenic regions. Although overall, LINEs occur mostly in gene-poor AT-rich regions of the human genome, this tendency is not observed here, as the LINE1 elements represent about 30% of the total sequence of this gene-rich region. Because many of the genes present in this region do not have any orthologs in the corresponding human cluster, it is likely that they arose from duplications that are specific to the rodent lineage. Also, because LINEs are present in both species, the appearance of LINEs must have predated human and mouse speciation, indicating that their presence in this region probably occured prior to the duplications leading to the Ifi200 cluster. These duplications also likely led to the duplication of intragenic LINE elements and thus explain, at least partially, the emergence of such a high percentage of LINE elements in this region. In addition, new intergenic LINE elements could arise from the high rate of recombination in mouse. The presence of such a high percentage of LINEs in this region may have been the driving force behind the duplications leading to the appearance of this cluster. The low amount of SINEs may be explained either by a

decline in their activity during the evolution of the cluster or by unexplained negative selection events.

	number of elements	length occupied	percentage of sequence
SINEs:	33	4153 bp	1.89%
B1s	19	2438 bp	1.11%
B2-B4	14	1715 bp	0.78%
IDs	0	0 bp	0.00%
MIRs	0	0 bp	0.00%
LINEs:	72	64982 bp	29.53%
LINE1	72	64982 bp	29.53%
LINE2	0	0 bp	0.00%
L3/CR1	0	0 bp	0.00%
LTR elements:	30	10875 bp	4.94%
MaLRs	5	900 bp	0.41%
ERVL	0	0 bp	0.00%
ERV_classI	0	0 bp	0.00%
ERV_classII	18	5986 bp	2.72%
DNA elements:	5	1223 bp	0.56%
MER1_type	3	611 bp	0.28%
MER2_type	0	0 bp	0.00%
Unclassified:	0	0 bp	0.00%
Total interspersed repeats:		81233 bp	36.92%
Small RNA:	0	0 bp	0.00%
Satellites	0	0 bp	0.00%
Simple repeats:	66	4316 bp	1.96%
Low complexity	33	1645 bp	0.75%

Table 3.18: Identification of the repeat elements on contig 1.

The analysis of the repeat elements in the 154284 base pairs of contig 2 shows that they make up to 36.56% of the sequence of this contig (table 3.19). This observation is similar to that observed above for contig 1, since the LINE elements here also contain only LINE1 family elements and represent ~70% of the total repeat

elements present. Only ~5% of the total repeat elements in this contig are SINE B1 and B2 elements. Thus, these two contigs share common repeat element distributions. Inspection of the position of the LINEs in this contig shows that they mainly are present in the region encompassing the "204-like", 202c and "203-like" genes (or pseudogenes) (see Fig. 3.25). These LINEs occur within the intronic and intergenic regions of these genes or pseudogenes at a percentage similar to contig 1, suggesting that they appeared during the evolution scale at a time similar to the time of appearance of the Ifi200 genes, and that multiple duplications led to the simultaneous appearance of genes and pseudogenes within this region.

	number of elements	length occupied	percentage of sequence
SINEs:	23	3049 bp	1.98%
B1s	7	787 bp	0.51%
B2-B4	15	2143 bp	1.39%
IDs	0	0 bp	0.00%
MIRs	1	119 bp	0.08%
LINEs:	52	40032 bp	25.95%
LINE1	52	40032 bp	25.95%
LINE2	0	0 bp	0.00%
LTR elements:	22	7783 bp	5.04%
MaLRs	3	435 bp	0.28%
Retroviral	15	6037 bp	3.91%
DNA elements:	2	454 bp	0.29%
MER1_type	2	454 bp	0.29%
MER2_type	0	0 bp	0.00%
Unclassified	2	1011 bp	0.66%
Total interspersed repeats		52329 bp	33.92%
Small RNA	0	0 bp	0.00%
Satellites	0	0 bp	0.00%
Simple repeats	47	2978 bp	1.93%
Low complexity	25	1096 bp	0.71%

Table 3.19: Identification of the repeat elements on contig 2.

III-1-8 Mouse Ifi200 gene cluster: a summary

In this dissertation research, seven genes were predicted in contig 1, and three genes or pseudogenes were predicted in contig 2 (table 3.20).

Gene or pseudogene	Contig position	Biological effects	Tissue distribution
Ifi203c (last 3 exons)	3843-10558	unknown	unknown
Ifi202b	24762-51851	unknown	unknown
Ifi203b	69310-91294	unknown	unknown
Ifi202a	98716-125808	retards cell proliferation	thymus, heart, muscle...
Ifi203a	143300-165288	unknown	thymus, bone marrow and spleen
Ifi204	170378-189955	retards cell proliferation and triggers myoblast fusion	myeloid tissues
Ifi201	199986-213271	unknown	unknown
Ifi205	unknown	unknown	myeloid tissues
"204-like"	18705-53100	unknown	unknown
Ifi202c	61679-79950	unknown	unknown
"203-like"	97371-104776	unknown	unknown

Table 3.20: Contig position, biological effect and tissue distribution of the genes and pseudogenes analyzed in this dissertation research. Ifi205 position remains to be established. The "204-like, Ifi202c and "203-like" genes or pseudogenes are positioned on contig 2.The Ifi203c, Ifi202b, Ifi203b, Ifi202a, Ifi203a, Ifi204 and Ifi201 genes are positioned on contig 1. Determination of the contig positions is described d above.

Because of the extensive sequence similarity between the Ifi202a and Ifi202b exonic and intronic regions, and between the Ifi203a and Ifi203b exonic and intronic regions, the sequence of these regions revealed the precise distribution pattern of the

genes present in this cluster. A DOTTER comparison between contig 1, and itself (Fig. 3.24) revealed that a genomic region of approximately ~60 kb, containing the Ifi202a and Ifi203a genes, was duplicated to form the region of the cluster that contains the Ifi202b and Ifi203b genes.

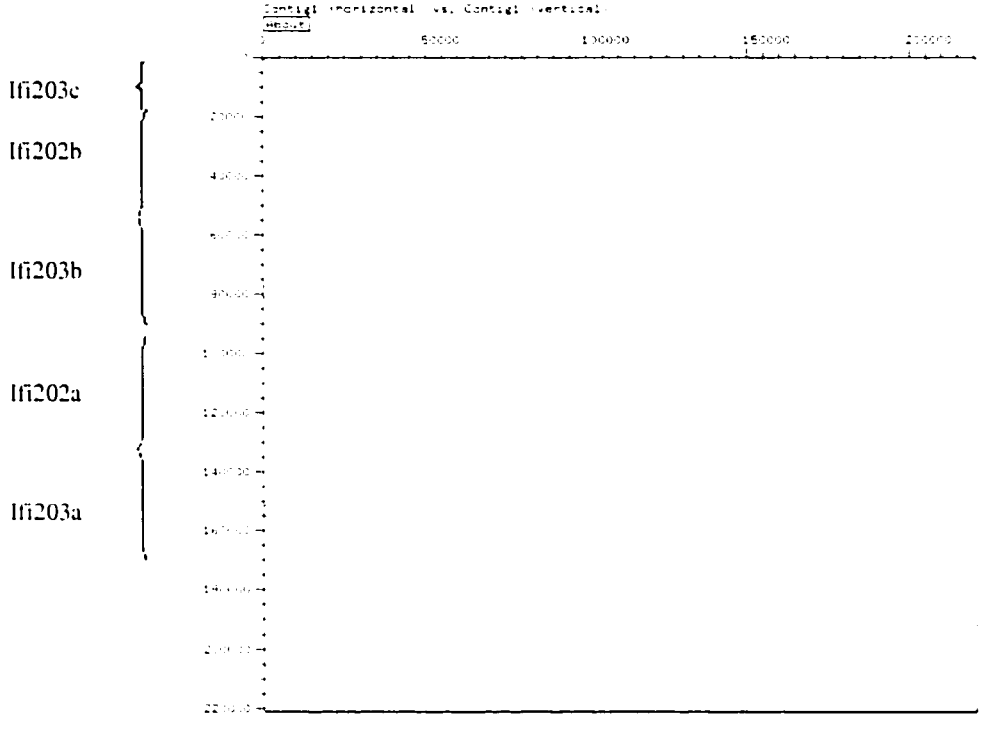


Fig. 3.24: Dot-plot analysis, using DOTTER, of contig 1 (horizontal) and itself (vertical). The positions of the Ifi202a, Ifi203a, Ifi202b, Ifi203b genes and the last three exons of the Ifi203c genes are indicated on the vertical axis.

Another DOTTER comparison between contig 1 and contig 2 (Fig. 3.25) revealed that a portion of the same genomic region, ~40 kb containing the Ifi202a gene and the first five exons of the Ifi203a gene was duplicated to form the cluster containing the "203-like" and Ifi202c pseudogenes. It appears therefore that the genomic region containing the Ifi202a and Ifi203a genes was duplicated at least twice during the evolution of this mouse Ifi200 gene cluster. Although the physical basis of this

duplication remains unknown, it could be related to the unusual amount of L1 repeat elements in this genomic region.

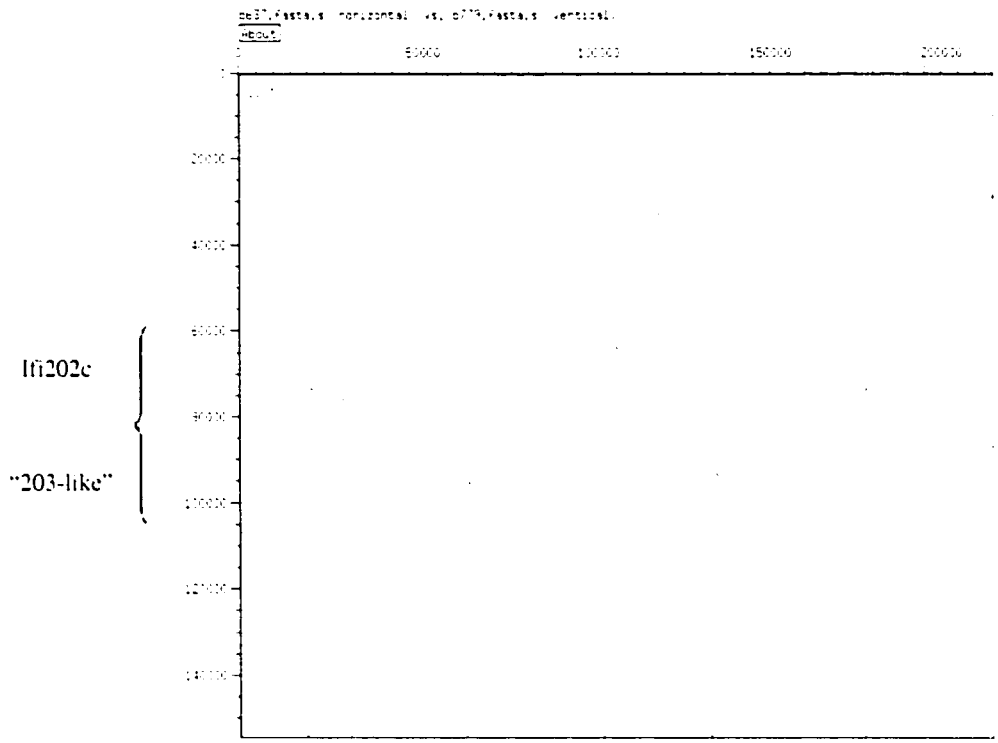


Fig. 3.25: Dot-plot analysis, using DOTTER, of contig 1 (horizontal) and contig 2 (vertical). The positions of the Ifi202c pseudogene and the "203-like" putative gene are indicated on the vertical axis.

III-1-9 Comparative analysis of the mouse and human Ifi200 gene clusters

The comprehensive annotation and analysis of a DNA sequence generally is based on two approaches. The first approach is based on the comparison of genomic sequences with other genomic or cDNA sequences or ESTs. Although this approach can find, and often identify, a high percentage of orthologous or paralogous genes in a

given sequence, it obviously is limited by the amount of data present in these databases. The second approach involves heuristic exon prediction programs. These programs are based on typical exon characteristics, such as codon usage, hexanucleotide distribution and consensus exon-intron splice sites. While they give a useful initial identification and characterization of a given sequence, they also have the tendency to produce large amounts of false positive predictions. Similarly, programs used for the prediction of promoters and other regulatory elements also give a high number of false positive predictions. To limit false positives, sequences regulating gene expression are predicted on very short sequences, where the probability of having regulatory sequences is strong.

The recent completion of the human genome sequence and the progress of the mouse genome sequence allow identification of new genes and regulatory elements based on their cross-species conservation. The premise of this approach is that, while non-coding (intronic and intergenic) regions exhibit variable levels of sequence similarity between species, coding regions and other functionally important regions are usually strongly conserved. The detailed analysis of sequences in conserved segments may help our understanding of the genomic organization of important genes and suggest candidate regulatory regions likely to carry important biological functions. Several comparative studies on a number of regions in the mouse and human genomes already showed the value of this approach. Footz *et al.* (2001) identified two putative genes in the cat eye syndrome critical region on human chromosome 22, in the absence of any EST hits, by comparing segments of human genomic sequences with segments of mouse genomic sequences from the region of conserved synteny on mouse chromosome 6. Also, the comparative genomic sequence analysis of the human and mouse cystic fibrosis transmembrane regulator (CFTR) genes by Ellsworth *et al.* (2000) proved to be a useful tool for the structural characterization of the CFTR gene in mouse and human and of potential sequences regulating their expression in both species. Finally, the comparative genomic sequence analysis by Mallon *et al.* (2000) of

the human and mouse Bpa/Str region located on the X chromosome of both species and encompassing a pair of murine X-linked dominant disorders aided the identification of at least four putative conserved genes in the region.

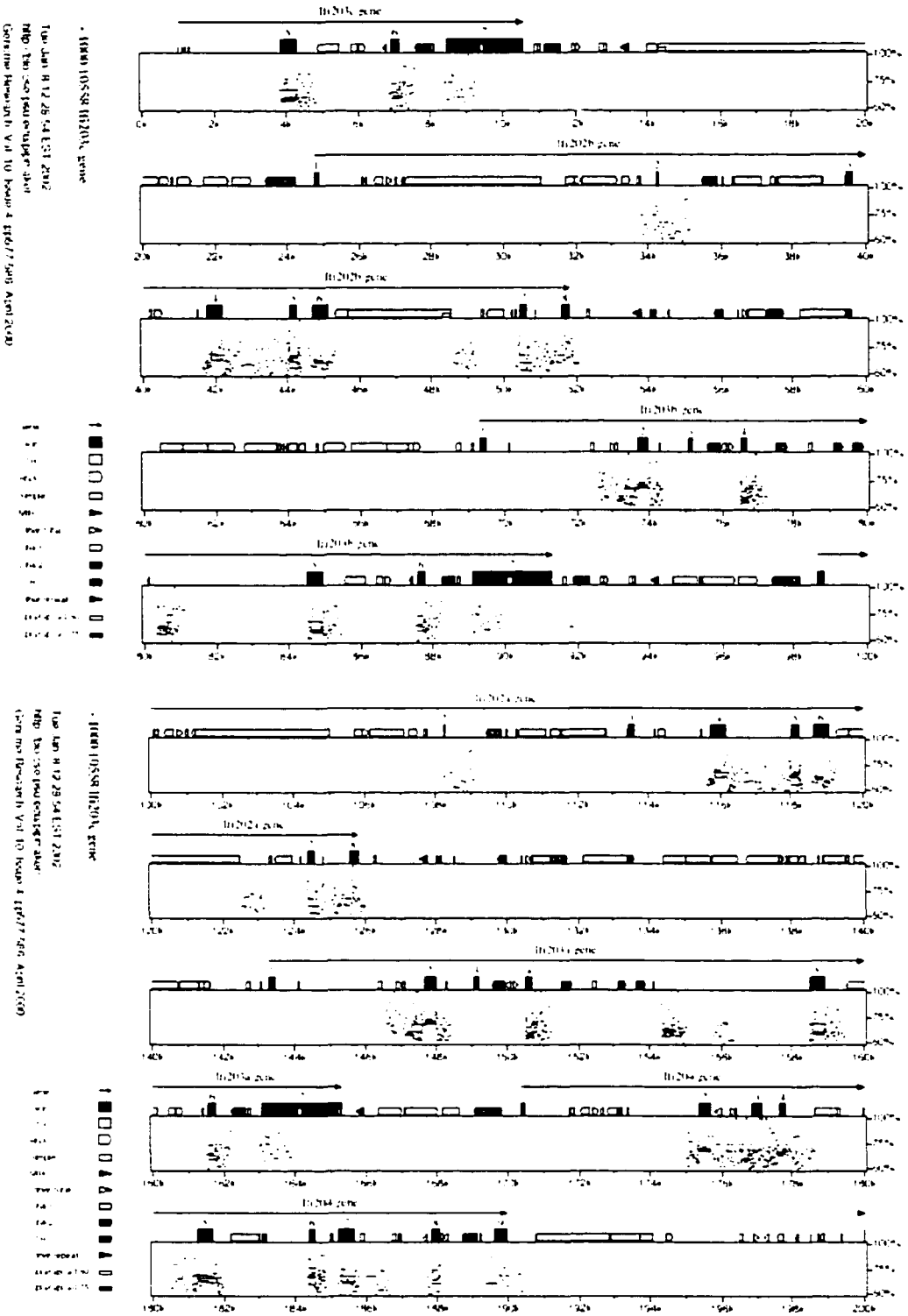
As discussed in the Introduction chapter, the human counterpart of the mouse Ifi200 gene cluster is made up of three genes whose expression is under complex controls. The first gene encodes the Myeloid Nuclear Differentiation Antigen (MNDA). Partial sequencing of the purified protein followed by cDNA cloning revealed that this protein of 403 amino acids includes one 200 amino acid domain (domain *a*). The amino terminus of this protein is homologous to the amino termini of the IFI-16, p204 and D3 proteins and these four proteins also share a potential nuclear localization sequence (AQKRK). In addition, MNDA has a short unique sequence in which 12 out of 23 amino acids are identical with those in a sequence of the interferon-inducible transcriptional repressor IRF-2 protein. The second gene encodes the 729 amino acid IFI-16 protein. IFI-16 includes two 200 amino acid domains (domains *a* and *b*). The two domains are separated by a unique 116 amino acid spacer domain, whereas, in the case of p202a and p204, the two domains (*a* and *b*) are contiguous. The amino terminus of this protein is homologous to the amino termini of the MNDA, p204 and D3 proteins. Also, IFI-16 shares with these proteins a potential nuclear localization signal (AQKRK). The genomic organization of the IFI-16 gene (Trapani *et al.*, 1994) has 10 exons covering at least 28 kb of genomic DNA. Each of the two domains (*a* and *b*) are encoded by specific exons: exons 4 and 5 for domain *a*, exons 7 and 8 for domain *b*. The 116 amino acid spacer domain is encoded by exons 5 and 6, indicating that exon 5 encodes for both a portion of the *a* domain and a portion of the spacer domain. The third gene encodes the 344 amino acid AIM2 protein. AIM2 includes only one 200 amino acid domain (domain *a*) and shares strong sequence similarity with 200-family proteins, such as IFI-16, MNDA and p204. Unlike the other members of the 200-family of proteins, which are primarily located in the nucleus (IFI-16, MNDA and p203) or are

detected both in the cytoplasm and the nucleus (p202), cell fractionation followed by western-blot analysis showed that AIM2 is primarily located in the cytoplasm (Choubey *et al.*, 2000). PsortII analysis further predicts that AIM2 is localized in cytoplasmic membranous organelles, including mitochondria.

The genes encoding MNDA, IFI-16 and AIM2 were mapped using various techniques, including pulse field gel electrophoresis and fluorescence *in situ* hybridization (FISH). The three genes are localized to human chromosome 1q21-23 in a region that is conserved in human and mouse. Initial analysis of this region using GENSCAN and BLAST indicates that the MNDA and IFI-16 genes are in the same orientation, whereas the AIM2 gene is in the opposite orientation, i.e., is located on the complementary strand of DNA. This location of the MNDA, IFI-16 and AIM2 genes reveals that the mouse Ifi200 gene cluster and its human counterpart are located within a large linkage group conserved between human and mouse.

III-1-9-1 Human/mouse cluster genomic sequence comparison

It is interesting to note that, while the mouse Ifi200 gene cluster consists of at least eight genes, two pseudogenes and maybe an additional Ifi200-related gene (the seven genes, the two pseudogenes and the "Ifi204-like" putative gene discussed in this dissertation research plus the Ifi205 gene, encoding the D3 protein), its human counterpart likely only contains three genes (MNDA, IFI-16 and AIM2). The region of genomic DNA encompassing these three genes has been previously sequenced (Genbank accession number AP002534). Although the reasons behind this structural discrepancy remain to be established, it is possible to compare the human and mouse cluster genomic sequences using the approach depicted in the percent identity plot (PIP) shown in Fig. 3.26.



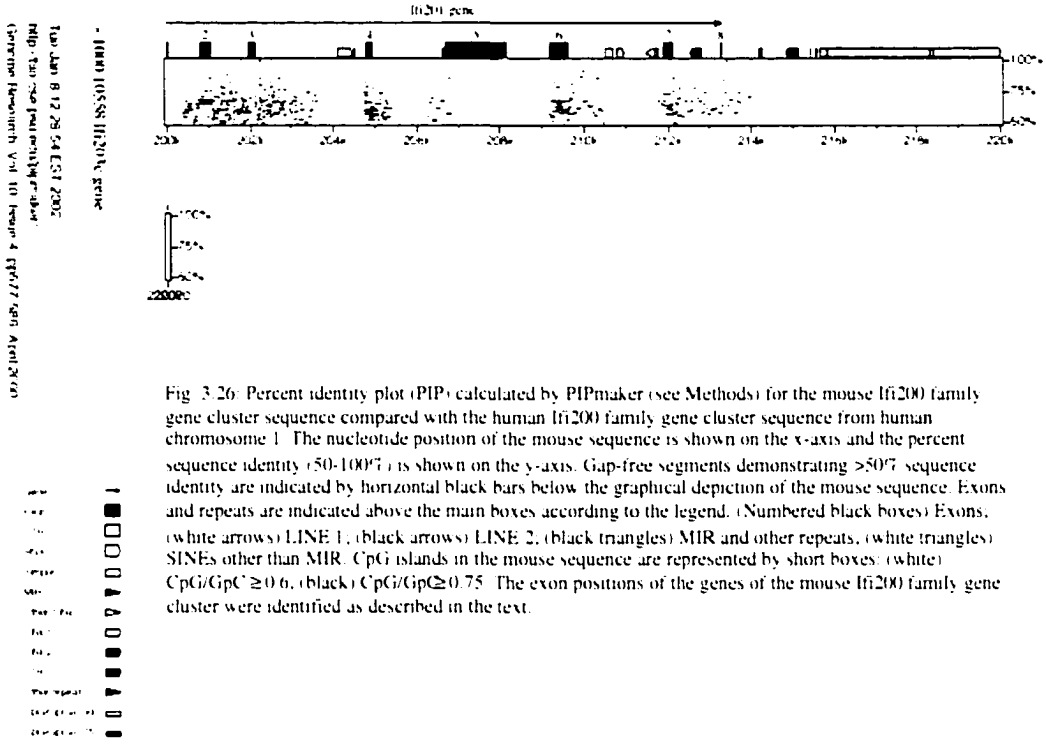


Fig. 3.26: Percent identity plot (PIP) calculated by PIPmaker (see Methods) for the mouse Ifi200 family gene cluster sequence compared with the human Ifi200 family gene cluster sequence from human chromosome 1. The nucleotide position of the mouse sequence is shown on the x-axis and the percent sequence identity (50-100%) is shown on the y-axis. Gap-free segments demonstrating >50% sequence identity are indicated by horizontal black bars below the graphical depiction of the mouse sequence. Exons and repeats are indicated above the main boxes according to the legend. (Numbered black boxes) Exons, (white arrows) LINE 1, (black arrows) LINE 2, (black triangles) MIR and other repeats, (white triangles) SINEs other than MIR. CpG islands in the mouse sequence are represented by short boxes: (white) CpG/GpC ≥ 0.6, (black) CpG/GpC≥0.75. The exon positions of the genes of the mouse Ifi200 family gene cluster were identified as described in the text.

This graphical overview facilitates identification of conserved exons and shows the arrangement of repeat elements and CpG islands. As presented in Fig. 3.26, the majority of the mouse 200 gene family exons from all seven different genes shows some degree of sequence conservation while there is an absence of conservation in 5' and 3' flanking, intronic and intergenic regions. The absence of conservation in the 5' flanking region of the genes of the mouse Ifi200 cluster is surprising, suggesting a mode of transcriptional regulation that may differ, at least partially, from that of the human Ifi200 cluster genes. Interruption in sequence alignments mainly is observed in repeat-rich regions. However, two small intronic regions did show sequence conservation with the human 200 family gene cluster, indicating either a duplication event or a biologically relevant sequence. The PIP indicates that all genes in the region have diverged at a similar rate (the amount of percent identity is similar for every aligned

segments). However, several exons, including exons 1 and 3 of Ifi202a, Ifi202b, Ifi203a and Ifi203b, as well as exon 1 of Ifi204, did not show any sequence conservation with the human sequence. In the case of Ifi202a and Ifi202b, significantly more sequence conservation is observed with the exons located downstream of exon 3 in these genes. This likely is due to the observation that although exon 3 is the first coding exon for Ifi202a and Ifi202b, the partially conserved 200 amino acid domains that are needed for the function of the proteins encoded by Ifi202a and Ifi202b (Wang *et al.*, 1999) are encoded by exons 4 and 5 for domain a and exons 6 and 7 for domain b. In the case of Ifi204, exons 2, 3 and 4 encoding domains that are located upstream from the conserved 200 amino acid domains (encoded by exons 5 and 6 and exons 7 and 8, for domains *a* and *b* respectively) exhibit some sequence conservation, suggesting important functional domains located at the amino terminus of p204. In the case of Ifi203a and b, although only exons 5 and 6 encodes the conserved 200 amino acid *b* domain, exons 2 and 4 exhibit some sequence conservation, again suggesting important functions for domains located at the amino terminus of the protein. There also is very little human-mouse sequence conservation of the non-coding exon 1 of Ifi204 and the coding exons 7 of Ifi203a, b and c, and exon 5 of Ifi201 with other exons in the 200 family genes. Finally, further examination of the PIP clearly reveals aspects of gene duplication, since many of the exons on the mouse 200 family gene cluster exhibit conserved sequences. For example, exon 2 of Ifi203a and Ifi203b shows three distinct homologous regions with a percent identity of approximately 75% (Fig. 3.26).

The human and mouse 200 family genomic sequences also were compared by a Dotplot comparative analysis (Fig. 3.27) (Schwartz *et al.*, 2000). A repetitive pattern was observed as human genes IFI16, MNDA and AIM2 show regions of sequence conservation with each of the seven genes of the mouse 200 family gene cluster. Three non-coding regions of the human cluster also exhibited a similar pattern and might correspond to putative pseudogenes present in the human cluster. Initial BLAST

analysis reveals that at least one of these putative pseudogenes also shares sequence identity with the human gene IFI16.

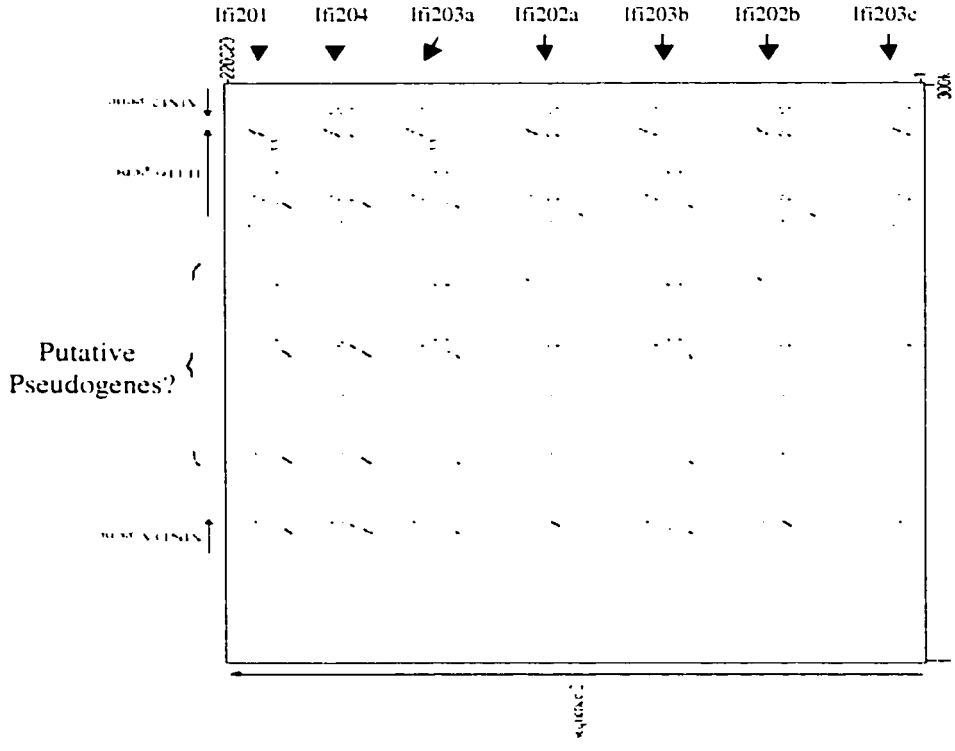
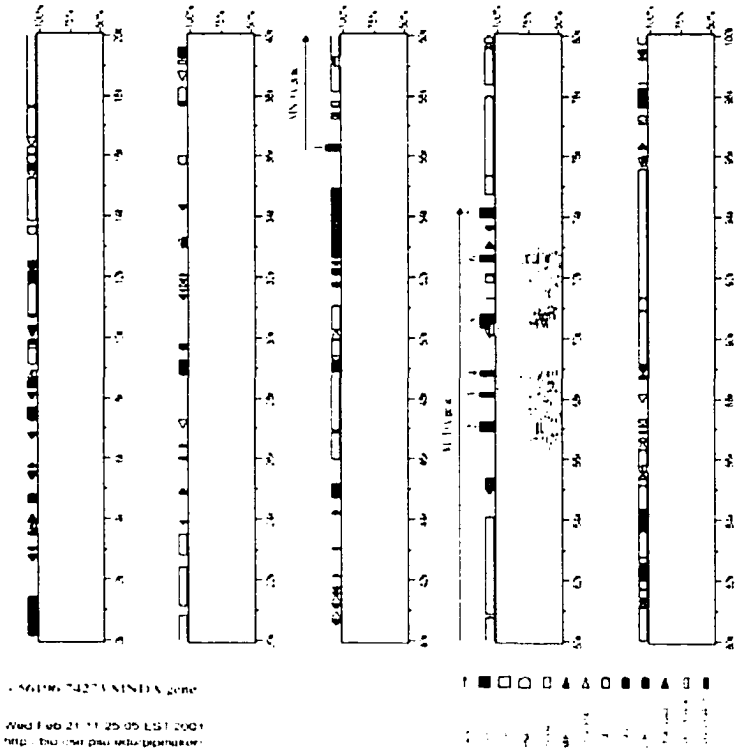
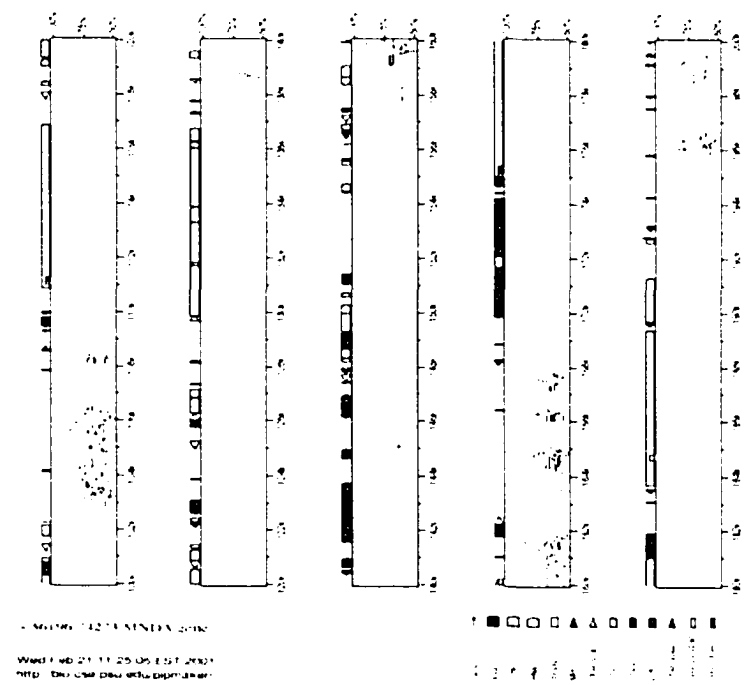


Fig. 3.27: Dotplot comparative scale of mouse and human cluster sequences. The Dotplot scale was generated using BLASTZ, a high stringency implementation of BLASTN (Schwartz *et al.*, 2000). Conservations in the Dotplot are represented with black dots. The genes of the mouse and human clusters are indicated on the x- and y-axes, respectively, and were identified using CROSSMATCH based on the available cDNA sequences.

The PIP (Fig. 3.26) also showed that only the exonic regions of the mouse cluster exhibited sequence conservation with regions of the human cluster. Therefore, another PIP was generated, in which the human and mouse sequences were aligned and the alignments converted into segments of percent identity relative to position in the human cluster. This PIP (Fig. 3.28) showed that the exonic regions of IFI16, MNDA and AIM2 are similar to exonic regions of each of the seven mouse 200 family genes.





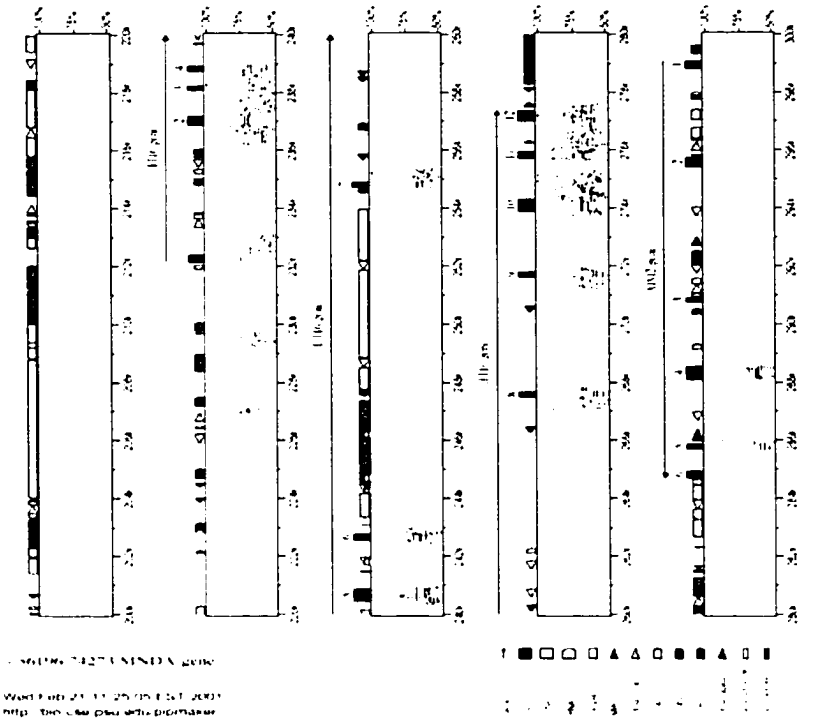


Fig. 3.28. Percent identity plot (PIP) calculated by PIPmaker (see Methods) for the human Ifi200 family gene cluster sequence compared with the mouse Ifi200 family gene cluster sequence. The nucleotide position of the human sequence is shown on the x-axis and the percent sequence identity (50-100%) is shown on the y-axis. Gap-free segments demonstrating >50% sequence identity are indicated by horizontal black bars below the graphical depiction of the human sequence. Exons and repeats are indicated above the main boxes according to the legend. (Numbered black boxes) Exons; (white arrows) LINE 1; (black arrows) LINE 2; (black triangles) MIR and other repeats; (white triangles) SINEs other than MIR. CpG islands in the mouse sequence are represented by short boxes: (white) CpG/GpC ≥ 0.6; (black) CpG/GpC ≥ 0.75. The exon positions of the genes of the human Ifi200 family gene cluster were identified by CROSSMATCH comparison with the cDNA sequences.

In the case of IFI-16, whose genomic organization has been reported (Trapani *et al.*, 1994), both the exons encoding the conserved 200 amino acid *a* and *b* domains, and exons 2, 3, 4, 9, 10 and 12 of IFI-16 show sequence conservation with the mouse Ifi200 gene cluster sequence. Interestingly, exons 5 and 6 of the IFI-16 gene, which encodes the 116 amino acid spacer domain unique to the IFI-16 protein, show sequence conservation with the mouse Ifi200 gene cluster sequence. In the case of AIM2, only exons 4 and 5 exhibit sequence conservation, suggesting that they encode a functionally important domain of the AIM2 protein. Finally, in the case of MNDA, every exon, except the first and the last exon, show sequence conservation with the mouse Ifi200

gene cluster sequence, suggesting that the first and the last exons of the MNDA gene likely are non-coding exons.

These data can be related to the presence of one or two partially conserved 200 amino acid domains (domains *a* and *b*) in all the proteins encoded by the mouse and human Ifi200 gene clusters. The amino acid composition of the *a* domain expressed in different proteins is highly conserved, as is that of the *b* domains. For example, the sequence similarities between the *a* domains of p202a and p204, and between the *b* domain of p202a and p204 are 55% and 50% respectively, a higher level of conservation than those seen between the *a* and *b* domains of p202a (27%) and between the *a* and *b* domains of p204 (34%) (Lengyel *et al.*, 1995). This suggests an evolutionary duplication of an ancestral gene encoding a unique 200 amino acid domain. Johnstone *et al.* (1998) mapped the transcriptional repression domain of IFI-16 to the 200 amino acid *a* and *b* domains and later further demonstrated that the functional interaction of IFI-16 with p53 required a conserved motif (MFHATVAT) present in the 200 amino acid domains of all mouse and human 200-family of proteins (Johnstone *et al.*, 2000). Koul *et al.* (2000) demonstrated that this same motif was sufficient for the self-association of the p202 protein. Thus, although the function of these conserved domains still remains to be clearly established, it appears that the presence of this signature domain in all the proteins encoded by the mouse and human Ifi200 genes most likely reflects important conserved functions. Therefore, considering that both the human and mouse proteins most likely assume the same type of functions by modulating the transcriptional activity of targeted genes, it is not surprising that the exons encoding the *a* and *b* domains of the different proteins of the mouse and human Ifi200 gene clusters are conserved between both species.

In addition to the two 200 amino acid domains, all members of the 200-family of proteins, except p202a and probably p202b, show sequence conservation at their amino termini (Johnstone and Trapani, 1999) with the presence of a putative nuclear

localization motif and a series of highly conserved leucine and basic residues which are thought to be involved in a homo- or heterodimerization process (Xie *et al.*, 1997). Although the absence of sequence conservation for the exons located upstream from exon 4 of the AIM2 gene remains to be explained, it is interesting to note that the only protein without a conserved amino terminus, p202a, does exhibit a significant amount of sequence conservation for the exons encoding domains *a* and *b*.

Taken together these data indicate that, although there is a high divergence rate between the mouse and human genes of the Ifi200 cluster, the observed sequence similarity between the mouse and human genes likely reflects some common functions as several domains are conserved.

III-1-9-2 Human/mouse cluster 5' flanking sequence comparison

The 5' flanking region of the MNDA and the IFI-16 genes have been examined by Kao *et al.* (1996) who identified two interferon alpha/beta-stimulated response element (ISRE) flanking a multiple transcription start site region identifying MNDA as a TATA-less interferon-regulated gene. Other DNA elements including a cluster of Myb sites, several Ets, an Ets-related PU.1 site and a Sp1 site also were identified within 600 bp of the transcription start sites. However, analysis of deletion mutants by Kao *et al.* (1997) indicated that the Sp1 site, but not the Ets or the PU.1 sites, contributed to the restricted expression of the MNDA gene. Trapani *et al.* (1994) identified two ISREs in the 5' flanking region of the IFI-16 gene that bear only limited sequence similarity to other previously described ISREs. Although MNDA can be induced by interferon alpha and IFI-16 can be induced by interferon gamma, two regions that could correspond to interferon gamma activation sites (GAS) were identified in the 5' flanking region of the IFI-16 gene. A consensus Ets binding site overlapping its ISRE occurs in the 5'

flanking region of the IFI-16 gene. Surprisingly, although they are located in the first 1000 bp immediately upstream from the first exon followed by the first exon of the MNDA and IFI-16 genes, none of the ISRE and GAS mentioned above were identified using MOTIF. Interestingly, all of these interferon-dependent motifs, as well as the ISREs in Ifi201, Ifi202a, Ifi202b and Ifi204, are located within the 5' untranslated exons of their respective genes, rather than upstream from the transcription start sites.

A CLUSTALW alignment was performed between the first 1000 bp immediately upstream from the first exon followed by the first 200 bp of the first exon of the Ifi202a, Ifi203a, Ifi204, IFI-16 and MNDA genes (Fig. 3.29).

```
IFI16_5_        ------------------------------------------------------------
MNDA_5_         TATGCTGCTGGTTTGGATTAAGACAATCAGCTTTATTCTTCTCTGTAGATGGCCTTGGAT
202a5           ------------------------------------------------------------
2045            ------------------------------------------------------------
203a5           ------------------------------------------------------------

IFI16_5_        ----------------------------CCTTTTCCACTCATGACCACATCCTGCTTGCA
MNDA_5_         CATTTTCCCCTATTAGCTGTTTTGGGATCCTTTTCCAGTCATGACCACATTTTGCTTACA
202a5           ------------------------AAGAGAGAAGCTAGAAAAAGAAAGAGAGAGGAGGGA
2045            ----------AAATGCAGGGGTATAAGAGAGAAGCTAGAAAAAG--AGAGAGAGGAGGGA
203a5           -------CTGGTTTATTGAACACACACCCCAAGACTGATTGAAAAGGGGCTCAGCTCAGA
                                                *                 *     *

IFI16_5_        GATCTG--CTGCTGGGTTAACACCACATCATCTGCTAGTTGATGCCATTTATTTGACATC
MNDA_5_         GATCAG--CTAGTTGGTAACCACAGCACTGTCTGCTAACTGATACCATTCATTTGACAAT
202a5           GAGAGA--GAAAAGGGAGGGAGGGAGAGGGAGAGAGAGAGAAGGCTAGAGAGAAAGAGAG
2045            GAGAGA--GAAAAGGGAGGGAGGGAGAGGGAGAGAGAGAGAGAAGGCTAGAGAGAAAGAG
203a5           TACAAGAGCTATTTTCCAGGGTCAGTTTATATAGGCAGAAAATCATAATGAGCTCATTAG
                 *                                 * *           *

IFI16_5_        CTGGGGTTACAGCACGCTAAAGGGCACCCACTGTCCTTAACAGAGAAAAATACACTGCTG
MNDA_5_         TCAGAGTTATGATACTTTGTAGTGCACCCACTGTTCTTCACAGGAAAAAAAA-ATTGCAG
202a5           GAAGAGAGAGAGAGGAGAGGAGAAGACAAAGAGAAAGGAAAGAGGAGAGAAGAGTGAGAG
2045            TAAGAGAGAGAGAAGAGAGGAGAAGACAAAGAGAAAGGAAAGAGGAGAGAAGAGTG----
203a5           CGGGTGAGGGGAGGGCTGTAGGGTGGTCAACCTTTTTCCAAGTCTTTTTTTTTCAAGCTG
                   * *                 *       *         *

IFI16_5_        AGCAGAGGATTAAACAGACTCCTCACACTACCGGAAGTTCAGTGCATTTTCCAGGAT--T
MNDA_5_         TACAAAGGGTTTATCAGGCCCCCTATCATTCCAGAGTTTCAGTGC-TGTTCCAGAAT--G
202a5           AAGTGAGAGTTAAGAGAACAA-GGAGTAAGAGAAGAGACAAGTAAGATAGCAAACGG--G
2045            -----AGAGTTAAGAGAACAAAGGAGTAAGAGAAGAGACAAGTAAGATAGCAAACAGAAG
203a5           TCCTCTATCTTTCTTTGCTCTGAGTGAAGGCTATGGGTTGGTAATTTCTGATGATGATGA
                          **

IFI16_5_        TTCAA--GTCTTCGATGCTGTCATGTTGAGAGTCCACTATCCACGTTTGTTTATCCAGGA
MNDA_5_         TTCAA--GGTTTCA-----GTCATGATATGAACCAACTATCCATA-------------A
202a5           CTGAACAGCCCTTTTTTATGGTCTTTACTGTTGCTAGGTAACTGGAGAGGA--GTTTAGCC
2045            CTGAACAGCCCTTTTTTATGGTCTTTACTGTTGCTAGGTAACTGGGGAGGA--GTTTAGCC
203a5           CTATTTAAGCCTCAGAATATGGTCAACTTGACCTTTCTTTCAGTCAAACT---TTTTCTT
                 *         *                       *  *
```

```
IFI16_5_        ATAATGCAGAAAA--TCTGAGCATTCGTGAATCTAATCATTGAGGTAAAGTAACCTAGTC
MNDA_5_         ATAATGCAAAAAG--CATGAGTAATCATGAATTTAATAATCAAGGTAAGGCAGCTCAGTT
202a5           TGAAGGTCAGAAG--CTTGGGCCATTGCCTACATGACT-TCTACCCATGCTTCTCTTGTG
2045            TGAAGGTCAGAAG--CTTGGGCCATTGCCTACATGACT-TCTACCCATGCTTCTCTTGTG
203a5           TTAATTTTAAAATGTTTTTATCTTCTTTTTTTCTTTTTCTTTTTATTAGATGTTTTCTTT
                  **      **     *                *     *                  *

IFI16_5_        AT---GGTCAAGAGGA-CAGCT--AGTGTTAAACTCC---TTCATATAAATTTAAGGAAG
MNDA_5_         ATTATACTCAAGAGAG-CAACT--AGTATGAAACTCT---TTCAAATTA-TTTGAATAAG
202a5           GGGTCTGTGGGGGGAGGTAACT--TAGGCAGGAGCCAGAGTTCCAGCAGCATAAGGGAAT
2045            GGGTCTGTGGGGGGAGGTAACT--TAGGCAGGAGCCCGAGTTCCAGCAGCATAAGGGAAT
203a5           ATTTACATTTAGGATGTTATCCCCTTTCCTGGTTTCCCCACTGAAGACTCCCTATCCCCT
                       *   *     * *              *     *

IFI16_5_        GATTTACATATT-CATAATGACTTTCTGA-TCTGGAAAGC-CCAGGCTTGTCAGTTATTA
MNDA_5_         GATTTACATATT-CATAATGACTTTCTGCCTCTGTCAAGCTCCAGGTGTATCAGTTATTA
202a5           GCCTACTGTGTCATGTAGCTGAATTATCACCCTTTGGCGT-TCAGACCTCAGCTCAACTG
2045            GCCTACTGTGTCATGTAGCTGAATTATCACCCTTTGGAGT-TCAGACCTCAGCTCAACTG
203a5           CCTCCCTCCATCTGCTCACGAACCAACCC-ACACTGGCTTCCTGGCCCTGGCATTCCCCA
                          *   *              *              *   *


IFI16_5_        ATAATTGGA-TGAATACT--TCAGGTAGTATTTAAACAAAT-AGCGATTTCATAGCAGAC
MNDA_5_         ATAACTGGG-TGAATATG--T-AGGTACTATTTAAACAAA--AACAATTTCACAGTAGAC
202a5           GAGACCAGCCTGTCTGTG--CAAAACCTCAAAACCACAATT-CTCTGGCTTCTGAAGGGC
2045            GAGACCAGCCTGTCTGTG--CAAGACCTCAAAACCACAATT-CTCTGGCTTCCAAAGGGC
203a5           ATACTGGCTGTGCCTACAGTCAAGCTGCTTTTGTGTCAAACAAGCTGGCAGGCATGAAAC
                        **  *       *              ***     *              *

IFI16_5_        ---ATTTCCCCAACCAGGAGATGCAGTAAAGTATTGGT--AGCAATAGATGCTC--CTAC
MNDA_5_         ---CCACCCCTATCCAAGAGATGAAGTCAAGTTCTGGT--AGCAATAGACATCT--CTAC
202a5           TTGATACTTGTAAACCCTGAGTGTATTAGACTCTTCCCTGGGCCTTGAGTACTTTACTAC
2045            TTGATACTTGTAAACCCTGAGTGTATTAGACTCTTCCCTGGGCCTTGAGTACTTTACTAC
203a5           ---AAAAGTATTACAGCTAAG-GTAGCAGGCTTCATCAATTCCAAGGAACAACAGACTCT
                               * *      *           *                 **

IFI16_5_        CAAGATCTCCTCCT--GCCACTGCCACTCTGTGCAC--AGGAAAGGAGGG-GAGAAGAAG
MNDA_5_         CAAGATCTCCT-----GCTCCTGCCACTCTGTGAAC--AGGAAAGGAGGGAGAAACTTTG
202a5           AATAAAAACTCC----ACCTTAGCTGTCCAGGACTCCTATCTAAGTAGAGCAGAACCTTG
2045            AATAAAAACTCC----ACCTTAGCTGTCCAGGACTCCTATCTAAGTAGAGCAGAACCCTG
203a5           GAGGAGCTCCCCTCTGACCACTTTCATACTGAACAAACAAAGAAGGGTGA-GGTGGCTGA
                 *  *   *          *           * *        *   ***

IFI16_5_        CCATGTTGCTTCTC-----TTATATTT-GAGCATCATTTAAGGGAATATAAGTCTCTAAT
MNDA_5_         GTGTTTTGCCTTTCCCAATTTATATTAAGATCCTTATTCAAGCGGCTACCGGTTTCTAAT
202a5           GTCAGCTAGTTTTAACACCTTGATCTGGTTGTACCATGTATGTGAAT-TTTGTGTCTAGT
2045            GTCAGCTAGTTTTAACACCTTGATCTGGTTGTACCATGTATGTGAAT-TTTGTGTCTAGT
203a5           GTGACTCGCTTGTATTTCCTTGTGATT-CATCTTCTTTCAGATGAGA-GTGATTCCTAGT
                          * *      **    *          *  *   *        *  *** *

IFI16_5_        GACTCACATGTAAGTCTGCTGGGTAAGAAAGTTCTAATGTTGGAACTCCTTTCAGCTATC
MNDA_5_         GACCTATATGTAAGTTAGCTTTATAAGAAAGTTCTGAGTTTGAAACTCCTTTCAGCAAGT
202a5           GGCCAG--TGTATATCTGTTTGCTATAAAAGTTGTAAGGTTAGGGCTT-TCTTATTCAGT
2045            GGCTAG--TGTACATCCGTTTGCTACAAAAGTTGTAAGGTTAGGACTT-TCTTATTTAGT
203a5           GACCCACCCCTAAGTCTACTGGACATAAAACATGCAGTGTTGCAATTTCTCTC-TCTCCT
                * *       **  *    *    *  ***  *      **     *  * *

IFI16_5_        GAATTTTGTAAGAAAAAAAACTTATTTGTGTTTATATACATTTATGAGATACAAGTGTAA
MNDA_5_         A---TTTGTAAGTCAAAAAAGCGAAGAAGAGGCAC---CATTTAC----TGTCATT-TCT
202a5           GAGATTCATCAGAAAAGAATGTGAGCC------TG----ACTAAGC---TGTGATT--TT
2045            GAGATTCATCAGAAAAGAATGTGAGCC------TG----ACTAAGC---TGTGATT--TT
203a5           GTG-TGTGTGTGTTT-GTGTGTGTGTG-----TG----TGTGTGTG---TGTGTGTA-TG
                     *  *  *                                    *     *
```

```
IFI16_5_        TTTCGTTACATGGAT-ATATGCCATATTGGTGAAGTCAGAGATTTC--AGTGTGCACATC
MNDA_5_         CCTTGTGACTGAGATTACTTACTACAAAGAGGAACTGACCACTTTC--TCTG---ACATC
202a5           TTTTTTCATAGTGAC--CACAGTGTTTACAGGAAGTTAAGATGTTT--CCAAA--GTCAG
2045            TTTTT-CATAGTGAC--CACAGTATTTACAGGAAGTTAAGATGTTT--CCAAA--GCCAG
203a5           CTTCTTTCTAGAAAT--TACATTAATAACAGGAAATAAAGTGCTCTATTTTAAGAGCTGG
                   *          *                  *** * *     *

IFI16_5_        ACCCGAAAAATGTTAACTGTACCCATTAAAGTAATTTCTCATCCCCCATTTCCCCTCACC
MNDA_5_         TGAAATAGAAAGCTACAAAGGCTGATAGGCGTGGCTTCAAAAT-----TGTTTCATGATT
202a5           GGAAATTGAAAGCTATGAACGAAACTGGGAGGAGGCCTGGCTGATTCAGCACTCCTCACC
2045            GGAAATTGAAAGCTATGAACGAAACTGGGAGTAGGCCTGGCTATTTCAGCACTCCTCACC
203a5           GGAAATCGAAAGCTATGAACAGTGCTGGGAGGAGCCCAATTTC----ATCTTAAGTCTCC
                        ** * **          *    *                        *

IFI16_5_        ACATGCTCCAGCCCTTGCCAGGAAACTGTTCA-TTTTCTCTGACTAACAGAAACGAAAGC
MNDA_5_         ACATAACTTCTTTCTAGTTAGGAATTAGGACAGTGTAAGAAGGCCATCTACAATCAAGAT
202a5           CAATGGAGACAGCACAGGGAGAGAC--AGCCAGTGCTGTTCAAGAAATGAAA--CAACTC
2045            CGAGGGAGACAGCACAGGGAGAGAC--AGCCAGTGCTGCTCAAGAAATGAAA--CAACTC
203a5           CCAGGAAGACAGCACACTGTAAGAC--ATCCAGCCCTGCTCAAGCTAGGAAC--TTACTT
                   *                    *      **                      *

IFI16_5_        TAAAAACACTGGTGGGAGGAGTCTCCACATTGTTTCCTACTCCATTT-TCTCTGGGGCAA
MNDA_5_         TGAGAGTGGCTCTAACAAGTG---CCATTTTTCCTTGTTAGCTTTCA-TTTCTCAGCCCT
202a5           TGAGAGTGTTGTAATCACTACCATCTTCCTTTACACCCAACTGTTCAGTTTCTCATTTAC
2045            TGAGAGTTTTGTAATCACTACCCTCTTCCTTTACACACAACTGTTCAGTTTCTCATTTAC
203a5           TAACCAGTTCTCCCGCCCCCCCCCC----------------------------------
                * *                  *

IFI16_5_        TAGCAGAATAGGA-GCAAGCCAGCACTAGTCAGCTAACTAAGTGACTCAACCAAGGCCTT
MNDA_5_         TTACAAGATTAAA-ATA-GTCTGCAGT--TTAATCTCTCCAAAGCTTTACGGACAGTGAT
202a5           TGACTTATCTGCCTACCTACTCAAGCCAAGCAGGCCACTTCTTGACCCGGTGAAGGTCTC
2045            TGACTTATCTGCCTACCTACTCAAGCCAAGCAGGCCACTTCTTGACCCGGTGAAGGTCTC
203a5           ---------------------------------------------------------

IFI16_5_        TTTTCCTTGTTATCTTTGCAGATACTTCAT--TTTCTTAGCGTTTCTGGAGATTACAACA
MNDA_5_         TCTGTCCTA--AACAAGACAGTGACTCCAGGATTTCTGAAGACTATTGTGGAAGA-AGCA
202a5           AGGATCTGTACATCACTGCAGAAATATCCAGGAAGGTGAGTTCTGTTTGTTTCTTAAACA
2045            AGGATCTGTACATCACTGCAGAAATATCCAGGAAGGTGAGTTCTGTTTGTTTCTTAAACA
203a5           ---------------------------------------------------------

IFI16_5_        TCC--TGCGGTTCCGTTTCTGGGAACTTTACTGATTTATCTCCCCCCTCACACAAATAAG
MNDA_5_         TCCATTAAGGTGAGATTTCTGGGAAGTTCT-----------------------------
202a5           TGTGATCAGAT-C--TTTCTGTGAAATATAAAGGACCATAGAAAAAAATCAGTGTTGAAT
2045            TGTGATCAGAT-C--TATCTGTGAAATATAAAGGACCATAGAAAAAAATCAGTGTTGAAT
203a5           ---------------------------------------------------------

IFI16_5_        CATTGATTCCTGCATTTCTGA
MNDA_5_         ---------------------
202a5           GGTGGAGGACCAAGGCT----
2045            GGTGGAGGACCA---------
203a5           ---------------------
```

Fig. 3.29: CLUSTALW alignment comparison between the first 1000 bp upstream from the first exon followed by the first 200 bp of the first exon of the IFI-16. MNDA.

Ifi202a, Ifi204 and Ifi203a genes. The different ISREs and GAS (see Text) are underlined.

No significant sequence similarity was observed between the 5' flanking regions of these genes. The different ISREs and GAS described by Choubey *et al.* (1989) for Ifi202a and Ifi204, by Trapani *et al.* (1994) for IFI-16 and by Kao *et al.* (1996) for MNDA and analyzed using MOTIF for Ifi203a were positioned according to their sequence on Fig. 3.29. Interestingly, they were contained in a very small area of the 5' flanking regions of each gene. In addition, ISREs described for the MNDA, Ifi202a, Ifi203a and Ifi204 genes, all induced by interferons alpha and beta, were located at similar positions in the 5' flanking region of each gene. This suggests that, despite a low level of sequence similarity in the region, the position of both ISREs in the 5' flanking region of the MNDA, Ifi202a and Ifi204 genes could play an important role in driving the protein interactions between the different elements of the interferon alpha/beta-induced transcription complex.

A CLUSTALW alignment also was performed between regions located upstream from the transcription start site in MNDA, Ifi202a and Ifi204 and downstream from the transcription start site in Ifi203a (Fig. 3.30).

```
202a5   - 138   CACAGTGTTTAGHGGHHGTAAGA---TGTTTCCAAAGTCAGGGAAATTGAAAGCTATG
2045    - 138   CACAGTATTTAGHGGHHGTAAGA---TGTTTCCAAAGCCAGGGAAATTGAAAGCTATG
203a5   + 60    TACATTAATAAGHGGHHHTAAAGTGCTCTATTTTAAGAGCTGGGGAAATCGAAAGCTATG
MNDA_5  - 88    CTTACTACAAAGHGGHHCTGACCA---CTTTCTCTGACATCTG-AAATAGAAAGCTACA
                              Ets sites                    ISRE sites
                    * *    * ***** * *       * * *        *  **** ********

202a5   - 82    AACGAAACTGGGAGGAGGCCTGGCTG
2045    - 82    AACGAAACTGGGAGTAGGCCTGGCTA
                          Ik-2 site
203a5   + 120   AACAGTGCTGGGAGGAGCCCAATTTC
                          Lyf-1 site
MNDA_5  - 32    AAGGCTGATAGGCGTGGCTTCAAAAT
                          Sp1 site
                **       * ** *  *
```

Fig. 3.30: CLUSTALW alignment comparison between regions located upstream from the initiation of transcription sites in Ifi202a, Ifi204, Ifi203a and MNDA. The positions of these regions, in base pairs, are shown related to the transcription start site position. The putative transcription factor binding sites are shown on the figure.

Although these regions are located at different positions of the 5' flanking regions of each gene, they all share some sequence similarity at positions that correspond to one ISRE and a putative Ets binding site (described by Kao *et al.* (1997) for MNDA. Interestingly, mutations of the putative Ets site by Kao *et al.* (1997) did not significantly alter promoter activity of the MNDA gene, indicating that this site likely is not involved in the activation of the MNDA gene, whereas mutations of the putative Sp1 site, located downstream from the ISRE of the MNDA gene (Fig. 3.30) reduced the overall promoter activity of the MNDA gene (Kao *et al.*, 1997), indicating that Sp1 is one of the factor involved in the specific expression of the MNDA gene. Surprisingly, the putative Ets site exhibits more sequence conservation between mouse and human than the Sp1 site despite the fact that Ifi204 and MNDA exhibit the same pattern of expression, restricted to cells of the myelomonocytic lineage (Briggs *et al.*, 1994a; Briggs *et al.*, 1994b; Gariglio *et al.*, 1998). However, a putative Ik-2 site and a putative Lyf-1 site were found in the 5' flanking region of the Ifi204 gene and Ifi203b gene, respectively, colinear to the Sp1 site in the 5' flanking region of the MNDA gene. Since Ik-2 and Lyf-1 are important transcription factors for hematopoietic cell differentiation (Kerhl, 1995), they could modulate Ifi204 and Ifi203b expression in hematopoietic cells, just as Sp1 does for MNDA gene expression. However, whether the putative Ets site is involved in the transcriptional activity of the Ifi202a, Ifi203a and Ifi204 genes remains to be established.

Another CLUSTALW alignment was performed between regions upstream from the transcription start sites in MNDA and IFI-16 (Fig. 3.31).

```
MNDA_5    - 88  CTTACTACAAAGAGGAACTGACCACTTTCTCTGACATCTGAAATAGAAAGCTACAAAGGC
IFI16_5   - 88  CAGCCCTTGCCAGGAAACTGTTCATTTTCTCTGACTAACAGAAACGAAAGCTAAAAACAC
                *   *       * *****  ** **********      **  ******** ***  *

MNDA_5    - 27  TGATAGGCGTGGCTTCAAAAT
IFI16_5   - 28  TGGTGGGAGGAGTCTCCACAT
                ** * ** *  *  ** * **
```

Fig. 3.31: CLUSTALW alignment comparison between regions located upstream from the transcription start sites in MNDA and IFI-16. The positions in base pairs, indicated in the right, are relative to the transcription start site position. The ISREs are underlined, and the Sp1 site is italicized.

MNDA is induced by interferons alpha/beta and IFI-16 by interferon gamma, and has a unique pattern of expression (see Introduction). However, this region contains two ISREs and one Sp1 site, which directly contributes to the restricted expression of the MNDA gene (Kao *et al.*, 1997) and exhibits sequence variations that could account for the differences in interferon response and constitutive expression of these two genes. Interestingly, the ISREs that account for interferon alpha/beta-induced expression of the MNDA and IFI-16 genes are in a region that appears to be relatively well conserved for both genes, even though each ISRE induces MNDA and IFI-16 gene expression in a different manner. Because of the relative sequence conservation in this region, it remains to be established whether the induction of MNDA and IFI-16 gene expression by these ISREs could play a role in establishing lineage/stage specific expression for both genes.

III-1-9-3 Human/mouse cluster repeat sequence comparison

Analysis of the repeat elements in the 300000 bps of the contig containing the human MNDA, IFI-16 and AIM2 genes (Genbank accession number AP002534), using RepeatMasker, showed that they represent ~52.3% of the sequence (Table 3.21). Similarly to the mouse Ifi200 gene cluster, LINE1 family elements represent an estimated ~64% of the repeat elements. Alu repeats, which represent an abundant repeat

element in human, make only ~10% of the total repeats present in this human contig, whereas ~20% of the repeat elements are LTRs. Inspection of the distribution of LINE1 elements throughout the human Ifi200 gene cluster, using PIPmaker (Fig. 3.27), shows that they are equally distributed and mostly present within intergenic and long intronic regions. This pattern of distribution of LINE1 elements is similar to the one observed throughout the mouse Ifi200 gene cluster, indicating that their appearance occured prior to mouse-human speciation. In addition, the low percentage of Alu repeats in the human Ifi200 gene cluster is similar the low percentage of B1 and B2 repeats in the mouse Ifi200 gene cluster. This observation suggests that these SINEs could have been selectively eliminated from this region before mouse-human speciation. Because SINEs, rather than LINEs, are present mainly in gene-rich regions, while LINEs mainly occur in gene-poor regions, it remains to be explained why such a high percentage of LINE1 elements are present in this region. It is possible that these LINE1 elements give a selective advantage to organism containing them in the Ifi200 gene cluster. A similar pattern of distribution of LINEs and SINEs also has been observed in a relatively small gene-rich region in mouse and human (Mallon *et al.*, 2000) and might represent a general mechanism of chromosomal rearrangements in these particular gene-rich regions with a relatively low G/C content. Finally, it is interesting to note that the percentage of sequences occupied by repeat elements in the human and mouse clusters is 52.37% and 39.63% respectively. This difference can be explained, at least partially, by the higher presence of LTR elements in the human cluster, and by noting that human Alu repeats are significantly longer than mouse B1/B2 repeats.

	number of elements	length occupied	percentage of sequence
SINEs:	92	20278 bp	6.76%
Alus	54	14841 bp	4.95%
MIRs	38	5437 bp	1.81%
LINEs:	91	101092 bp	33.70%
LINE1	63	93714 bp	31.24%

LINE2	23	6058 bp	2.02%
L3/CR1	5	1320 bp	0.44%
LTR elements:	38	30219 bp	10.07%
MaLRs	17	10876 bp	3.63%
ERVL	2	780 bp	0.26%
ERV_classI	14	11395 bp	3.80%
ERV_classII	5	7168 bp	2.39%
DNA elements:	16	3224 bp	1.07%
MER1_type	8	1017 bp	0.34%
MER2_type	8	2207 bp	0.74%
Unclassified:	0	0 bp	0.00%
Total interspersed repeats:		154813 bp	51.60%
Small RNA:	1	54 bp	0.02%
Satellites	1	180 bp	0.06%
Simple repeats:	28	1370 bp	0.46%
Low complexity:	22	688 bp	0.23%

Table 3.21: Identification of the repeat elements on the contig (Genbank accession number AP002534) corresponding to the human Ifi200 gene cluster.

III-1-10 Evolutionary aspects of the mouse Ifi200 gene cluster

The extent of sequence conservation between Ifi202a and Ifi202b and that between Ifi203a and Ifi203b indicates that duplication events might have involved one or more of these genes quite recently. In this regard, the respective positions of Ifi202a and b, and Ifi203a and b in the mouse cluster suggest the possibility of a dual duplication event, leading to the conjoint appearance of Ifi202b and Ifi203b. It has been observed that the Ifi203b and Ifi202a genes exhibit a reciprocal pattern of expression in various mouse strains, indicating that p203b and p202a might have a similar function in a mouse strain in which only one of them is expressed (Gribaudo *et al.*, 1999). The synonymous and non-synonymous substitution rates of the coding regions were determined, according to the method of Ina (1995), for the two pairs of paralogous genes (Ifi202a

and Ifi202b, Ifi203a and Ifi203b, respectively) (see table 3.22). Synonymous substitution rates reflect the rates of evolutionary divergence of nucleic acid sequences (Oeltjen *et al.*, 1997) while non-synonymous substitution rates reflect the rates of protein evolution (Li and Graur, 1991). As expected, the rate of synonymous substitution is higher than the rate of non-synonymous substitution, reflecting functional constraints on non-synonymous substitution. Furthermore, the ratio between the two rates is more pronounced for the Ifi202a and Ifi202b pair of genes (see Ka/Ks), suggesting stronger constraints on non-synonymous substitutions for these two genes. However, no substantial variations were observed between the synonymous substitution rates (~2.4 fold range of variation) and the non-synonymous substitution rates (~4.5 fold range of variation), indicating that Ifi202b and Ifi203b probably appeared, if not simultaneously, at least in a very close period of time during the evolution of the cluster.

	Synonymous ($\times 10^7$)	Non-Synonymous ($\times 10^7$)	Ka/Ks
Ifi202a/Ifi202b	0.908	0.390	2.32
Ifi203a/Ifi203b	2.132	1.766	1.20
range of variation	2.34	4.52	

Table 3.22: Synonymous and non-synonymous substitution rates, calculated by the method of Ina (1995). (Ks) Substitution rate per synonymous site; (Ka) Substitution rate per non-synonymous site.

III-1-11 Closing the gap between the two contigs

The mouse Ifi200 gene cluster contains two contigs (contig 1 and contig 2), encoding three genes from the 203 subfamily (Ifi203a, b and c) (Kingsmore *et al.*, 1989), as well as three genes from the 202 subfamily (Wang *et al.*, 1999), two genes from the 204 subfamily (Choubey *et al.*, 1989), a "203-like" putative pseudogene and the Ifi201 gene. In addition, the genomic region containing the Ifi202a and Ifi203a

genes was duplicated at least twice during the evolution of this mouse Ifi200 gene cluster, and the extent of conservation between the duplicated regions is over 99% (see above).

The mutual orientation and the position of contig 1 and contig 2, as well as the size and sequence of the gap between the two contigs, remain to be established. A BLAST analysis of the mouse assembly database, based on whole genome shotgun (WGS) trace data currently available at http://www.ensembl.org/Mus_musculus indicates that contig 1 and contig 2 are separated by a gap of ~23 kb located between the 3' end of contig 1 and the 5' end of contig 2. However, a BLAST analysis of the current mouse WGS trace database (see http://www.trace.ensembl.org) indicates that the traces mapped to the gap region share over 99% sequence identity with traces mapped, during this dissertation research, to three other regions located on contigs 1 and 2, and corresponding to the genomic regions at the 5' end of the "203-like", Ifi203a, and Ifi203b genes. Thus, the trace data mapped to the gap region by WGS assembly may actually map to any of the three other conserved regions. Interestingly, the analysis of the mouse Ifi200 gene cluster sequence generated by WGS assembly indicates that the conserved regions on contigs 1 and 2 are superposed, and therefore misassembled. In addition, the WGS trace data (available at http://trace.ensembl.org) are generated from RPCI-23 female (C57BL/6J) mouse BAC libraries that differs from the one used during this dissertation research (mouse strain SV129 provided by Genome Systems, Inc.). Single nucleotide polymorphisms (SNPs), relatively frequent between two mouse strains, may occur in the gap region between contigs 1 and 2, and are impossible to differentiate with base pair substitutions between extremely well conserved duplicated regions. Thus, closing the gap between contigs 1 and 2 requiring additional data will be the subject of additional future studies.

III-1-12 Two-dimensional dot-plot comparison of the human and mouse Ifi200 clusters

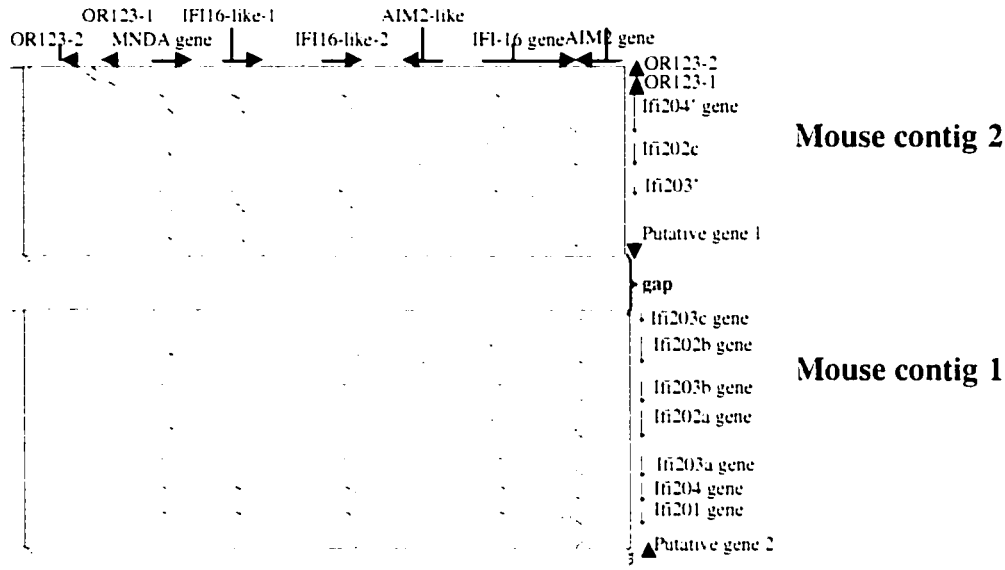


Fig. 3.32a: Dot-plot comparison of the mouse and human Ifi200 clusters. Each of the genes are represented on the x-axis (human genes) and the y-axis (mouse genes) according to their positions in the clusters. (OR123) Putative olfactory receptor pseudogene; (IFI16-like) Putative pseudogene sharing sequence similarity with the IFI-16 gene; (AIM2-like) Putative pseudogene sharing sequence similarity with the AIM2 gene. The directions of transcription are marked by an arrow for each gene. The two contigs from the mouse cluster are separated by a gap of unknown size.

The dot-plot analysis of the human and mouse 200 family genomic sequences shown in Fig. 3.32a reveals a repetitive pattern, as the human IFI-16, MNDA and AIM2 genes contain regions which are similar to each of the seven genes of the mouse 200 family gene cluster. Three regions of the human cluster, located between the MNDA and IFI-16 genes, also share the same pattern, indicating the additional presence of three pseudogenes within the human cluster. A BLAST analysis of this region confirms that two of these pseudogenes share sequence homology with the human gene IFI-16. The third of these pseudogenes is located on the complementary DNA strand and shares sequence homology with the human gene AIM2. In addition, genomic regions downstream from the murine 203-like and Ifi201 genes are conserved in the human cluster and share sequence similarity with the Ifi202a, Ifi202b and Ifi204 cDNAs. These

conserved sequences, which may represent additional Ifi200 genes, are indicated as putative gene 1 and putative gene 2 in Fig. 3.32a.

III-1-13 Postulated evolution of the Ifi200 gene cluster

Comparison of the mouse and human Ifi200 gene sequences indicates greater sequence similarity between paralogs than between orthologs, therefore suggesting that a number of mechanisms, including unequal crossover between repeating units, gene conversion, and gene amplification, may have occurred during the evolution of the cluster (for review, see Liao, 1999). The mouse paralogs were more conserved than they were in human, and one or a combination of these concerted evolution mechanisms likely took place in the mouse lineage, after the human/mouse speciation. In addition, the presence of several pseudogenes in both the mouse and human chromosome 1 genomic regions suggests that several of the genes became non-functional over evolutionary time. Thus, as described in the model developed by Brunner *et al.* (1986), several duplicated genes that initially underwent homogenization by gene conversion, somehow escaped gene conversion, and evolved independently. A postulated evolutionary pathway for the mouse and human Ifi200 cluster is proposed in Fig. 3.33a. As can be seen, it is likely that a primordial Ifi200 gene encoded a protein containing only a type *a* domain. Gene duplication and gene conversion then occurred prior to the human/mouse speciation, and resulted in a cluster of genes similar, in terms of structure and organization, to the human cluster. Additional duplications in the mouse cluster, which likely occurred after speciation, then resulted in the multiple gene copies that are unique to the mouse lineage. The subsequent accumulation of deleterious mutations, which transformed selected genes into pseudogenes in the mouse and human lineages, resulted in the formation of the MNDA gene in the human cluster, in which exons encoding a type *b* domain in the ancestor protein became non-functional through mutations. Although cis-acting

elements may influence the concerted evolution of DNA sequences (Hibner *et al.*, 1991; Thompson-Stewart *et al.*, 1994), their effect on the evolution of the genes present in the mouse and human clusters can only be speculated. Finally, according to this postulated evolutionary pathway, it is predicted that two genes from the mouse cluster are yet to be discovered (Fig. 3.33a), and one of these two genes may be the D3 gene which codes for a protein containing only a type *a* domain (Tannenbaum *et al.*, 1993).

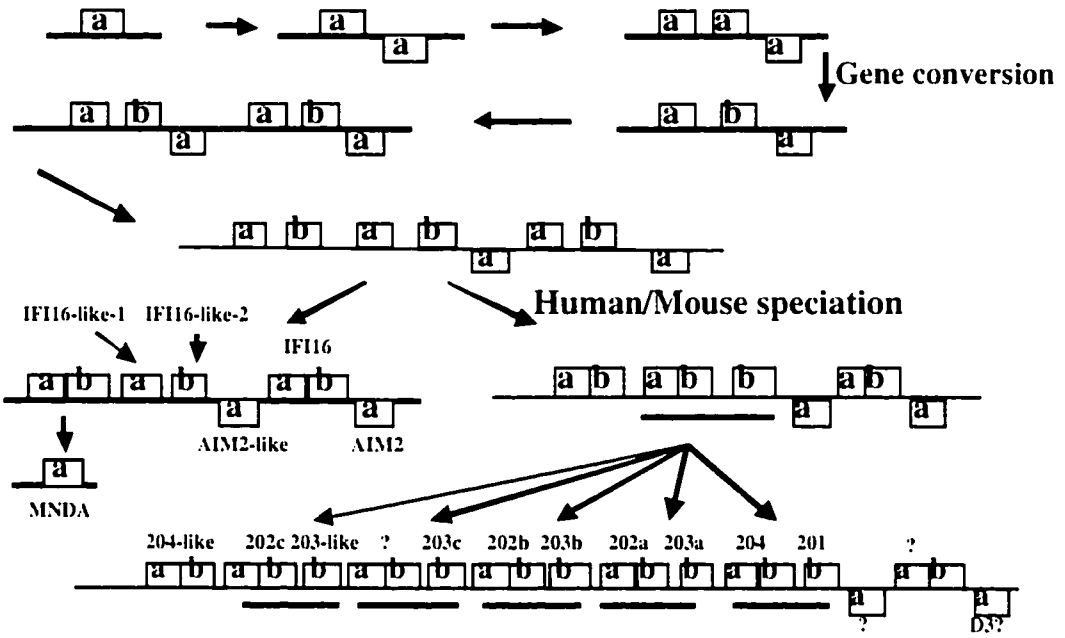


Fig. 3.33a: Postulated evolution of the Ifi200 gene cluster. Each of the Ifi gene encodes for a protein containing either a type *a* and type *b* 200-amino acid conserved domain, or only one of these two domains. These *a* and *b* domains are indicated by a box "a" or a box "b", according to the domain(s) they encode for. An "?" indicates that the gene structure and characteristics remain to be established. The "D3?" indicates the existence of a putative D3 gene. The black line underneath portion of the mouse genomic region corresponds to the possible unit of duplication.

III-2- Genetic aspects of bone mineral density

Bone mineral density is the strongest determinant of osteoporotic fracture, and 60 to 70% of the normal variability in human bone density is genetically determined. The identification of genes involved in bone density regulation will aid in understanding pathways regulating bone density and factors influencing diseases such as osteoporosis. Multiple environmental factors, such as nutrition and exercise, can effect the adult skeletal system by altering the production of endogenous factors, such as hormones and mineral levels, which modulate bone cell function at the molecular level. Because of the complex interactions between genetic factors and environmental factors, studies of genes that regulate bone mineral density have been difficult. The main strategies for identification of such genes are based on finding evidence of an association between altered bone density and polymorphic genetic markers. This is done by analyzing genetic markers throughout the genome to identify Quantitative Trait Loci (QTL; see Introduction) which may be concentrated in specific regions (candidate loci) or near specific genes of interest (candidate genes). Through association studies, several candidate genes for bone density regulation have been identified (see Introduction). Also, genome-wide scans of appropriate human population or inbred strains of mice have been widely successful in identifying QTLs. However, none of the genes responsible for these QTLs have been identified. In addition, although one might expect candidate genes identified by association studies to be located within a QTL, only one gene (tumor necrosis factor receptor 2) is located in a QTL, and another gene (osteocalcin) is located near to a QTL.

Thus, identifying the genes involved in regulation of bone density is still at an early stage. Furthermore, many of the genes have to be thoroughly studied before understanding the possible role of their gene product at the molecular level and how the abnormalities related to these genes can lead to osteoporosis.

To accelerate the identification of genes present in QTL of bone density, new strategies involving the genomic resources currently available should be used. For instance, the draft working sequence of the human genome should be useful to identify the predicted genes in the QTL regions, to facilitate selecting candidate genes for further studies. In addition, the sequence of syntenic regions between human and mouse chromosomes and the availability of the sequence of the mouse genome will provide a convenient model for identifying genes of QTL of bone density. The next challenge relates to the ability to analyze the functionally important regions (5' flanking region, exon/intron borders, coding regions and 3' flanking region) of each candidate gene within genomic sequences.

Five mouse BAC clones (rp23-21118, rp23-395h6, rp23-9p22, rp23-145f9 and rp23-77a8), which map to a QTL located on chromosome 1q have been sequenced and annotated during this dissertation research. Each candidate gene is discussed according to its putative ability to regulate bone mineral density. A map of this region is shown in Fig. 3.1, and Table 3.23 lists each of the clones, its predicted genes and its size.

Clone	Genbank Accession Number	Size (bp)	Genes present in the clone
rp23-145f9	AC091521	198810	CD150/SLAM,CD48/BCM1 BC011154/19A24, AK002863
rp23-77a8	AC091523	235508	Ly108, CD84
rp23-21118	AC083893	244345	none
rp23-395h6	AC084821	211748	mCAR, myelin P0, ApoA-II, NDUFS2, beta4Gal-T3, Ppox, USP23, DEDD, Nit1, Adamts4, Fcer1g, L27' (pseudogene), RIKEN1, RIKEN2, RIKEN3
rp23-9p22	AC079446	223103	none

Table 3.23: summary of the genes present in the mouse BAC clones sequenced during this dissertation research. See text for details.

III-2-1 Mouse BAC clone rp23-145f9

The mouse BAC clone rp23-145f9 (Genbank accession number AC091521) is 198810 bp in length and is located on mouse chromosome 1q22. It contains two putative genes (Genbank accession numbers BC011154 and AK002863) and two known genes, BCM1 (Genbank accession number X17501) and SLAM (Genbank accession number NM_013730), that are both involved in the lymphocytic activation of T cells in mouse.

III-2-1-1 Mouse CD150/SLAM gene

Besides the T cell receptor, numerous cell surface molecules participate in the interaction of T cells with antigen-presenting cells such as lymphocytes and macrophages. One of these molecules is CD2, a protein expressed on the surface of both T cells and natural killer cells (Davis and van der Merwe, 1996). In addition to CD2, optimal T cell activation for proliferation and cytokine production requires the involvement of co-stimulatory signals provided by accessory molecules (Jenkins, 1994). One T cell co-stimulatory molecule that is a member of the CD2 subfamily of the Ig superfamily, SLAM (signaling lymphocytic activation molecule) has been described as a transmembrane glycoprotein expressed on the surface of T, B, natural killer and dendritic cells (Davis and van der Merwe, 1996). The extracellular domain of SLAM is the receptor for measles virus (Tatsuo *et al.*, 2000), and monoclonal antibodies (mAbs) directed against human SLAM (CD150) enhance antigen-specific proliferation and cytokine production by human $CD4^+$ T cells (Cocks *et al.*, 1995). Human SLAM is associated with a small cytoplasmic adaptor protein named SAP that has been identified as being responsible for the X-linked lymphoproliferative disease (XLP) (Coffey *et al.*,

1998; Sayos *et al.*, 1998). XLP is characterized by an increased susceptibility to primary Epstein-Barr virus (EBV) infection, leading to an uncontrolled proliferation of virus-containing B cells and reactive cytotoxic T cells (Purtilo, 1985). It has been suggested that the absence of SAP in XLP patients affects T/B cells interactions induced by SLAM, leading to an uncontrolled B cell proliferation caused by EBV infection (Sayos *et al.*, 1998).

The human and mouse CD150 genes have seven exons in a genomic region that is highly homologous, in terms of sequence and exon/intron organization, between mouse and human (Wang *et al.*, 2001). AN ALIGN comparison between the amino acid sequences of the mouse and human SLAM proteins indicates that they share 51.2% sequence identity (Fig. 3.32).

```
               10         20         30         40         50
SLAMhu MDPKGLLSLTFVLFLSLAFGASYGTGGRMMNCPKILRQLGSKVLLPLTYE-RINKSMNKS
       ::::: ::  ..:::::::  :::::: .:.:: ::..::. . :::: : .:::::.:::
SLAMmu MDPKGSLSWRILLFLSLAFELSYGTGGGVMDCPVILQKLGQDTWLPLTNEHQINKSVNKS
               10         20         30         40         50         60

      60         70         80         90        100        110
SLAMhu IHIVVTMAKSLENSVENKIVSLDPSEAGPPRYLGDRYKFYLENLTLGIRESRKEDEGWYL
       ..:.::::: :  .. ..:::::.: :... : .: : :.:  .:::.: :  .:.:.:::::
SLAMmu VRILVTMATSPGSKSNKKIVSFDLSKGSYPDHLEDGYHFQSKNLSLKILGNRRESEGWYL
               70         80         90        100        110        120

     120        130        140        150          160        170
SLAMhu MTLEKNVSVQRFCLQLRLYEQVSTPEIKVLNKTQEN--GTCTLILGCTVEKGDHVAYSWS
       ...:.:::::.:: ::.::::::: ::::::::::::  :::.:.:.:::.:::::.::::
SLAMmu VSVEENVSVQQFCKQLKLYEQVSPPEIKVLNKTQENENGTCSLLLACTVKKGDHVTYSWS
               130        140        150        160        170        180

       180        190        200        210        220        230
SLAMhu EKAGTHPLNPANSSHLLSLTLGPQHADNIYICTVSNPISNNSQTFS-PWPGCRTDPS-ET
       ..:::: :. :: :::: .::. :: :.:: ::.:::::. :.:::.    .:. . : :.
SLAMmu DEAGTHLLSRANRSHLLHITLSNQHQDSIYNCTASNPVSSISRTFNLSSQACKQESSSES
               190        200        210        220        230        240

         240        250          260        270        280        290
SLAMhu KPWAVYAGLLGGVIMILIMV--VILQLRRRGKTNHYQTTVEKKSLTIYAQVQKPGPLQKK
       .::  :. .  ::..:.:.:  .:.....:.::.:: :  ::.:::::::::::: :
SLAMmu SPWMQYTLVPLGVVIIFILVFTAIIMMKRQGKSNHCQPPVEEKSLTIYAQVQKSG-----
               250        260        270        280        290

           300        310        320        330
SLAMhu LDSFPAQDPCTTIYVAATEPVPESVQETNSITVYASVTLPES
       . :.:     ..:. ..       .   ..  :  ..
SLAMmu VRSMPHLAGVSVIFRTGFL-----IAALHTTMVLQGLL---E
         300        310            320
```

Fig.3.32: ALIGN comparisons between the amino acid sequences of the human (top) and mouse (bottom) SLAM proteins.

A GENSCAN analysis, followed by a BLAST search, of the BAC clone rp23-145f9 revealed that the genomic segment at position 152526-185145 corresponds to the mouse CD150 gene (Fig. 3.33).

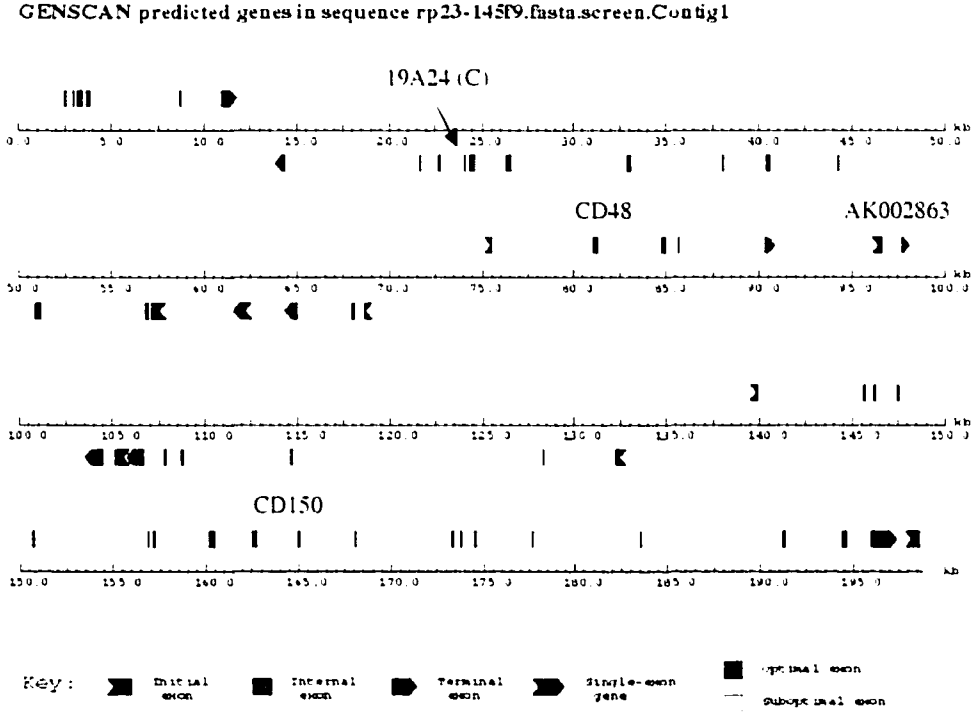
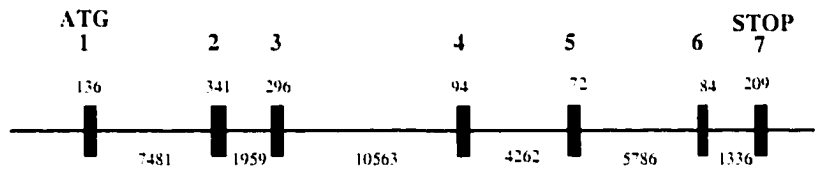


Fig. 3.33: GENSCAN analysis of the mouse BAC clone rp23-145f9. The genes are indicated above the sequence. (C) The gene is located on the complementary strand.

A CROSSMATCH analysis between the mouse CD150 cDNA and the segment at position 152526-185145 confirms that the mouse gene is made of seven exons and indicates that it spans a genomic region of approximately 33 kb (Fig. 3.34 and table 3.24). The results of the Neural Network splice site prediction program analysis confirmed the presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of this genomic region and the CD150 cDNA (table 3.25).

CD150/SLAM gene:



CD150 protein:



Fig. 3.34: Structural organization of the CD150/SLAM gene and its encoded protein. (Top) Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines. (Bottom) The dark gray box indicates the presence of an extracellular Ig-like V domain. The clear gray box corresponds to an extracellular Ig-like C2 domain. The white boxes correspond to the signal sequence (exon 1) and the transmembrane region (exon 4). The black boxes corresponds to the cytoplasmic CY1, CY2 and CY3 regions, encoded by exons 5, 6, and 7, respectively. The last two domains (CY2 and CY3) are absent from the splice variant form of CD150.

Exon number	Contig position	Exon length (bp)
1	152526-152663	137
2	160144-160485	341
3	162444-162740	296
4	173302-173397	95
5	177658-177731	73
6*	183506-183601	95
7*	184936-185145	209

Table 3.24: contig position and length of the exons of the mouse CD150 gene. The fifth exon codes for the entire cytoplasmic domain (CY1) of an alternatively spliced mRNA form of CD150. (*) These two exons are not transcribed for the mRNA splice variant of CD150.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
1/1	152665 ggaacagGTgagtg	1/2	160143 ttccctcAGgtgga
2/2	160487 ctttatgGTaaaaaa	2/3	162443 cccgacAGaacag

	162742		173301
3/3	tcctcagGTgatca	3/4	tttttcAGaatcg
	173399		177657
4/4	agacaagGTaggct	4/5	ttcttcAGgtaaa
	177733		183505
5/5	atcagggGTacgtt	5/6	ttgtgtAGcctcaa
	183603		184935
6/6	tgtccagGTgacata	6/7	gcccacAGgaaccaa

Table 3.25: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 152587, in the first exon. Castro *et al.* (1999) characterized the mouse CD150 gene and revealed that the first exon actually encodes a 24 amino acid signal peptide that precedes an extracellular region of 221 amino acid. The extracellular domain consists of an incomplete Ig-like V domain lacking intrachain disulfide bonds and one membrane-proximal Ig-like C_2 domain with two disulfide bonds. The extracellular domains are encoded by exons 2 and 3, respectively. The transmembrane region is encoded by exon 4, and each of the three cytoplasmic regions (CY1, CY2 and CY3) are encoded by exons 5, 6 and 7, respectively (Fig.3.34). The ORF of CD150 is 1232 bp with a 149 bp 5' and a 60 bp 3' UTR. Castro *et al.* (1999) also characterized an alternatively spliced mRNA, in which the fifth exon (CY1) encodes the entire cytoplasmic domain of the corresponding protein. Splicing of the next exon (exon 6) to an internal 5' splice site situated in exon 5 gives rise to the cytoplasmic domain of the CD150 protein.

Initial analysis of the first 1000 bp located immediately upstream from the first exon of CD150 with MOTIF shows that the 5' flanking region of CD150 contains several potential binding sites for transcription factors that are important for the development of cells of the hematopoietic lineages. These regions include the

CCAAT/enhancer-binding protein b (C/EBP-b), c-Myb, c-Ets, NF-AT, MZF1, GATA1, GATA2, GATA3, Ik-1, Ik-2, Lyf-1 and Pbx-1 transcription factor binding sites. The promoter prediction by NNPP predicted two potential TATA-containing promoter regions located 254-304 bp and 62-112 bp upstream from the first exon of CD150, respectively.

III-2-1-2 Mouse CD48/BCM1 gene

CD2 is a 55 to 60 kD protein expressed on the surface of T cells and natural killer cells, that participates in the interaction of T cells with antigen-presenting cells (Davis and van der Merwe, 1996). One ligand for CD2 in mice is BCM1 (Kato *et al.*, 1992). BCM1 is a glycosylphosphatidylinositol-anchored molecule whose expression is restricted to antigen-presenting lymphocytes and macrophages and to dendritic cells (Kato *et al.*, 1992). The human homolog of BCM1, CD48, does not bind CD2 with high affinity, as the major ligand for CD2 in human is CD58 (Selvaraj *et al.*, 1987). CD48, CD58 and BCM1 are structurally related and are members of the CD2 subfamily of the Ig superfamily (Wong *et al.*, 1990). Although the function of CD2 and of its ligands remains uncertain, Gonzalez-Cabrero *et al.* (1999) revealed, by generating mice deficient in CD48 expression ($CD48^{-/-}$ mice), that $CD4^+$ T cells from $CD48^{-/-}$ mice were defective in activation, suggesting that CD48 plays a significant role in the T cell activation process.

The mouse BCM1 gene, previously characterized by Wong *et al.* (1990), is located on a genomic region structurally conserved between mouse and human. In addition, Wong *et al.* (1990) demonstrated that the BCM1 locus was structurally related to another locus located on mouse chromosome 3 and containing the CD2 and LFA3 genes. LFA3, a glycosylphosphatidylinositol-anchored antigen of the CD2 subfamily of the Ig superfamily, is involved in adhesion reactions between T cells and accessory cells

(Tiefenthaler *et al.*, 1987). These data suggest that the duplication of a chromosome region including the precursors of the genes for BCM1, CD2 and LFA3 gives rise to the linkage groups now observed. Because of the sequence similarity between LFA3 and BCM1, Wong *et al.* (1990) suggested the existence of a gene encoding a recognizer molecule in linkage with BCM1 on mouse chromosome 1. According to the present data, this gene might correspond to the neighboring CD150 gene.

A GENSCAN analysis, followed by a BLAST search, of the clone rp23-145f9 revealed that the genomic segment at position 67401-90612 corresponds to the mouse BCM1 gene (Fig. 3.33). A CROSSMATCH analysis between this segment and the mouse BCM1 cDNA indicates that this gene is made up of four exons and spans a region of approximately 23 kb (table 3.26 and Fig. 3.35).

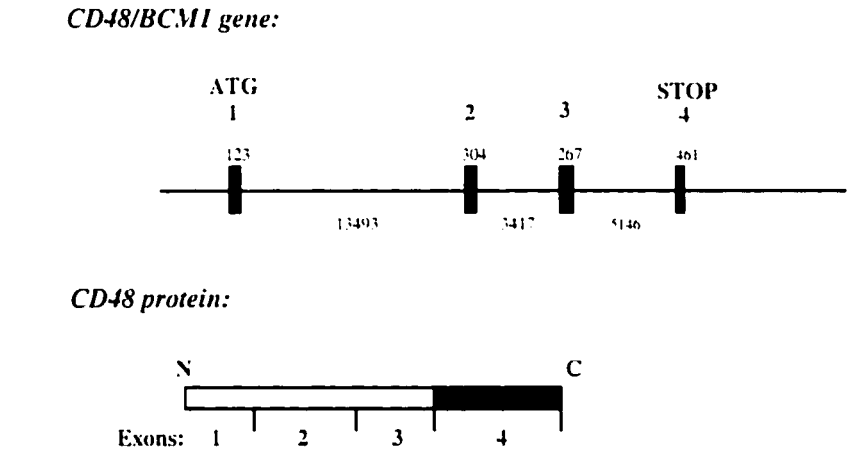


Fig. 3.35: Structural organization of the CD48/BCM1 gene and its encoded protein. (Top) Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines. (Bottom) The white box corresponds to the extracellular domain. The black box corresponds to the glycosylphosphatidylinositol anchor, encoded by exon 4.

The results of the Neural Network splice site prediction program analysis confirmed the presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of this genomic region and the CD48 cDNA (table 3.27).

Exon number	Contig position	Exon length (bp)
1	67401-67524	123
2	81019-81321	302
3	84739-85005	266
4	90154-90612	458

Table 3.26: contig position and length of the exons of the mouse CD48 gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
	67526		81018
1/1	tttcaagGTagctt	1/2	cttggcagAGgtcat
	81323		84738
2/2	gtatttgGTgagtt	2/3	aaaaatcAGatcct
	85007		90153
3/3	gatctagGTaagaa	3/4	cttctttAGccaga

Table 27: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 67452, in the first exon. The first 51 bp of the BCM1 ORF encodes a signal peptide. The glycosylphosphatidylinositol anchor is encoded by the region located at position 90161-90230 that corresponds roughly to the fourth exon followed by a 382 bp 3' UTR. The extracellular domain, consisting of an Ig-like V domain without a disulfide bond and an Ig-like C2 domain, is encoded by the 3' end of exon 1, exon 2, exon 3 and the 5' end of exon 4 (Wong *et al.*, 1990) (Fig. 3.35).

An ALIGN comparison between the amino acid sequences of the mouse BCM1 protein and its human counterpart indicates that they share 52.2% sequence identity (Fig. 3.36).

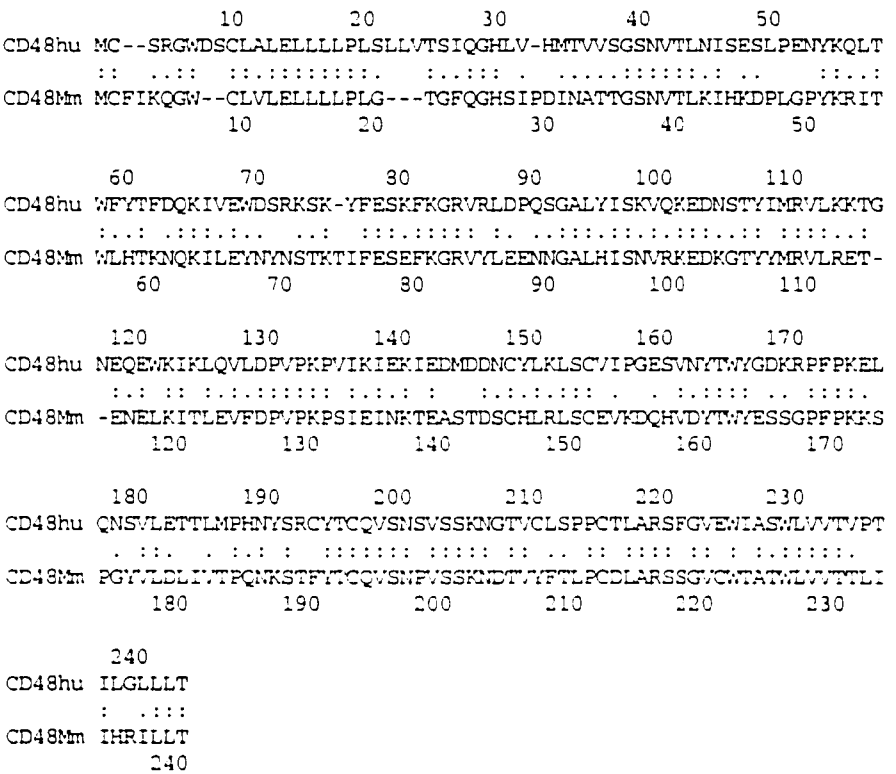


Fig.3.36: ALIGN comparison between the amino acid sequences of the human (top) and mouse (bottom) CD48 proteins.

Analysis of the first 1000 bp located upstream from the first exon of BCM1, using MOTIF, reveals that the 5' flanking region of BCM1 contains potential binding sites for transcription factors that are important for the development of hematopoietic cells, including c-Ets, MZF1, GATA1, GATA2, GATA3, c-Myb transcription factor binding sites, as well as a potential c-Myc/Max heterodimer binding site 121-132 bp upstream from the first exon. The promoter prediction by NNPP predicted a single potential promoter region located 612-662 bp upstream from the first exon of the BCM1 gene.

III-2-1-3 Mouse BC011154 gene

A GENSCAN analysis, followed by a BLAST search, of the mouse clone rp23-145f9 revealed that the genomic segment at position 34288-19843, on the complementary strand, corresponds to an uncharacterized mouse gene with the Genbank accession number BC011154, that shares sequence similarity with the human 19A24 cDNA (Fig. 3.33). A CLUSTALW analysis between the mouse BC011154 cDNA and the human 19A24 cDNA confirmed that both genes share 53.3% sequence identity (Fig. 3.37) and a BLAST analysis against the EST database dbEST confirmed that the mouse BC011154 gene was indeed expressed, and therefore not a pseudogene.

```
m19A24      --------------------CCTGAAAACGCTATGGCTCGTTTCTCAAC
h19A24      GCGGCCGCGAATTCGGCACGAGCAGAGAGCAATATGGCTGGTTCCCCAAC
                                 . ** * .  ******* *** * ****

m19A24      GTACATCATCTTTACCTCTGTCCTCTGTCAGCTAACAGTCACAGCAGCTT
h19A24      ATGCCTCACCCTCATCTATATCCTTTGGCAGCTCACAGGGTCAGCAGCCT
             . * *** * * * ** * **** ** ***** ****   ******* *

m19A24      CTGGAACTCTGAAGAAGGTGGCCGGTGCCCTTGATGGATCTGTGACATTC
h19A24      CTGGACCCGTGAAAGAGCTGGTCGGTTCCGTTGGTGGGGCCGTGACTTTC
            ***** *  ****  ** *** **** ** *** ***  * ***** ***

m19A24      ACTCTGAATATCACTGAAATAAAGGTTGACTATGTTGTATGGACGTTCAA
h19A24      CCCCTGAAGTCCAAAGTAAAGCAAGTTGACTCTATTGTCTGGACCTTCAA
             * *****   **  * **   * ******* * **** ***** *****

m19A24      CACATTCTTTCTTGCCATGGTAAA---AAAAGACGGCGTTA-CAT-----
h19A24      CACAACCCCTCTTGTCACCATACAGCCAGAAGGGGGCACTATCATAGTGA
            ****  *  ***** **   ** *   * ***  ***  ** ***

m19A24      CACAAAGTAGTAACAAAGAAAGGATAGTCTTTCCAGATGGACTCTACTCC
h19A24      CCCAAAATCGTAATAGGGAGAGAGTAGACTTCCCAGATGGAGGCTACTCC
            * **** * **** *  ** **  *** *** *********  *******

m19A24      ATGAAGCTCAGCCAATTGAAGAAGAATGACTCTGGAGCCTACCGTGCAGA
h19A24      CTGAAGCTCAGCAAACTGAAGAAGAATGACTCAGGGATCTACTATGTGGG
             ********** ** ************** **   ****  **  *

m19A24      GATTTACAGTACATCGAGTCAGGCTTCCTTAATCCAGGAGTATGTGCTGC
h19A24      GATATACAGCTCATCACTCCAGCAGCCCTCCACCCAGGAGTACGTGCTGC
            *** *****  ****    ***    ***  * ********* *******

m19A24      ATGTCTACAAGCATTTGTCAAGGCCCAAGGTCACCATAGATCGGCAAAGC
h19A24      ATGTCTACGAGCACCTGTCAAAGCCTAAAGTCACATTGGGTCTGCAGAGC
            ********* ****  ****** *** ** *****  * * ** *** ***

m19A24      AACAAGAATGGCACCTGCGTAATCAATCTGACATGTTCCACGGATCAGGA
h19A24      AATAAGAATGGCACCTGTGTGACCAATCTGACATGCTGCATGGAACATGG
            ** ************** ** * ************ * ** *** ** *
```

```
m19A24      CGGGGAGAATGTAACCTACAGCTGGAAAGCTGTGGGGCAGGGGGACAATC
h19A24      GGAAGAGGATGTGATTTATACCTGGAAGGCCCTGGGGCAAGCAGCCAATG
             *  *** **** *  ** * ****** **  ******* *  * ****

m19A24      AGTTTCATGATGGTGCCACCCTCTCCATCGCCTGGAGATCAGGAGAGAAA
h19A24      AGTCCCATAATGGGTCCATCCTCCCCATCTCCTGGAGATGGGGAGAAAGT
            ***  *** ****  *** **** ***** *********  ***** *

m19A24      GACCAGGCCTTAACATGCATGGCCAGGAATCCAGTCAGCAACAGTTTCTC
h19A24      GATATGACCTTCATCTGCGTTGCCAGGAACCCTGTCAGCAGAAACTTCTC
            **   * **** *  *** * ******** ** *******  *  *****

m19A24      AACCCCCGTCTTTCCCCAGAAGCTCTGTGAAGATGCTGCCACGGATCTAA
h19A24      AAGCCCCATCCTTGCCAGGAAGCTCTGTGAAGGTGCTGCTGATGACCCAG
            ** **** ** ** **  ************** ******    ** * *

m19A24      CTTCACTCAGGGGCATCCTATACATCCTGTGCTTCTCAGCAGTGCTCATC
h19A24      ATTCCTCCATGGTCCTCCTGTGTCTCCTGTTGGTGCCCCTCCTGCTCAGT
             ***   ** ** * **** *   ******   *  *     ******

m19A24      CTATTTGCTGTCTTGCTGACTATTTTTCATACTACGTGGATAAAGAAAAG
h19A24      CTCTTTGTACTGGGGCT-ATTTCTTTGGTTTCTGAAGAGAGAGAGACAAG
            ** ****   *   *** * *  ***   * **     ** * *** ***

m19A24      AAAAGAAAA--GAAGACCAGAAGAAGATGCACCAAACACATTTTATTCCA
h19A24      AAGAGTACATTGAAGAGAAGAAGAGAGTGGACATTTGTCGGGAAACTCC-
            ** ** * *  *****  ******   ** **      *     * ***

m19A24      CTGTGCAGATCCCCAAAGTGGTAAAGAGTCCCAGCTCCCTGCCTGCAAAG
h19A24      TAACATATGCCCCCATTCTGGAGAGAACACAGAGTACG--A----CACAA
                  *   *****   ***  *  *  *  **  *       ** *

m19A24      CCACTCGTGCCAAGGTCATTAAGCTTTGAAAATGTTATCTAGATGACAGC
h19A24      TCCCTCACACTAATAGAAC-AATCCTAAAGGAAG--ATCCAGCAAATACG
             * ***   * **    *  ** * *  *  * *  *** **   * *

m19A24      ACTCCGTCCTCTCCAGAAAAAAACAAAACAAAACAAAACTGCAAAACAAA
h19A24      GTTTACTCCACTGTGGAAATACCGAAAAAGATGGAAAATCCCCACTCACT
              *   *** **   **** *   **** *   ****   * *  **

m19A24      CAAAACCTCACCATTCTGAGCAGA--AATGAAAACTTCTGTCAAAGACTG
h19A24      G-----CTCACGATGCCAGACACACCAAGGCTATTTGCCTATGAGAATGT
                  ***** ** *    ** *  ** *  *  * *     *  *

m19A24      AATGTAAATGTCTCCCCGCCAGACCCACATGTTTGCATCCAGACAGAGAA
h19A24      TATCTAGACAGCAGTGCACTGCCCCTAAGTCTCTGCTCAAAAAAAAAACA
             ** ** *   *    * *    ** *  * * ***    * * * *  *

m19A24      GTTCA-GGCTCAAGGGCTTTCAGGGTACAATGACTCTTGGGGACTGGGAT
h19A24      ATTCTCGGCCCAAAGAA----AACATACAGG-------------------
             ***  *** *** *      *   ****

m19A24      ACAGCCCATTCAGACATTCTGACTAAGAATCTCTCTATGCCCATCCTGTC
h19A24      ----------------------------------------------------

m19A24      TCCGGTAATGATGAATGAGGCTGTGCTCATAGGCAGTGGACTAATTTGTC
h19A24      ----------------------------------------------------

m19A24      TGATGAAGAAATTTTTGACAAAGCGCAGCATCTAGGCTGTGGCACAGCCA
h19A24      ----------------------------------------------------
```

```
m19A24      TGTTCGCTGTTTTTTATTCACATTTATAGCTGTATGTAAATAATCTCAATC
h19A24      --------------------------------------------------

m19A24      ATTCAGTTCCCAGATAAAAGACGCATAGACGCATAAAAATTAAAAAAAAA
h19A24      --------------------------------------------------

m19A24      AAAAAAAA
h19A24      --------
```

Fig. 3.37: CLUSTALW comparison between the cDNA sequences of human 19A24 and mouse BC011154. (m19A24) mouse sequence. (h19A24) human sequence.

In addition, an ALIGN comparison between the amino acid sequence deduced from the coding region of the mouse BC011154 gene and the human 19A24 protein indicate that they share 39.7% sequence identity (Fig.3.38).

```
              10         20         30         40         50         60
19A24h MAGSPTCLTLIYILWQLTGSAASGPVKELVGSVGGAVTFPLKSKVKQVDSIVWTFNTTPL
       ::   : . .  .: ::: .:::: .:...:.. :.::: :.    .:: .:::::::  :
19A24m MARFSTYIIFTSVLCQLTVTAASGTLKKVAGALDGSVTFTLNITEIKVDYVVWTFNTFFL
              10         20         30         40         50         60

              70         80         90        100        110        120
19A24h VTIQPEGGTIIVTQNRNRERVDFPDGGYSLKLSKLKKNDSGIYYVGIYSSSLQQPSTQEY
       . .. .:   ...:. :.::. :::: ::.:::.::::::: : . :::.: :    :::
19A24m AMVKKDG---VTSQSSNKERIVFPDGLYSMKLSQLKKNDSGAYRAEIYSTSSQASLIQEY
                70         80         90        100        110

             130        140        150        160        170        180
19A24h VLHVYEHLSKPKVTMGLQSNKNGTCVTNLTCCMEHGEEDVIYTWKALGQAANESHNGSIL
       :::::.:::.:::::.  ::::::::: :::: .. :.: :.:::.::. :. :.:. :
19A24m VLHVYKHLSRPKVTIDRQSNKNGTCVINLTCSTDQDGENVTYSWKAVGQGDNQFHDGATL
       120        130        140        150        160        170

             190        200        210        220        230        240
19A24h PISWRWGESDMTFICVARNPVSRNFSSPILARKLCEGAADDPDSSMVLLCLLLVPLLLSL
        :.:: ::.:... :.::::::: .::.:.. .:::: :: :  :   .: .:    .: :
19A24m SIAWRSGEKDQALTCMARNPVSNSFSTPVFPQKLCEDAATDLTSLRGILYILCFSAVLIL
       180        190        200        210        220        230

               250        260        270        280        290
19A24h F--VLGLF--LWFLKRERQEEYIEEKKRVDI-CRETPNICPHSGENTEYDTIPHTNRTIL
       :  .: .:   :. ::....   .  :.. . : .     :.:::... .   :   ..
19A24m FAVLLTIFHTTWIKKRKEKKTRRRCTKHILFHCAD-----PQSGKESQLPACKATRAKVI
       240        250        260        270            280        290

         300        310        320        330
19A24h KEDPANTVYSTVEIPKKMENPHSLLTMPDTPRLFAYENVI
       :                                      .
19A24m K--------------------------------------L
```

Fig. 3.38: ALIGN comparison between the amino acid sequences of the human (top) and mouse (bottom) 19A24 proteins.

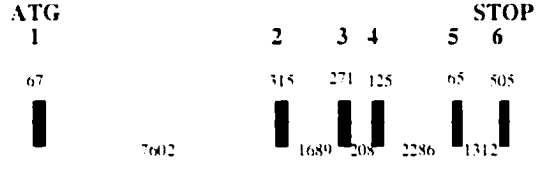
The human 19A24 gene does not have any known function. However, a BLASTP analysis revealed some sequence similarity between the human 19A24 protein and the extracellular Ig-like domain of another lymphocytic antigen termed Ly9. The Ly9 glycoprotein is expressed on the cell surface of B and T lymphocytes and, similarly to the CD48 and CD150 proteins, consists of four Ig-like domains with the structural features of the CD2 subfamily. Sayos *et al.* (2001) demonstrated that Ly9, in a similar fashion to CD150, was able to recruit SAP, the product of the gene mutated in X-linked lymphoproliferative disease. The Ly9 gene contains ten exons, and each Ig-like domain is encoded by a separate exon (Tovar *et al.*, 2000). In addition, the Ly9 gene has been mapped in a genomic region located in proximity to the region encompassing the CD48 and CD150 antigens in mouse and in human (Doudney *et al.*, 2001). The 19A24 gene may therefore be a member of a large family of genes, including CD48, CD150 and Ly9, that are closely related to each other in terms of structure and location.

A CROSSMATCH analysis between the genomic segment at position 34288-19843 and the mouse 19A24 cDNA reveals that 19A24 consists of six exons and spans a genomic region of approximately 15 kb (table 3.28 and Fig. 3.39). The results of the Neural Network splice site prediction program analysis confirmed the presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of this genomic region and the 19A24 cDNA (table 3.29).

It is likely that each putative Ig-like domain of 19A24 is encoded by a separate exon, in a similar manner to that observed for the CD150 and Ly9 genes. A TMHMM analysis to search for putative transmembrane domains (see www.cbs.dtu.dk/services/TMHMM-2.0/) inside the 19A24 protein indicates the presence of a potential transmembrane domain at amino acid position 228-250 (Fig. 3.40). This domain is encoded by exon 4, and similarly to the CD150 and CD48

proteins, probably corresponds to a single transmembrane domain anchoring the 19A24 protein into the membrane. The predicted transmembrane domain in the amino terminal region of the 19A24 protein (Fig. 3.40) most likely reflects the presence of a signal peptide and suggests that the 5' end of exon 1 encodes for a signal domain.

BC011154/19A24 gene:



19A24 protein:

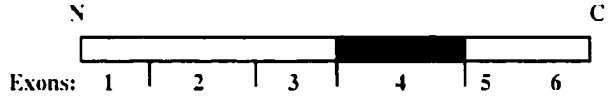


Fig. 3.39: Structural organization of the BC011154/19A24 gene and its encoded protein. (Top) Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines. (Bottom) The black box corresponds to the transmembrane domain as predicted by TMHMM.

Exon number	Contig position	Exon length (bp)
1	34288-34221	67
2	26614-26304	310
3	24616-24344	272
4	24133-24011	122
5	21722-21660	62
6	20344-19843	501

Table 3.28: Contig position and length of the exons of the mouse 19A24 gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
	34219		26615
1/1	ctaacagGTgagtc	1/2	tgtttgcAGtcaca
	26302		24613
2/2	gtctacaGTaagca	2/3	tctttcacAGcattt
	24342		24132

3/3	ctgtgaagGTgacag	3/4	cctttaaAGatgct
	24009		21721
4/4	agaaaagGTaaaac	4/5	ttgttaatAGaaaag
	21658		20343
5/5	ccaaagtgGTaagaa	5/6	tgcctgcAGgtaaa

Table 3.29: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

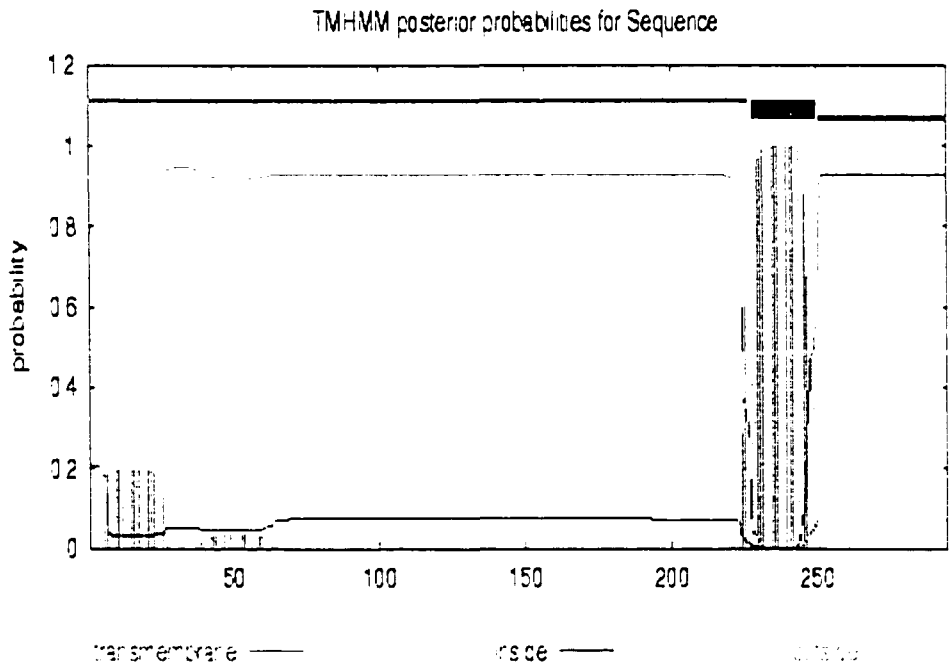


Fig. 3.40: TMHMM search for transmembrane domain of the 19A24 protein. The putative transmembrane domain is located at amino acid position 228-250.

A putative translation start site is located at position 34276, in the first exon. Initial analysis of the first 1000 bp located immediately upstream from the first exon of the mouse 19A24 gene, using MOTIF, indicates that the 5' flanking region of 19A24 contains potential binding sites for the transcription factors MZF1, NF-AT, c-Ets, Sp1, c-Myb, Pbx-1, Lyf-1, Ik-2, GATA1, GATA2, GATA3, which are involved in the early development and maturation of lymphocytes, as well as a potential binding site for E2F transcription factor (489-496 bp upstream from the first exon). The promoter prediction

by NNPP predicted two potential promoter regions, 703-753 bp and 330-380 bp upstream from the first exon of the mouse 19A24 gene, respectively.

III-2-1-4 Mouse AK002863 gene

A GENSCAN analysis, followed by a BLAST search, of the mouse clone rp23-145f9 indicates the presence of a potential single-exon gene at position 96427-96855, that shares over 99% sequence similarity with the cDNA sequence corresponding to the Genbank accession number AK002863 (Fig. 3.33). A CROSSMATCH analysis between the mouse clone rp23-145f9 sequence and the AK002863 cDNA indicates that this potential single-exon gene lacks the first 25 bp of the AK002863 cDNA sequence. This discrepancy may be explained by sequencing mistakes or the existence of a gene slightly different from the AK002863 gene. A putative translation start site (ATG) is located at position 96443. A search of the PROSITE protein profile database did not indicate the presence of any particular motif in the putative protein encoded by the AK002863 gene. PsortII predicted that the putative AK002863 protein is nuclear. The analysis of the 1000 bp located at position 95427-96427, using MOTIF, indicates the presence of multiple potential binding sites for transcription factors. Among them are potential binding sites for c-Myb, Oct-1, CREB, AP-1, Lyf-1, NF-E2, c-Ets and E2F transcription factors. The promoter prediction by NNPP predicted a single potential promoter region, 796-846 bp upstream from the 5' end of the putative AK002863 single-exon gene.

III-2-1-5 Repeat elements in the mouse clone rp23-145f9

The analysis of repeat elements in the 198810 bp of the mouse clone rp23-145f9, using Repeatmasker, shows that they make up to 45.35% of the sequence of this

clone (table 3.30). LINE and LTR elements represent ~42.5% and ~38% of the total repeat elements present in this contig. The LINE elements here are almost exclusively of the LINE1 type. The SINE B1 and B2 elements represent only ~4% of the total repeat elements present in this mouse clone. This mouse clone rp23-145f9 maps to a portion of a cluster of genes encoding glycoproteins exposed on the surface of hematopoietic cells. As described above for the mouse Ifi200 gene cluster, this region, also on mouse chromosome 1, similarly has a high percentage of LINE1 elements and a low percentage of SINE B1 and B2 elements. Whether this trend can be related to this particular mouse genomic region (1q21-23), or the existence of two distinct clusters, whose genomic organization appears to be based on the duplication of an ancestor gene, remains to be established.

	number of elements	length occupied	percentage of sequence
SINEs:	75	9750 bp	4.90%
B1s	34	3989 bp	2.01%
B2-B4	25	4086 bp	2.06%
IDs	2	90 bp	0.05%
MIRs	14	1585 bp	0.80%
LINEs:	37	38384 bp	19.31%
LINE1	33	36933 bp	18.58%
LINE2	3	1299 bp	0.65%
L3/CR1	1	152 bp	0.08%
LTR elements:	88	34437 bp	17.32%
MaLRs	40	12327 bp	6.20%
ERVL	5	4100 bp	2.06%
ERV_classI	3	696 bp	0.35%
ERV_classII	18	9332 bp	4.69%
DNA elements:	6	1069 bp	0.54%
MER1_type	5	694 bp	0.35%
MER2_type	0	0 bp	0.00%
Unclassified:	2	722 bp	0.36%
Total interspersed repeats		84362 bp	42.43%
Small RNA:	4	333 bp	0.17%
Satellites:	0	0 bp	0.00%

Simple repeats:	68	3935 bp	1.98%
Low complexity:	27	1906 bp	0.96%

Table 3.30: Identification of the repeat elements on the mouse clone rp23-145f9.

III-2-2 Mouse BAC clone rp23-77a8

The mouse BAC clone rp23-77a8 (Genbank accession number AC091523) is located on mouse chromosome 1q21-23 and contains 235508 bps. A single contig, totaling ~378 kb was generated by merging BAC clones rp23-77a8 and rp23-145f9 (Fig. 3.41).

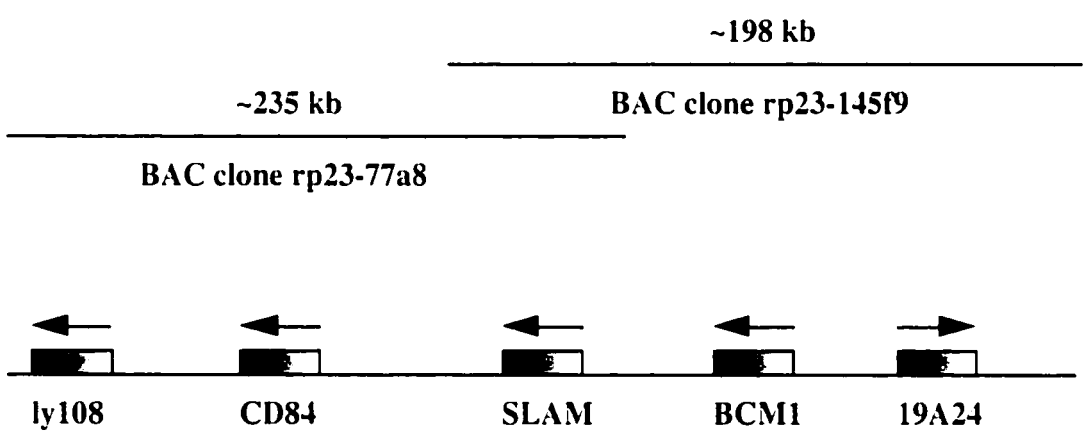


Fig. 3.41: Position of the genes on the single contig, generated by merging BAC clones rp23-145f9 and rp23-77a8. The arrows indicate the direction of transcription.

The clone rp23-77a8 contains two known genes (Ly108 and CD84) that both are involved in the interaction between lymphocytes and T cells in mouse. Ly108 and CD84 belong to the CD2 subfamily of the Ig superfamily, which consists of multiple members, including, besides Ly108 and CD84, CD2 (Davis and van der Merwe, 1996), CD58 (Seed, 1987), BCM1/CD48 (Wong *et al.*, 1990), 2B4 (Boles *et al.*, 1999), Ly9 (Sandrin *et al.*, 1996), SLAM/CD150 (Castro *et al.*, 1999) and NAIL (Kubin *et al.*,

1999). All of these genes result in proteins which are characterized by a typical pattern of Ig-like folds in their extracellular domain, a transmembrane domain and a cytoplasmic domain whose size may vary because of the presence of a splice variant (Castro *et al.*, 1999). Members of the CD2 family are predominantly expressed on hematopoietic cells and play an active role in the immune response, as the interaction of these surface molecules with their receptors may contribute indirectly to signal transduction or promote cell-cell interaction. The genes encoding CD2 family members are clustered on mouse chromosomes 1 and 3 and on human chromosome 1. The genomic organization of the portion of the mouse chromosome 1 cluster, containing the Ly108, CD84, SLAM and BCM1 genes, was revealed by the sequence of the contig generated by merging BAC clones rp23-145f9 and rp23-77a8 (Fig. 3.41).

III-2-2-1 Mouse Ly108 gene

The Ly108 gene was recently characterized by Peck and Ruley (2000). It contains at least 8 exons that produce an ORF of 993 bps encoding a protein of 331 amino acids. In addition, there is a 5' UTR of 171 bps and a 3' UTR of 1281 bps. The predicted protein contains a putative 30 amino acid hydrophobic signal domain and a 23 amino acid hydrophobic transmembrane domain. The extracellular domain of Ly108 is made of two Ig-like domains. The amino-terminal Ig domain lacks any disulfide bonds and thus resembles a V-like fold, whereas the second Ig domain is characteristic of a truncated C2-like domain with two disulfide bonds. A splice variant was characterized by EST analysis (Peck and Ruley, 2000). The putative protein encoded by this splice variant is identical to the Ly108 protein through amino acid 327, at which point the two sequences diverge, with the deletion of four amino acids and the addition of 24 amino acids to the cytoplasmic domain. In addition, Ly108 transcripts were detected in all but the most highly differentiated mouse B and T cells (Peck and Ruley, 2000), raising the

possibility of a role for the Ly108 protein as a B and T cell co-stimulatory factor, similar to the role played by SLAM/CD150.

A GENSCAN analysis, followed by a BLAST search, of the mouse BAC clone rp23-77a8 reveals that the genomic segment at position 159881-186212 corresponds to the mouse Ly108 gene (Fig. 3.42).

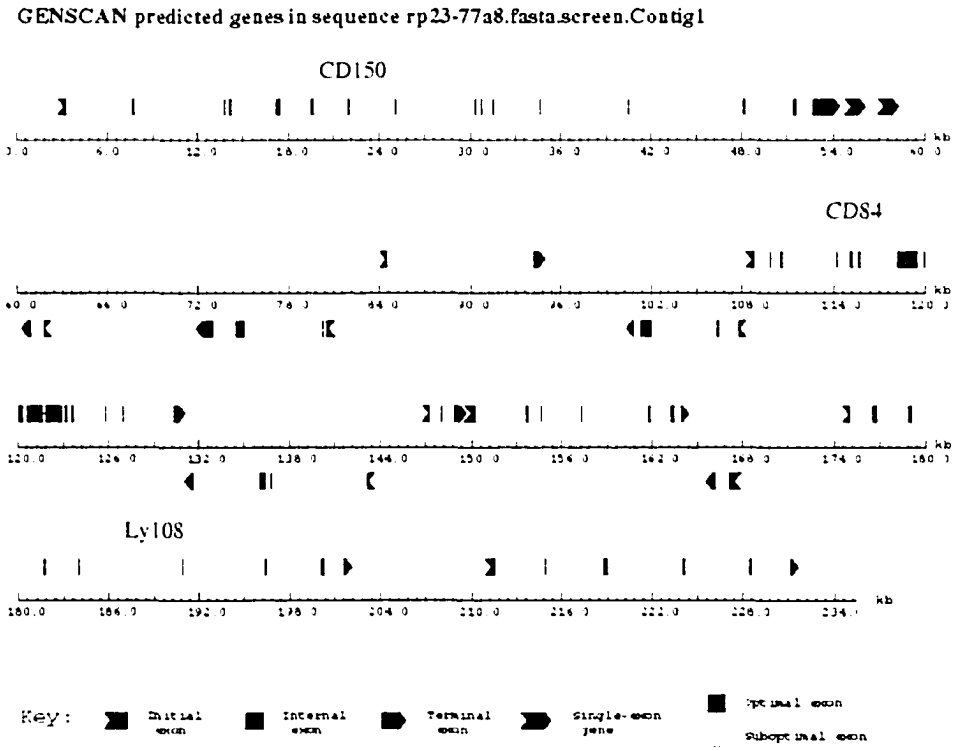
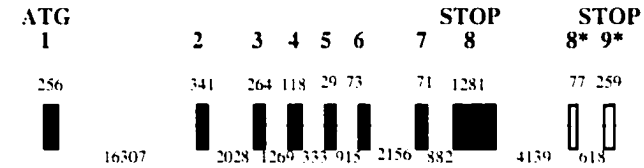


Fig. 3.42: GENSCAN analysis of the mouse BAC clone rp23-77a8. The genes are indicated above their respective sequences.

A CROSSMATCH analysis between the BAC clone rp23-77a8 and the Ly108 cDNA indicates that the Ly108 gene contains eight exons (Fig.3.43). It further reveals that the putative 30 amino acid hydrophobic signal domain located at the amino terminus of the Ly108 protein is encoded by the first exon of the Ly108 gene, and the 23 amino acid hydrophobic transmembrane domain is encoded by exon 4 of the Ly108 gene.

Ly108 gene:



Ly108protein:

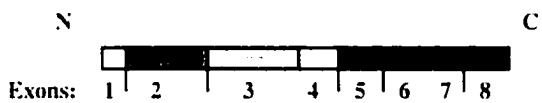


Fig. 3.43: Structural organization of the Ly108 gene and its encoded protein. (Top) Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines. The two alternatively spliced exons (8* and 9*) are indicated by a white box. (Bottom) The dark gray box indicates the presence of an extracellular Ig-like V domain. The clear gray box corresponds to an extracellular Ig-like C2 domain. The white boxes correspond to the signal sequence (exon 1) and the transmembrane region (exon 4). The black box corresponds to the cytoplasmic regions encoded by exons 5, 6, 7 and 8. The isoform of the Ly108 protein contains an additional cytoplasmic region encoded by exons 8* and 9*, instead of the cytoplasmic region encoded by exon 8.

Another CROSSMATCH analysis between the BAC clone rp23-77a8 and the cDNA of the splice variant of the Ly108 gene (Genbank accession number AF248636) indicates that the splice variant is encoded by the genomic segment at position 159881-191296 and consists of nine exons (Fig.3.43). Although the first seven exons of the Ly108 splice variant are similar to the first seven exons of the Ly108 gene, its last two exons are unique, confirming that the two isoforms of Ly108 differs in their cytoplasmic domains (table 3.31). The Ly108 gene spans a genomic region of approximately 26 kb, and contains a large ~16 kb first intron (table 3.31 and Fig.3.43). In addition, although the Ly108 gene contains a large 3' UTR corresponding to the last exon, there is no such UTR at the 3' end of the Ly108 splice variant. Also, the results of the Neural Network splice site prediction program analysis confirmed the presence of

potential splice sites at the exon-intron borders indicated by the ALIGN comparison of this genomic region and the Ly108 cDNA (table 3.32).

Exon number	Contig position	Exon length (bp)
1	159858-160136	278
2	176443-176784	341
3	178813-179076	263
4	180346-180459	113
5	180796-180825	29
6	181740-181813	73
7	183969-184040	71
8	184923-186204	1281
8*	190343-190419	76
9*	191037-191297	260

Table 3.31: contig position and length of the exons of the mouse Ly108 gene. 8* and 9* indicates the 8th and 9th exons of the splice variant of the Ly108 gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
1/1	160138 cggctcagGTaaaaa	1/2	176442 gtgtttacAGggagt
2/2	176786 ggtctttgGTgagtcc	2/3	178812 tttatgtAGaacgac
3/3	179078 ctgcaaagGTaatgt	3/4	180345 ttcttttcAGgggttc
4/4	180461 gagaagagGTaagttt	4/5	180795 catttcccAGgttctc
5/5	180827 catccagGTgagcat	5/6	181739 aaccatacAGagtcct
6/6	181815 ccaatgcagGTaagtcc	6/7	183968 ttgtttacAGgaaatg
7/7	184042 tccagagaGTaaatcc	7/8	184922 tttctcctAGgccgaa
7/7	184042 tccagagaGTaaatcc	7/8*	190342 ttttcttgtAGgaaaca
8*/8*	190421 cctgaagGTaaatg	8*/9*	191036 ttccccctAGgaagag

Table 3.32: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 160052, in the first exon. The putative polyadenylation site described by Peck and Ruley (2000) is located at position 186192, in the last exon. The putative termination of translation (TAA) is located at position 184935, indicating that the last four amino acids of the Ly108 protein, deleted in the splice variant, are encoded by the last exon. Initial analysis of the 1000 bp located immediately upstream from the first exon of Ly108, using MOTIF, indicates the presence of potential binding sites for transcription factors that are important for the development and function of hematopoietic cells, including MZF1, GATA, c-Ets, c-Myb, Ik-2, Pbx1, STAT and NF-AT binding sites. In addition, the 5' flanking region of Ly108 contains potential binding sites for AP-1, MyoD, p53, E2F and NF-kB transcription factors. The promoter prediction program NNPP predicted three potential promoter regions, located 839-889 bp, 507-557 bp and 498-548 bp, respectively, upstream from the initiation of transcription. BLAST analysis demonstrated that a protein sharing sequence homology with a mouse Ly108 protein isoform, coded by the first 8 exons of the Ly108 gene, is present in human, and therefore may correspond to the putative human ortholog of the mouse Ly108 protein. An ALIGN comparison between this mouse Ly108 protein isoform and its putative human counterpart indicates that they share 41.5% sequence identity (Fig. 3.44).

```
                                10        20        30        40
ly108h M---------------LWLFQSLLFVFCFGPGNVVSQSSLTPLMVNGILGESVTLPLEFP
       :                .:::    .::::.: :. :::::  : ..:::::::::::..:
ly108m MAVSRAPAPDSACQRMVWLFP---LVFCLGSGSEVSQSSSDPQLMNGVLGESAVLPLKLP
               10        20          30        40        50

          50         60          70        80        90        100
ly108h AGEKVNFITWLFN-ETS--LAFIVP-HETKSPEIHVTNPKQGKRLNFTQSYSLQLSNLKM
       ::. .:.: : .. :.:   :...   . .::.:  :. :  ::::.::::::::.::: :
ly108m AGKIANIIIWNYEWEASQVTALVINLSNPESPQIMNTDVK--KRLNITQSYSLQISNLTM
        60        70        80        90         100       110

             110       120       130       140       150       160
```

```
ly108h EDTGSYRAQISTKTSAKLS-SYTLRILRQLRNIQVTNHSQLFQNMTCELHLTCSVEDADD
       ::::: :::.:: :  .. .: ::.....: :...::.. :..: ::..::.: ... ..
ly108m ADTGSYTAQITTKDSEVITFKYILRVFERLGNLETTNYTLLLENGTCQIHLACVLKNQSQ
         120       130       140       150       160       170

              170       180       190       200       210       220
ly108h NVSFRWEALGNTLSSQPNLTVSWDPRISSEQDYTCIAENAVSNLSFSVSAQKLCEDVKIQ
       .:: .:.: ::   . ::.:. :::: :..: :.: :.:::::::: :::.:.::. : .
ly108m TVSVEWQATGNISLGGPNVTIFWDPRNSGDQTYVCRAKNAVSNLSVSVSTQSLCKGV-LT
         180       190       200       210       220       230

              230       240          250       260       270
ly108h YTDTKMILFMVSGICIVFGFIILLL---LVLRKRRDSLSLSTQRTQGPESARNLEYVSVS
          . . ::.. : :. . :....   . . ::: :: :..:.   :::... .  . :
ly108m NPPWNAVWFMTT-ISIISAVILIFVCWSIHVWKRRGSLPLTSQH---PESSQSTDGPG-S
          240       250       260       270          280

       280       290       300       310       320       330
ly108h PTNNTVYASVTHSNRETEIWTPRENDTITIYSTINHSKESKPTFSRATALDNVV
       : .::::::::.  .: .:  : .::..:::: .:::.:.. .
ly108m P-GNTVYAQVTRPMQEMKIPKPIKNDSMTIYSIVNHSREAEYS----------
     290       300       310       320       330
```

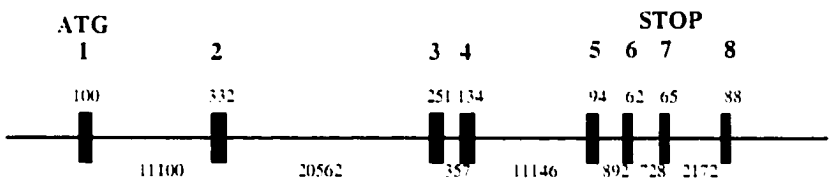
Fig.3.44: ALIGN comparison between the amino acid sequences of the human (top) and mouse (bottom) Ly108 proteins.

III-2-2-2 Mouse CD84 gene

CD84 was first discovered by de la Fuente *et al.* (1997) as a leukocyte antigen predominantly expressed by human mature B cells and monocytes. Its murine counterpart, also named CD84, was characterized by de la Fuente *et al.* (1999) as a new member of the CD2 family of cell surface molecules. The mouse CD84 gene has an ORF encoding a 329 amino acid protein displaying an amino acid sequence identity of 57.3% with the human CD84 protein (de la Fuente *et al.*, 1999). The sequence of the mouse CD84 begins with a 21 amino acid signal domain, followed by an extracellular domain made of two Ig-like sub-domains characteristic of the CD2 family, a transmembrane domain and a 87 amino acid cytoplasmic domain. The CD84 transcripts are predominantly expressed on the surface of B lymphocytes and macrophages. Two major mRNA species were detected in mouse hematopoietic tissues, suggesting the existence of a splice variant. Interestingly, Palou *et al.* (2000) showed, by screening a

peripheral blood leukocyte cDNA library, that five different human CD84 isoforms were obtained, differing in their 3' sequence. The corresponding proteins, differing in their cytoplasmic domains, were generated by several mechanisms including alternative splicing events, reading frameshift, use of a cryptic splice site and absence of splicing. Sayos *et al.* (2001) showed that the mouse CD84 protein, similarly to the cell surface receptor Ly9, can recruit the X-linked lymphoproliferative disease protein SAP via an interaction between the SAP protein and the cytoplasmic domain of CD84. Whether the existence of potential isoforms of the mouse CD84 protein differing in their cytoplasmic domains can alter their interaction with SAP remains to be established. Finally, Martin *et al.* (2001) showed that CD84, although preferentially expressed on B lymphocytes and macrophages, also was expressed on thymocytes and T cells, and that CD84 can bind to itself *in vitro*. Furthermore, CD84 can act as a co-stimulatory factor by enhancing interferon gamma secretion in human lymphocytes (Martin *et al.*, 2001).

CD84 gene:



CD84 protein:

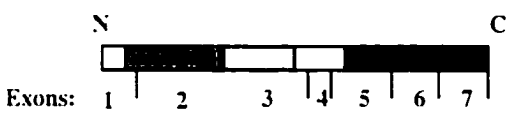


Fig. 3.45: Structural organization of the CD84 gene and its encoded protein. (Top) Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines. (Bottom) The dark gray box indicates the presence of an extracellular Ig-like V domain. The clear gray box corresponds to an extracellular Ig-like C2 domain. The white boxes correspond to the signal sequence (exon 1) and the transmembrane region (exon 4). The black box corresponds to the cytoplasmic region, encoded by exons 5, 6, and 7.

A GENSCAN analysis, followed by a BLAST search, of the clone rp23-77a8, reveals that the segment at position 82948-131030 corresponds to the mouse CD84 gene (table 3.33 and Fig.3.42). A CROSSMATCH analysis between the clone rp23-77a8 and the mouse CD84 cDNA indicates that the CD84 gene is made of 8 exons, spans a genomic region of approximately 50 kb and contains large intronic regions (introns 1, 2 and 4 are ~11 kb, ~20 kb and ~11 kb, respectively, in length) (Fig.3.45). The results of the Neural Network splice site prediction program analysis confirmed the presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of this genomic region and the CD84 cDNA (table 3.34).

Exon number	Contig position	Exon length (bp)
1	82948-83048	100
2	94148-94480	332
3	115042-115293	251
4	115650-115784	134
5	126930-127024	94
6	127916-127978	62
7	128706-128771	65
8	130943-131031	88

Table 3.33: contig position and length of the exons of the mouse CD84 gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
1/1	83050 acaaacctGTgagttc	1/2	94147 tatcctcAGggtctg
2/2	94482 tatctaccGTaagtta	2/3	115041 ccccttatAGgtcgac
3/3	115295 atgtacagGTaaccag	3/4	115649 ctgtttccAGacactc
4/4	115786 cctggaagGTaaaac	4/5	126929 tatttgAGcagatg
5/5	127026 agtccaagGTaagctg	5/6	127915 tcttttccAGatgctg
6/6	127980 ctgagaagGTaaatcc	6/7	128705 cttacacAGatgaag

	128773		130942
7/7	ttgtctagGTgagcat	7/8	gctcaccAGgtgatt

Table 3.34: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 83002, in the first exon, indicating that the CD84 gene contains a short 54 bp 5' UTR. A putative translation termination site (TAG) is located at position 128768, in the 7th exon, indicating that the last exon is a non-coding exon, and that the CD84 gene contains a 91 bp 3' UTR. An ALIGN comparison between the amino acid sequences of the mouse and human CD84 proteins indicates that they share 56.2% sequence identity (Fig. 3.46).

```
               10         20         30         40         50         60
CD84hu MAQHHLWILLLCLQTWPEAAGHDSEIFTVNGILGESVTFPVNIQEPRQVKIIAWTSKTSV
       :::.::::: .:::::: :::::::.. ..:::::::::::: .::::::.. :::::.::
CD84mu MAQRHLWIWFLCLQTWSEAAGKDADPVVMNGILGESVTFLLNIQEPKKIDNIAWTSQSSV
               10         20         30         40         50         60

               70         80         90        100        110        120
CD84hu AYVTPGDSETAPVVTVTHRNYYERIHALGPNYNLVISDLRMEDAGDYKADINTQADPYTT
       :.. :: ...   ::.:. .:  ::. .  .:.::: :::::::: :::::: . .  :
CD84mu AFIKPGVNKAE--VTITQGTYKGRIEIIDQKYDLVIRDLRMEDAGTYKADINEENEE-TI
               70          80         90        100        110

              130        140        150        160        170        180
CD84hu TKRYNLQIYRRLGKPKITQSLMASVNSTCNVTLTCSVEKEEKNVTYNWSPLGEEGNVLQI
       :: : :.::::::  ::::::::..:.:.:::.:::::::::::.:::.:::.::.:::::
CD84mu TKIYYLHIYRRLKTPKITQSLISSLNNTCNITLTCSVEKEEKDVTYSWSPFGEKSNVLQI
       120        130        140        150        160        170

              190        200        210        220        230        240
CD84hu FQTPEDQELTYTCTAQNPVSNNSDSISARQLCADIAMGFRTHHTGLLSVLAMFFLLVLIL
        ..: ::.:::::::::::::::::.....: :.:   .:. .:. : . ::..:::.::
CD84mu VHSPMDQKLTYTCTAQNPVSNSSDSVTVQQPCTDTP-SFHPRHAVLPGGLAVLFLLILIP
       180        190        200        210         220        230

              250        260        270        280        290        300
CD84hu SSVFLFRLFKRRQGRIFPEGSCLNTFTKNPYAASKKTIYTYIMASRNTQPAESRIYDEIL
         .::::::.:::. ::  :..           .:::::.:. .  :::.::.::::::::
CD84mu MLAFLFRLYKRRRDRIVLEAD----------DVSKKTVYAVV--SRNAQPTESRIYDEIP
        240        250                260         270        280

              310        320        330        340
CD84hu QSKVLPSKEEPVNTVYSEVQFADKMGKASTQDSKPPGTSSYEIVI
       :::.:  :..::.:.:: :::...:: ... .: . : . . :::.
CD84mu QSKMLSCKKDPVTTIYSSVQLSEKMKETNMKDRSLPKALGNEIVV
          290        300        310        320
```

Fig.3.46: ALIGN comparison between the amino acid sequences of the human (top) and mouse (bottom) CD84 proteins.

The different sub-domains of the mouse CD84 protein have been described by de la Fuente *et al.* (1999). The 21 amino acid signal domain is encoded by the first exon and the first 15 bp of the second exon. The first Ig-like sub-domain is encoded by the remainder of the second exon. The second Ig-like sub-domain is encoded by the third exon and the first 24 bp of the fourth exon. The transmembrane domain is encoded by exon 4, whereas the cytoplasmic tail is encoded by the last 41 bp of exon 4, and exons 5, 6 and 7. Initial analysis of the first 1000 bp located upstream from the transcription start site, using MOTIF, indicates several interesting features. First, the 5' flanking region of the mouse CD84 gene contains potential binding sites for transcription factors that are important for the maturation of lymphocytes, including c-Ets, Oct-1, NF-E2, GATA1, GATA2, GATA3, NF-AT, Lyf-1, MZF1, p300, c-Myb and CCAAT/enhancer binding protein (C/EBP) potential binding sites. In addition, the 5' flanking region of CD84 contains potential MyoD, CRE-binding protein 1/c-Jun (CRE-BP1/c-Jun) heterodimer, AP-1, CREB, PAR-type vitellogenin promoter-binding protein (VBP), RAR-related orphan receptor alpha2 (RORalpha2), zinc finger with interaction domain (ZID) protein, sterol regulatory element-binding protein 1 (SREBP-1) and Thing1/E47 (Th1/E47) heterodimer binding sites. The presence of a potential SREBP-1 binding site is surprising, considering that this transcription factor normally controls the expression of genes encoding proteins essentials for cholesterol biosynthesis/uptake and fatty acid biosynthesis (Yokoyama *et al.*, 1993). Thing1 was first identified by Hollenberg *et al.* (1995) as a basic helix-loop-helix transcription factor containing a repression activity and appears to be essential for placentation and cardiac morphogenesis (Riley *et al.*, 1998). ZID was first identified by Bardwell and Treisman (1994) as a POZ domain-

containing transcription factor. The POZ domain is a conserved motif present in a large family of proteins that acts to inhibit the interaction of the associated finger regions of these proteins with DNA (Bardwell and Treisman, 1994). Interestingly, the POZ domain is fused to RARalpha (retinoic acid receptor alpha) as a result of a variant t(11;17) chromosomal translocation occuring in a small subset of acute promyelocytic leukemia, where it acts as a potent oncogene (Dong *et al.*, 1996). Finally, the promoter prediction program NNPP predicted two potential promoter region in the 5' flanking region of the mouse CD84 gene, 461-511 bp and 182-232 bp, respectively, upstream from the initiation of transcription.

III-2-2-3 Repeat elements in the mouse clone rp23-77a8

The analysis of repeat elements in the mouse clone rp23-77a8, using Repeatmasker, shows that they make up to 46.92% of the sequence of this clone (table 3.35). As in the clone rp23-145f9, the repeat elements mainly are LINE and LTR elements which represent ~30% and ~46% of the total repeat elements present in the clone rp23-77a8. The LINE elements are almost exclusively LINE1 elements. SINEs B1 and B2 elements are ~12.4% of the total repeat elements present in this mouse clone. The total percentages of repeat elements present in the clones rp23-77a8 and rp23-145f9 are very similar (~45% of the total sequence in both clones), the percentage of LINE elements present in the clone rp23-77a8 is lower than that present in the clone rp23-145f9, and the percentage of LTR elements present in the clone rp23-77a8 is higher than that present in the clone rp23-145f9.

	number of elements	length occupied	percentage of sequence
SINEs:	103	13753 bp	5.84%
B1s	43	5052 bp	2.15%
B2-B4	43	6983 bp	2.97%
IDs	4	258 bp	0.11%

MIRs	13	1460 bp	0.62%
LINEs:	51	33220 bp	14.11%
LINE1	38	30269 bp	12.85%
LINE2	12	2893 bp	1.23%
L3/CR1	1	58 bp	0.02%
LTR elements:	80	51035 bp	21.67%
MaLRs	36	14769 bp	6.27%
ERVL	6	10202bp	4.33%
ERV_classI	2	662 bp	0.28%
ERV_classII	16	18660 bp	7.92%
DNA elements:	14	2272 bp	0.96%
MER1_type	11	1726 bp	0.73%
MER2_type	2	430 bp	0.18%
Unclassified:	4	1344 bp	0.57%
Total interspersed repeats		101624 bp	43.15%
Small RNA:	0	0 bp	0.00%
Satellites:	0	0 bp	0.00%
Simple repeats:	101	6234 bp	2.65%
Low complexity:	31	2490 bp	1.06%

Table 3.35: Identification of the repeat elements on the mouse clone rp23-77a8.

III-2-3 Mouse BAC clone rp23-21118

The mouse BAC clone rp23-21118 on mouse chromosome 1q22 (Genbank accession number AC083893) is 244345 bp in length. A GENSCAN analysis, followed by a BLAST search, on the genomic sequence of the clone rp23-21118 revealed that this clone does not harbor any known or putative genes (Fig.3.47). The library sample was initially contaminated with another BAC clone of 233219 bp, that was subsequently named rp23-21118b (Genbank accession number AC092498). A GENSCAN analysis and a BLAST search on of the clone rp23-21118b revealed that it contains a gene encoding a small GTPase of the Ras family, Rab11a, that has been mapped to mouse

chromosome 9C (Bhartur *et al.*, 2000). Thus, it is very likely that the clone rp23-21118b does not correspond to any genomic region of the QTL located on mouse chromosome 1q. In addition, two STS markers (D4Mit322 and D4Mit275) match with the mouse BAC clone rp23-21118. However, these markers have been mapped to mouse chromosome 4. Based on this data, it appears that the mouse BAC clone rp23-21118 does not belong to mouse chromosome 1q region but rather to mouse chromosome 4. The position of rp23-21118 therefore is different from the one that was assumed at the beginning of this dissertation research.

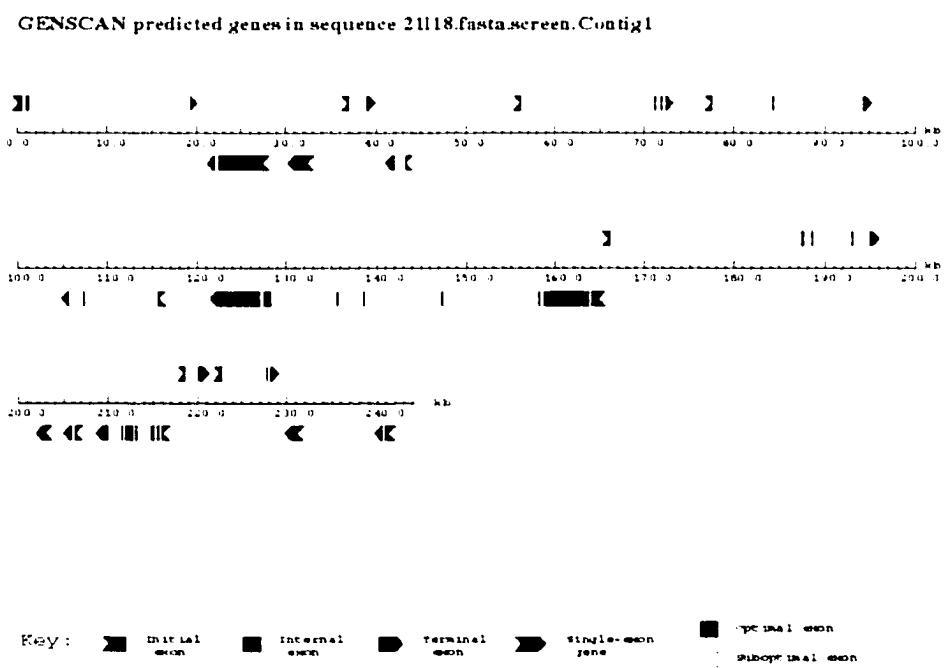


Fig.3.47: GENSCAN analysis of the mouse BAC clone rp23-21118. The genes predicted by GENSCAN are repeat elements or false positives.

The analysis of the repeat elements in the 244345 bp of the clone rp23-21118, using Repeatmasker, indicates a very high percentage of repeat elements, as more than half of the total sequence of this clone is made of repeat DNA (table 3.36). Most of the repeat elements correspond to LINE1 elements and, to a lesser extent, LTR elements.

SINE elements are only 1.2% of the total repeat DNA present in this clone, confirming that there is a high percentage of LINE1 elements and a low percentage of SINEs in several clones mapped to the same genomic region (mouse chromosome 1q21-23). In this particular case, however, the high percentage of LINE elements can be correlated to the total absence of genes, as LINE elements have been reported to being predominant in gene-poor regions (Lander *et al.*, 2001).

	number of elements	length occupied	percentage of sequence
SINEs:	23	2938 bp	1.20%
B1s	12	1310 bp	0.54%
B2-B4	11	1628 bp	0.67%
IDs	0	0 bp	0.00%
MIRs	0	0 bp	0.00%
LINEs:	96	96164 bp	39.36%
LINE1	93	95906 bp	39.25%
LINE2	2	185 bp	0.08%
L3/CR1	1	73 bp	0.03%
LTR elements:	59	26543 bp	10.86%
MaLRs	34	10253 bp	4.20%
ERVL	0	0 bp	0.00%
ERV_classI	1	67 bp	0.03%
ERV_classII	9	10334 bp	4.23%
DNA elements:	8	1630 bp	0.67%
MER1_type	5	547 bp	0.22%
MER2_type	0	0 bp	0.00%
Unclassified:	3	552 bp	0.23%
Total interspersed repeats		127827 bp	52.31%
Small RNA:	4	333 bp	0.17%
Satellites:	0	0 bp	0.00%
Simple repeats:	85	4904 bp	2.01%
Low complexity:	36	1439 bp	0.59%

Table 3.36: Identification of the repeat elements on the mouse clone rp23-21118.

III-2-4 Mouse BAC clone rp23-395h6

III-2-4-1 Mouse BAC clone rp23-395h6: a preview

The mouse clone rp23-395h6 (Genbank accession number AC084821) maps to a gene-rich mouse chromosome 1q21-24 region. A total of fourteen genes, including three hypothetical genes, and one processed pseudogene have been predicted on mouse BAC clone rp23-395h6, using GENSCAN (Fig. 3.48) and BLAST.

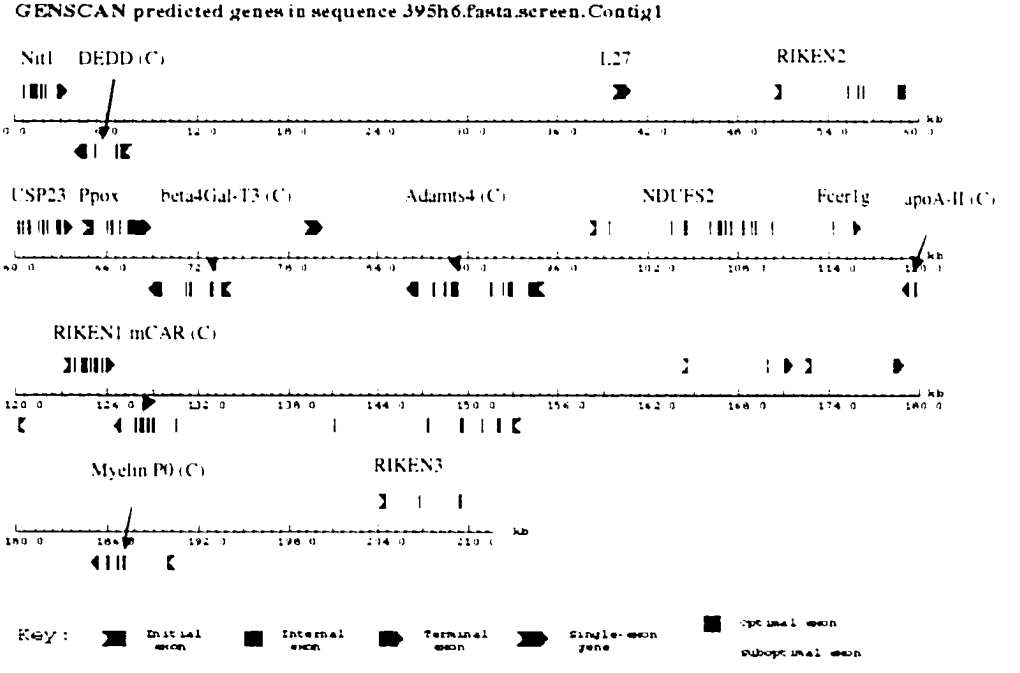


Fig 3.48. GENSCAN analysis of the mouse BAC clone rp23-395h6. (C) Complementary strand.

All 14 genes likely are expressed as ESTs corresponding to them were found in the dbEST database. The genes present on mouse clone rp23-395h6, their positions and their biological activities, are summarized on table 3.37 and Fig.3.49.

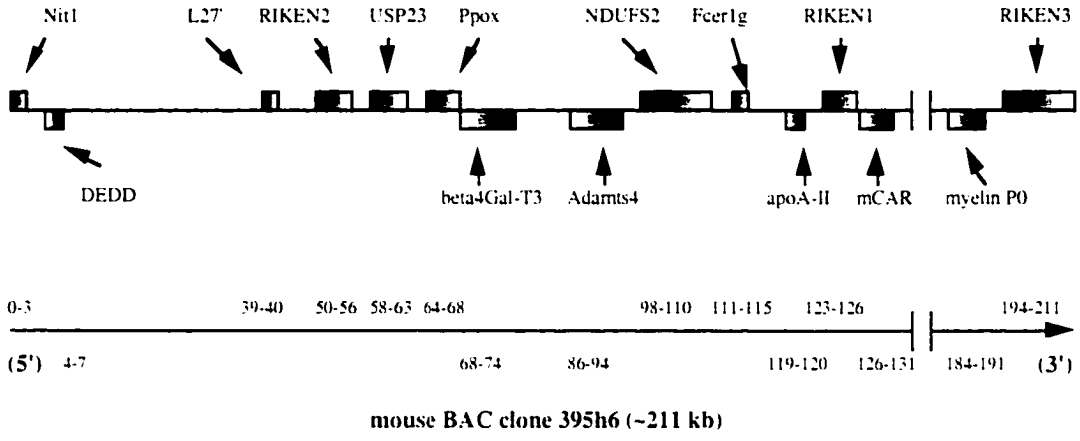


Fig. 3.49: The position of each gene and pseudogene on the mouse BAC clone rp23-395h6 is indicated by a full square on the upper lane and by the corresponding position (in kb) on the lower lane. The clone is represented by a plain line and is not represented in its totality on (a) and (b). (a) The upper squares correspond to genes transcribed in the 5'-3' direction, and the lower squares correspond to genes transcribed in the 3'-5' direction. (b) The numbers correspond to the position on the clone 395h6 (in kb) of each gene represented on (a), and are positioned according to the position of the corresponding gene in (a). The 5' and 3' directions are indicated.

Gene	Position (kb)	Biological activities
nit1	0-3	nitrilase
DEDD	4-7 (C)	CD95-mediated apoptosis
L27' (pseudogene)	39-40	ribosomal protein
RIKEN2	50-56	unknown
USP23	58-63	ubiquitin-specific protease
Ppox	64-68	protoporphyrinogen oxidase
beta4Gal-T3	68-74 (C)	beta-galactosyltransferase
Adamts4	86-94 (C)	aggrecanase-1
NDUFS2	98-110	mitochondrial complex I subunit
Fcer1g	111-115	IgE receptor gamma subunit
apoA-II	119-120 (C)	HDL metabolism

RIKEN1	123-126	eukaryotic porin?
mCAR	126-131 (C)	cellular detoxification
myelin P0	184-191 (C)	myelin formation
RIKEN3	194-211	succinate dehydrogenase of cytochrome b

Table 3.37: Clone position, and biological activities of the genes present on mouse clone rp23-395h6.
(C) complementary strand.

Several genes present on the clone rp23-395h6 are in close proximity to each other. The 3' ends of the Ppox and beta4Gal-T3 genes are separated by 98 bp only, whereas the 3' ends of the RIKEN1 and mCAR genes are separated by only 190 bp. The poly(A) signals of the Ppox and beta4Gal-T3 genes are located at position 68516 and 68652, respectively, and are therefore separated by only 136 nucleotides. Although the existence of bidirectional poly(A) signals has been demonstrated in prokaryotes (Swartzman *et al.*, 1990) as well as in eukaryotes (Irniger *et al.*, 1991), it appears that, in the case of the Ppox and beta4Gal-T3 genes, the gene transcription ends at an unidirectional poly(A) signal.

In the subsequent sections, each gene and its encoded protein, their structural and functional features, will be described according to the data currently available.

III-2-4-2 Mouse CAR gene

Various organic compounds, also known as xenochemicals or xenobiotics, include dietetary compounds, environmental pollutants and pharmaceutical drugs require detoxification to ensure their proper elimination from the body. Hepatic microsomal cytochrome P450s (encoded by the CYP gene superfamily) are the most important enzymes catalyzing the metabolic detoxification. The efficiency of this process is enhanced because many foreign chemicals selectively induce, via activation of gene

transcription, the CYP genes required for their metabolism (Gonzalez, 1989). The transcriptional activation of CYP genes by xenobiotics occurs mostly in the liver and is mediated by the interaction of ligand-nuclear receptor complexes with enhancer sequences located upstream from the CYP gene promoters (Waxman, 1999). There are two classes of xenobiotics that induce the different sets of CYP genes: polycyclic aromatic hydrocarbons, inducing CYP1A and CYP1B, and phenobarbital, inducing CYP2A, CYP2B, CYP2C, CYP2H and CYP3A (Zelko and Negishi, 2000). Phenobarbital most effectively induces the CYP2B gene, and a phenobarbital-response element (PBRE) was first identified upstream from the CYP2B gene (Honkakoski *et al.*, 1998). The orphan nuclear receptor CAR, which was originally described as a constitutive activator of an empirical set of retinoic acid response elements (Choi *et al.*, 1997), has been shown to bind to the PBRE as a heterodimer with the retinoid X receptor alpha (Sueyoshi *et al.*, 1999), following the phenobarbital-induced translocation of the CAR receptor from the cytoplasm to the nucleus (Kawamoto *et al.*, 1999). The nuclear translocation of the CAR receptor depends on a leucine-rich peptide located at its carboxy-terminus (Zelko *et al.*, 2001). Unlike other nuclear receptors, CAR binds to its response element and activates transcription in the absence of ligand (Choi *et al.*, 1997). Its constitutive activity is inhibited by the steroids androstanol and androstenol, because these compounds promote the release of the co-factor steroid co-activator SRC-1 from the putative ligand-binding domain of CAR (Forman *et al.*, 1998), raising the interesting possibility of a regulatory mechanism repressing the constitutive activity of CAR in order to acquire the phenobarbital responsiveness. Gene-knockout studies in mice by Wei *et al.* (2000) further established that the nuclear receptor CAR activated by phenobarbital was able to enhance the expression of the CYP2B gene, and that CAR was a central component in mediating the specific xenobiotic induction of drug metabolism.

The mouse CAR (mCAR) gene, characterized by Choi *et al.* (1997), is encoded by nine exons and the mCAR protein is very similar, in terms of structural organization, to proteins belonging to a small subgroup of the nuclear hormone receptor superfamily, including the mammalian vitamin D receptor and an orphan receptor from *Xenopus laevis*. In addition, an ALIGN comparison between the mouse and human CAR proteins indicates that they share 57.5% sequence identity (Fig. 3.50).

```
                          10         20         30         40         50
CARhum MAS--------REDEL--RNCVVCGDQATGYHFNALTCEGCKGFFRRTVSKSIGPTCPFA
       :..         :.:   ::::::::.:::::::.:::::::::::::::::::.::: ::::
CARmus MTAMLTLETMASEEEYGPRNCVVCGDRATGYHFHALTCEGCKGFFRRTVSKTIGPICPFA
               10         20         30         40         50         60

               60         70         80         90        100        110
CARhum GSCEVSKTQRRHCPACRLQKCLDAGMRKDMILSAEALALRRAKQAQRRAQQTPVQLSKEQ
       : :::::.:::::::::::::::..:::::::::::::::::::::.::::::... .::...:
CARmus GRCEVSKAQRRHCPACRLQKCLNVGMRKDMILSAEALALRRARQAQRRAEKASLQLNQQQ
               70         80         90        100        110        120

              120        130        140        150        160        170
CARhum EELIRTLLGAHTRHMGTMFEQFVQFRPPAHLFIHHQPLPTLAPVLPLVTHFADINTFMVL
       .::.. :::::::::: .:.::::::.:::.::::.:.   .::::::.:::::::::::
CARmus KELVQILLGAHTRHVGPLFDQFVQFKPPAYLFMHHRPFQPRGPVLPLLTHFADINTFMVQ
              130        140        150        160        170        180

              180        190        200        210        220        230
CARhum QVIKFTKDLPVFRSLPIEDQISLLKGAAVEICHIVLNTTFCLQTQNFLCGPLRYTIEDGA
       :.::::::::.:::: .::::::::::::::: :: ::::::::::.::.::::: : .::..
CARmus QIIKFTKDLPLFRSLTMEDQISLLKGAAVEILHISLNTTFCLQTENFFCGPLCYKMEDAV
              190        200        210        220        230        240

              240        250        260        270        280        290
CARhum RVGFQVEFLELLFHFHGTLRKLQLQEPEYVLLAAMALFSPDRPGVTQRDEIDQLQEEMAL
       ..::: :::: ..::: .:. :.::::::::.:: :::::
CARmus HAGFQYEFLESILHFHKNLKGLHLQEPEYVLMAATALFSP--------------------
              250        260        270        280

              300        310        320        330        340
CARhum TLQSYIKGQQRRPRDRFLYAKLLGLLAELRSINEAYGYQIQHIQGLSAMMPLLQEICS
                                          :. .:                :
CARmus -----------------------------------GFCMQ----------------S
```

Fig. 3.50: ALIGN comparison between the amino acid sequences of the human (top) and mouse (bottom) CAR proteins.

Typically, the protein encoded by mCAR contains a very short amino-terminal region, followed by a DNA-binding domain, a hinge region and a ligand binding/dimerization domain. As a consequence of alternative splicing, the mCAR gene

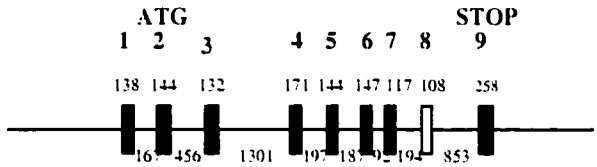
encodes two proteins. Both proteins are identical, except for the ligand binding/dimerization domains where a 107 bp out-of-frame deletion generates a truncated variant containing only 6 new amino acids instead of the last 78 residues of the mCAR protein (Choi *et al.*, 1997). A CROSSMATCH analysis between the mCAR gene and the mCAR cDNA indicates that the truncated splice variant actually lacks the region coded by the 8th exon of the mCAR gene. Interestingly, the mCAR gene shares significantly less than the typical 90% homology with its human counterpart hCAR (Baes *et al.*, 1994) (Fig. 3.50), suggesting that mCAR and hCAR are derived from distinct genes encoding two CAR isoforms (Choi *et al.*, 1997).

A GENSCAN analysis, followed by a BLAST search, of the 211748 base pairs-BAC clone rp23-395h6 reveals that the position 131490-126682 on the complementary strand corresponds to the 9 exon-mouse CAR gene (table 3.38 and Fig. 3.48). It also indicates that the mCAR gene spans a genomic region of approximately 5 kb (Fig. 3.51). The results of the Neural Network splice site prediction program analysis also confirmed the presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of the mCAR genomic region and the mCAR cDNA (table 3.39).

Exon number	Contig position	Exon length (bp)
1	131490-131352	138
2	131182-131040	142
3	130581-130451	130
4	129146-128977	169
5	128779-128640	139
6	128449-128304	145
7	128211-128095	116
8	127899-127794	105
9	126936-126682	254

Table 3.38: contig position and length of the exons of the mouse CAR gene. A truncated variant of mCAR does not contain the region encoded by the 8th exons because of an out-of-frame deletion.

mCARgene:



mCAR protein:

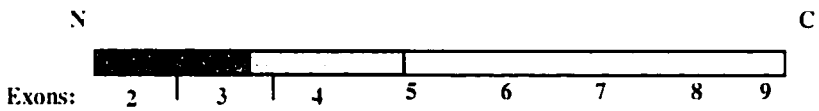


Fig. 3.51: Structural organization of the mCAR gene and its encoded protein. (Top) Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines. The alternatively spliced exon (8) is indicated by a white box. (Bottom) The dark gray box indicates the presence of a DNA-binding domain. The clear gray box corresponds to the hinge region. The white box corresponds to the ligand binding/dimerization domain. The 8th exon (white box) is deleted in the splice variant of mCAR because of an out-of-frame deletion, generating a truncated variant.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
	131350		131183
1/1	ggaaacagGTaggtaa	1/2	ccacattcAGgagacc
	131038		130582
2/2	tcttcagGTgaatgc	2/3	ctatctcacAGacgaac
	130449		129147
3/3	aaagacaGTgagttg	3/4	ctgacacAGtgatac
	128975		128780
4/4	gttcaagGTgagaac	4/5	actctacAGcctccg
	128638		128450
5/5	tcttccgGTgagtag	5/6	gacacagAGgtccct
	128302		128212
6/6	gtccatgGTgagatg	6/7	ctctcacAGcagggt
	128093		127900
7/7	tcccctgGTgaggat	7/8	ctctctcAGacagac
	127792		126937
8/8	aaagtcgGTagggag	8/9	cctccaaATgtttct

Table 3.39: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 131177, eight base pairs downstream from the 5' end of the second exon, indicating that the first exon is a non-coding exon. Initial analysis of the first 1000 bp located immediately upstream from the 5' end of the first exon, using MOTIF, reveals that the 5' flanking region of mCAR contains potential binding sites for the ROR alpha 1 receptor and the aryl hydrocarbon (Ah) receptor nuclear translocation homodimer (Arnt), 937-949 bp and 299-314 bp, respectively, upstream from the 5' end of the first exon. The ROR alpha 1 receptor is a retinoic acid receptor-related orphan receptor, isolated by Giguere *et al.* (1994), which binds DNA as a monomer and constitutively activates transcription from a large subset of hormone response elements. The Arnt homodimer is a component of the DNA-binding form of the Ah receptor and was isolated by Hoffman *et al.* (1991). The Ah receptor binds various environmental pollutants, such as dioxin, and the ligand-bound receptor activates CYP 1a1 gene transcription through interaction with specific DNA sequences, termed the xenobiotic-responsive elements (XREs) (Reyes *et al.*, 1992). The Arnt protein contains a basic helix-loop-helix motif, and therefore may be responsible for interacting with the XRE (Reyes *et al.*, 1992). The presence of a putative binding site for Arnt in the 5' flanking region of mCAR suggests that the transcription of the mCAR gene, itself encoding a receptor involved in the xenobiotic induction of drug metabolism, is under direct control of nuclear receptors activated by the presence of xenochemicals. The mCAR receptor therefore may be part of a transcriptional cascade involving several nuclear receptors activating the transcription of downstream genes, in response to the presence of environmental pollutants. In addition, the presence of a putative binding site for the ROR alpha 1 receptor may reflect a basal and constitutive transcriptional activity of the mCAR gene that does not depends on the

presence of xenobiotic elements. Finally, the promoter prediction by NNPP predicted one potential TATA box-containing promoter located 690-740 bp upstream from the 5' end of the first exon of the mCAR gene.

III-2-4-3 Mouse myelin protein zero (P0) gene

The myelin sheath is a specialized membranous organelle that serves as an electrical insulator and efficiently enhances the conduction velocity of axons in the vertebrate nervous system. Synthesized by oligodendrocytes in the central nervous system and by Schwann cells in the peripheral nervous system, this organelle consists of plasma membrane extensions that are repeatedly and tightly wrapped around axons to form compact myelin. Compaction is mediated by specialized myelin proteins. In the peripheral nervous system, the most abundant of these proteins is myelin protein zero (P0) (Roomi *et al.*, 1978), a transmembrane glycoprotein belonging to the Ig superfamily and accounting for more than 50% of the proteins of the peripheral nervous system myelin (Lemke et al., 1988). Several studies have demonstrated that myelin protein zero is essential for peripheral nerve development and function. The phenotypes of mice with mutations in the P0 gene are reminiscent of human neuropathies, such as Charcot-Marie-Tooth disease type 1B, Dejerine-Sottas syndrome and congenital hypomyelination (Hayasaka *et al.*, 1995; Kulkens *et al.*, 1993; Su *et al.*, 1993), and include hypomyelination, loss of motor control, tremors and abnormal myelin sheath (Giese *et al.*, 1992). Heterologously overexpressed P0 protein accumulates at membrane contact sites between neighboring cells and induces homophilic adhesion (D'Urso *et al.*, 1990). It has been suggested that both the carbohydrate structure of P0 (Filbin and Tennekoon, 1991), and the protein backbone within the extracellular Ig-like domain of P0 (Zhang *et al.*, 1996) were involved in these interactions, implying that P0 holds together adjacent membrane layers of myelin because of its homophilic adhesion

capacity. X-ray crystallographical data revealed that the extracellular domains of P0 are assembled as tetramers, adhering to P0 tetramers on the opposing myelin layer (Shapiro *et al.*, 1996). In addition, the cytoplasmic domain of P0 has been suggested to be involved in the compaction of intracellular surfaces of membrane layers, thus adding another level of compaction of the myelin sheath (Ding and Brunden, 1994).

During Schwann cells development, the P0 gene is coordinately induced with genes encoding other myelin-specific proteins, such as myelin basic protein and PMP-22 (Lemke, 1988). This regulation of the P0 gene reflects a combination of transcriptional and translational controls. The 5' flanking region of the P0 gene contains most of the regulatory elements required for tissue-specific and developmentally accurate P0 expression (Brown and Lemke, 1997). In addition, the overexpression of the P0 protein in peripheral nerves causes a dose-dependent dismyelinating neuropathy, manifested by delayed nerve development, that reveals the importance of precisely regulated gene dosage for normal myelination (Wrabetz *et al.*, 2000).

A GENSCAN analysis, followed by a BLAST search, of the mouse clone rp23-395h6 indicates that the genomic segment at position 191305-184608 on the complementary strand corresponds to the 6 exon-mouse myelin P0 gene (Fig. 3.48). A CROSSMATCH analysis between the mouse clone rp23-395h6 and the P0 gene cDNA indicates that this gene has six exons, covering approximately 7 kb of genomic DNA (table 3.40 and Fig. 3.52).

Myelin P0 gene:



Fig. 3.52: Structural organization of the myelin P0 gene. Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines.

In addition, the results of the Neural Network splice site prediction program analysis also confirmed the presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of the P0 genomic region and the P0 cDNA (table 3.41).

Exon number	Contig position	Exon length (bp)
1	191305-189879	1426
2	187166-187000	166
3	186774-186561	213
4	186029-185894	135
5	185753-185693	60
6	185585-184608	977

Table 3.40: contig position and length of the exons of the mouse P0 gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
1/1	189877 tctttggGTaagtga	1/2	187167 tgtttacAGtgctgt
2/2	186998 catttcgGTgagtgc	2/3	186775 tcctcatAGatcttc
3/3	186559 gaaaaagGTgtgaga	3/4	186030 ccctcctAGtgccca
4/4	185892 ggctcagGTaaggggg	4/5	185754 ttgccacAGtgccat
5/5	185691 gcggcagGTtagtag	5/6	185586 tccccgcAGacgcca

Table 3.41: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

The putative translation start site (ATG) is located at position 189946, in the first exon, and the putative termination site (TAG) is located at position 185485, in the last exon (Fig.3.52). An ALIGN comparison between the mouse and human myelin protein zero indicates that they share 35.7% sequence identity (Fig. 3.53).

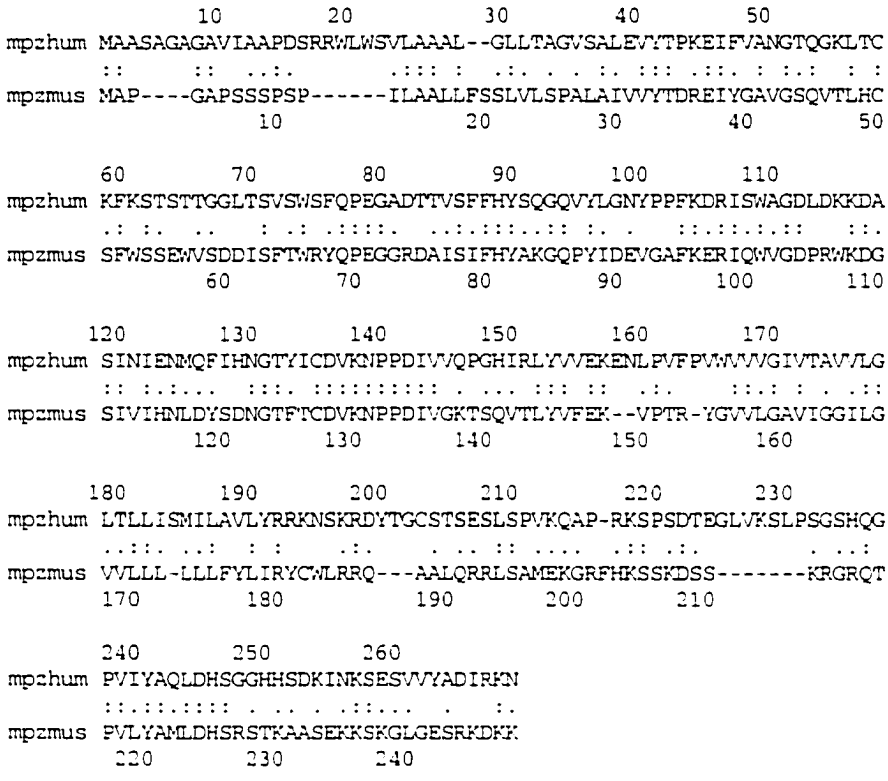


Fig. 3.53: ALIGN comparison between the amino acid sequences of the human (top) and mouse (bottom) myelin protein zero.

Initial analysis of the 1000 bp immediately upstream from the initiation of transcription, using MOTIF, indicates the presence of several potential transcription factor binding sites. Brown and Lemke (1997) determined that a potent transcriptional regulatory region was present from position -350 to +45, and that the basal P0 promoter was identified as two transcription factor binding sites, a Sp1 site and a NF-Y site, located at positions -48 to -59 and -66 to -79, respectively. None of these two sites were detected, using MOTIF, in the 5' flanking region of the P0 gene. However, MOTIF predicted other potential binding sites for transcription factors in the genomic region at position -350 to +45, such as MZF1, C/EBP, Oct-1, Lyf-1 and Pbx-1 binding sites. In addition, two potential promoter region were predicted by NNPP, located 395-445 bp and 20-70 bp, respectively, upstream from the initiation of transcription. The presence of

a putative NF-Y CCAAT box binding site 66-79 bp upstream from the initiation of transcription suggests that the promoter region located 20-70 bp upstream from the P0 gene may correspond to the actual promoter.

III-2-4-4 Mouse apoA-II gene

Plasma levels of high density lipoproteins (HDL) are inversely correlated to the risk of artery disease. HDLs protect against artery disease by removing cholesterol from peripheral tissues and transporting it to the liver (Tall, 1990). The two major apolipoproteins associated with HDL are apolipoproteins A-I (apoA-I) and A-II (apoA-II). In human, apoA-I and apoA-II constitute approximately 60% and 20%, respectively, of the protein mass of HDL (Tall, 1990). Although the role of apoA-I in maintaining HDL structure is well established (Tall, 1990), the role of apoA-II is still unclear. Unlike apoA-I, apoA-II is not required for assembly of HDL (Deeb *et al.*, 1990). However, apoA-II has been shown to affect several aspects of HDL metabolism, including cholesterol transport to the liver (Barbaras *et al.*, 1987), HDL composition (Zhong *et al.*, 1994) and HDL levels (Zhong *et al.*, 1994). In addition, Warden *et al.* (1993) showed that the overexpression of apoA-II in transgenic mice on both chow and atherogenic diets promoted the development of atherosclerosis, suggesting a significant role for apoA-II in regulating HDL metabolism and plasma cholesterol levels.

A GENSCAN analysis, followed by a BLAST search, of the BAC clone rp23-395h6 indicates that the genomic segment at position 120437-119156 on the complementary strand corresponds to the mouse apoA-II gene (Fig. 3.48). It also indicates that this gene spans a genomic region of approximately 1.3 kb (table 3.42 and Fig. 3.54). In addition, the results of the Neural Network splice site prediction program analysis also confirmed the presence of potential splice sites at the exon-intron borders

indicated by the ALIGN comparison of the apoA-II genomic region and the apoA-II cDNA (table 3.43).

ApoA-II gene:

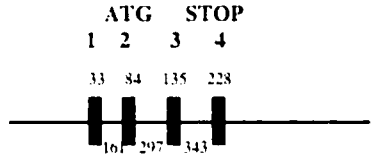


Fig. 3.54: Structural organization of the ApoA-II gene. Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines.

Exon number	Contig position	Exon length (bp)
1	120437-120404	33
2	120241-120159	82
3	119859-119727	132
4	119382-119156	226

Table 3.42: contig position and length of the exons of the mouse apoA-II gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
	120402		120242
1/1	agaatcgGTatgctg	1/2	tctcaccAGcagcac
	120157		119860
2/2	ctggaagGTgaggtt	2/3	tttgtgtAGgagctt
	119725		119383
3/3	aggtcaaGTaagtct	3/4	gttccccAGggcata

Table 3.43: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 120211, in the second exon, indicating that the first exon is a non-coding exon. An ALIGN

comparison between the mouse and human Apo A-II proteins indicates that they share 58.8% sequence identity (Fig. 3.55).

```
               10         20         30         40         50         60
apohum MKLLAATVLLLTICSLEGALVRRQAKEPCVESLVSQYFQTVTDYGKDLMEKVKSPELQAE
       ::::: ..:::::::::::::::::  : ..:: .:::::.:.::::::::.:. :.:..
apomus MKLLAMVALLVTICSLEGALVKRQADGPDMQSLFTQYFQSMTEYGKDLVEKAKTSEIQSQ
               10         20         30         40         50         60

               70         80         90         100
apohum AKSYFEKSKEQLTPLIKKAGTELVNFLSYFVELGTQPA--TQ
       .:.::::..:::::..::: :::::: ...:  .::  ..
apomus VKAYFEKTHEQLTPLVRSAGTSLVNFFSSLMNLEEKPAPAAK
               70         80         90        100
```

Fig.3.55: ALIGN comparison between the amino acid sequences of the human (top) and mouse (bottom) apo A-II proteins.

Analysis of the 1000 bp located immediately upstream from the 5' end of the apoA-II gene, using MOTIF, reveals several interesting features. First, besides the presence of potential binding sites for ubiquitous transcription factors, such as Oct-1, GATA and p53 binding sites, a potential binding site for sterol regulatory element-binding protein 1 (SREBP-1) is located 54-64 bp upstream from the 5' end of the first exon of apoA-II. SREBP-1 is a basic-helix-loop-helix-leucine zipper protein that controls the expression of genes encoding proteins essentials for cholesterol biosynthesis/uptake and fatty acid biosynthesis (Yokoyama *et al.*, 1993). The presence of a potential binding site for SREBP-1 may be related to the fact that apoA-II may play a significant role in regulating HDL metabolism and plasma cholesterol levels (Warden *et al.*, 1993). Surprisingly, a potential interferon-stimulated response element (ISRE) is located 340-354 bp upstream from the 5' end of the apoA-II gene. The mouse apoA-II gene is not known to be transcriptionally regulated by interferons, and further experiments are needed to determine whether it is. Finally, the promoter prediction program NNPP predicted three potential promoter regions in the 5' flanking region of the apoA-II gene, 467-517 bp, 403-453 bp and 208-258 bp upstream from the 5' end of the apoA-II gene, respectively. Analysis of the 5' flanking region of the apoA-II gene,

using MOTIF, revealed the presence of a potential Sp1 GC box-binding site, 443-452 bp upstream from the first exon, and a potential NF-Y CCAAT-box binding site, 466-476 bp upstream from the first exon, suggesting that the predicted promoter located 467-517 bp upstream from the apoA-II gene may correspond to the actual promoter.

III-2-4-5 Mouse NDUFS2 gene

The oxidative phosphorylation system of the mitochondria uses the products of approximately 60 nuclear genes and 13 mitochondrial genes to generate ATP. These proteins are organized into five large complexes: electron-transport chain complexes I-IV and ATP synthase complex (complex V) (Hatefi, 1985). Complex I, also known as nicotinaminide adenine dinucleotide (NADH): ubiquinone oxidoreductase is the largest complex. Studies from bovine heart mitochondria have identified at least 35 nuclear-encoded and 7 mitochondrial-encoded complex I subunits (Skehel *et al.*, 1998). Complex I is embedded in the inner mitochondrial membrane and serves to dehydrogenate NADH and to shuttle electrons to coenzyme Q. This electron transport generates a proton gradient across the inner mitochondrial membrane and provides a proton motive force used to synthesize ATP. The 35 nuclear-encoded proteins from bovine heart complex I identified so far consists of 3 flavoproteins, 7 iron-sulfur (Fe-S) proteins, 24 hydrophobic proteins (Galante and Hatefi, 1979), and a 17.2 kD subunit that remains to be identified (Skehel *et al.*, 1998). The flavoproteins and Fe-S proteins protrude from the inner mitochondrial membrane into the mitochondrial matrix (Grigorieff, 1998), forming an "arm" which binds and transfers electrons to NADH. The hydrophobic proteins are embedded in the phospholipid bilayer of the inner mitochondrial membrane and apparently mediate proton translocation from the mitochondrial matrix into the intermembrane space (Belogrudov and Hatefi, 1994). The NDUFS2 gene encodes a 49 kD Fe-S protein that is part of the eukaryotic

mitochondrial complex I (Smeitink *et al.*, 1998). In humans, this protein is one of the seven nuclear-encoded counterparts of the *E. coli* Nuo proteins that appear to form a small form of complex I, consisting of 14 subunits (Weidner *et al.*, 1993). Although the role of the NDUFS2 protein still remains to be established, mutational detection studies in enzymatic complex I deficient patients by Loeffen *et al.* (2001) revealed three missense mutations in the NDUFS2 gene, indicating that the NDUFS2 protein is essential for maintaining the integrity of the human mitochondrial complex I.

The human NDUFS2 cDNA was cloned and characterized by Loeffen *et al.* (1998). The open reading frame of the human NDUFS2 gene consists of 1392 bp, coding for 463 amino acids, and showing 96% sequence identity with the corresponding bovine translation product.

NDUFS2 gene:

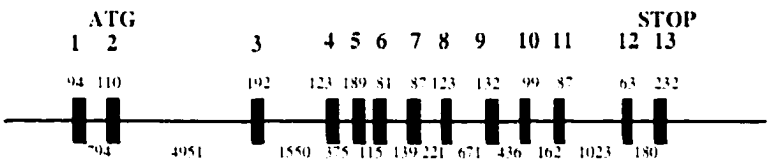


Fig. 3.56: Structural organization of the NDUFS2 gene. Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines

A GENSCAN analysis, followed by a BLAST search, of the mouse clone rp23-395h6 indicates that the genomic segment at position 98439-110668 corresponds to the mouse NDUFS2 gene (Fig. 3.48). A CROSSMATCH analysis between the mouse NDUFS2 cDNA (Genbank accession number BC003898) and the mouse clone rp23-395h6 reveals that the mouse NDUFS2 is encoded in 13 exons (table 3.44 and Fig. 3.56). In addition, the results of the Neural Network splice site prediction program analysis also confirmed the presence of potential splice sites at the exon-intron borders

indicated by the ALIGN comparison of the NDUFS2 genomic region and the NDUFS2 cDNA (table 3.45).

Exon number	Contig position	Exon length (bp)
1	98439-98533	94
2	99330-99436	106
3	104390-104580	190
4	106133-106253	120
5	106630-106817	187
6	106935-107012	77
7	107155-107240	85
8	107463-107582	119
9	108258-108387	129
10	108826-108921	95
11	109087-109170	83
12	110197-110254	57
13	110440-110668	228

Table 3.44: contig position and length of the exons of the mouse NDUFS2 gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
1/1	98535 cccagcagGTgaggag	1/2	99329 attccccAGaggtgc
2/2	99434 tggaatgGTgagtgc	2/3	104389 cacttctcAGatgtgg
3/3	104582 tctgcagGTgtggag	3/4	106132 tttccccAGgccctt
4/4	106255 atccgagGTatgttt	4/5	106629 gtctcacAGtgctct
5/5	106819 gcaccagGTgagtga	5/6	106934 gcttctagAGgacctа
6/6	107014 ggaggagGTgagaca	6/7	107154 cgtccatAGatgctg
7/7	107242 gattcagGTaggaga	7/8	107462 tgccctatAGtggggt
8/8	107584 acgatagGTaagatt	8/9	108257 tcatttcAGgtacct
9/9	108389 gatgaagGTtggcct	9/10	108825 ttccttcAGacgtcc
	108923		109086

10/10	tcctaagGTgagggg	10/11	cttggacAGggagag
	109172		110196
11/11	ccacctgGTaagaca	11/12	tttctgcAGgctggt
	110256		110439
12/12	atcatagGTatgaga	12/13	acctttAGgtaccc

Table 3.45: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 99387, in the second exon, indicating that the first exon is non-coding. The mouse NDUFS2 protein consists of 412 amino acids, 51 amino acids shorter than its human counterpart. An ALIGN comparison between the amino acid sequences of the mouse and human NDUFS2 proteins reveals that the mouse NDUFS2 protein lacks a region corresponding to the first 51 amino acids of the human NDUFS2 protein, but, besides this region, the two proteins are highly conserved (Fig. 3.57).

```
mouse  M----------------------------------------------------YPSKETAH
       :                                                    ::::::::
human  MAALRALCGFRGVAAQVLRPGAGVRLPIQPSRGVRQWQPDVEWAQQFGGAVMYPSKETAH
               10        20        30        40        50        60

      10        20        30        40        50        60
mouse  WKPPPWNDVDILKEKAVTNMTLNFGPQHPAAHGVLRLVLELSGEMVRKCDPHIGLLHRGT
       ::::::::::  :.  : :.:::::::::::::::::::.:::::::::::::::::::::
human  WKPPPWNDVDPPKDTIVKNITLNFGPQHPAAHGVLRLVMELSGEMVRKCDPHIGLLHRGT
               70        80        90       100       110       120

      70        80        90       100       110       120
mouse  EKLIEYKTYLQALPYFDRLDYVSMMCNEQAYSIAVEKLLNIQPPPRAQWIRVLFGEITRI
       ::::::::::::::::::::::::::::::::.:::::::.:::::::::::::::::::.
human  EKLIEYKTYLQALPYFDRLDYVSMMCNEQAYSLAVEKLLNIRPPPRAQWIRVLFGEITRL
               130       140       150       160       170       180

     130       140       150       160       170       180
mouse  LNHIMAVTTHALDIGAMTPFFWMFEEREKMFEFYERVSGARMHAAYIRPGGVHQDLPLGL
       :::::::::::::.:::::::.:::::::::::::::::::::::::::::::::::::::
human  LNHIMAVTTHALDLGAMTPFFWLFEEREKMFEFYERVSGARMHAAYIRPGGVHQDLPLGL
               190       200       210       220       230       240

     190       200       210       220       230       240
mouse  LDDIYEFSKNFSLRIDEVEEMLTNNRIWRNRTVDIGVVTAEDALNYGFSGVMLRGSGIQW
       .:::::::::::::.::.::.:::::::::::::.:::::::.:::::::::::::::::::
human  MDDIYQFSKNFSLRLDELEELLTNNRIWRNRTIDIGVVTAEEALNYGFSGVMLRGSGIQW
               250       260       270       280       290       300

     250       260       270       280       290       300
```

```
mouse  DLRKTQPYDVYDQVEFDVPIGSRGDCYDRYLCRVEEMRQSLRIIEQCLNKMPPGEIKVDD
       :::::::::::::::::::.:::::::::::::::::::::::: ::::::::::::::
human  DLRKTQPYDVYDQVEFDVPVGSRGDCYDRYLCRVEEMRQSLRIIAQCLNKMPPGEIKVDD
              310       320       330       340       350       360

     310       320       330       340       350       360
mouse  AKVSPPKRAEMKTSMESLIHHFKLYTEGYQVPPGATYTAIEAPKGEFGVYLVSDGSSRPY
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
human  AKVSPPKRAEMKTSMESLIHHFKLYTEGYQVPPGATYTAIEAPKGEFGVYLVSDGSSRPY
              370       380       390       400       410       420

     370       380       390       400       410
mouse  RCKIKAPGFAHLAGLDKMSKGHMLADVVAIIGTQDIVFGEIDR
       ::::::::::::::::::::::::::::::::::::::::.::
human  RCKIKAPGFAHLAGLDKMSKGHMLADVVAIIGTQDIVFGEVDR
              430       440       450       460
```

Fig. 3.57: ALIGN comparison between the amino acid sequences of the mouse and human NDUFS2 proteins.

A MOTIF analysis on the first 1000 bp located upstream from the 5' end of the NDUFS2 gene reveals potential binding sites for ubiquitous transcription factors, such as GATA, p53, c-Ets, NF-kB, AP-1, CREB, c-Myb, E2F and Oct-1 binding sites. Because of the important place of this protein in the energetics metabolism of the cell, it is expected that the transcriptional activity of the corresponding NDUFS2 gene is highly regulated by one or more of the transcription factors mentioned above. A search for the promoter region, using NNPP, on the first 1000 bp located upstream from the 5' end of the NDUFS2 gene did not predict any TATA-containing promoter region, suggesting that the NDUFS2 gene transcriptional activity is initiated from a TATA-less promoter.

III-2-4-6 Mouse beta4Gal-T3 gene

Enzymatic glycosylation of proteins and lipids is an essential process that involves a very large number of glycosyltransferases catalyzing the synthesis of complex oligosaccharides and glycoconjugates (Kleene and Berger, 1993). Glycosyltransferases have high donor and acceptor substrate specificities, and are

usually limited to catalysis of only one glycosydic linkage. The UDP-Gal:beta-Glc NAc beta-1,4-galactosyltransferase (beta4Gal-T1) was the first glycosyltransferase to be isolated and cloned (Shaper *et al.*, 1986), but a comparison with the EST database led to the identification of five novel beta4Gal-T genes with similarities to the beta4Gal-T1 gene, designated beta4Gal-T2 to -T6 (Almeida *et al.*, 1997; Schwientek *et al.*, 1998; Sato *et al.*, 1998; Nomura *et al.*, 1998). All seven members of the beta4Gal-T family of proteins have specificity exclusively for the donor substrate UDP-Gal, and all transfer Gal in a beta-1,4-linkage to similar acceptor sugars: GlcNAc, Glc and Xyl. The fine specificity of the beta4Gal-T family of proteins for different types of glycoconjugates and branch points of oligosaccharide structures has not been determined. However, Almeida *et al.* (1997) demonstrated that beta4Gal-T3 was not very efficient in using several different glycoprotein substrates but was rather very efficient in using glycolipid substrates, suggesting that beta4Gal-T3 may be involved in glycolipid biosynthesis.

The human beta4Gal-T3 cDNA,cloned and characterized by Almeida *et al.* (1997), consists of six exons. Also, the predicted coding region of human beta4Gal-T3 has a single translation start site (ATG) placed immediately upstream from a sequence encoding a potential hydrophobic transmembrane domain. The predicted coding sequence yields a type II transmembrane glycoprotein with an amino-terminal cytoplasmic domain of 4 residues, a transmembrane domain of 18 residues and a catalytic domain of 371 residues (Almeida *et al.*, 1997). The genomic organization of this gene is identical to that of the human beta4Gal-T1 and beta4Gal-T2 genes, suggesting that it is part of a homologous gene family that arose through gene duplication.

A GENSCAN analysis, followed by a BLAST search, of the mouse BAC clone rp23-395h6 revealed that the genomic segment at position 74328-68638 on the complementary strand corresponds to the mouse beta4Gal-T3 gene (Fig. 3.48). This gene is encoded by seven exons and spans a genomic segment of approximately 5.7 kb

(table 3.46 and Fig. 3.58). In addition, the results of the Neural Network splice site prediction program analysis also confirmed the presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of the beta4Gal-T3 genomic region and the beta4Gal-T3 cDNA (table 3.47).

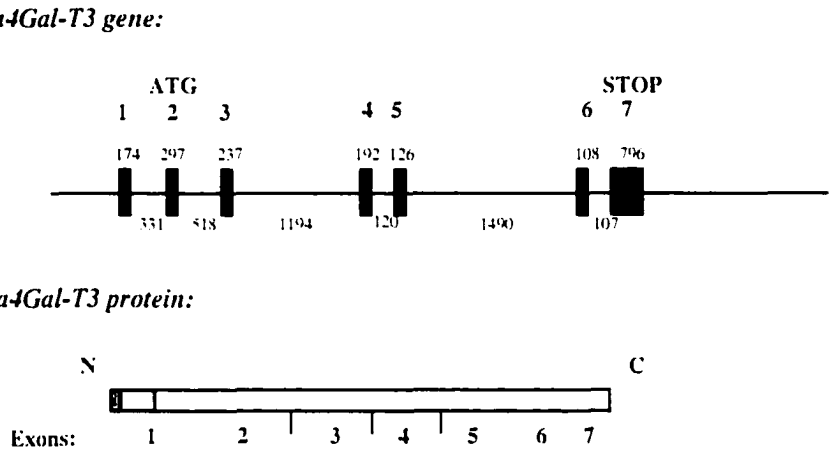


Fig. 3.58: Structural organization of the beta4Gal-T3 gene and its encoded protein. (Top) Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines. (Bottom) The dark gray box indicates the presence of a putative cytoplasmic domain. The clear gray box corresponds to a putative transmembrane domain. The white box corresponds to the putative catalytic domain.

Exon number	Contig position	Exon length (bp)
1	74328-74154	174
2	73820-73526	294
3	73006-72771	235
4	71574-71384	190
5	71261-71139	122
6	69645-69541	104
7	69421-68638	783

Table 3.46: contig position and length of the exons of the mouse beta4Gal-T3 gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
	74152		73821
1/1	aacacacGTacgtgt	1/2	tgcttccAGaaatcc

	73524		73007
2/2	ttcttagGTgagtag	2/3	taccaacAGtgggtc
	72769		71575
3/3	ccaccagGTatcccc	3/4	atcttctAGgctgga
	71382		71262
4/4	gatacagGTagaggg	4/5	tgtccttAGcctccc
	71137		69646
5/5	ctaccagGTcagact	5/6	ttctcacAGggtccg
	69539		69422
6/6	cccacagGTaggaaa	6/7	caccctcAGatttga

Table 3.47: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 73785, in the second exon, indicating that the first exon is non-coding and that the coding region of the mouse beta4Gal-T3 gene consists of six exons. The mouse beta4Gal-T3 protein consists of 395 amino acids and an ALIGN comparison between the amino acid sequences of the mouse and human beta4Gal-T3 proteins indicate that they share 96.5% sequence identity (Fig. 3.59).

```
               10         20         30         40         50         60
mouse  MLRRLLERPCTLALLVGSQLAVMMYLSLGGFRSLSALFGRDPGPTFDYSHPHDVYSNLSH
       :::::::::::::::::::::::::::::::::::::::: ::::::::.:::::::::
human  MLRRLLERPCTLALLVGSQLAVMMYLSLGGFRSLSALFGRDQGPTFDYSHPRDVYSNLSH
               10         20         30         40         50         60

               70         80         90        100        110        120
mouse  LPAAPGAAGAPPAQALPYCPERSPFLVGPVSVSFSPVPSLAEIVERNPRVESGGRYRPAG
       ::.:::  : :  :.:::::::::.:::::::::::::::::::::::: ::::::::
human  LPGAPG--GPPAPQGLPYCPERSPLLVGPVSVSFSPVPSLAEIVERNPRVEPGGRYRPAG
                 70         80         90        100        110

              130        140        150        160        170        180
mouse  CEPRSRTAIIVPHRAREHHLRLLLYHLHPFLQRQQLAYGIYVIHQAGNGTFNRAKLLNVG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
human  CEPRSRTAIIVPHRAREHHLRLLLYHLHPFLQRQQLAYGIYVIHQAGNGTFNRAKLLNVG
      120        130        140        150        160        170

              190        200        210        220        230        240
mouse  VREALRDEEWDCLFLHDVDLLPENDHNLYVCDPRGPRHVAVAMNKFGYSLPYPQYFGGVS
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
human  VREALRDEEWDCLFLHDVDLLPENDHNLYVCDPRGPRHVAVAMNKFGYSLPYPQYFGGVS
      180        190        200        210        220        230

              250        260        270        280        290        300
```

```
mouse  ALTPDQYLKMNGFPNEYWGWGGEDDDIATRVRLAGMKISRPPTSVGHYKMVKHRGDKGNE
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
human  ALTPDQYLKMNGFPNEYWGWGGEDDDIATRVRLAGMKISRPPTSVGHYKMVKHRGDKGNE
      240       250       260       270       280       290

               310       320       330       340       350       360
mouse  ENPHRFDLLVRTQNSWTQDGMNSLTYRLLARELGPLYTNITADIGTDPRGPRAPSGPRYP
       :::::::::::::::::::::::::.::::::::::::::::::::::::::::::::::::
human  ENPHRFDLLVRTQNSWTQDGMNSLTYQLLARELGPLYTNITADIGTDPRGPRAPSGPRYP
      300       310       320       330       340       350

               370       380       390
mouse  PGSSQAFRQEMLQRRPPARPGPLPTANHTAPRGSH
       :::::::::::::::::::::::: :::::: ::::
human  PGSSQAFRQEMLQRRPPARPGPLSTANHTALRGSH
      360       370       380       390
```

Fig. 3.59: ALIGN comparison between the amino acid sequences of the mouse and human beta4Gal-T3 proteins.

A TMHMM analysis to search for a putative transmembrane domain inside the mouse beta4Gal-T3 protein indicates the presence of a potential transmembrane domain located immediately downstream from the translation start site (ATG), at amino acid position 12-31 (Fig. 3.60).

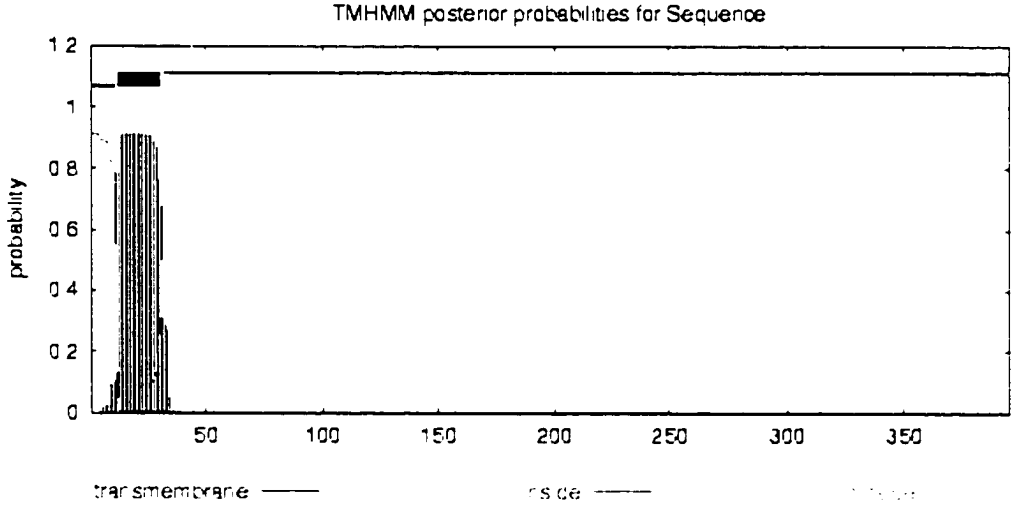


Fig. 3.60: TMHMM search for transmembrane domain of the mouse beta4Gal-T3 protein. The putative transmembrane domain is located at amino acid position 12-31.

Taken together, these data reveal that the mouse and human beta4Gal-T3 proteins share a high level of similarity, in terms of protein structure and protein sequence. A MOTIF analysis on the first 1000 bp located upstream from the 5' end of the mouse beta4Gal-T3 gene indicates several potential binding sites for transcription factors, such as AP-1, GATA, p45, E2F, C/EBP and p300 binding sites, in addition to potential CREB binding sites and a potential glucocorticoid receptor binding site (275-293 bp upstream from the 5' end of the beta4Gal-T3 gene). Interestingly, cAMP and glucocorticoids have been shown to control several aspects of glucose metabolism (Hanson and Reshef, 1997; Yoon *et al.*, 2001), and may have an important role in regulating the transcription of an enzyme involved in the glycosylation of proteins and lipids, and in the biosynthesis of oligosaccharides. Finally, the promoter prediction program NNPP did not predict any TATA-containing promoter region in the 1000 bp located immediately upstream from the beta4Gal-T3 gene, suggesting that the promoter for the mouse beta4Gal-T3 gene is a TATA-less promoter.

III-2-4-7 Mouse Ppox gene

Protoporphyrinogen oxidase (Ppox) is the penultimate enzyme in the heme biosynthetic pathway, catalyzing the six-electron oxidation of protoporphyrinogen to protoporphyrin. In mammalian cells, this enzyme is associated with the inner mitochondrial membrane and utilizes O_2 as the terminal electron acceptor (Dailey, 1990). The mouse enzyme has been purified (Dailey and Karr, 1987). It has a molecular weight of ~55 kDa and contains a flavin moiety. Ppox, as well as the other enzymes of the heme biosynthetic pathway, is induced several fold during erythroid differentiation (Sassa, 1976).

Ppox is of particular interest for two distinct reasons. In human, a deficiency in Ppox leads to a rare genetic disorder known as variegate porphyria (Deybach *et al.*,

1981). This is an autosomally dominant disorder most common in South Africa and characterized by a chronic photodermatitis and several neurovisceral abnormalities, including abdominal pain, tachycardia and psychiatric disturbances. Also, in plants, Ppox is the cellular target of several herbicides, including diphenyl ethers and pyrazole phenyl ethers (Matringe *et al*, 1992). These herbicides inhibit Ppox, resulting in a cellular accumulation of protoporphyrinogen which is subsequently converted to protoporphyrin by plasma membrane peroxidases. The accumulation of the photosensitizer free porphyrin leads to severe photochemical damages to the plant exposed to visible radiation.

The mouse Ppox cDNA also was cloned and characterized by Dailey *et al.* (1995). The cDNA sequence contains an open reading frame of 1431 base pairs, encoding a 477 amino acid translation product. Two mRNA species were isolated from mouse hepatoma cells with a respective length of 1.8 kb and 3.6 kb. According to Dailey *et al.* (1995), the longer mRNA species corresponds to a second downstream polyadenylation site. Mouse and human protoporphyrinogen oxidases have 88.7% sequence identity and regions near the amino- and carboxy-termini are the most homologous (Dailey *et al.*, 1995) (Fig. 3.61). Also, the deduced amino acid sequence of mouse Ppox gives rise to a protein without either a typical mitochondrial targeting sequence or a computer-predicted membrane domain, even though the enzyme is located on the cytoplasmic side of the inner mitochondrial membrane (Dailey *et al.*, 1995).

```
               10        20        30        40        50        60
ppoxhu MGRTVVVLGGGISGLAASYHLSRAPCPPKVVLVESSERLGGWIRSVRGPNGAIFELGPRG
       :::::.:::::::::::::::: :.: ::::.:::.:.:::::::.:: .::::::::::
ppoxmu MGRTVIVLGGGISGLAASYHLIRGPSPPKVILVEGSKRLGGWIRSIRGSDGAIFELGPRG
               10        20        30        40        50        60

               70        80        90       100       110       120
ppoxhu IRPAGALGARTLLLVSELGLDSEVLPVRGDHPAAQNRFLYVGGALHALPTGLRGLLRPSP
       :::::::::::::::::::.:::::::::::::::::::::.:: ::.::::::::::
ppoxmu IRPAGALGARTLLLVSELGLESEVLPVRGDHPAAQNRFLYVGGTLHPLPSGLRGLLRPSP
               70        80        90       100       110       120

              130       140       150       160       170       180
ppoxhu PFSKPLFWAGLRELTKPRGKEPDETVHSFAQRRLGPEVASLAMDSLCRGVFAGNSRELSI
       :::::::::::::: ::::::::::::::::::::::::::::::::::::::::::::
```

```
ppoxmu PFSKPLFWAGLRELLKPRGKEPDETVHSFAQRRLGPEVASLAMDSLCRGVFAGNSRELSI
              130       140       150       160       170       180

              190       200       210       220       230       240
ppoxhu RSCFPSLFQAEQTHRSILLGLLLGAGRTPQPDSALIRQALAERWSQWSLRGGLEMLPQAL
       :::::::::::::::::::::::::::..:::::.::::: ::::::::::::::.:::::
ppoxmu RSCFPSLFQAEQTHRSILLGLLLGAGQSPQPDSSLIRQARAERWSQWSLRGGLEVLPQAL
              190       200       210       220       230       240

              250       260       270       280       290       300
ppoxhu ETHLTSRGVSVLRGQPVCGLSLQAEGRWKVSLRDSSLEADHVISAIPASVLSELLPAEAA
       ..::.:.::.:: :::::::::: :::::::: ::::::::.:::::::: ::.::::::::
ppoxmu HNHLASKGVTVLSGQPVCGLSLQPEGRWKVSLGDSSLEADHIISAIPASELSKLLPAEAA
              250       260       270       280       290       300

              310       320       330       340       350       360
ppoxhu PLARALSAITAVSVAVVNLQYQGAHLPVQGFGHLVPSSEDPGVLGIVYDSVAFPEQDGSP
       :::: ::.: ::::::::::::.:: ::::::::::::::: ::::::::::::::::.:
ppoxmu PLARILSTIKAVSVAVVNLQYRGACLPVQGFGHLVPSSEDPTVLGIVYDSVAFPEQDGNP
              310       320       330       340       350       360

              370       380       390       400       410       420
ppoxhu PGLRVTVMLGGSWLQTLEASGCVLSQELFQQRAQEAAATQLGLKEMPSHCLVHLHKNCIP
       :.:::::::::: ::: :.:.:  :: :::::.:::::::::::::: ::::::::::::::
ppoxmu PSLRVTVMLGGYWLQKLKAAGHQLSPELFQQQAQEAAATQLGLKEPPSHCLVHLHKNCIP
              370       380       390       400       410       420

              430       440       450       460       470
ppoxhu QYTLGHWQKLESARQFLTAHRLPLTLAGASYEGVAVNDCIESGRQAAVSVLGTEPNS
       :::.::::::.:: ::::::.:::::::::::::::::::::::::::::.::::: ::
ppoxmu QYTIGHWQKLDSAMQFLTAQRLPLTLAGASYEGVAVNDCIESGRQAAVAVLGTESNS
              430       440       450       460       470
```

Fig.3.61: ALIGN comparison between the amino acid sequences of the human (top) and mouse (bottom) protoporphyrinogen oxidases.

A GENSCAN analysis, followed by a BLAST search, of the BAC clone rp23-395h6 revealed that the genomic segment at position 64368-68536 corresponds to the mouse Ppox gene (Fig. 3.48). A CROSSMATCH analysis between the BAC clone rp23-395h6 and the mouse Ppox cDNA indicates that this gene consists of 13 exons, encompassing a genomic region of approximately 4.1 kb (table 3.48 and Fig. 3.62). In addition, the results of the Neural Network splice site prediction program analysis also confirmed the presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of the Ppox genomic region and the Ppox cDNA (table 3.49).

Ppox gene:

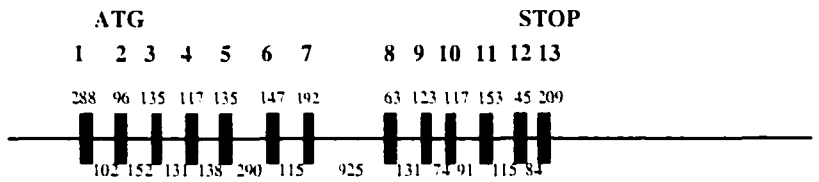


Fig. 3.62: Structural organization of the Ppox gene. Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines

Exon number	Contig position	Exon length (bp)
1	64368-64656	288
2	64761-64855	94
3	65007-65141	134
4	65274-65389	115
5	65528-65660	132
6	65955-66099	144
7	66216-66406	190
8	67334-67394	60
9	67529-67647	118
10	67725-67835	110
11	67933-68082	149
12	68201-68243	42
13	68330-68536	206

Table 3.48: contig position and length of the exons of the mouse Ppox gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
1/1	64658 taaatcaGTaagtgt	1/2	64760 ttttcccAGgcctta
2/2	64857 tcctaagGTgagtcc	2/3	65006 tggcgccAGgtgatc
3/3	65143 gctcctgGTgagcgt	3/4	65273 gagaacaAGgtttct
4/4	65391 gcctcagGTaacacc	4/5	65527 tccatgcAGgggggct
5/5	65662 acctgagGTgacatt	5/6	65954 tactcttaAGgtggca
6/6	66101 ggcgcagGTgagggg	6/7	66215 tcctttAGgacaga

	66408		67333
7/7	ctggaaggGTaggaga	7/8	ctcactctAGgtgtct
	67396		67528
8/8	gcttcagGTaatggg	8/9	tgctctccAGagctca
	67649		67724
9/9	tgtccagGTataaga	9/10	ttctccaAGggattt
	67837		67932
10/10	agtgactGTgagagg	10/11	tcctcgtAGgtgatg
	68084		68200
11/11	acacaagGTgagtca	11/12	tctccttAGaactgt
	68245		68329
12/12	aaactagGTaagctg	12/13	accttcAGactcag

Table 3.49: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 64768, in the second exon, indicating that the 5' UTR is approximately 400 base pair in length. It is possible that there is an additional upstream sequence because the procedure used to generate the cDNA sequence does not usually produce full length 5' UTR sequences (Dailey *et al.*, 1995). A putative translation termination site is located at position 68469, in the last exon, and a putative polyadenylation site (AGTAAA) is located 47 bp downstream from the termination site. Initial analysis of the first 1000 bp located upstream from the 5' end of the first exon, using MOTIF, indicates the presence of potential binding sites for transcription factors, such as AP-1, STAT, GATA, E2F and c-Ets binding sites. In addition, a potential barbiturate-inducible element (also called a "Barbie box" sequence) is located 335-349 bp upstream from the first exon. The Barbie box sequences are present in the 5' flanking region of all eukaryotic and prokaryotic genes encoding barbiturate-inducible proteins, such as the detoxifying cytochrome P450 (CYP) enzyme gene (Liang *et al.*, 1995) whose expression is regulated by porphyrinogenic agents (e.g. phenobarbital) (Salonpaa *et al.*, 1997). Interestingly, all CYP enzymes contain heme as a

prosthetic group. Porphyrinogenic agents regulate CYP enzymes activity by regulating their gene expression, and likely by modulating the expression of enzymes involved in the heme biosynthetic pathway, such as protoporphyrinogen oxidase. Finally, the promoter prediction program NNPP predicted a single TATA-containing promoter region, located 923-973 bp upstream from the 5' end of the first exon of the Ppox gene. However, because of the large distance in bp between this putative promoter region and the first exon of the Ppox gene, it remains to be established whether the Ppox gene transcriptional activity is mediated by this putative promoter region, or whether the Ppox gene promoter is a TATA-less promoter and this putative promoter region a "false positive".

III-2-4-8 Mouse USP23 gene

For many short-lived eukaryotic proteins, conjugation to the protein ubiquitin is an obligatory step in their degradation. Ubiquitin is synthesized as a fusion protein, fused to itself or to a ribosomal protein. Deubiquitinating enzymes cleave the precursor fusion protein in order to release free ubiquitin. Ubiquitin-conjugating enzymes ligate ubiquitin to proteins via a covalent linkage between the carboxy-terminus Gly residue of ubiquitin and lysine epsilon-amino groups of the acceptor proteins. In many cases, ubiquitination of a protein leads to its degradation by the multienzymatic 26S proteasome complex, releasing free amino acids and re-usable ubiquitin. A growing number of cellular processes are linked to protein degradation by ubiquitin, including regulation of the cell cycle, class I antigen processing, signal transduction pathways and receptor-mediated endocytosis (for review, see Hochstrasser, 1996).

There are two classes of deubiquitinating enzymes, all of which are characterized by highly conserved His and Cys domains: the ubiquitin C-terminal hydrolases (UCH) and the ubiquitin-specific proteases (USP). Both classes are involved in ubiquitin re-

cycling, but the USP enzyme activity is directed towards disassembling various forms of polyubiquitin or ubiquitin-like protein complexes (Wilkinson, 1997). USP23 is a member of the USP class of deubiquitinating enzymes. Northern-blot analysis by Smith and Southan (2000) revealed that USP23 was ubiquitously expressed in various tissues.

The mouse and human USP23 cDNAs were cloned by Smith and Southan (2000). Both cDNAs encode a 62 kD protein containing highly conserved His and Cys domains characteristic of the protease family of ubiquitin-specific processing proteases (Smith and Southan, 2000). The mouse and human share 96.5% amino acid sequence identity, suggesting that they arose from duplication of a common ancestor (Smith and Southan, 2000) (Fig. 3.63).

```
               10         20         30         40         50         60
usp23h MPQASEHRLGRTREPPVNIQPRVGSKLPFAPRARSKERRNPASGPNPMLRPLPPRPGLPD
       :::::::::::::::::::::::.:::::.:.:: :::::::::::. ::: :::::::::: ::
usp23m MPQASEHRLGRTREPPVNVQPRVGAKIPFPPRARSKERRNPVPGPNSMLRPLPPRPGPPD
               10         20         30         40         50         60
               70         80         90        100        110        120
usp23h ERLKKLELGRGRTSGPRPRGPLRADHGVPLPGSPPPTVALPLPSRTNLARSKSVSSGDLR
       :::::::::::::::: :::::::::::::::::::::.:::::::::::::::::::::::::
usp23m ERLKKLELGRGRTSGSRPRGPLRADHGVPLPGSPPPAVALPLPSRTNLARSKSVSSGDLR
               70         80         90        100        110        120
              130        140        150        160        170        180
usp23h PMGIALGGHRGTGELGAALSRLALRPEPPTLRRSTSLRRLGGFPGPPTLFSIRTEPPASH
       ::::::::::.:::::::::::::::::::::::::::::::::::::::::.::::::.::
usp23m PMGIALGGHRGAGELGAALSRLALRPEPPTLRRSTSLRRLGGFPGPPTLLSIRTEPPTSH
              130        140        150        160        170        180
              190        200        210        220        230        240
usp23h GSFHMISARSSEPFYSDDKMAHHTLLLGSGHVGLRNLGNTCFLNAVLQCLSSTRPLRDFC
       ::::::::: ::::::::::::::::::::::::::::::::::::::::::::::::::::
usp23m GSFHMISARPSEPFYSDDKMAHHTLLLGSGHVGLRNLGNTCFLNAVLQCLSSTRPLRDFC
              190        200        210        220        230        240
              250        260        270        280        290        300
usp23h LRRDFRQEVPGGGRAQELTEAFADVIGALWHPDSCEAVNPTRFRAVFQKYVPSFSGYSQQ
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
usp23m LRRDFRQEVPGGGRAQELTEAFADVIGALWHPDSCEAVNPTRFRAVFQKYVPSFSGYSQQ
              250        260        270        280        290        300
              310        320        330        340        350
usp23h DAQEFLKLLMERLHLEINRRGRRAPPILANGPVPSPPRRGG-ALLEEPELSDDDRANLMW
       ::::::::::::::::::::::::::::::::.:::::::::: :: ::::::::::::::
usp23m DAQEFLKLLMERLHLEINRRGRRAPPILASGPVPSPPRRGGGALHEEPELSDDDRANLMW
              310        320        330        340        350        360
     360        370        380        390        400        410
usp23h KRYLEREDSKIVDLFVGQLKSCLKCQACGYRSTTFEVFCDLSLPIPKKGFAGGKVSLRDC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
usp23m KRYLEREDSKIVDLFVGQLKSCLKCQACGYRSTTFEVFCDLSLPIPKKGFAGGKVSLRDC
              370        380        390        400        410        420
     420        430        440        450        460        470
usp23h FNLFTKEEELESENAPVCDRCRQKTRSTKKLTVQRFPRILVLHLNRFSASRGSIKKSSVG
       :.::::::::::::::::::::::::::::::::::::::::::::::::.::::::::::::
usp23m FSLFTKEEELESENAPVCDRCRQKTRSTKKLTVQRFPRILVLHLNRFSTSRGSIKKSSVG
```

```
              430        440        450        460        470        480
     480       490        500        510        520        530
usp23h VDFPLQRLSLGDFASDKAGSPVYQLYALCNHSGSVHYGHYTALCRCQTGWHVYNDSRVSP
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
usp23m VDFPLQRLSLGDFASDKAGSPVYQLYALCNHSGSVHYGHYTALCRCQTGWHVYNDSRVSP
              490        500        510        520        530        540
     540       550        560
usp23h VSENQVASSEGYVLFYQLMQEPPRCL
       :::::::::::::::::::::: :::
usp23m VSENQVASSEGYVLFYQLMQEPLRCL
              550        560
```

Fig. 3.63: ALIGN comparison between the amino acid sequences of the human (top) and mouse (bottom) Usp23 proteins.

A GENSCAN analysis, followed by a BLAST search, of the mouse clone rp23-395h6 reveals that the segment at position 58559-63326 corresponds to the mouse USP23 gene (Fig. 3.48). A CROSSMATCH analysis between the BAC clone rp23-395h6 and the mouse USP23 cDNA indicates that the USP23 gene consists of 12 exons and spans a genomic region of approximately 4.7 kb (table 3.50 and Fig. 3.64). In addition, the results of the Neural Network splice site prediction program analysis also confirmed the presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of the USP23 genomic region and the USP23 cDNA (table 3.51).

USP23 gene:

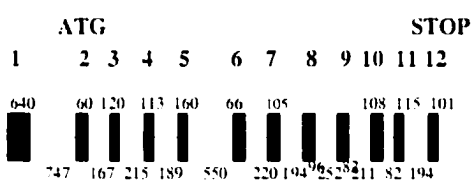


Fig. 3.64: Structural organization of the USP23 gene. Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines

Exon number	Contig position	Exon length (bp)
1	58559-59179	620
2	59928-59987	59
3	60154-60274	120
4	60491-60602	111

5	60793-60951	158
6	61503-61567	64
7	61788-61891	103
8	62088-62174	86
9	62436-62514	78
10	62727-62834	107
11	62918-63032	114
12	63227-63326	99

Table 3.50: contig position and length of the exons of the mouse USP23 gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
	59181		59927
1/1	caagatgGTgaggac	1/2	actttctAGgctcac
	59989		60153
2/2	aaatacaGTaagagc	2/3	cttccccAGtgcttc
	60276		60490
3/3	acagaagGTgggcaa	3/4	aaaccaacAGcctttg
	60604		60792
4/4	gatacagGTggggaa	4/5	tggctgtAGccagca
	60953		61502
5/5	agttgagGTaagggt	5/6	ctgtggcAGtgatga
	61569		61787
6/6	attgtggGTgggtat	6/7	tctacctAGacctgt
	61893		62087
7/7	ccccaagGTgggatt	7/8	tccccctAGaaagga
	62176		62435
8/8	tgccccaGTatgtgg	8/9	tttatttAGgtgtgt
	62516		62726
9/9	gtgctccGTatatcc	9/10	agggcttAGatctga
	62836		62917
10/10	aaagcggGTgagtct	10/11	aacccgcAGgaagcc
	63034		63226
11/11	actcccgGTgagaat	11/12	tcttctcAGcgtttc

Table 3.51: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 58579, in the first exon, indicating that the 5' UTR is only 20 bp. The unusually short 5' UTR may be related to the close proximity of the Ppox gene whose 3' end is located approximately 1000 bp upstream from the 5' end of the USP23 gene. A putative termination site is located at position 63314, in the last exon, indicating that the 3' UTR is only 12 bp. Initial analysis of the 1000 bp located immediately upstream from the 5' end of the USP23 gene, using MOTIF, reveals several potential binding sites for transcription factors, such as GATA, Sp1, AP-1, p300, E2F binding site and a potential Ras-responsive element binding protein 1 binding site. The NNPP promoter prediction program predicted four potential promoter regions, located 820-870 bp, 598-648 bp, 252-302 bp and 196-246 bp, respectively, upstream from the 5' end of the USP23 gene. The analysis of the 5' flanking region of USP23, using MOTIF, indicates the presence of a potential TATA box, 277-291 bp upstream from the 5' end of USP23, suggesting that the promoter region located 252-302 bp upstream from USP23 may correspond to the actual promoter region.

III-2-4-9 Mouse DEDD gene

Apoptosis plays a crucial role in development and homeostasis of organisms, as it eliminates individual cells which are no longer needed or became greatly damaged (Steller, 1995). One of the signaling pathways that triggers cell apoptosis is that of the "death receptor" CD95 (Nagata, 1997). CD95 belongs to the death receptor subfamily of the tumor necrosis factor receptor superfamily. CD95 is characterized by the presence of a "death domain" within its cytoplasmic domain, and has been shown to trigger apoptosis upon binding of its specific ligand. The first event which can be detected during the apoptosis cascade mediated by CD95 is the recruitment of an adaptor molecule, FADD, through the homophilic interaction between the death domains

of CD95 and FADD (Boldin *et al.*, 1996). FADD, in turn, recruits the protease procaspase-8 through homophilic interaction between the "death effector domains" (DED) of FADD and procaspase-8 (Boldin *et al.*, 1995). The aggregation of caspase-8 molecules leads to activation of this protease by self-processing (Medema *et al.*, 1997). Caspase-8 then triggers downstream apoptosis events by activating downstream effectors such as caspase-3 (Scaffidi *et al.*, 1997). The effector caspases cleave many cellular substrates, including structural proteins, signaling proteins and regulators of DNA replication or transcription, underlying many of the biochemical events of apoptosis (Thornberry and Lazebnik, 1998). One caspase substrate is the inhibitor of caspase-activated DNAse (ICAD) (Enari *et al.*, 1998; Sakahira *et al.*, 1998). ICAD binds to the caspase-activated DNAse (CAD) and keeps it in the cytoplasm. The cleavage of ICAD by caspases allows CAD to translocate to the nucleus and digest DNA.

DEDD (DED-containing DNA-binding protein) was identified through EST database screening (Stegh *et al.*, 1998). It contains an amino-terminal DED with homology to the DED of FADD and caspase-8. The mouse DEDD protein shares 98.7% sequence identity with its human counterpart (Fig. 3.65).

```
                10        20        30        40        50        60
deddhu MAGLKRRASQVWPEEHGEQEHGLYSLHRMFDIVGTHLTHRDVRVLSFLFVDVIDDHERGL
       ::::::::::::::.:::::::::::::::::::::::::::::::::::::::::::::
deddmu MAGLKRRASQVWPEERGEQEHGLYSLHRMFDIVGTHLTHRDVRVLSFLFVDVIDDHERGL
                10        20        30        40        50        60

                70        80        90       100       110       120
deddhu IRNGRDFLLALERQGRCDESNFRQVLQLLRIITRHDLLPYVTLKRRRAVCPDLVDKYLEE
       ::::::::::::::::::::::::::::::::::::::::::::::.:::::::::::::::
deddmu IRNGRDFLLALERQGRCDESNFRQVLQLLRIITRHDLLPYVTLKKRRAVCPDLVDKYLEE
                70        80        90       100       110       120

               130       140       150       160       170       180
deddhu TSIRYVTPRALSDPEPRPPQPSKTVPPHYPVVCCPTSGPQMCSKRPARGRATLGSQRKRR
       :::::::::::::::::::::::::::::::::::: ::::::::::.:::::::::::::
deddmu TSIRYVTPRALSDPEPRPPQPSKTVPPHYPVVCCPTSGSQMCSKRPARGRTTLGSQRKRR
               130       140       150       160       170       180

               190       200       210       220       230       240
deddhu KSVTPDPKEKQTCDIRLRVRAEYCQHETALQGNVFSNKQDPLERQFERFNQANTILKSRD
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
deddmu KSVTPDPKEKQTCDIRLRVRAEYCQHETALQGNVFSNKQDPLERQFERFNQANTILKSRD
```

```
              190       200       210       220       230       240

              250       260       270       280       290       300
deddhu LGSIICDIKFSELTYLDAFWRDYINGSLLEALKGVFITDSLKQAVGHEAIKLLVNVDEED
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
deddmu LGSIICDIKFSELTYLDAFWRDYINGSLLEALKGVFITDSLKQAVGHEAIKLLVNVDEED
              250       260       270       280       290       300

              310
deddhu YELGRQKLLRNLMLQALP
       ::::::::::::::::::
deddmu YELGRQKLLRNLMLQALP
              310
```

Fig. 3.65: ALIGN comparison between the amino acid sequences of the human (top) and mouse (bottom) DEDD proteins.

On the DNA level, DEDD is 94.1% identical between human and mouse (Stegh *et al.*, 1998). The DEDD mRNA is widely expressed in various tissues. The DEDD protein contains two putative nuclear localization signals and binds strongly to DNA. DEDD induces weak apoptosis through interaction with FADD and caspase-8, and translocates to the nucleus after CD95 stimulation and caspase activation. In addition, overexpressed DEDD co-localizes exclusively to the nucleolus with UBF, a basal factor required for RNA polymerase I transcription, and recombinant DEDD inhibits transcription of rDNA *in vitro*. These observations suggest that DEDD may serve to shut down transcription during CD95-mediated apoptosis.

A GENSCAN analysis, followed by a BLAST search, of the BAC clone rp23-395h6 reveals that the segment at position 7239-4168 on the complementary strand corresponds to the mouse DEDD gene (Fig. 3.48). The DEDD gene consists of four exons and spans a genomic region of approximately 3 kb (table 3.52 and Fig. 3.66). In addition, the results of the Neural Network splice site prediction program analysis also confirmed the presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of the DEDD genomic region and the DEDD cDNA (table 3.53).

DEDD gene:



DEDD protein:

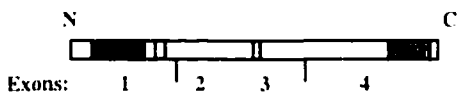


Fig. 3.66: Structural organization of the DEDD gene. (Top) Sizes, in base pairs, are given below the introns and above the exons. The exons are indicated by black boxes and the connecting introns by horizontal lines. (Bottom) The black box corresponds to the DED domain, the clear gray boxes to two putative nuclear localization signals. The dark gray box indicates a domain with sequence homology with some DNA-binding proteins (see Stegh *et al.*, 1998).

Exon number	Contig position	Exon length (bp)
1	7239-6887	352
2	6704-6597	107
3	5334-5188	146
4	4699-4168	531

Table 3.52: contig position and length of the exons of the mouse DEDD gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
1/1	6885 cgagctgGTgaggcc	1/2	6705 ctcctgcAGtgtgcc
2/2	6595 aaaacagGTgagagg	2/3	5335 ttctgctcAGtgcctc
3/3	5185 acatgtgGTaaggaa	3/4	4700 ttctcccAGatatca

Table 3.53: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 7212, in the first exon, and a putative termination site is located at position 4324, in the last exon. Initial

analysis of the 1000 bp located immediately upstream from the first exon, using MOTIF, indicates the presence of potential binding sites for transcription factors, such as GATA, STAT, AP-1, Oct-1, c-Myb, p300 transcription factors, and a putative binding site for p53, 820-829 bp upstream from the 5' end of the DEDD gene. The tumor suppressor gene p53 is one of the major elements involved in programmed cell death (Sjöström and Bergh, 2001), and one of the mechanisms triggered during p53-mediated apoptosis may involve the transcriptional regulation of DEDD by p53. The promoter prediction program NNPP predicted two potential promoter regions, located 702-752 bp and 57-107 bp upstream from the 5' end of the DEDD gene. The analysis of the first 1000 bp located upstream from the DEDD gene, using MOTIF, predicted the presence of a potential CAAT box, 726-737 bp upstream from the 5' end of the DEDD gene, suggesting that the promoter region 702-752 bp upstream from the 5' end of DEDD may correspond to the actual promoter.

III-2-4-10 Mouse Nit1 gene

Nitriles are organic compounds containing a carbon-nitrogen triple bond. Some nitriles, such as cyanoglycosides are synthesized by a wide range of plants, while others, such as acetonitrile, are commercially manufactured (Bork and Koonin, 1994). Some microorganisms can utilize nitrile as a carbon and nitrogen source (Kobayashi *et al.*, 1998), and the microbial degradation of nitrile involves two distinct enzymes, nitrilase and nitrile hydratase. Nitrilase cleaves nitriles and organic amides to the corresponding carboxylic acids plus ammonia (Kobayashi *et al.*, 1998). Nitrilase also is involved in the biosynthesis of the plant hormone indole-3-acetic acid, from indole-3-acetonitrile (Bartling *et al.*, 1994).

The mouse Nit1 gene, along with its human counterpart NIT1, is a member of an uncharacterized mammalian gene family, with sequence homology to the bacterial and plant nitrilases (Pekarsky *et al.*, 1998). Although the role of this enzyme is still unclear, it has been postulated that nitrilases in mammalian cells are involved in the detoxification process induced by highly toxic nitrile-based compounds, such as acetonitrile (Pekarsky *et al.*, 1998).

The mouse Nit1 cDNA was isolated by Pekarsky *et al.* (1998). It encodes a 323 amino acid translation product which shares 84.1% amino acid sequence identity with its human counterpart, NIT1 (Fig. 3.67).

```
               10        20        30        40        50        60
nit1hu MLGFITRPPHRFLSLLCPGLRIPQLSVLCAQPRPRAMAISSSSCELPLVAVCQVTSTPDK
       ::::::::::.   ::: : :. .  :::.:::::.:. :::: ::::::::::::::.:
nit1mu MLGFITRPPHQ---LLCTGYRLLRTPVLCTQPRPRTMS-SSTSWELPLVAVCQVTSTPNK
               10        20        30        40        50

               70        80        90       100       110       120
nit1hu QQNFKTCAELVREAARLGACLAFLPEAFDFIARDPAETLHLSEPLGGKLLEEYTQLAREC
       :.:::::::::.:::::::::::::::::::::.::::: ::::::.: :: .:.:::::
nit1mu QENFKTCAELVQEAARLGACLAFLPEAFDFIARNPAETLLLSEPLNGDLLGQYSQLARES
         60        70        80        90       100       110

              130       140       150       160       170       180
nit1hu GLWLSLGGFHERGQDWEQTQKIYNCHVLLNSKGAVVATYRKTHLCDVEIPGQGPMCESNS
       :.:::::::::::::::::.:::::::::::::::.:::.:::::::::::::::::: :::
nit1mu GIWLSLGGFHERGQDWEQNQKIYNCHVLLNSKGSVVASYRKTHLCDVEIPGQGPMRESNY
        120       130       140       150       160       170

              190       200       210       220       230       240
nit1hu TMPGPSLESPVSTPAGKIGLAVCYDMRFPELSLALAQAGAEILTYPSAFGSITGPAHWEV
       : :: .:: ::.:::::::.:::.:::::::::::: :::::::::: :::::.:::::::
nit1mu TKPGGTLEPPVKTPAGKVGLAICYDMRFPELSLKLAQAGAEILTYSSAFGSVTGPAHWEV
        180       190       200       210       220       230

              250       260       270       280       290       300
nit1hu LLRARAIETQCYVVAAAQCGRHHEKRASYGHSMVVDPWGTVVARCSEGPGLCLARIDLNY
       ::::::::.::::.:::::::::: ::::::::::::::::::::::::::::::::..
nit1mu LLRARAIESQCYVIAAAQCGRHHETRASYGHSMVVDPWGTVVARCSEGPGLCLARIDLHF
        240       250       260       270       280       290

              310       320
nit1hu LRQLRRHLPVFQHRRPDLYGNLGHPLS
       :.:.:::::::::::::::.::::::
nit1mu LQQMRQHLPVFQHRRPDLYGSLGHPLS
        300       310       320
```

Fig.3.67: ALIGN comparison between the amino acid sequences of the human (top) and mouse (bottom) nit1 proteins.

The mouse Nit1 gene is approximately 3.6 kb in length and contains eight exons. Northern analysis revealed that different mouse tissues showed different levels of expression of Nit1 mRNA, and the highest level of expression was in mouse liver and kidney. In addition, alternatively spliced Nit1 mRNAs have been observed. Some of the transcripts lack exon 2 and encode a 323 amino acid protein, while some alternative transcripts containing exon 2 encodes a shorter, 290 amino acid protein, starting with methionine 34 (Pekarsky *et al.*, 1998). These observations suggests that different mRNA species are generated by alternative splicing and alternative initiating methionines, respectively.

A GENSCAN analysis, followed by a BLAST search, of the BAC clone rp23-395h6 reveals that the genomic segment at position 521-3284 corresponds to the 3' end of the mouse Nit1 gene (Fig. 3.48). A CROSSMATCH analysis between the BAC clone rp23-395h6 and the mouse Nit1 cDNA further reveals that the genomic segment at position 521-3284 encompasses the last six exons of the Nit1 gene and that the first exon of the Nit1 gene is not located on the BAC clone rp23-395h6 (table 3.54). The analysis of the Nit1 gene therefore is limited to the determination of its exon-intron borders (Fig. 3.68). However, the sequences of the first exon and of the first 778 bp located upstream from the first exon were extracted as trace data from the whole genome shotgun sequencing analysis of the mouse genome (see http//:www.ensembl.org/Mus_musculus. The complete analysis of the Nit1 gene was not possible because of the presence of gaps in the whole genome shotgun assembly of Nit1, but analysis of the 778 bp located immediately upstream from the first exon of Nit1, using MOTIF, indicates the presence of numerous potential transcription factor binding sites, among them MZF1, GATA, c-Myb, USF, Lyf-1, E2F, STAT, p53, NF-E2, c-Myc and CREB. Also, the search for promoter, using NNPP, did not predict any promoter region in the 778 bp located upstream from the first exon. Finally, the results of the Neural Network splice site prediction program analysis also confirmed the

presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of the Nit1 genomic region and the Nit1 cDNA (table 3.55).

Nit1 gene (last 6 exons):

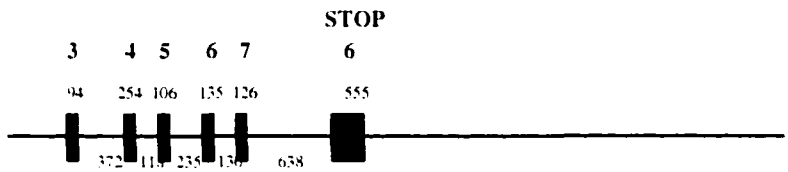


Fig. 3.68: Structural organization of the last six exons of the Nit1 gene. Sizes, in base pairs, are given below the introns and above the exons. The exons are indicated by black boxes and the connecting introns by horizontal lines.

Exon number	Contig position	Exon length (bp)
2	521-615	94
3	990-1241	251
4	1362-1466	104
5	1702-1835	133
6	1966-2091	125
7	2732-3284	552

Table 3.54: contig position and length of the last six exons of the mouse Nit1 gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
1/1	tcttcatGTaggacc	1/2	cttcttcAGgctggg
2/2	617 agcccagGTaacata	2/3	989 tgtcattcAGgcccag
3/3	1243 ttgccagGTacaagg	3/4	1361 ccccccccAGggaatg
4/4	1468 agcaaggGTgagacg	4/5	1701 tctctctAGgatcag
5/5	1837 tggcaagGTgggagt	5/6	1965 tcatttcAGgttggt
6/6	2093 tgggagGTaagaga	6/7	2731 catcctcAGgtgctg

Table 3.55: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

III-2-4-11 Mouse Adamts4 gene

Aggrecan is the major proteoglycan of cartilages. It carries a large number of negatively charged groups resulting in high osmotic pressure in the tissue, allowing aggrecan to hydrate the framework of cartilage collagen fibrils and providing the tissue with its properties of compressibility and elasticity. Aggrecan is a large chondroitin sulfate proteoglycan that accounts for about 10% of the dry weight of cartilage. Aggrecan monomers interact with hyaluronan and usually are found as part of a large aggregate containing 10-100 monomers per hyaluronan molecules (Hardingham *et al.*, 1992).

Aggrecan degradation is an important factor in the erosion of articular cartilage in arthritic diseases. It is one of the first matrix components to undergo measurable loss in cartilage degradation associated with arthritis. This degradation involves the proteolysis of the aggrecan molecule near its amino-terminus. Two major sites of proteolytic cleavage have been identified within the molecule. One site, between Asn^{341} and Phe^{342} is cleaved by matrix metalloproteinase (Flannery *et al.*, 1992), the other site between Glu^{373} and Ala^{374} is attributed to aggrecanase-1 (ADAMTS4) (Tortorella *et al.*, 1999). Several other sites of cleavage by aggrecanase-1 have been identified (Tortorella *et al.*, 2000a). The aggrecan products generated by aggrecanase-1 occur in the synovial fluid from patients with arthritis, suggesting that this enzyme may be important in arthritic diseases involving cartilage degradation.

Aggrecanase-1 (ADAMTS4) is a member of the *a* *d*isintegrin *a*nd *m*etalloprotease with *t*hrombo*s*pondin motifs (ADAMTS) protein family. All members of the ADAMTS family contain a common amino-terminal sequence domain, a

metalloproteinase domain and a disintegrin-like domain (Tang and Hong, 1999). The carboxy-terminus of ADAMTS proteins contains a various number of thrombospondin motifs, which are important for aggrecan binding and cleavage (Tortorella *et al.*, 2000b). The amino-terminus of the tissue inhibitors of metalloproteinase protein-3 (TIMP-3) is a strong inhibitor of human aggrecanase-1 (Kashiwagi *et al.*, 2001), providing a starting point for developing a treatment against arthritic diseases.

The human ADAMTS4 cDNA was cloned by Tortorella *et al.* (1999). This cDNA contains a 2511 base pair open reading frame, encoding a 837 amino acid translation product, and a signal sequence followed by a propeptide domain that precedes the catalytic domain. The catalytic domain of ADAMTS4 is followed by a disintegrin-like domain and a carboxy-terminal domain containing a thrombospondin type I motif, similar to those present in other members of the ADAMTS protein family, such as ADAMTS1 (Tortorella *et al.*, 1999). The presence of a signal sequence and a propeptide suggests that ADAMTS4 is synthesized as a zymogen and is cleaved to remove the propeptide domain and generate the mature protein (Tortorella *et al.*, 1999). Northern and RT-PCR analysis indicates that ADAMTS4 is present in brain, lung and heart tissues, with very low levels in placenta and skeletal muscle tissues (Tortorella *et al.*, 1999).

A GENSCAN analysis, followed by a BLAST search, of the BAC clone rp23-395h6 reveals that the genomic region at position 94678-86350 on the complementary strand corresponds to a gene whose coding regions have sequence similarity with the human ADAMTS4 cDNA (Fig. 3.48). A CROSSMATCH analysis between the coding sequence of this putative gene, as defined by GENSCAN, and the BAC clone rp23-395h6 reveals that the coding region of this putative gene consists of 9 exons (table 3.56 and Fig. 3.69). Also, the results of the Neural Network splice site prediction program analysis also confirmed the presence of potential splice sites at the exon-intron borders

indicated by the ALIGN comparison of this putative Adamts4 gene and its coding sequence generated by GENSCAN (table 3.57).

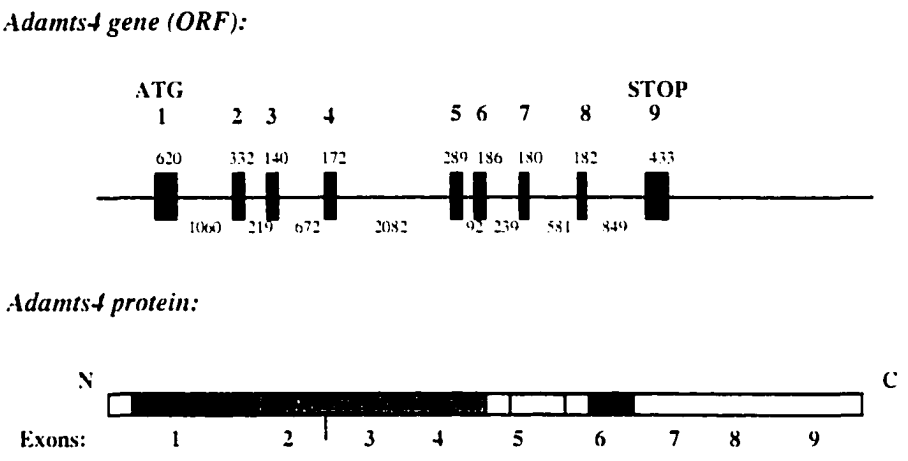


Fig. 3.69: Structural organization of the ORF of the putative Adamts4 gene and its encoded protein. (Top) Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines. (Bottom) The dark gray box indicates the presence of a catalytic metalloprotease domain. The clear gray box corresponds to a disintegrin-like domain. The white box in the first exon corresponds to the signal sequence. The black box in exon 1 corresponds to the propeptide domain, and the black box in exon 6 corresponds to the thrombospondin type I motif.

Exon number	Contig position	Exon length (bp)
1	94678-94058	620
2	92990-92667	323
3	92440-92308	132
4	91634-91464	170
5	89380-89094	286
6	89000-88814	186
7	88573-88398	175
8	87809-87634	175
9	86776-86350	426

Table 3.56: contig position and length of the exons of the coding region of the putative Adamts4 gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
	94056		92991
1/1	aaccaagGTaggcaa	1/2	ctcctcaAGcgcttc
	92665		92441
2/2	ccggcagGTaaggcc	2/3	tctccacAGgacctg

	92306		91635
3/3	gaactggGTaaggaa	3/4	ctcttatAGgccatg
	91462		89381
4/4	ggttatgGTaagctg	4/5	ccctttcAGggcact
	89092		89001
5/5	cttcaatGTgagacc	5/6	cccttgcAGgttcct
	88812		88574
6/6	ggctcagGTgagggg	6/7	tgcctgtAGcattga
	88396		87810
7/7	gccccggGTgaggac	7/8	gctcctcAGgtggca
	87632		86777
8/8	aattcagGTcctttc	8/9	tcctcacAGgtatgg

Table 3.57: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

An ALIGN comparison between the amino acid sequence deduced from the coding sequence of this putative gene and the amino acid sequence of the human ADAMTS4 protein indicates that they share 90.2% amino acid sequence identity (Fig. 3.70), confirming that this region likely encodes the mouse counterpart of the human ADAMTS4 protein, Adamts4.

```
               10         20         30         40         50         60
human  MSQTGSHPGRGLAGRWLWGAQPCLLLPIVPLSWLVWLLLLLLASLLPSARLASPLPREEE
       ::: : :: :::.:.::   ::::      ::  . :  ::::: ::  :  ::::::::::
mouse  MSQMGLHPRRGLTGHWLQRFQPCL-----PLHTVQWRRLLLLAFLLSLAWPASPLPREEE
               10         20              30         40         50

               70         80         90        100        110
human  IVFPEKLNGS-VLPGSGAPARLLCRLQAFGETLLLELEQDSGVQVEGLTVQYLGQAPELL
       :::::::::: .::::::.:::::: :: :::: :::::::: ::::::::::::::::::.:
mouse  IVFPEKLNGSSILPGSGVPARLLYRLPAFGEMLLLELEQDPGVQVEGLTVQYLGQAPEML
          60         70         80         90        100        110

     120        130        140        150        160        170
human  GGAEPGTYLTGTINGDPESVASLHWDGGALLGVLQYRGAELHLQPLEGGTPNSAGGPGAH
       :::::::::::::::::::::::::::::::::::::::::::::::::. :::::::::
mouse  GGAEPGTYLTGTINGDPESVASLHWDGGALLGVLQYRGAELHLQPLEGGALNSAGGPGAH
         120        130        140        150        160        170

     180        190        200        210        220        230
human  ILRRKSPASGQGPMCNVKAPLGSPSPRPRRAKRFASLSRFVETLVVADDKMAAFHGAGLK
       ::::::::::.::::.:::: :::::  ::.:::::::::::::::::::::::::.:::
mouse  ILRRKSPASSQGPMCTVKAPSGSPSPISRRTKRFASLSRFVETLVVADDKMAAFHGTGLK
         180        190        200        210        220        230
```

```
     240       250       260       270       280       290
human  RYLLTVMAAAAKAFKHPSIRNPVSLVVTRLVILGSGEEGPQVGPSAAQTLRSFCAWQRGL
       :::::::::::::::::::::::::::::.::::::::::::::::::::.:::::
mouse  RYLLTVMAAAAKAFKHPSIRNPVNLVVTRLVILGSGQEGPQVGPSAAQTLRSFCTWQRGL
         240       250       260       270       280       290

     300       310       320       330       340       350
human  NTPEDSDPDHFDTAILFTRQDLCGVSTCDTLGMADVGTVCDPARSCAIVEDDGLQSAFTA
       :::.:::::::::::::::::::::::::::::::::::::::::::::::::::::::
mouse  NTPNDSDPDHFDTAILFTRQDLCGVSTCDTLGMADVGTVCDPARSCAIVEDDGLQSAFTA
         300       310       320       330       340       350

     360       370       380       390       400       410
human  AHELGHVFNMLHDNSKPCISLNGPLSTSRHVMAPVMAHVDPEEPWSPCSARFITDFLDNG
       :::::::::::::::::: .:::  ..:::::::::::::::::::::::::::::::::
mouse  AHELGHVFNMLHDNSKPCTNLNGQGGSSRHVMAPVMAHVDPEEPWSPCSARFITDFLDNG
         360       370       380       390       400       410

     420       430       440       450       460       470
human  YGHCLLDKPEAPLHLPVTFPGKDYDADRQCQLTFGPDSRHCPQLPPPCAALWCSGHLNGH
       :::::::::::::::::.:::::::::::::::::::: ::::::::::::::::::::
mouse  YGHCLLDKPEAPLHLPATFPGKDYDADRQCQLTFGPDSSHCPQLPPPCAALWCSGHLNGH
         420       430       440       450       460       470

     480       490       500       510       520       530
human  AMCQTKHSPWADGTPCGPAQACMGGRCLHMDQLQDFNIPQAGGWGPWGPWGDCSRTCGGG
       :::::::::::::::: .:::::::::::.:::.:::.:::::::::::::::::::::
mouse  AMCQTKHSPWADGTPCGSSQACMGGRCLHVDQLKDFNVPQAGGWGPWGPWGDCSRTCGGG
         480       490       500       510       520       530

     540       550       560       570       580       590
human  VQFSSRDCTRPVPRNGGKYCEGRRTRFRSCNTEDCPTGSALTFREEQCAAYNHRTDLFKS
       :::::::::::::::::::::::::::::::.:: ::::::::::::::::::::::::
mouse  VQFSSRDCTRPVPRNGGKYCEGRRTRFRSCNTENCPHGSALTFREEQCAAYNHRTDLFKS
         540       550       560       570       580       590

     600       610       620       630       640       650
human  FPGPMDWVPRYTGVAPQDQCKLTCQAQALGYYYVLEPRVVDGTPCSPDSSSVCVQGRCIH
       :::::::::::::::.:::::::::.:::::::::::.:::::::.:::::::::::
mouse  FPGPMDWVPRYTGVAPRDQCKLTCQARALGYYYVLEPRVADGTPCSPDTSSVCVQGRCIH
         600       610       620       630       640       650

     660       670       680       690       700       710
human  AGCDRIIGSKKKFDKCMVCGGDRSGCSKQSGSFRKFRYGYNVVTIPAGATHILVRQQGN
       :::::::::::::::::::::: : ::::::::.:::::::.:::::::::::::::::.
mouse  AGCDRIIGSKKKFDKCMVCGGDGSRCSKQSGSFKKFRYGYSDVVTIPAGATHILVRQQGG
         660       670       680       690       700       710

     720       730       740       750       760       770
human  PGHRSIYLALKLPDGSYALNGEYTLMPSPTDVVLPGAVSLRYSGATAASETLSGHGPLAQ
       : .::::::::: ::::::::::::::::::::.:::::::::::::::::::::::::
mouse  SCLKSIYLALKLSDGSYALNGEYTLMPSPTDVVLPGAVSLRYSGATAASETLSGHGPLAQ
         720       730       740       750       760       770

     780       790       800       810       820       830
human  PLTLQVLVAGNPQDTRLRYSFFVPRPTPSTPRPTPQDWLHRRAQILEILRRRPWAGRK
       :::::::::::::..:::::::::::::.::::::: :::::.::::.::.:::.:::::::
mouse  PLTLQVLVAGNPQNARLRYSFFVPRPVPSTPRPPPQDWLQRRAEILKILRKRPWAGRK
         780       790       800       810       820       830
```

Fig. 3.70: ALIGN comparison between the amino acid sequences of the human ADAMTS4 protein (top) and of its putative mouse counterpart protein (bottom)

A BLAST analysis of the BAC clone rp23-395h6, using the mouse dbEST database, indicates that the genomic region at position 91808-91381 on the complementary strand has sequence homology with a mouse Adamts1 5' EST. In addition, Hurskainen *et al.* (1999) determined that the ADAMTS4 gene was located on human chromosome 1q21-23, in a region structurally conserved between human and mouse, and corresponding to the approximate location of the mouse clone rp23-395h6.

The 5' flanking region of human ADAMTS4 was characterized by Mizui *et al.* (2000). They showed that the region between -383 and +10 relative to the transcription start site was necessary for full promoter activity, and that this region contained one Sp1 and three AP-2 sites. In addition, they showed that the region between -726 and -384 appears to contain silencer elements, and that the nuclear factor I (NF-I) site at -441 to -429 was involved in negative regulation of the human ADAMTS4 promoter activity in chondrocytes. Analysis of the 4000 bp located immediately upstream from the putative translation start site (ATG) of Adamts4 (at position 94678), using MOTIF, indicates the presence of numerous potential transcription factor binding sites, among them are GATA, NF-kB, p53, STAT, N-Myc, MyoD, c-Myb, Oct-1, NF-E2 and CREB potential binding sites. Interestingly, the MOTIF search also revealed NF-I binding sites at positions 99-116 bp and 1364-1347 bp upstream from the translation start site, and a TATA box located at position 1139-1153 bp upstream from the ATG translation start site. Thus, it is likely that the mouse Adamts4 gene and the human ADAMTS4 gene have similar modes of transcriptional regulation. The NF-I site located 1364-1347 bp upstream from the ATG triplet corresponds to the functional NF-I binding site, and the transcriptional start site of Adamts4 is probably located in a region approximately 920 bp upstream from the translation start site.

III-2-4-12 Mouse Fc epsilon RI gamma subunit gene

Acute allergic reactions are triggered by three major classes of proinflammatory mediators: granule-associated amines (e.g., histamine and serotonin), newly-synthesized arachidonic acid metabolites (e.g., leukotriene C_4 and prostaglandin D_2) and proinflammatory vasoactive cytokines (e.g., tumor-necrosis factor alpha and interleukin-6) (Galli, 1993). These mediators are released from sensitized mast cells upon activation through the antigen-mediated cross-linking of their cell surface IgE receptors/Fc epsilon RI (Blank *et al.*, 1989; Galli, 1993). IgE receptor/Fc epsilon RI is a tetrameric complex consisting of an IgE-binding alpha subunit, a beta subunit and two disulfide-linked gamma subunits (Blank *et al.*, 1989). It is a member of a family of immune receptors that includes the T cell receptor, the B cell receptor and two IgG receptors, Fc gamma RI and Fc gamma RIII (Cambier, 1995). Both beta and gamma subunits contains immunoreceptor tyrosine-based activation motifs (ITAMs) within their cytoplasmic domains that allow interaction with protein-tyrosine kinases and their substrates via their Src homology 2 domains (Tamir and Cambier, 1998). The interaction of IgE receptors with their ligand triggers a cascade of events, leading to the subsequent tyrosine phosphorylation of downstream substrates, and their direct implication in the physiology of acute allergic reactions (Tamir and Cambier, 1998).

The cDNA for the gamma subunit of the mast cell receptor for IgE (Fc epsilon RI) was cloned in mouse by Ra et al. (1989) and the gene subsequently located on human chromosome 1q23, in a region conserved between mouse and human (Le Coniat *et al.*, 1990). An ALIGN comparison between the mouse and human IgE gamma subunits indicates that they share 88.4% sequence identity (Fig. 3.71). Residues 1-17 encodes a putative signal peptide, and hydrophobic segments in the mouse protein corresponds to a transmembrane domain that encompasses residues 24-44 (Ra *et al.*, 1989). The open reading frame is 258 bp in length, encompassing a 86 amino acid

translation product, becoming a 69 amino acid protein after removal of the signal peptide (Ra *et al.*, 1989).

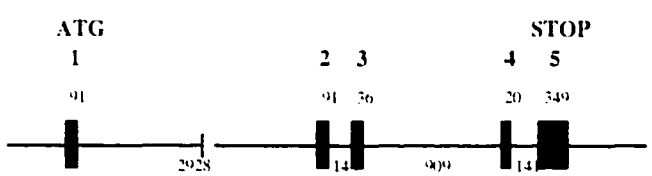
```
               10         20         30         40         50         60
fceghu MIPAVVLLLLLLVEQAAALGEPQLCYILDAILFLYGIVLTLLYCRLKIQVRKAAITSYEK
       :: ::.:.:::::::::::::::::::::::::::::::::::::::::::::.:: ::
fcegmu MISAVILFLLLLVEQAAALGEPQLCYILDAVLFLYGIVLTLLYCRLKIQVRKAAIASREK
               10         20         30         40         50         60

               70         80
fceghu SDGVYTGLSTRNQETYETLKHEKPPQ
       .:.::::::.::.:::::::::::::
fcegmu ADAVYTGLNTRSQETYETLKHEKPPQ
               70         80
```

Fig. 3.71: ALIGN comparison between the amino acid sequences of the human (top) and mouse (bottom) Fcer1g proteins.

A GENSCAN analysis, followed by a BLAST search, of the BAC clone rp23-395h6 reveals that the genomic segment at position 111240- 115950 corresponds to the mouse Fc epsilon RI gamma subunit gene (Fcer1g) (Fig. 3.48). A CROSSMATCH analysis between the mouse BAC clone rp23-395h6 and the Fcer1g cDNA reveals that this gene consists of five exons, spanning a genomic region of approximately 4.7 kb (table 3.58 and Fig. 3.72).

Fcer1g gene:



Fcer1g protein:

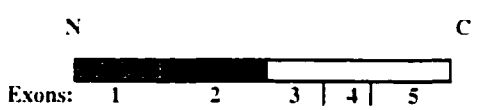


Fig. 3.72: Structural organization of the Fcer1g gene and its corresponding protein. (Top) Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines. (Bottom) The black box corresponds to the putative transmembrane domain, and the dark gray box corresponds to a signal peptide.

In addition, the results of the Neural Network splice site prediction program analysis also confirmed the presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of the Fcer1g gene and its cDNA (table 3.59).

Exon number	Contig position	Exon length (bp)
1	111240-111330	90
2	114259-114350	91
3	114496-114531	35
4	115440-115460	20
5	115600-115950	350

Table 3.58: contig position and length of the exons of the mouse Fcer1g gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
	111332		114258
1/1	caagcagGTaagggg	1/2	tcccctcAGccgccc
	114352		114495
2/2	actcaagGTaaggca	2/3	ccactccAGatccag
	114533		115439
3/3	ccgtgagGTatggag	3/4	tctctgtAGaaagca
	115462		115599
4/4	ctacacgGTgagtga	4/5	ctcctgcAGggcctg

Table 3.59: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 111281, in the first exon, and a putative termination site is located at position 115659, in the last exon. The signal peptide is encoded by the first exon, and the transmembrane domain of the receptor is encoded by nucleotides located at position 114278-114340, corresponding approximately to the second exon. Initial analysis of the 1000 bp located immediately upstream from the 5' end of the Fcer1g gene, using MOTIF, reveals potential binding sites for transcription factors that are important for the development of hematopoietic cells, such as c-Ets, GATA, c-Myb, MZF1, Ik-2, NF-AT binding sites, and potential

binding sites for p53, E2F and N-Myc transcription factors. A search in the possible promoter region, using NNPP, reveals the presence of four potential promoters at positions 559-609 bp, 350-400 bp, 317-367 bp and 214-264 bp upstream from the 5' end of the gene. Analysis of the same region, using MOTIF, indicates the presence of a potential TATA box, 377-391 bp upstream from the 5' end of the gene, suggesting that a TATA promoter likely is used in initiating gene expression of the Fcer1g gene.

III-2-4-13 Mouse L27' processed pseudogene

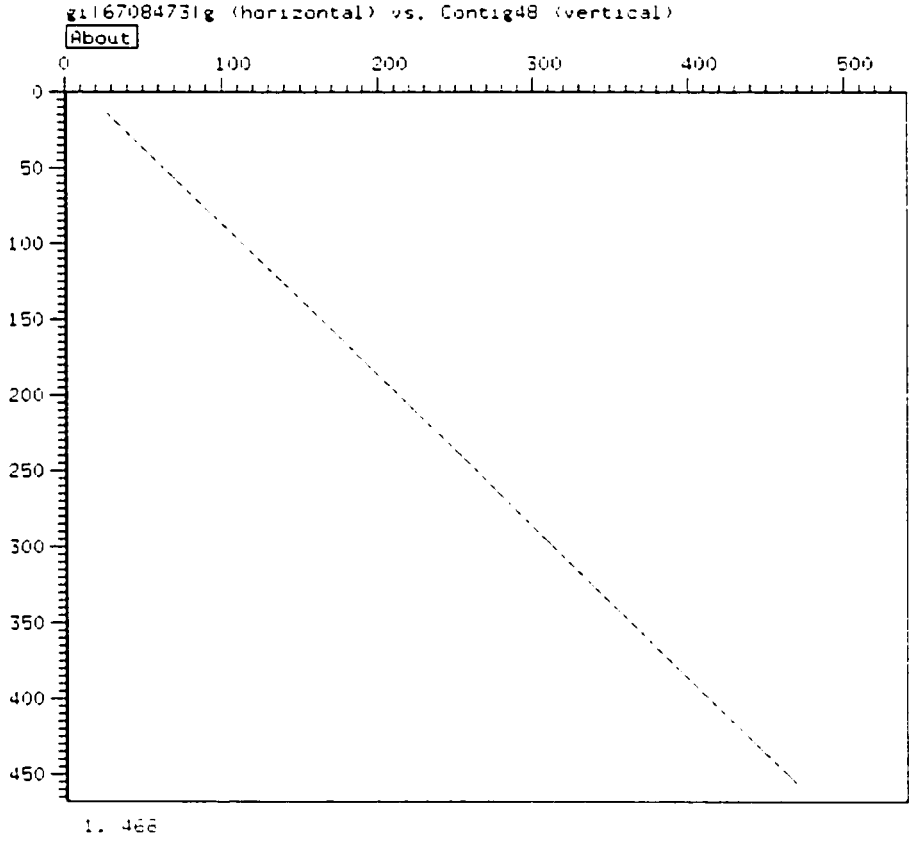


Fig. 3.73: Dot-plot analysis, using DOTTER, between the genomic sequence at position 39990-40457 (horizontal) and the mouse L27' cDNA (vertical).

A GENSCAN analysis, followed by a BLAST search, of the BAC clone rp23-395h6 revealed that the genomic segment at position 39990-40457 has 86.1% sequence homology with the mouse L27' ribosomal protein cDNA (Fig. 3.48). The L27' gene was cloned by Belhumeur *et al.* (1987), is present at approximately 15 copies of the intron-containing gene (Kusuda *et al.*, 1999) per genome, and is highly homologous to the yeast L29 ribosomal protein gene. It has been suggested that L27' is involved in peptidyl transferase activity and 50S subunit assembly (Lotti *et al.*, 1987; Wower *et al.*, 1998).

A DOTTER comparison between the genomic sequence at position 39990-40457 and the mouse L27' cDNA sequence reveals that the two sequences are colinear, and that the genomic sequence lacks any intronic sequence (Fig. 3.73). This lack of introns, characteristic of a large population of ribosomal processed pseudogenes (Harrison *et al.*, 2002) suggests that this genomic sequence corresponds to a non-functional L27' processed pseudogene.

III-2-4-14 Mouse RIKEN1 gene

A GENSCAN analysis, followed by a BLAST search, of the mouse BAC clone rp23-395h6 revealed the presence of an hypothetical gene at position 123083-126492 (Fig. 3.48). This hypothetical gene shares over 99% sequence similarity with a cDNA sequence (Genbank accession number AK012264) from the RIKEN full length enriched mouse cDNA library (Carninci and Hayashizaki, 1999). This hypothetical gene therefore was named the RIKEN1 gene. A CROSSMATCH analysis between the cDNA sequence of RIKEN1 and the genomic segment at position 123083-126492 reveals that the RIKEN1 gene consists of 10 exons (table 3.60 and Fig. 3.74). In addition, the results of the Neural Network splice site prediction program analysis also confirmed the presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of the RIKEN1 genomic region and its cDNA (table 3.61).

RIKEN1 gene:

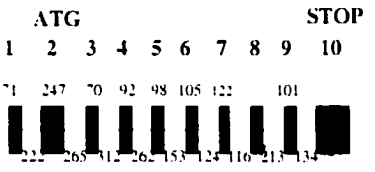


Fig. 3.74: Structural organization of the hypothetical RIKEN1 gene. Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines

Exon number	Contig position	Exon length (bp)
1	123083-123152	69
2	123374-123623	249
3	123888-123955	67
4	124267-124359	92
5	124621-124722	101
6	124875-124980	105
7	125104-125226	122
8	125342-125418	76
9	125631-125733	102
10	125867-126492	625

Table 3.60: contig position and length of the exons of the putative RIKEN1 gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
	123154		123373
1/1	tgagatgGTaagggc	1/2	ttattgcAGgtctag
	123625		123887
2/2	tgcaaagGTgagacc	2/3	tcccttcAGatgtat
	123957		124266
3/3	tttccagGTgctccc	3/4	tttcctcAGgtggct
	124361		124620
4/4	cacagagGTgagggc	4/5	ccctaacAGgtgttc
	124724		124874
5/5	cttccagGTgaccca	5/6	cccacccAGacacag
	124982		125103
6/6	gaatcagGTggggaa	6/7	ccctgactAGtgatca
	125228		125341

7/7	tactcagGTacgaga	7/8	accctgcAGctgtcc
	125420		125630
8/8	tgaacagGTaagact	8/9	gcttctcAGgttcag
	125735		125866
9/9	tttagagGTgagggt	9/10	ccccggcAGgcttgg

Table 3.61: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 123508, indicating that the first exon and the 5' end of the second exon are transcribed into a 5' UTR. The analysis of the first 1000 bp located upstream from the 5' end of the gene, using MOTIF, reveals several interesting features. First, potential binding sites for MZF1, GATA, c-Rel, NF-kB, p53, c-Ets, Lyf-1, NF-AT, Oct-1, c-Myb, STAT and E2F are present in the 5' flanking region of the RIKEN1 gene, suggesting that the transcription of this gene is highly regulated. In addition, a potential interferon-stimulated response element is present 362-376 bp upstream from the 5' end of the gene, suggesting that this hypothetical gene may be responsive to interferons. The NNPP promoter prediction program predicted two putative promoter regions at positions 218-268 bp and 191-241 bp upstream from the transcription start site. A search of the Pfam protein profile database indicated the presence of a putative eukaryotic porin domain in the carboxy terminus of the putative protein encoded by the RIKEN1 gene, similar to the domain found in outer mitochondrial membrane protein porins. Interestingly, a BLASTP search of the putative RIKEN1 protein reveals significant sequence similarity between the RIKEN1 protein and the mouse mitochondrial import receptor subunit Tom40 protein porin. An ALIGN comparison between the two protein sequences indicates that they have 53.8% sequence identity (Fig. 3.75). Tom40 is a member of the mitochondrial outer membrane translocation machinery, one of the membrane-embedded components, along with Tom5, Tom6 and Tom7, that form the general import pore which facilitates

the translocation of preproteins across the outer membrane (Rapaport *et al.*, 2001; Suzuki *et al.*, 2000). Whether the RIKEN1 protein is a component of a membranous translocation machinery remains to be established.

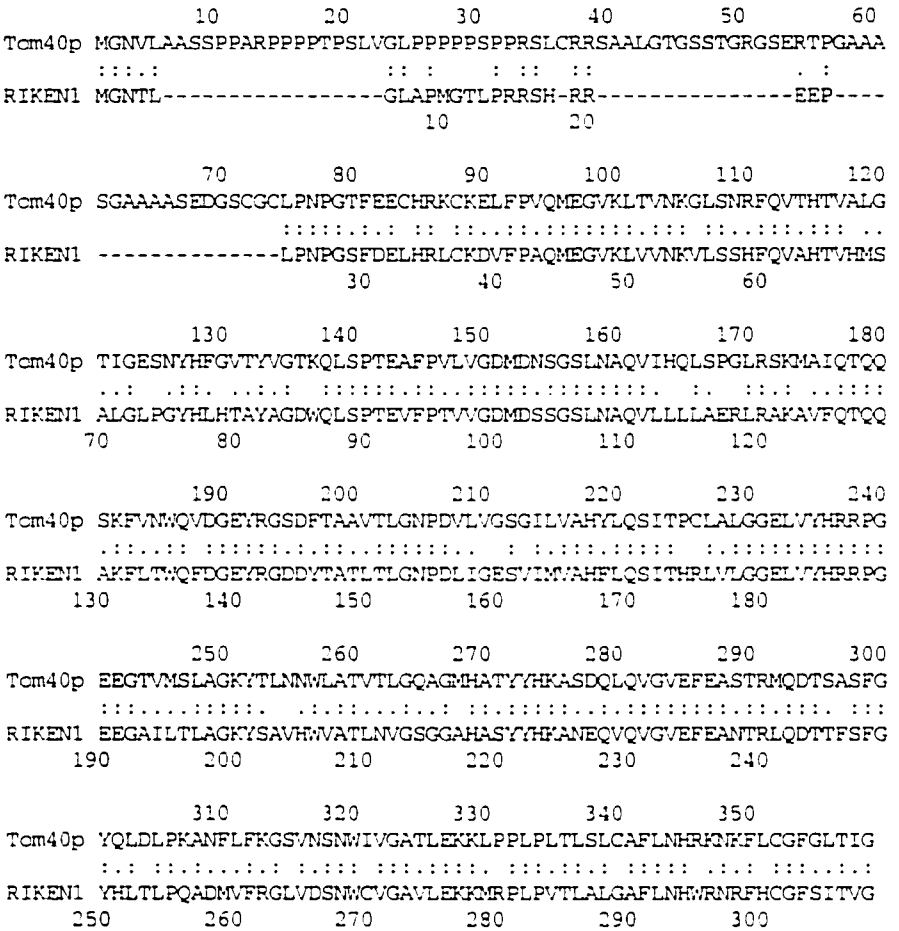


Fig. 3.75: ALIGN comparison between the amino acid sequences of the mouse Tom40 protein (top) and of the protein encoded by the hypothetical RIKEN1 gene (bottom).

III-2-4-15 Mouse RIKEN2 gene

A GENSCAN analysis, followed by a BLAST search, of the mouse clone rp23-395h6 reveals the presence of a second hypothetical gene at position 50546-56964 (Fig. 3.48). This hypothetical gene has over 99% sequence similarity with a cDNA sequence

(Genbank accession number NM_025388) from the RIKEN mouse cDNA library (Carninci and Hayashizaki, 1999), and was named the RIKEN2 gene. A CROSSMATCH analysis between the RIKEN2 cDNA sequence and the sequence of the genomic segment at position 50546-56964 indicates that the RIKEN2 gene consists of 7 exons (table 3.62 and Fig. 3.76). In addition, the results of the Neural Network splice site prediction program analysis also confirmed the presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of the RIKEN2 genomic region and its cDNA (table 3.63).

RIKEN2 gene:

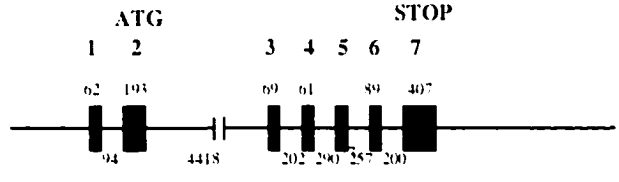


Fig. 3.76: Structural organization of the hypothetical RIKEN2 gene. Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines

Exon number	Contig position	Exon length (bp)
1	50546-50608	62
2	50702-50895	193
3	55313-55380	67
4	55582-55645	63
5	55935-56011	76
6	56268-56358	90
7	56558-56965	407

Table 3.62: contig position and length of the exons of the putative RIKEN2 gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
	50610		50701
I/1	cacacccGTaggtca	1/2	ttcccacAGtaaaga
	50897		55312
2/2	gatccggGTcagttt	2/3	gtttgcAGtatgtg

	55382		55581
3/3	gacccgGTaagcag	3/4	ctgttgcAGgtggtt
	55647		55934
4/4	gtttgaaGTgagtgt	4/5	ttcctccAGattcct
	56013		56267
5/5	tgtacagGTaggact	5/6	tttctcacAGgggtgg
	56360		56557
6/6	cctggggGTgagttt	6/7	gcattgcAGctgggt

Table 3.63: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 50772, in the second exon. The analysis of the first 1000 bp located upstream from the transcription start site, using MOTIF, revealed the presence of potential binding sites for several transcription factors, such as Lyf-1, p53, GATA, NF-AT, Ik-1, Ik-2, AP-1, NRF-2 and c-Ets binding sites. The NNPP promoter prediction program did not predict any promoter in the first 1000 bp upstream from the RIKEN2 gene, suggesting that this gene has a TATA-less promoter. Finally, a BLASTP search on the amino acid sequence of the putative RIKEN2 protein, deduced from the RIKEN2 cDNA sequence, indicates that the RIKEN2 protein corresponds to a hypothetical protein, CGI-126, that is evolutionarily conserved between human and *Caenorhabditis elegans* (Lai *et al.*, 2000). The CGI-126 protein exhibits sequence conservation with a *C. elegans* hypothetical protein encoded by the C40H1.6 gene in chromosome III (Wilson *et al.*, 1994). An ALIGN comparison between the CGI-126 and C40H1.6 proteins indicates that they share 68.3% homology (Fig. 3.77).

```
                10        20        30        40        50        60
RIKEN2 MADEATRRVVSEIPVLKTNAGPRDRELWVQRLKEEYQSLIRYVENNKNSDNDWFRLESNK
       : ::::.  .. ::. ::.:.::: .:::.:::::::..:  :.::::. : :::.:::::.
Hypot. M-DDATKSSLKAIPLCKTKASPRDGDLWIERLKEEYEAIIAAVQNNKDCDRDWFQLESNE
                10        20        30        40        50

                70        80        90       100       110       120
RIKEN2 EGTRWFGKCWYIHDFLKYEFDIEFEIPITYPTTAPEIAVPELDGKTAKMYRGGKICLTDH
       .::.:::::::::.:.:::::.::.:::::::.::::::.:::::::::::::::::..:
```

```
Hypot. RGTKWFGKCWYFHNMVKYEFDVEFDIPITYPVTAPEIALPELDGKTAKMYRGGKICLSEH
      60        70        80        90       100       110

              130       140       150       160
RIKEN2 FKPLWARNVPKFGLAHLMALGLGPWLAVEVPDLIQKGVIQHKEKCSQ
       :::::::::.:::::.:: .:::::::::.:::::.::.:: :
Hypot. FKPLWARNTPKFGIAHAFALGLGPWMAVEIPDLIEKGLIQPKA----
     120       130       140       150       160
```

Fig. 3.77: ALIGN comparison between the amino acid sequences of the hypothetical proteins CGI-126 (bottom) and encoded by the hypothetical RIKEN2 gene (top).

HI-2-4-16 Mouse RIKEN3 gene

A GENSCAN analysis, followed by a BLAST search, of the mouse clone rp23-395h6 reveals the presence of a third hypothetical gene at position 194924-209533 (Fig. 3.48), which has over 99% sequence identity with a cDNA (Genbank accession number NM_025321) from the RIKEN mouse cDNA library (Carninci and Hayashizaki, 1999). This hypothetical gene was named RIKEN3. A CROSSMATCH analysis between the RIKEN3 cDNA and the mouse clone rp23-395h6 reveals that the 3' end of the RIKEN3 gene is not present on the mouse clone rp23-395h6, and that the genomic segment at position 194924-209533 encompasses only the first 5 exons of the RIKEN3 gene (table 3.64 and Fig. 3.78). Also, the results of the Neural Network splice site prediction program analysis also confirmed the presence of potential splice sites at the exon-intron borders indicated by the ALIGN comparison of the RIKEN3 genomic region and its cDNA (table 3.65).

RIKEN3 gene (first five exons):

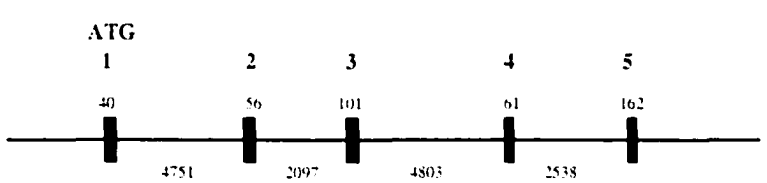


Fig. 3.78: Structural organization of the first five exons of the hypothetical RIKEN3 gene. Sizes, in base pairs, are given below the introns and above the exons.The exons are indicated by black boxes and the connecting introns by horizontal lines

Exon number	Contig position	Exon length (bp)
1	194924-194968	44
2	199715-199771	56
3	201868-201969	101
4	206772-206833	61
5	209371-209533	162

Table 3.64: contig position and length of the first 5 exons of the putative RIKEN3 gene.

	Donor site sequence		Acceptor site sequence
Exon/Intron	Sequence	Intron/Exon	Sequence
	194970		199714
1/1	tgctgagGTgacttt	1/2	ctcttgcAGacatgt
	199773		201867
2/2	tcagaaaGTaagttc	2/3	ttattttAGtgctgc
	201971		206771
3/3	tctacaaGTaagggc	3/4	tctttgtAGatggtc
	206835		209370
4/4	agtggagGTatgtat	4/5	ctctctcAGgggtct

Table 3.65: Neural Network splice site prediction sequences: exons and introns are numbered according to their positions and the numbers on top of the donor and acceptor sites correspond to their contig positions. The numbered bases are underlined.

A putative translation start site (ATG) is located at position 194948, in the first exon, suggesting that the 5' UTR is a short region. Initial analysis of the 1000 bp located immediately upstream from the transcription start site, using MOTIF, indicates the presence of several potential transcription factor binding sites, among them are potential GATA, NF-AT, Sp1, c-Myb, AP-1, Gfi-1, NF-kB, c-Rel and c-Ets binding sites. The search for a promoter, using NNPP, predicted 5 potential promoter regions at positions 758-808 bp, 750-800 bp, 318-368 bp, 55-105 bp and 38-98 bp upstream from the transcription start site. A CCAAT box was located, using MOTIF, 685-696 bp upstream from the transcription start site, suggesting that the promoter region is located 318-368 bp upstream from the 5' end of RIKEN3 . A search of the PROSITE protein profile database with the hypothetical RIKEN3 predicted amino acid sequence revealed

two succinate dehydrogenase cytochrome b subunit signature sites, located at amino acid positions 50-74 and 127-140. A BLASTP analysis of the putative RIKEN3 protein revealed sequence homology with the human large subunit of cytochrome b (cybL) in the succinate-ubiquinone oxidoreductase (complex II) of mitochondria (Hirawake *et al.*, 1999). An ALIGN comparison between the mouse RIKEN3 protein and the human cybL protein indicates that these two proteins have 79.3% sequence identity (Fig. 3.79) and suggests that the RIKEN3 gene may correspond to the mouse ortholog of the human cybL gene. Since Hirawake *et al.* (1997) reported that the human cybL gene was mapped to chromosome 1q21, in a region conserved between mouse and human that corresponds to that of mouse BAC clone rp23-395h6, it is likely that this region encodes the mouse ortholog of the human cybL gene.

```
                10        20        30        40        50        60
RIKEN3 MAAFLLRHVSRHCLRAHLNAQRCIRNAAPLGTTAKEEMERFWKKNTSSNRPLSPHLTIYK
       :::.::::::.:::::::.. : :::::.::::::::::::::.:: .:::::::::.:::.
human  MAALLLRHVGRHCLRAHFSPQLCIRNAVPLGTTAKEEMERFWNKNIGSNRPLSPHITIYS
                10        20        30        40        50        60

                70        80        90       100       110       120
RIKEN3 WSLPMALSVCHRGSGIALSGGVSLFGLSALLLPGNFESYLMFVKSLCLGPTLIYSAKFVL
       :::::::.:.::::.:::::::.::::::::::::: .:::::::::.::..::::.:
human  WSLPMAMSICHRGTGIALSAGVSLFGMSALLLPGNFESYLELVKSLCLGPALIHTAKFAL
                70        80        90       100       110       120

               130       140       150       160
RIKEN3 VFPLMYHSLNGIRHLLWDLGKGLAIPQVWLSGVAVVVLAVLSSGGLAAL
       :::::::. :::::::.:::::::: ::::.. :::.:.::.::::: ::::.
human  VFPLMYHTWNGIRHLMWDLGKGLKIPQLYQSGVVVLVLTVLSSMGLAAM
               130       140       150       160
```

Fig. 3.79: ALIGN comparison between the amino acid sequences of the human cybL protein (bottom) and of the protein encoded by the hypothetical RIKEN3 gene (top).

III-2-4-17 Repeat elements in the mouse BAC clone rp23-395h6

A Repeatmasker analysis of the repeat elements in the 211771 bp of the mouse clone rp23-395h6 indicates a relatively low percentage of repeat elements, as 38.13% of

the total sequence of this clone consists of repeat elements (table 3.66). Most of the repeat elements correspond to SINEs and, to a lesser extent, LTR elements. LINE elements make only 2.96% of the total repeats present in this clone, in contrast to what has been observed in the mouse clone rp23-21118, in which LINE elements represent about 40% of the total sequence of the clone. The high percentage of LINE elements in the mouse clone rp23-21118 was correlated to the total absence of genes, but in the case of the clone rp23-395h6, the high percentage of SINE elements can be correlated to the presence of a high number of genes. This is consistent with the overall observation that SINE elements in the human genome mostly are present in gene-rich regions (Lander *et al.*, 2001).

	number of elements	length occupied	percentage of sequence
SINEs:	319	42666 bp	20.15%
B1s	170	20082 bp	9.48%
B2-B4	129	20491 bp	9.68%
IDs	10	723 bp	0.34%
MIRs	10	1370 bp	0.65%
LINEs:	16	3033 bp	1.43%
LINE1	7	658 bp	0.31%
LINE2	8	2309 bp	1.09%
L3/CR1	1	66 bp	0.03%
LTR elements:	62	25940 bp	12.25%
MaLRs	23	6718 bp	3.17%
ERVL	1	486 bp	0.23%
ERV_classI	4	2260 bp	1.07%
ERV_classII	19	9737 bp	4.60%
DNA elements:	7	949 bp	0.45%
MER1_type	6	805 bp	0.38%
MER2_type	1	144 bp	0.07%
Unclassified:	4	769 bp	0.36%
Total interspersed repeats		73357 bp	34.64%
Small RNA:	1	96 bp	0.05%
Satellites:	0	0 bp	0.00%
Simple repeats:	99	5569 bp	2.63%

Low complexity:	31	1758 bp	0.83%

Table 3.66: Identification of the repeat elements on the mouse clone rp23-395h6.

III-2-5 Mouse BAC clone rp23-9p22

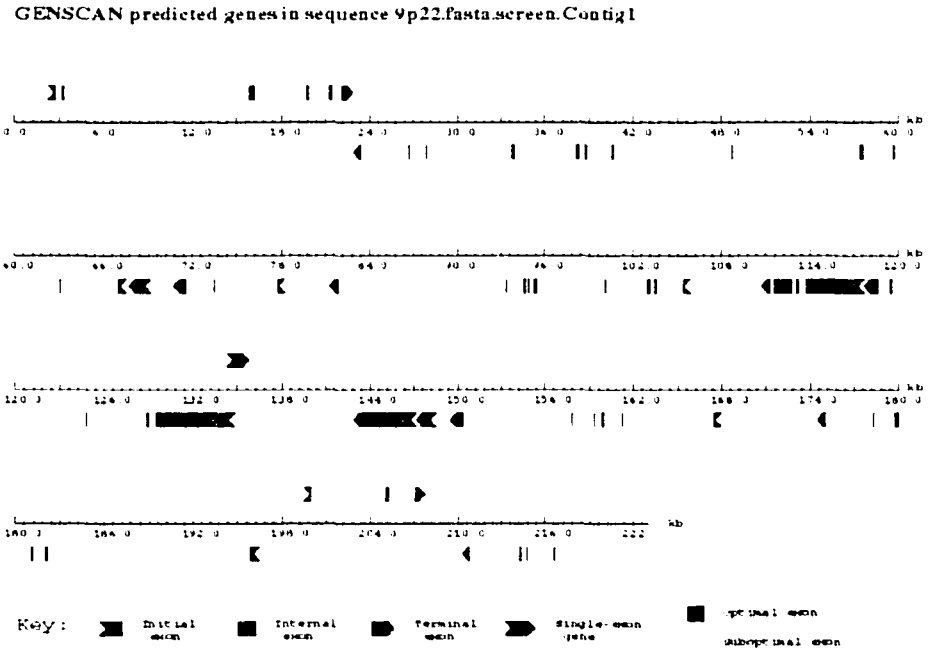


Fig. 3.80: GENSCAN analysis of the mouse BAC clone rp23-9p22. The genes predicted by GENSCAN are repeat elements or false positives.

The mouse BAC clone rp23-9p22 (Genbank accession number AC079446), originally mapped to mouse chromosome 1q22 consists of 223103 bp. Both GENSCAN analysis and BLAST searches versus the nr and dbEST databases, reveal that this clone does not encode any known or putative genes (Fig. 3.80). In addition, STS markers contained into the mouse BAC clone rp23-9p22 have been mapped to mouse chromosome X (see http://www.ensembl.org/Mus_musculus). Based on this data, it appears that the mouse BAC clone rp23-9p22 does not belong to mouse

chromosome 1q region but rather to mouse chromosome X. The position of rp23-9p22 therefore would be different from the one that was assumed at the beginning of this dissertation research.

The Repeatmasker analysis of the repeat elements in the 223103 bp of the clone rp23-9p22 indicates that 53.37% of the total sequence of this clone is made of repeat elements (table 3.67). Most of the repeat elements correspond to LINE1 elements and, to a lesser extent, LTR elements. SINE elements are only 6.72% of the total repeat elements present in this clone, confirming the above observation that regions with a high percentage of LINE1 elements and a low percentage of SINEs in this region of mouse chromosome 1q21-23 encode few or no genes. Interestingly, the repeat density observed for the BAC clone rp23-9p22 is similar to that observed for the BAC clone rp23-21118, as their high percentage of LINE elements can be correlated to the total absence of genes (Lander *et al.*, 2001).

	number of elements	length occupied	percentage of sequence
SINEs:	59	8005 bp	3.59%
B1s	26	3268 bp	1.46%
B2-B4	27	4334 bp	1.94%
IDs	5	356 bp	0.16%
MIRs	1	47 bp	0.02%
LINEs:	77	86310 bp	38.69%
LINE1	77	86310 bp	38.69%
LINE2	0	0 bp	0.00%
L3/CR1	0	0 bp	0.00%
LTR elements:	42	20045 bp	8.98%
MaLRs	15	7292 bp	3.27%
ERVL	1	476 bp	0.21%
ERV_classI	2	618 bp	0.28%
ERV_classII	15	8531 bp	3.82%
DNA elements:	0	0 bp	0.00%
MER1_type	0	0 bp	0.00%
MER2_type	0	0 bp	0.00%
Unclassified:	1	116 bp	0.05%
Total interspersed repeats		114476 bp	51.31%

Small RNA:	0	0 bp	0.00%
Satellites:	0	0 bp	0.00%
Simple repeats:	73	3668 bp	1.64%
Low complexity:	17	941 bp	0.42%

Table 3.67: Identification of the repeat elements on the mouse clone rp23-9p22.

III-2-6 Fine mapping of the QTL of bone mineral density

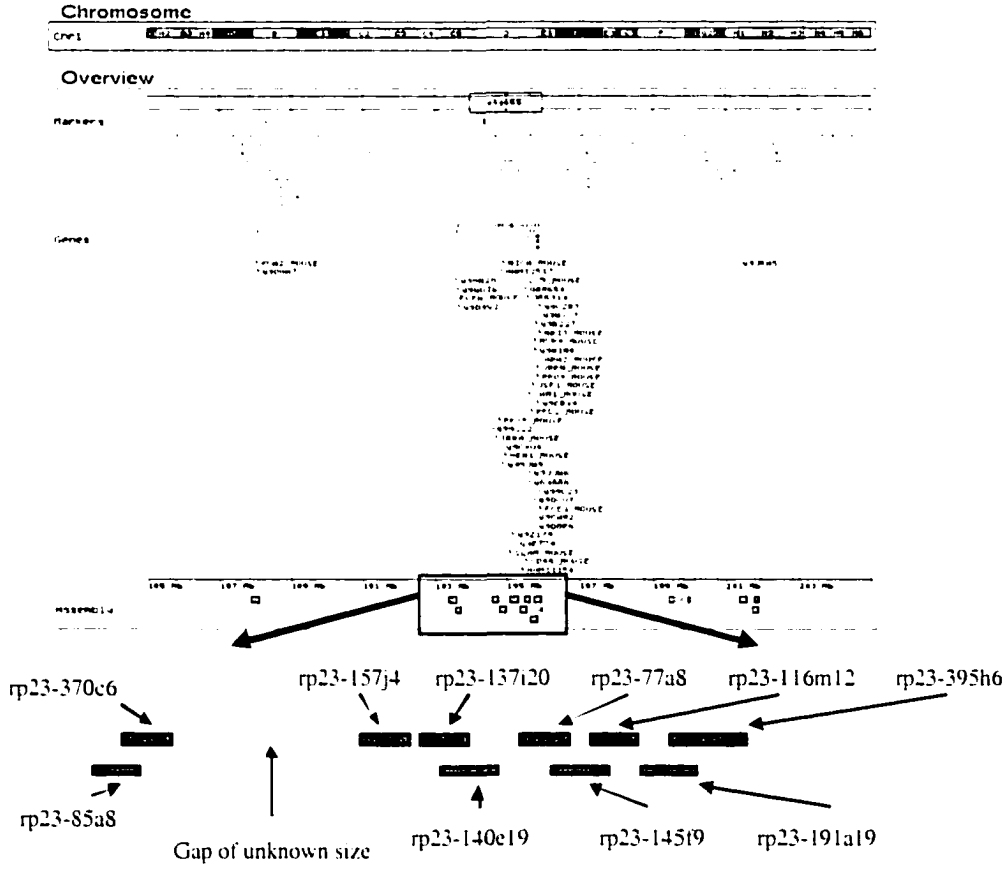


Fig. 3.81: Corrected map of the QTL region of bone density.

A new map of the QTL of bone density was generated based on the sequencing data of the mouse BAC clones contained in this region (Fig. 3.81). The chromosome

location of the QTL of bone density was first extracted from the mouse ensembl network (http://www.ensembl.org/Mus_musculus), and then confirmed based on the data generated in our laboratory (Fig. 3.81). The QTL region consists of two 1.4-Mb- and 0.37-Mb contigs. The Ifi200 cluster of genes is located ~380 kb upstream from the 0.37 Mb contig, and the 1.4-Mb- and 0.37-Mb contigs are separated by a gap of ~850 kb, based on assembly data generated from whole genome shotgun trace data at http://www.ensembl.org/Mus_musculus.

Chapter IV

Conclusion

In this dissertation research, nine BAC clones which tentatively were mapped to mouse chromosome 1q were sequenced using the double stranded random shotgun approach and analyzed.

The sequences of four BAC clones (mgs1-68e20, mgs1-166n14, rp22-225p5 and mgs1-423c02) generated two 240-kb- and 154-kb-contigs containing a cluster of at least ten genes encoding a series of interferon-activatable protein, termed the "200 family" of proteins. Five previously identified genes (Ifi201, Ifi202a, Ifi202b, Ifi203 - renamed Ifi203b, and Ifi204) were localized on the 240-kb contig, in addition to two new Ifi200 family members, Ifi203a and Ifi203c. A pseudogene, Ifi202c, was localized on the 154-kb contig, along with a putative pseudogene, "203-like", and a putative new gene, "204-like". A putative splice variant of "204-like", exhibiting strong sequence similarity with the mRNA of an additional member of the "200 family" of proteins, D3, was discovered, using *in silico* analysis of the "204-like" sequence. The Ifi202a and Ifi202b genes, and the Ifi203a and Ifi203b genes, showed 99.65% and 99.62% sequence similarity between their exonic, and intronic, sequences, respectively. At the protein level, the translation product of Ifi203b (p203b) differed from the translation product of Ifi203a (p203a) by only one amino acid out of 408, while the translation product of Ifi202a (p202a) differed from that of Ifi202b (p202b) by 7 amino acids out of 445. An additional 144-nucleotide long putative coding segment was found in the sequence of intron 4 of Ifi203a and Ifi203b, raising the possibility of an alternatively spliced form of the Ifi203a and Ifi203b genes.

At the transcriptional level, the genes of the mouse Ifi200 cluster exhibit different modes of transcriptional regulation. Comparison of the mouse Ifi200 gene

cluster with its human counterpart indicates an absence of conservation in the non-coding regions.

Sequencing the numerous repetitive regions on mouse BAC clone mgs1-423c02 was extremely complex because it contained large strongly conserved sequences. Thus, a series of consensus DNA sequences were generated from the "walking" sequence reads off the repeat spanning subclones and added into the Phrap assembly. The resulting assembled data were confirmed by experimentally generating a series of nested deleted clones by minitransposon random insertion (Chatterjee and Coren, 1997). These deletion clones then were end-sequenced with specifically designed custom primers to confirm that this complex region had been assembled correctly.

In contrast to the ten genes contained in the mouse Ifi200 cluster, the human Ifi200 gene cluster has only three genes (IFI-16, MNDA and AIM2) and several putative pseudogenes (see above). The organization of these Ifi genes into clusters may have regulatory significance, as it does, for example, for the clustering of globin genes (Hardison and Miller, 1993). These clusters also may offer an evolutionary advantage because they could facilitate rapid crossing over and gene conversion events that may serve to expand a gene family, as it has been proposed for the histone gene cluster (Matsuo, 2000). It is interesting to note that the contiguous region containing the seven genes from the mouse Ifi200 gene cluster has a gene encoding a protein containing the two *a* and *b* domains (Ifi202a, Ifi202b and Ifi204), followed by a gene encoding a protein containing one of the two domains (Ifi203c, Ifi203b, Ifi203a and Ifi201). In contrast, the human Ifi200 gene cluster exhibits a slightly different genomic organization, as it consists of only one gene encoding for a protein containing the two *a* and *b* domains (IFI-16), surrounded by two genes encoding for a protein containing one of the two domains (MNDA and AIM2). Although the functional significance, if any, of this genomic arrangement is not known, it may reflect a selective advantage or adaptation to ecological aspects that are more important to mice than to humans.

How can the smaller size of the human Ifi200 gene cluster, compared to the mouse, be explained? Comparative analysis of human chromosome 19 with syntenic regions in mouse (Dehal *et al.*, 2001) reveals striking differences between single-copy human chromosome 19 genes, which overwhelmingly are conserved in mouse, and genes organized into tandem familial clusters. Thus, the clustered genes in this region of chromosome 19 in humans and mice differ greatly in number, coding capacity, and organization. As an example, the *Kruppel*-type zinc finger proteins which code for putative transcription factors are clustered in eleven different locations in human chromosome 19, and most of these clusters contain highly similar genes that appear to have arisen by duplications of ancestral genes. Interestingly, many of these human and related mouse gene clusters contain very different sets of orthologous genes, suggesting that if their ancestral genes were duplicated, they may have been lost or selected independently in each cluster during the ~70-100 million years since the human-mouse speciation (O'h Uigin and Li, 1992)).

Several comparative studies of the mouse and human olfactory receptor (OR) genes (Dehal *et al.*, 2001; Lane *et al.*, 2001) reveal that a substantial fraction of human OR genes have lost their capacity to encode functional proteins and therefore have become pseudogenes. It has been suggested that the diminished number of OR genes in humans compared to mouse, resulted in the decline in both the breadth and discriminatory power of the human olfactory system (Dehal *et al.*, 2001). A similar phenomenon also may have occurred in the Ifi200 gene cluster. Here it is known that the human MNDA, IFI-16 and AIM2 genes have a very narrow pattern of expression and are restricted to cells of the myelomonocytic lineage in a stage-dependent manner. In contrast, the genes of the mouse Ifi200 gene cluster appear to have a wider range of expression in different cell lineages (see Introduction) as, for example, Ifi202a is widely expressed in mouse and can be induced in various organs, including spleen, heart and brain. It therefore is possible that the proteins encoded by Ifi202a, and most likely

Ifi202b, exert functions in non-myeloid tissues that are similar to those exerted, in human, by proteins that are not encoded by the human Ifi200 gene cluster. In this regard, it is interesting to note that the human IFI-16, MNDA and AIM2 genes also show significant sequence similarities to the mouse Ifi204 gene, whose expression similarly is restricted to myelomonocytic cells, and thus can be considered as a direct neighboring ortholog of the human genes.

In addition, as has been shown, in the case of a *Kruppel*-type zinc finger protein cluster (Dehal *et al.*, 2001), highly similar duplicate regions often display distinct patterns of tissue-specific expression, suggesting that the newly duplicated genes might acquire new functions. The strong sequence similarity between Ifi202a and Ifi202b, and between Ifi203a and Ifi203b, respectively, could be explained by recent gene conversions, and the presence of extremely similar but non identical proteins might provide one path for rapid functional diversity in a tissue-dependent manner.

During this dissertation research, five BAC clones in the mouse chromosome 1q21-23 QTL bone density region also were sequenced and merged with the sequence of other BAC clones to create two contigs of approximately 1.4 Mb and 0.37 Mb containing at least 40 genes. The analysis of the genes present in each contig revealed several new candidate genes that could be involved in bone metabolism. For example, Forus *et al.* (1998) identified the apo A-II gene (on mouse BAC clone rp23-395h6) as an important element in the recurrent amplification of the genomic region 1q21-22 in bone tumors, suggesting that apoAII may have an important role in bone metabolism. However, this dissertation research indicated the presence of numerous other genes in this QTL region, encoding, among others, the chaperone prefoldin (Vainberg *et al.*, 1998), the Junctional Adhesion Molecule JAM present in tight junctions (Martin-Padura *et al.*, 1998), the cytoplasmic coatomer protein (De la Vega and Stockert, 1999), the lymphocytic marker Ly-9 (Mathieson *et al.*, 1980), nicastrin, a protein implicated in

Alzheimer's disease (Fagan *et al.*, 2001), PEA-15, a protein involved in apoptosis (Estelles *et al.*, 1999), the soluble lectin intelectin (Komiya *et al.*, 1998), the serum amyloid P-component (de Haas, 1999), the IgE receptor alpha subunit (Blank *et al.*, 1989), the Duffy blood group antigen (Cartron and Colin, 2001) and calsequestrin, the major calcium-binding protein of the sarcoplasmic reticulum of skeletal and cardiac muscles (Yano and Zarain-Herzberg, 1994).

In addition, at least five genes in this QTL region encode putative calcium-ATPases that can be grouped into families, based on their subcellular location in the plasma membrane calcium-ATPases (PMCAs) and the sarco(endo)plasmic reticulum calcium-ATPases (SERCAs). Both are multigene families that play an important role in maintaining intracellular calcium concentration. The SERCAs pump calcium ions into internal stores, such as the endoplasmic reticulum or sarcoplasmic reticulum, while the PMCAs pump calcium ions outside the cells into the extracellular spaces (for review, see Shull, 2000). SERCAs mostly are involved in muscle contractions and predominate in skeletal and cardiac muscles (for review, see East, 2000; Shull, 2000), but the role of PMCAs is much less obvious. Mineralization of bones involves the laying down of extracellular matrix, calcium transport into the matrix and formation of hydroxyapatite. These processes are regulated by cells that line the surface of bone, including bone-resorbing osteoclasts and bone-forming osteoblasts. One possible role for PMCAs, suggested by Abramowitz and Suki (1996), is that the efflux of calcium from bone cells to the mineralizing matrix involves osteoblastic PMCAs. There are at least four genes that encode the human PMCAs which map to 12q21-23, 3p25-36, Xq28 and 1q25-37. The QTL of BMD at mouse 1q21-23 is in a region of conserved linkage between mouse and human, but no known PMCAs map to this region. However, the putative calcium-ATPases, encoded by genes located in this region of mouse chromosome 1, do contain domains that are similar to calcium-ATPase domains, and therefore may have functions similar to those of the bone mineralization PMCAs. Thus, the putative

calcium-ATPase genes are primary candidates for BMD regulation in the QTL located on mouse 1q21-23.

Interestingly, a gene involved in a metabolic disorder known as absorptive hypercalciuria (AH) has been mapped to this 1q23.3-24 region by Reed *et al.* (1999). AH is a biological syndrome marked by an excretion in the urine of more than 0.1 mmol/kg/24 hours of calcium and caused by an intestinal hyperabsorption of calcium, a defective reabsorption of calcium by the renal tubule or an increased bone resorption (Audran and Legrand, 2000). Pietschmann *et al.* (1992) previously demonstrated that patients with AH displayed a significantly lower lumbar bone density compared with healthy subjects, and therefore were more at risk for vertebral factures. Thus, the putative calcium-ATPases whose genes are located in this region may be involved in maintaining calcium homeostasis inside the bone cell, and a defect in this calcium-ATPase-encoding gene could be involved in bone readsorption .

The mouse BAC clones mgs1-68e20 and mgs1-166n14, and the mouse BAC clones rp22-225p5 and mgs1-423c02 encompassing the Ifi200 gene cluster, are located in the neighborhood of the 0.37-Mb contig containing the QTL of bone mineral density (see Fig. 3.81). These Ifi200 genes encode inteferon-activatable proteins with cell growth regulation and inflammation-associated functions (see Introduction). When the best characterized "200 family" proteins, p202a and p204, are overexpressed, they retard cell proliferation *in vitro* (see Introduction). Interestingly, the nearby QTL region associated with BMD encodes several genes that are involved in immune- and inflammation-associated functions. These genes include the IgE receptors alpha and gamma subunit (Fcer1a and Fcer1g, respectively), the Duffy blood group antigen, the serum amyloid P-component (Sap), which is a constituent of amyloid deposits in secondary amyloidosis, a chronic inflammatory condition (de Haas, 1999), and a series of membrane proteins from immune cells involved in antigen recognition (CD150,

CD48, 19A24, Ly108, CD84, Ly9 and 2B4). Interestingly, one effect of T cells-derived interferon gamma is to induce the enhanced expression of MHC class II antigens on accessory cells, such as monocytes and macrophages, which either do not express class II antigens constitutively or express them at a very low level (Liu and Janeway, 1990). An additional effect of interferons is to increase expression of the IgE receptor gamma subunit on mast cells (Woolhiser *et al.*, 2001). Thus, this region of chromosome 1, which is highly conserved in humans and mice, contains a complex gene set which have immunomodulatory and cell growth regulatory functions, and whose expression is influenced by cytokines. Further studies should establish whether there is a complex of genes in this region whose transcription is regulated by common elements -i.e. cytokines, and whose translation products are involved in immunomodulatory functions.

Finally, molecular phylogeny can be considered as the ultimate tool for determining evolutionary relationships. As the full genomes of several major eukaryotes become available, the phylogenic trees of these different species, as well as an accurate understanding of how the cells and genomes of these species evolved, can be answered through comparison and integration of the analysis of a large number of independent molecular loci. These comparative genomic studies have confirmed the earlier observations that were obtained from ribosomal RNA phylogenic studies (Field *et al.*, 1988), such as the partition of the cellular living world into three domains (archaea, bacteria and eucarya), as well as the monophyly of each domains (Sicheritz-Ponten and Andersson, 2001; Woese, 1987). In addition, these comparative genomic studies also reveal an unexpected degree of genome plasticity, and the wider than expected occurrence of horizontal gene transfer. These events likely are a major evolutionary force in both prokaryotes and eukaryotes (Ball and Cherry, 2001). These events also could be the main reason behind intermixing of species from different domains in various phylogenic trees (Doolittle, 1999). In recent studies, Brown *et al.* (2001)

reduced the risk of error linked to the analysis of individual protein trees by analyzing widely distributed genes as concatenated protein sequence data sets, rather than individual protein sequences. In this approach, the accumulation of sampled gene loci lowers the statistical error rate in the data. During this dissertation research, a similar quantitative genetic analysis was applied to the genes present on the two mouse chromosome 1 contigs sequenced. The amino acid sequences of the coded proteins of the orthologous human genes therefore were systematically compared, using BLAST, with the amino acid sequences of their orthologs in other species. The results of this analysis from mouse, rat, cow, fish, fly, worm, plant, yeast and *E. coli*, and their resulting percentage of sequence identity are plotted in table 3.68. The resulting data for each species then were combined to draw a comparison of the average percentage of coding sequence identity in this region between humans and each of the species depicted in table 3.68. As shown on table 3.68, while prokaryotes and yeast are further separated from humans, in terms of coding sequence identity, vertebrates, such as rat, mouse, and cow, are the closest and share an average of approximately 70% of their coding sequences with humans. These data of course confirms that vertebrates, such as rat, mouse and cow, are closer to humans, in term of gene coding sequences, than are lower eukaryotes, such as worms, and prokaryotes (Dacks and Doolittle, 2001).

The proteins coded by genes present in this region may be classified in two categories. First, this region encodes proteins involved in housekeeping activities, such as protoporphyrinogen oxidase, nitrilase, or beta-1,4-galactosyltransferase, that appeared early in the evolution. In addition, this region also encodes proteins present only in eukaryote that are involved in regulatory activities, such as the ubiquitin-specific protease Usp23, or the nuclear receptor CAR, or that are involved in eukaryote cell-specific functions, such as the lymphocyte antigen Ly-9, or the IgE gamma subunit Fcer1g. According to Rubin *et al.* (2000), many eukaryote-only gene encoded proteins have extracellular domains involved in cell-cell and cell-substrates contacts. Intelectin,

	Rattus norvegicus	Mus musculus	Bos taurus	Fugu rubripes	Drosophila melanogaster	C. elegans	A. thaliana	S. cerevisiae	E. coli
19A24		39.7							
CD48	49.8	52.2							
CD84		56.2							
CD150		51.2	65.5						
Ly-9	*64.0*	53.1							
2B4	36.8	48.2							
Fcer1g	90.7	88.4	93.0	60.0					
Fcer1a	45.7	47.1	26.8	27.7					
Intln		81.2							
Nectin-like	*96.0*	87.3		42.7					
Usf1	98.4	98.4		49.0					
SAP	69.3	68.3		40.5					
PEA-15	100	99.2		79.2					
nhlh1		92.5		64.6					
DEDD	98.4	98.7		37.7					
Myelin P0	35.3	35.7	31.6	32.6					
Apo A-II	55.9	58.8	50.0	24.8					
JAM	71.8	68.1	74.6	40.7					

	Rattus norvegicus	Mus musculus	Bos taurus	Fugu rubripes	Drosophila melanogaster	C. elegans	A. thaliana	S. cerevisiae	E. coli
Duffy	*61.8*	60.8	62.8						
Timm23	95.7	96.2				31.0		24.8	
nicastrin	*90.9*	77.1		52.9	26.7	20.3	26.1		
kcnj9	99.0	93.6	47.1	67.5	38.8	33.9			
kcnj10	*96.0*	99.2	36.3	70.6	31.4	27.3			
Adamts4	91.7	90.2	95.9	45.2	36.6	40.1			
CAR	76.5	57.5	34.8	37.1	17.9	29.1			
Pxf	85.4	90.3		39.3	30.1	30.5	27.7	12.2	
Casq1	70.3	92.7		67.6	19.1	31.3	22.8		
Usp23	98.5	96.5		45.6	23.4	18.7	17.2	15.3	
Prefoldin 2	*95.4*	93.5		55.3	44.6	31.8	37.3	26.9	
Nit1	*89.0*	84.1		33.2	31.7	23.6	24.5	19.4	31.4
ppox	*86.1*	88.7		55.0	38.0		25.3	24.0	19.0
NDUFS2	*92.7*	92.4	88.8	87.7	71.8	62.8	60.0		30.0
betagal	*70.9*	83.5	83.6	56.1	31.9	21.7	17.3	17.2	14.2
	79.0	**76.4**	**60.8**	**50.7**	**34.0**	**30.9**	**28.6**	**19.9**	**23.6**

Table 3.68: phylogenic analysis of the proteins coded by the genes present on the two 0.37- and 1.4-Mb mouse contigs.Each proteins were compared with the proteins coded by the human orthologs of the mouse genes, using BLASTP and ALIGN, and the percentage of sequence identity plotted on the table. The data in italics under the rat column were obtained by comparing rat ESTs with the corresponding segment of the human coding sequence. The data in bold indicates the average percent of sequence identity in this region for each species. (19A24) Sayos *et al.*, 2001; (CD48) Wong *et al.*, 1990; (CD84) de la Fuente *et al.*, 1997; (CD150) Wang *et al.*, 2001; (Ly-9) Mathieson *et al.*, 1980; (2B4) Boles *et al.*, 1999; (Fcer1g) Ra *et al.*, 1989; (Fcer1a) Blank *et al.*, 1989; (Intln: intelectin) Komiya *et al.*, 1998; (Nectin-like: nectin-like protein 1) Tachibana *et al.*, 2000; (Usf1: upstream stimulatory factor 1) Halle *et al.*, 1995; (SAP: serum amyloid-P component) de Haas, 1999; (PEA-15: astrocytic phosphoprotein 15) Estelles *et al.*, 1999; (nhlh1) Murdoch et al., 1999; (DEDD: death effector domain-containing protein) Stegh *et al.*, 1998; (myelin P0: myelin protein zero) Roomi *et al.*, 1978; (Apo A-II: apolipoprotein A-II) Tall, 1990; (JAM: Junction Adhesion Molecule) Martin-Padura *et al.*, 1998; (Duffy: blood group antigen

CAR and JAM, as well as the immunoglobulin-related 19A24, CD48, CD84, CD150, Ly-9 and 2B4 proteins are involved in such contacts. Also, these immunoglobulin-related proteins are part of a large protein family generated by gene duplication of a common ancestor (Wong *et al.*, 1990). The observation that these genes only are present in higher eukaryotes (table 3.68), confirms the idea that gene duplication, as well as gene loss, were major forces in speciation (Ball and Cherry, 2001, Rubin *et al.*, 2000).

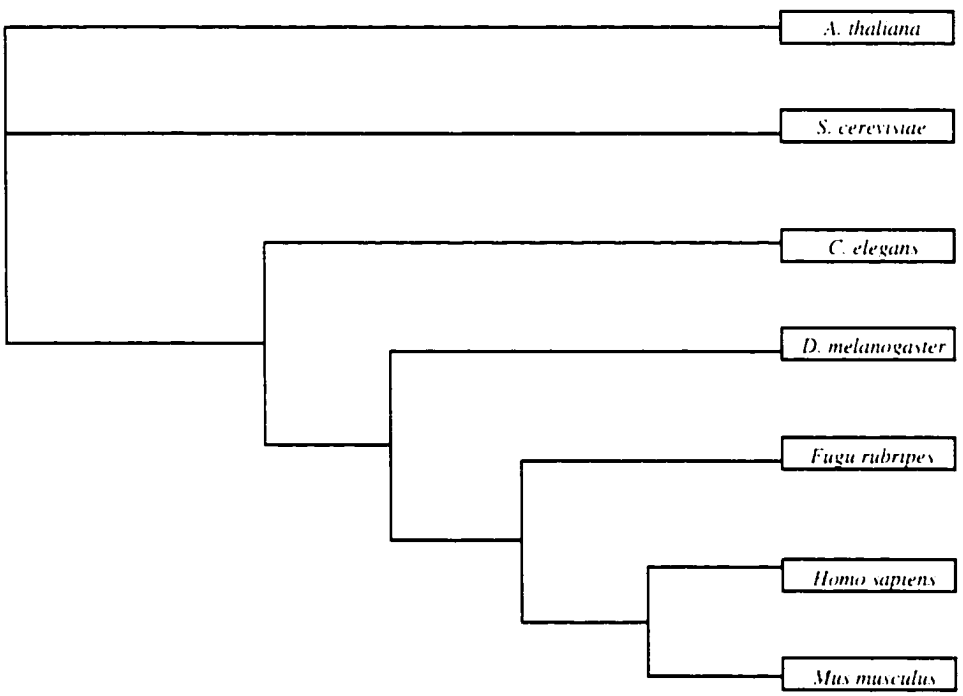


Fig. 3.82: Phylogenic tree for the concatenated set of pxf, Usp23, prefoldin subunit 2, NDUFS2 and beta-1,4-galactosyltransferase from human, mouse, fish, fly, worm, plant and yeast.

To further investigate the above evolutionary hypothesis, a phylogenic tree was built for the proteins described in table 3.68 that are widely distributed throughout evolution using the strategy described by Brown *et al.* (2001). Briefly, the protein sequences of the human pxf, Usp23, prefoldin subunit 2, NDUFS2 and beta-1,4-galactosyltransferase, and of their orthologs in mouse, fish, fly, worm, plant and yeast,

were manually concatenated for each species and aligned using CLUSTALW. Then, a phylogenic tree was created using TreeView (Fig. 3.82). These studies confirms that, in the case of worm, fly, fish, mouse and human, the successive grades of complexity that appeared in metazoans relied mostly on the progressive evolution of an already existing system, rather than the generation of a new one (Fig. 3.82) (Rubin *et al.*, 2000). In addition, the early demarcation of plants, as shown in Fig. 3.82, confirms the data presented by Adoutte *et al.* (2000) which indicated that plants and fungi bifurcated early from the metazoan tree of life. Finally, the demarcations of fish, mouse and human are in accord with the vertebrate phylogenic tree presented by Pollock *et al.* (2000). It must be noted that several proteins present in this region in lower organisms are part of large protein families (such as kcnj, Adamts, Usp, and galactosyltransferase), and that the determination of true orthologs determined by the "highest-hit" method (i.e. BLAST) therefore may not be truly accurate. This is because the existence of paralogs within the same species (Xie and Ding, 2000) might have led to new functions, in contrast to orthologs which retain the same function during evolution (Fitch, 1970). Although this may be true in cases when the compared species are not phylogenically distant, and when the orthologous relationships are simple, domain shuffling and domain sharing may lead to complex orthologous relationships, and the possibility of different functions among orthologs. For example, orthologs of Adamts4, a protein involved in the erosion of human articular cartilage (Tortorella *et al.*, 1999) are present in lower organisms (table 3.68). Thus, although Adamts4 (one of the gene studied on mouse chromosome 1) contains a mosaic of domains that may have been evolutionary conserved, these domains clearly perform different functions than those carried out by their early ancestors. Thus, it has been suggested that domain shuffling and domain sharing among proteins may be an important factor in determining the function of a protein during evolution (Li *et al.*, 2001).

The rapid increase in the availability of large amounts of DNA sequence data from a variety of organisms will dictate a continuing reevaluation of the evolutionary relationships between species (Adoutte *et al.*, 2000). This will be an ongoing process as the sequences of the genomes of additional animals, such as rat, mouse, cow or fish, become available. Therefore, over the next few years, as this new data emerges, additional evolutionary relationships will emerge to provide a deeper analysis of the complex relationships that developed between species throughout evolution.

Bibliography

Abramowitz J. and Suki W.N. (1996) Ca-ATPase and bone cell mineralization. *Miner. Electrolyte Metab.* 22: 336-344

Adams J.W., Kaufman R.E., Kretschmer P.J., Harrison M. and Nienhuis A.W. (1980) A family of long reiterated DNA sequences, one copy of which is next to the human beta globin gene. *Nucleic Acids Res.* 8: 6113-6128

Adoutte A., Balavoine G., Lartillot N., Lespinet O., Prud'homme B. and de Rosa R. (2000) The new animal phylogeny: reliability and implications. *Proc. Nat. Acad. Sci.* 97: 4453-4456

Almeida R., Amado M., David L., Levery S.B., Holmes E.H., Merkx G., van Kessel A.G., Hassan H., Bennett E.P. and Clausen H. (1997) A family of human beta4-galactosyltransferases: cloning and expression of two novel UDP-galactose:beta-n-acetylglucosamine beta-1,4-galactosyltransferases, beta4Gal-T2 and beta4Gal-T3. *J. Biol. Chem.* 272: 31979-31992

Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410

Anderson S. (1981) Shotgun DNA sequencing using cloned DNAse I-generated fragments. *Nucleic Acids Res.* 9: 3015-3027

Anderson S. *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290: 457-465

Amati B., Dalton S., Brooks M.W., Littlewood T.D., Evan G.I. and Land H. (1992) Transcriptional activation by the human c-Myc oncoprotein in yeast requires interaction with Max. *Nature* 359: 423-426

Angel P. and Karin M. (1991) The role of Jun, Fos and the AP-1 complex in cell proliferation and transformation. *Biochim. Biophys. Acta* 1072: 129-157

Audi L., Garcia-Ramirez M. and Carrascosa A. (1999) Genetic determinants of bone mass. *Horm. Res.* 51: 105-123

Audran M. and Legrand E. (2000) Hypercalciuria. *Joint Bone Spine* 67: 509-515

Avery O.T., MacLeod C.M. and McCarty M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.* 79: 137-158

Baes M., Gulick T., Choi H.-S., Martinoli M.G., Simha D. and Moore D.D. (1994) A new orphan member of the nuclear hormone receptor superfamily that interacts with a subset of retinoic acid response elements. *Mol. Cell. Biol.* 14: 1544-1552

Baeuerle P.A. and Henkel T. (1994) Function and activation of NF-kappa-B in the immune system. *Annu. Rev. Immunol.* 12: 141-179

Ball C.A. and Cherry J.M. (2001) Genome comparisons highlight similarity and diversity within the eukaryotic kingdoms. *Curr. Opin. Chem. Biol.* 5: 86-89

Barbaras R., Puchois P., Fruchart J.C. and Ailhaud G. (1987) Cholesterol efflux from cultured adipose cells is mediated by LpAI particles but not by LpAI:AII particles. *Biochem. Biophys. Res. Commun.* 142: 63-69

Bardwell V.J. and Treisman R. (1994) The POZ domain: a conserved protein-protein interaction motif. *Genes Dev.* 8: 1664-1677

Bartling D., Seedorf M., Schmidt R.C. and Weiler E.W. (1994) Molecular characterization of two cloned nitrilases from Arabidopsis thaliana: key enzymes in biosynthesis of the plant hormone indole-3-acetic acid. *Proc. Nat. Acad. Sci.* 91: 6021-6025

Beckmann H., Chen J.L., O'Brien T. and Tjian R. (1995) Coactivator and promoter-selective properties of RNA polymerase I TAFs. *Science* 270: 1506-1519

Belhumeur P., Paterno G.D., Boileau G., Claverie J.M. and Skup D. (1987) Isolation and characterization of a murine cDNA clone highly homologous to the yeast L29 ribosomal protein gene. *Nucleic Acids Res.* 15: 1019-1029

Bell S.P., Learned R.M., Jantzen H.M. and Tjian R. (1988) Functional cooperativity between transcription factor UBF1 and SL1 mediates human ribosomal RNA synthesis. *Science* 241: 1192-1197

Belogrudov G. and Hatefi Y. (1994) Catalytic sector of complex I (NADH:ubiquinone oxidoreductase): subunit stoichiometry and substrate-induced conformation changes. *Biochemistry* 33: 4571-4576

Benech P., Vigneron M., Peretz D., Revel M. and Chebath J. (1987) Interferon-responsive regulatory elements in the promoter of the human 2', 5'-oligo(A) synthetase gene. *Mol. Cell. Biol.* 7: 4498-4504

Berget S.M., Moore C. and Sharp P.A. (1977) Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Nat. Acad. Sci.* 74: 3171-3175

Bhartur S.G., Calhoun B.C., Woodrum J., Kurkjian J., Iyer S., Lai F. and Goldenring J.R. (2000) Genomic structure of murine Rab11 family members. *Biochem. Biophys. Res. Commun.* 269: 611-617

Blackwood E.M. and Eisenman R.N. (1991) Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. *Science* 251: 1211-1217

Blank U., Ra C., Miller L., White K., Metzger H. and Kinet J.-P. (1989) Complete structure and expression in transfected cells of high-affinity IgE receptor. *Nature* 337: 187-189

Bodenteich A., Chissoe S., Wang Y.F. and Roe B.A. (1993) "Shotgun cloning as the strategy of choice to generate templates for high-throughput dideoxynucleotide sequencing", in "Automated DNA sequencing and analysis techniques" (Venter, C. ed.) Academic Press, London

Boldin M.P., Goncharov T.M., Goltsev Y.V. and Wallach D. (1996) Involvement of MACH, a novel MORT1/FADD-interacting protease, in Fas/APO-1-and TNF receptor-induced cell death. *Cell* 85: 803-815

Boldin M.P., Varfolomeev E.E., Pancer Z., Mett I.L., Camonis J.H. and Wallach D. (1995) A novel protein that interacts with the death domain of Fas/APO1 contains a sequence motif related to the death domain. *J. Biol. Chem.* 270: 7795-7798

Boles K.S., Nakajima H., Colonna M., Chuang S.S., Stepp S.E., Bennett M., Kumar V. and Mathew P.A. (1999) Molecular characterization of a novel human natural killer cell receptor homologous to mouse 2B4. *Tissue Antigens* 54: 27-34

Bork P. and Koonin E.V. (1994) A new family of carbon-nitrogen hydrolases. *Protein Sci.* 3: 1244-1246

Botstein D., White R.L., Skolnick M. and Davis R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphism. *Am. J. Hum. Genet.* 32: 314-331

Brenner S., Jacob F. and Meselson M. (1960) An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* 190: 576-581

Brett D. *et al.* (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* 474: 83-86

Briggs R.C., Briggs J.A., Ozer J., Sealy L., Dworkin L.L., Kingsmore S.F., Seldin M.F., Kaur G.P., Athwal R.S. and Dessypris E.N. (1994a) The human myeloid cell nuclear differentiation antigen gene is one of at least two related interferon-inducible genes located on chromosome 1q that are expressed specifically in hematopoietic cells. *Blood* 83: 2153-2162

Briggs J.A., Burrus G.R., Stickney B.D. and Briggs R.C. (1992) Cloning and expression of the human myeloid cell nuclear differentiation antigen: regulation by interferon alpha. *J. Cell. Biochem.* 49: 82-92

Briggs R., Dworkin L., Briggs J., Dessypris E., Stein J., Stein G. and Lian J. (1994b) Interferon alpha selectively affects expression of the human myeloid cell nuclear differentiation antigen in late stage cells in the monocytic bu' not the granulocytic lineage. *J. Cell. Biochem.* 54: 198-206

Briggs R.C., Kao W.Y., Dworkin L.L., Briggs J.A., Dessypris E.N. and Clark J. (1994c) Regulation and specificity of MNDA expression in monocytes, macrophages, and leukemia/B lymphoma cell lines. *J. Cell. Biochem.* 56: 559-567

Brown J.R., Douady C.J., Italia M.J., Marshall W.E. and Stanhope M.J. (2001) Universal trees based on large combined protein sequence data sets. *Nat. Genet.* 28: 281-285

Brown A.M. and Lemke G. (1997) Multiple regulatory elements control transcription of the peripheral myelin protein zero gene. *J. Biol. Chem.* 272: 28939-28947

Brunner A.M., Schimenti J.C. and Duncan C.H. (1986) Dual evolutionary modes in the bovine globin locus. *Biochemistry* 25: 5028-5035

Buckingham M. (1994) Muscle: the regulation of myogenesis. *Curr. Opin. Genet. Dev.* 4: 745-751

Burge C. and Karlin S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268: 78-94

Burrus G.R., Briggs J.A. and Briggs R.C. (1992) Characterization of the human myeloid cell nuclear differentiation antigen: relationship to interferon-inducible proteins. *J. Cell. Biochem.* 48: 190-202

Buskin J.N. and Hauschka S.D. (1989) Identification of a myocyte nuclear factor that binds to the muscle-specific enhancer of the mouse muscle creatine kinase gene. *Mol. Cell. Biol.* 9: 2627-2640

Cambier J.C. (1995) Antigen and Fc receptor signaling. *J. Immunol.* 15: 3281-3285

Carninci P. and Hayashizaki Y. (1999) High-efficiency full-length cDNA cloning. *Methods Enzymol.* 303: 19-44

Cartron J.P. and Colin Y. (2001) Structural and functional diversity of blood group antigens. *Transfus. Clin. Biol.* 8: 163-199

Castro A.G., Hauser T.M., Cocks B.G., Abrams J., Zurawski S., Churakova T., Zonin F., Robinson D. *et al.* (1999) Molecular and functional characterization of mouse signaling lymphocytic activation molecule (SLAM): differential expression and responsiveness in Th1 and Th2 cells. *J. Immunol.* 163: 5860-5870

Chargaff E., Visher E., Doniger R., Green C. and Misani F. (1949) The composition of the deoxypentose nucleic acids of thymus and spleen. *J. Biol. Chem.* 177: 405-416

Chatterjee P.K. and Coren J.S. (1997) Isolating large nested deletions in bacterial and P1 artificial chromosomes by *in vivo* P1 packaging of products of Cre-catalyzed recombination between the endogenous and a transposed *loxP* site. *Nucleic Acids Res.* 25: 2205-2212

Chebath J., Benech P., Revel M. and Vigneron M. (1987) Constitutive expression of (2'-5') oligo A synthetase confers resistance to picornavirus infection. *Nature* 330: 587-588

Chellappan S.P., Hiebert S., Mudryj M., Horowitz J.M. and Nevins J.R. (1991) The E2F transcription factor is a cellular target for the Rb protein. *Cell* 65: 1053-1061

Chissoe S.L., Wang Y.F., Clifton S.W., Ma N., Sun H.J., Lobsinger J.S., Kenton S.M., White J.D. and Roe B.A. (1991) Strategies for rapid and accurate DNA sequencing. *Methods: Companion Methods Enzymol.* 3: 55-65

Chissoe S. *et al.* (1995) Sequence and analysis of the human ABL gene, the BCR gene, and regions involved in the Philadelphia chromosomal translocation. *Genomics* 27: 67-82

Choi H.-S., Chung M., Tzameli I., Simha D., Lee Y.-K., Seol W. and Moore D.D. (1997) Differential transcativation by two isoforms of the orphan nuclear hormone receptor CAR. *J. Biol. Chem.* 272: 23565-23571

Choubey D. and Gutterman J.U. (1996) The interferon-inducible growth inhibitory p202 protein: DNA binding properties and identification of a DNA binding domain. *Biochem. Biophys. Res. Commun.* 221: 396-401

Choubey D. and Lengyel P. (1992) Interferon action: nucleolar and nucleoplasmic localization of the interferon-inducible 72-kD protein that is encoded by the Ifi204 gene from the gene 200 cluster. *J. Cell Biol.* 116: 1333-1341

Choubey D. and Lengyel P. (1993) Interferon action: cytoplasmic and nuclear localization of the interferon-inducible 52 kD protein that is encoded by the Ifi202 gene from the gene 200 cluster. *J. Interferon Res.* 13: 43-52

Choubey D. and Lengyel P. (1995) Binding of an interferon-inducible protein (p202) to the retinoblastoma protein. *J. Biol. Chem.* 11: 6134-6140

Choubey D., Shyr-Jiann L., Datta B., Gutterman J.U. and Lengyel P. (1996) Inhibition of E2F-mediated transcription by p202. *EMBO J* 15: 5668-5678

Choubey D., Snoddy J., Chaturvedi V., Toniato E., Opdenakker G., Thakur A., Samanta H., Engel D.E. and Lengyel P. (1989) Interferons as gene activators. Indications for repeated gene duplication during the evolution of a cluster of interferon-activatable genes on murine chromosome 1. *J. Biol. Chem.* 264: 17182-17189

Choubey D., Walter S., Geng Y. and Xin H. (2000) Cytoplasmic localization of the interferon-inducible protein that is encoded by the AIM2 (absent in melanoma) gene from the 200-gene family. *FEBS Lett.* 474: 38-42

Chow L.T., Gelinas R.E., Broker T.R. and Roberts R.J. (1977) An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* 12: 1-8

Chu W.M., Ballard R., Carpick B.W. and Schmid C.W. (1998) Potential Alu function: regulation of the activity of double-stranded Rna-activated PKR. *Mol. Cell. Biol.* 18: 58-68

Ciccarone V.C., Chrivia J., Hardy K.J. and Young H.A. (1990) Identification of enhancer-like elements in human IFN-gamma genomic DNA. *J. Immunol.* 144: 725-730

Cocks B.G., Chang C.-C. J., Carballido J.M., Yssel H., de Vries J.E. and Aversa G. (1995) A novel receptor involved in T-cell activation. *Nature* 376: 260

Coffey A.J., Brooksbank R.A., Brandau O., Oohashi T., Howell G.R., Bye J.M., Cahn A.P., Durham J., Heath P., Wray P. *et al.* (1998) Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2 domain-encoding gene. *Nat. Genet.* 20: 129

Craxton M. (1991) Linear amplification sequencing: a powerful method for sequencing DNA. *Methods: Companion Methods Enzymol.* 3: 20-26

Crick F.H.C., Burnett L., Brenner S. and Watts-Tobin R.J. (1961) General nature of the genetic code for proteins. *Nature* 192: 1227-1232

Cummings S.R., Kelsey J.L., Nevitt M.C. and O'Dowd K.J. (1985) Epidemiology of osteoporosis and osteoporotic fracture. *Epidemiol. Rev.* 7: 178-208

Dacks J.B. and Doolittle W.F. (2001) Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help. *Cell* 107: 419-425

Dailey H.A. (1990) *in* Biosynthesis of Heme and Chlorophylls. (Dailey H.A. Ed.), pp. 123-161, McGraw-Hill, New York

Dailey T.A., Dailey H.A., Meissner P. and Prasad A.R.K. (1995) Cloning, sequence and expression of mouse protoporphyrinogen oxidase. *Arch. Biochem. Biophys.* 324: 379-384

Dailey H.A. and Karr S.W. (1987) Purification and characterization of murine protoporphyrinogen oxidase. *Biochemistry* 26: 2697-2701

Daniels D.R. and Deininger P.L. (1985) Repeat sequence families derived from mammalian tRNA genes. *Nature* 317: 819-822

Datta B., Li B., Choubey D., Nallur G. and Lengyel P. (1996) p202, an interferon-inducible modulator of transcription, inhibits transcriptional activation by the p53 tumor suppressor protein, and a segment from the p53-binding protein 1 that binds to p202 overcomes this inhibition. *J. Biol. Chem.* 271: 27544-27555

Datta B., Min W., Burma S. and Lengyel P. (1998) Increase in p202 expression during skeletal muscle differentiation: inhibition of MyoD protein expression and activity by p202. *Mol. Cell. Biol.* 18: 1074-1083

Davis S.J. and van der Merwe P.A. (1996) The structure and ligand interactions of CD2: implications for T-cell function. *Immunol. Today* 17: 177

Dawson M.J., Elwood N.J., Johnstone R.W. and Trapani J.A. (1998) The IFN-inducible nucleoprotein IFI 16 is expressed in cells of the monocyte lineage, but is rapidly and markedly down-regulated in other myeloid precursor populations. *J. Leuk. Biol.* 64: 546-554

Dawson M.J. and Trapani J.A. (1995) IFI16 gene encodes a nuclear protein whose expression is induced by interferons in human myeloid leukaemia cell lines. *J. Cell. Biochem.* 57: 39-51

Dawson M.J., Trapani J.A., Briggs R.C., Nicholl J.K., Sutherland G.R. and Baker E. (1995) The closely linked genes encoding the myeloid nuclear differentiation antigen (MNDA) and IFI16 exhibit contrasting haematopoietic expression. *Immunogenetics* 41: 40-43

Deeb S.S., Takata K., Peng R., Kajiyama G. and Albers J.J. (1990) A splice-junction mutation responsible for familial apolipoprotein AII deficiency. *Am. J. Hum. Genet.* 46: 822-827

DeGregori J., Kowalik T. and Nevins J.R. (1995) Cellular targets for activation by the E2F-1 transcription factor include DNA synthesis- and G_1/S-regulatory genes. *Mol. Cell. Biol.* 15: 4215-4224

de Haas C.J. (1999) New insights into the role of serum amyloid P-component, a novel lipopolysaccharide-binding protein. *F.E.M.S. Immunol. Med. Microbiol.* 26: 197-202

Dehal P., Predki P., Olsen A.S., Kobayashi A., Folta P. *et al.* (2001) Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* 293: 104-111

De la Fuente M.A., Pizcueta P., Nadal M., Bosch J. and Engel P. (1997) CD84 leukocyte antigen is a new member of the Ig superfamily. *Blood* 90: 2398-2405

De la Fuente M.A., Tovar V., Pizcueta P., Nadal M., Bosch J. and Engel P. (1999) Molecular cloning, characterization, and chromosomal localization of the mouse homologue of CD84, a member of the CD2 family of cell surface molecules. *Immunogenetics* 49: 249-255

De la Vega L.A. and Stockert R.J. (1999) The cytoplasmic coatomer protein COPI. A potential translational regulator. *J. Biol. Chem.* 274: 31135-31138

Deng H.W., Li J., Li J.L., Johnson M., Gong G. *et al.* (1998) Change of bone mass in postmenopausal Caucasian women with and without hormone replacement therapy is associated with vitamin D receptor and estrogen receptor phenotypes. *Hum. Genet.* 103: 576-585

Deybach J.C., de Verneuil H. and Nordman Y. (1981) The inherited enzymatic defect in porphyria variegata. *Hum. Genet* 58: 425-428

DeYoung K.L., Ray M.E., Su Y.A., Anzick S.L., Johnstone R.W., Trapani J.A., Meltzer P.S. and Trent J.M. (1997) Cloning a novel member of the human interferon-inducible gene family associated with control of tumorigenicity in a model of human melanoma. *Oncogene* 15: 453-457

Ding Y. and Brunden K.R. (1994) The cytoplasmic domain of myelin glycoprotein P0 interacts with negatively charged phospholipids bilayers. *J. Biol. Chem.* 269: 10764-10770

Dohi Y., Iki M., Ohgushi H., Gojo S., Tabata S., Kajita E., Nishino H. and Yonemasu K. (1998) A novel polymorphism in the promoter region for the human osteocalcin gene: the possibility of a correlation with bone mineral density in postmenopausal Japanese women. *J. Bone Miner. Res.* 13: 1633-1639

Dong S., Zhu J., Reid A., Strutt P., Guidez F., Zhong H.-J., Wang Z.-Y., Licht J., Waxman S., Chomienne C. *et al.* (1996) Amino-terminal protein-protein interaction motif (POZ-domain) is responsible for activities of the promyelocytic leukemia zinc finger-retinoic acid receptor-alpha fusion protein. *Proc. Nat. Acad. Sci.* 93: 3624-3629

Doolittle W.F. (1999) Phylogenetic classification and the universal tree. *Science* 284: 2124-2128

Doudney K., Murdoch J.N., Paternotte C., Bentley L., Gregory S., Copp A.J. and Stanier P. (2001) Comparative physical and transcript maps of approximately 1 Mb around loop-tail, a gene for severe neural tube defects on distal mouse chromosome 1 and human chromosome 1q22-23. *Genomics* 72: 180-192

D'Urso D., Brophy P.J., Staugaitis S.M., Gillespie C.S., Frey A.B., Stempak J.G. and Colman D.R. (1990) Protein zero of peripheral nerve myelin: biosynthesis, membrane insertion, and evidence for homotypic interaction. *Neuron* 2: 449-460

D'Souza S., Xin H., Walter S. and Choubey D. (2001) The gene encoding p202, an interferon-inducible negative regulator of the p53 tumor suppressor, is a target of p53-mediated transcriptional repression. *J. Biol. Chem.* 276: 298-305

Dyson N., Howley P.M., Munger K. and Harlow E. (1989) The human papilloma virus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product. *Science* 243: 934-937

East J.M. (2000) Sarco(endo)plasmic reticulum calcium pumps: recent advances in our understanding of structure/function and biology (Review). *Mol. Membr. Biol.* 17: 189-200

Eddy S.R. (1995) Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3:114-20.

Enari M., Sakahira H., Yokoyama H., Okawa K., Iwamatsu A. and Nagata S. (1998) A caspase-activated DNAse that degrades DNA during apoptosis, and its inhibitor ICAD. *Nature* 391: 43-50

Ellsworth R.E., Jamison D.C., Touchman J.W., Chissoe S.L., Braden V. V. *et al.* (2000) Comparative genomic sequence analysis of the human and mouse cystic fibrosis transmembrane conductance regulator genes. *Proc. Nat. Acad. Sci.* 97: 1172-1177

Engel D.A., Snoddy J., Toniato E. and Lengyel P. (1988) Interferons as gene activators: close linkage of two interferon-activatable murine genes. Virology 166: 24-29

Estelles A., Charlton C.A. and Blau H.M. (1999) The phosphoprotein protein PEA-15 inhibits Fas- but increases TNF-R1-mediated caspase-8 activity and apoptosis. *Dev. Biol.* 216: 16-28

Fagan R., Swindells M., Overington J. and Weir M. (2001) Nicastrin, a presenilin-interacting protein, contains an aminopeptidase/transferrin receptor superfamily domain. *Trends Biochem. Sci.* 26: 213-214

Fakharzadeh S.S., Trusko S.P. and George D. (1991) Tumorigenic potential associated with enhanced expression of a gene that is amplified in a mouse tumor cell line. *EMBO J* 10: 1566-1569

Farmer G., Bargonetti J., Zhu H., Friedman P., Prywes R. and Prives C. (1992) Wild-type p53 activates transcription in vitro. *Nature* 358: 83-86

Ferre-D'Amare A.R., Prendergast G.C., Ziff E.B. and Burley S.K. (1993) Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. *Nature* 363: 38-46

Fickett J.W. and Hatzigeorgiou A.G. (1997) Eukaryotic promoter recognition. *Genome Res.* 7: 861-878

Field K.G., Olsen G.J., Lane D.J., Giovannoni S.J., Ghiselin M.T., Raff E.C., Pace N.R. and Raff R.A. (1988) Molecular phylogeny of the animal kingdom. *Science* 239: 748-753

Fiers W. *et al.* (1981) Complete nucleotide sequence of SV40 DNA. *Nature* 273: 113-120

Filbin M.T. and Tennekoon G.I. (1991) The role of complex carbohydrates in adhesion of the myelin protein, P0. *Neuron* 7: 845-855

Fitch W. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* 19: 99-113

Flannery C.R., Lark M.W. and Sandy J.D. (1992) Identification of a stromelysin cleavage site within the interglobular domain of human aggrecan. Evidence for proteolysis at this site *in vivo* in human articular cartilage. *J. Biol. Chem.* 267: 1008-1014

Flati V., Frati L., Gulino A., Martinotti S. and Toniato E. (2001) The murine p202 protein, an IFN-inducible modulator of transcription, is activated by the mitogen platelet-derived growth factor. *J. Interferon Cytokine Res.* 21: 99-103

Footz T.K., Brinkman-Mills P., Banting G., Maier S.A., Ali Riazi M. *et al.* (2001) Analysis of the Cat Eye Syndrome critical region in humans and the region of conserved synteny in mice: a search for candidate genes at or near the human chromosome 22 pericentromere. *Genome Res.* 11: 1053-1070

Forman B.M., Tzameli I., Choi H.S., Chen J., Simha D., Seol W., Evans R.M. and Moore D.D. (1998) Androstane metabolites bind to and deactivate the nuclear receptor CAR-beta. *Nature* 395: 543-544

Forus A., Berner J.M., Meza-Zepeda L.A., Saeter G., Mischke D., Fodstad O. and Myklebost O. (1998) Molecular characterization of a novel amplicon at 1q21-q22 frequently observed in human sarcomas. *Br. J. Cancer* 78: 495-503

Fuchs T., Glusman G., Horn-Saban S., Lancet D. and Pilpel Y. (2001) The human olfactory subgenome: from sequence to structure to evolution. *Hum. Genet.* 108: 1-13

Galante Y.M. and Hatefi Y. (1979) Purification and molecular and enzymic properties of mitochondrial NADH dehydrogenase. *Arch. Biochem. Biophys.* 192: 559-568

Gall J.G. (1981) Chromosome structure and the C-value paradox. *J. Cell Biol.* 91: 3s-14s

Galli S.J. (1993) New concepts about the mast cell. *N. Engl. J. Med.* 328: 257-265

Gariglio M., De Andrea M., Lembo M., Ravotto M., Zappador C., Valente G. and Landolfo S. (1998) The murine homolog of the HIN 200 family, Ifi204, is constitutively expressed in myeloid cells and selectively induced in the monocyte/macrophage lineage. *J. Leuk. Biol.* 64: 608-614

Geng Y., D'Souza S., Xin H., Walter S. and Choubey D. (2000) p202 levels are negatively regulated by serum growth factors. *Cell Growth Diff.* 11: 475-483

Giese K.P., Martini R., Lemke G., Soriano P. and Schachner M. (1992) Mouse P0 gene disruption leads to hypomyelination, abnormal expression of recognition molecules, and degeneration of myelin and axons. *Cell* 71: 565-576

Giguere V., Tini M., Flock G., Ong E., Evans R.M. and Otulakowski G. (1994) Isoform-specific amino-terminal domains dictate DNA-binding properties of ROR alpha, a novel family of orphan hormone nuclear receptors. *Genes Dev.* 8: 538-553

Gilbert G., Marchionni M. and McKnight G. (1986) On the antiquity of introns. *Cell* 46: 151-154

Gonzalez F.J. (1989) The molecular biology of cytochrome P450s. *Pharmacol. Rev.* 40: 243-288

Gonzales-Cabrero J., Wise C.J., Latchman Y., Freeman G.J., Sharpe A.H. and Reiser H. (1999) CD48-deficient mice have a pronounced defect in $CD4^+$ T cell activation. *Proc. Nat. Acad. Sci.* 96: 1019-1023

Gregory T.R. and Hebert P.D. (1999) The modulation of DNA content: proximate causes and ultimate consequences. *Genome Res.* 9: 317-324

Gribaudo G., Ravaglia S., Guandalini L., Rira L., Gariglio M. and Landolfo S. (1997) Molecular cloning and expression of an interferon-inducible protein encoded by gene 203 from the gene 200 cluster. *Eur. J. Biochem.* 249: 258-264

Gribaudo G., Riera L., Hertel L. and Landolfo S. (1999) *In vitro* and *in vivo* expression analysis of the interferon-inducible 203 gene. *J. Interferon Cyt. Res.* 19: 129-136

Grigorieff N. (1998) Three-dimensional structure of bovine NADH:ubiquinone oxidoreductase (complex I) at 22 A in ice. *J. Mol. Biol.* 277: 1033-1046

Gutterman J.U. and Choubey D. (1999) Retardation of cell proliferation after expression of p202 accompanies an increase in $p21^{WAF1/CIP1}$. *Cell Growth Diff.* 10: 93-100

Hall B.D. and Spiegelman S. (1964) Sequence complementarity of T_2-DNA and T_2-specific RNA. *Proc. Nat. Acad. Sci.* 47: 137-146

Halle J.P., Stelzer G., Goppelt A. and Meisteremst M. (1995) Activation of traanscription by recombinant upstream stimulatory factor 1 is mediated by a novel positive cofactor. *J. Biol. Chem.* 270: 21307-21311

Hanson R.W. and Reshef L. (1997) Regulation of phosphoenolpyruvate carboxykinase (GTP) gene expression. *Annu. Rev. Biochem.* 66: 581-611

Harada H., Taniguchi T. and Tanaka N. (1998) The role of interferon regulatory factors in the interferon system and cell growth control. *Biochimie* 80: 641-650

Hardingham T.E., Fosang A.J. and Dudhia J. (1992) *in* Articular Cartilage and Osteoarthritis (Kuettner, K.E., Schleyerbach R., Peyton J.G. and Hascall V.C., eds) pp. 5-20, Raven Press, New York

Hardison R. and Miller W. (1993) Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. *Mol. Biol. Evol.* 10: 73-102

Harper J.W., Adami G.R., Wei N., Keyomarsi K. and Elledge S.J. (1993) The p21 Cdk-interacting protein Cip1 is a potent inhibitor of G1 cyclin-dependent kinases. *Cell* 75: 805-816

Harris M., Nguyen T.V., Howard G.M., Kelly P.J. and Eisman J.A. (1998) Genetic and environmental correlations between bone formation and bone mineral density: a twin study. *Bone* 22: 141-145

Harrison P.M., Hegyi H., Balasubramanian S., Luscombe N.M., Bertone P., Echols N., Johnson T. and Gerstein M. (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* 12: 272-280

Hasayaka K., Himoro M., Sato W., Takada G., Uyemara K., Shimizu N., Bird T.D., Conneally P.M. and Chance P.F. (1995) Charcot-Marie-Tooth neuropathy type 1B is associated with mutations of the myelin P0 gene. *Nat. Genet.* 11: 119-120

Hatefi Y. (1985) The mitochondrial electron transport and oxidative phosphorylation system. *Annu. Rev. Biochem.* 54: 1015-1069

Henriksson M. and Luscher B. (1996) Proteins of the Myc network: essential regulators of cell growth and differentiation. *Adv. Cancer Res.* 68: 109-182

Hershey A.D. and Chase M. (1952) Independent functions of viral proteins and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* 36: 39-56

Hertel L., De Andrea M., Azzimonti B., Rolle A., Gariglio M. and Landolfo S. (1999) The interferon-inducible 204 gene, a member of the Ifi200 family, is not involved in the

antiviral state induction by IFN-alpha, but is required by the mouse cytomegalovirus for its replication. *Virology* 262: 1-8

Hertel L., Rolle S., De Andrea M., Azzimonti B., Osello R., Gribaudo G., Gariglio M. and Landolfo S. (2000) The retinoblastoma protein is an essential mediator that links the interferon-inducible 204 gene to cell-cycle regulation. *Oncogene* 19: 3598-3608

Hibner B.L., Burke W.D. and Eickbush T.H. (1991) Sequence identity in an early chorion multigene family is the result of localized gene conversion. *Genetics* 128: 595-606

Hirawake H., Taniwaki M., Tamura A., Amino H., Tomitsuka E. and Kita K. (1999) Characterization of the human SDHD gene encoding the small subunit of cytochrome b (cybS) in mitochondrial succinate-ubiquinone oxidoreductase. *Biochim. Biophys. Acta* 1412: 295-300

Hirawake H., Taniwaki M., Tamura A., Kojima S. and Kita K. (1997) Cytochrome b in human complex II (succinate-ubiquinone oxidoreductase): cDNA cloning of the components in liver mitochondria and chromosome assignment of the genes for the large (SDHC) and small (SDHD) subunits to 1q21 and 11q23. *Cytogenet. Cell Genet.* 79: 132-138

Hochstrasser M. (1996) Ubiquitin-dependent protein degradation. *Annu. Rev. Genet.* 30: 405-439

Hoffman E.C., Reyes H., Chu F.F., Sander F., Conley L.H., Brooks B.A. and Hankinson O. (1991) Cloning of a factor required for activity of the Ah (dioxin) receptor. *Science* 252: 954-958

Hollenberg S.M., Sternglanz R., Cheng P.F. and Weintraub H. (1995) Identification of a new family of tissue-specific basic helix-loop-helix proteins with a two-hybrid system. *Mol. Cell. Biol.* 15: 3813-3822

Holley R.W., Apgar J., Everett G.A., Madision J.T., Marquisee M., Merrell S.H., Penswick J.R. and Zamir A. (1965) Structure of a ribonucleic acid. *Science* 147: 1462-1465

Honkakoski P., Moore R., Washburn K.A. and Negishi M. (1998) Activation by diverse xenochemicals of the 51-base pair phenobarbital-responsive enhancer module in the CYP2B10 gene. *Mol. Pharmacol.* 53: 597-601

Horowitz J.M., Yandell D.W., Park S.H., Canning S., Whyte P., Buchkovich K., Harlow E., Weinberg R.A. and Dryja T.P. (1989) Point mutational inactivation of the retinoblastoma antioncogene. *Science* 243: 937-940

Hovanessian A.G. (1991) Interferon-induced and double-stranded RNA-activated enzymes: a specific protein kinase and 2', 5'-oligoadenylate synthetase. *J. Interferon Res.* 11: 199-205

Hurskainen T.L., Hirohata S., Seldin M.F. and Apte S.S. (1999) ADAM-TS5, ADAM-TS6 and ADAM-TS7, novel members of a new family of zinc metalloproteases. *J. Biol. Chem.* 274: 25555-25563

Ina Y. (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* 40: 190-226

Inaoka T., Bilbe G., Ishibashi O., Tezuka K., Kumegawa M. and Kokubo T. (1995) Molecular cloning of human cDNA for cathepsin K: novel cysteine proteinase predominantly expressed in bone. *Biochem. Biophys. Res. Commun.* 206: 89-96

Innis M.A., Myambo K.B., Gelfand D.H. and Brown M.A. (1988) DNA sequencing with *Thermus aquaticus* DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. *Proc. Nat. Acad. Sci.* 85: 9436-9440

Irniger S., Egli C.M. and Braus G.H. (1991) Different classes of polyadenylation sites in the yeast *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 11: 3060-3069

Isaacs A. and Lindenmann J. (1957) Virus interference. I. The interferon. *Proc. R. Soc. Lond. B. Biol. Sci.* 147: 258-267

Jacob F. and Monod J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3: 318-356

James G.L., Goldstein J.L., Pathak R.K., Anderson R.G. and Brown M.S. (1994) PxF, a prenylated protein of peroxisomes. *J. Biol. Chem.* 269: 14182-14190

Jenkins M.K. (1994) The ups and downs of T cell costimulation. *Immunity* 1: 443

Johnson M.R., Polymeropoulos M.H., Vos H.L., Ortiz de Luna R.I. and Francomano C.A. (1996) A nonsense mutation in the cathepsin K gene observed in a family with pycnodysostosis. *Genome Res.* 6: 1050-1055

Johnstone R.W., Kerry J.A. and Trapani J.A. (1998) The human interferon-inducible protein, IFI 16, is a repressor of transcription. *J. Biol. Chem.* 273: 17172-17177

Johnstone R.W., Kershaw M.H. and Trapani J.A. (1998) Isotypic variants of the interferon-inducible transcriptional repressor IFI 16 arise through differential mRNA splicing. *Biochemistry* 37: 11924-11931

Johnstone R.W. and Trapani J.A. (1999) Transcription and growth regulatory functions of the HIN-200 family of proteins. *Mol. Cell. Biol.* 19: 5833-5838

Johnstone R.W., Wei W., Greenway A. and Trapani J.A. (2000) Functional interaction between p53 and the interferon-inducible nucleoprotein IFI 16. *Oncogene* 19: 6033-6042

Kao W.Y., Dworkin L.L., Briggs J.A. and Briggs R.C. (1996) Characterization of the human myeloid cell nuclear differentiation antigen gene promoter. *Biochim. Biophys. Acta* 1308: 201-204

Kao W.Y., Briggs J.A., Kinney M.C., Jensen R.A. and Briggs R.C. (1997) Structure and function analysis of the human myeloid cell nuclear differentiation antigen promoter: evidence for the role of Sp1 and not of c-Myb or PU.1 in myelomonocytic lineage-specific expression. *J. Cell. Biochem.* 65: 231-244

Karlsson S. and Nienhuis A.W. (1985) Developmental regulation of human globin genes. *Annu. Rev. Biochem.* 54: 1071-1108

Kashiwagi M., Tortorella M., Nagase H. and Brew K. (2001) TIMP-3 is a potent inhibitor of aggrecanase 1 (ADAM-TS4) and aggrecanase 2 (ADAM-TS5). *J. Biol. Chem.* 276: 12501-12504

Kato K., Koyanagi M., Okada H., Takanashi T., Wong Y.-W., Williams A.F., Okumura K. and Yagita H. (1992) CD48 is a counter-receptor for mouse CD2 and is involved in T cell activation. *J. Exp. Med.* 176: 1241-1249

Kawamoto T., Sueyoshi T., Zelko I., Moore R., Washburn K. and Negishi M. (1999) Phenobarbital-responsive nuclear translocation of the receptor CAR in induction of the CYP2B gene. *Mol. Cell. Biol.* 19: 6318-6322

Kerhl J.H. (1995) Hematopoeitic lineage commitment: role of transcription factors. *Stem Cells* 13: 223-241

Kingsmore S.F., Snoddy J., Choubey D., Lengyel P. and Seldin M.F. (1989) Physical mapping of a family of interferon-activated genes, serum amyloid P-component, and alpha-spectrin on mouse chromosome 1. *Immunogenetics* 30: 169-174

Klenow H., Overgaard-Hansen K. and Patkar S.A. (1971) Proteolytic cleavage of native DNA polymerase into two different catalytic fragments. Influence of assay conditions on the change of exonuclease activity and polymerase activity accompanying cleavage. *Eur. J. Biochem.* 22: 371-381

Kobayashi M., Goda M. and Shimizu S. (1998) Nitrilase catalyzes amide hydrolysis as well as nitrile hydrolysis. *Biochem. Biophys. Res. Commun.* 253: 662-666

Koller D.L., Econs M.J., Morin P.A., Christian J.C., Hui S.L., Parry P., Curran M.E., Rodriguez L.A., Conneally P.M., Joslyn G., Peacock M., Johnston C.C. and Foroud T. (2000) Genome screen for QTLs contributing to normal variation in bone mineral density and osteoporosis. *J. Clin. Endocrinol. Metab.* 85: 3116-3120

Komiya T., Tanigawa Y. and Hirohashi S. (1998) Cloning of the novel gene intelectin, which is expressed in intestinal paneth cells in mice. *Biochem. Biophys. Res. Commun.* 251: 759-762

Koul D., Lapushin R., Xu H.J., Mills G.B., Gutterman J.U. and Choubey D. (1998) p202 prevents apoptosis in murine AKR-2B fibroblasts. *Biochem. Biophys. Res. Commun.* 247: 379-382

Kubin M.Z., Parshley D.L., Din W., Waugh J.Y., Davis-Smith T., Smith C.A., Macduff B.M., Armitage R.J., Chin W., Cassiano L., Borges L., Petersen M., Trinchieri G. and Goodwin R.G. (1999) Molecular cloning and biological characterization of NK cell activation-inducing ligand, a counterstructure for CD48. *Eur. J. Immunol.* 29: 3466-3477

Kleene R. and Berger E.G. (1993) The molecular and cell biology of glycosyltransferases. *Biochim. Biophys. Acta* 1154: 283-325

Kulkens T., Bolhuis P.A., Wolterman R.A., Kemp S., te Nijenhuis S., Valentijn L.J., Hensels G.W., Jennekens F.G., de Visser M., Hoogendijk J.E. *et al.* (1993) deletion of the serine 34 codon from the major peripheral myelin protein zero gene in Charcot-Marie-Tooth disease type 1B. *Nat. Genet.* 5: 35-39

Kusuda J., Hirai M., Tanuma R., Hirata M. and Hashimoto K. (1999) Genomic structure and chromosome location of RPL27A/Rpl27a, the genes encoding human and mouse ribosomal protein L27A. *Cytogenet. Cell Genet.* 85: 248-251

Lai C.H., Chou C.Y., Ch'ang L.Y., Liu C.S. and Lin W. (2000) Identification of novel human genes evolutionarily conserved in *Caenorhabditis elegans* by comparative proteomics. *Genome Res.* 10: 703-713

Lander E. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921

Lane R.P., Cutforth T., Young J., Athanasiou M., Friedman C., Rowen L., Evans G., Axel R., Hood L. and Trask B.J. (2001) Genomic analysis of orthologous mouse and human olfactory receptor loci. *Proc. Nat. Acad. Sci.* 98: 7390-7395

Langdahl B.L., Knudsen J.Y., Jensen H.K., Gregersen N. and Eriksen E.F. (1997) A sequence variation: 713-8delC in the transforming growth factor-beta 1 gene has higher prevalence in osteoporotic women than in normal women and is associated with very

low bone mass in osteoporotic women and increased bone turnover in both osteoporotic and normal women. *Bone* 20: 289-294

Le Coniat M., Kinet J.-P. and Berger R. (1990) The human genes for the alpha and gamma subunits of the mast cell receptor for immunoglobulin E are located on human chromosome band 1q23. *Immunogenetics* 32: 183-186

Lembo D., Angeretti A., Benefazio S., Hertel L., Gariglio M., Novelli F. and Landolfo S. (1995) Constitutive expression of the interferon-inducible protein p202 in NIH 3T3 cells affects cell cycle progression. *J. Biol. Regul. Homeost. Agents* 9: 42-46

Lembo M., Sacchi C., Zappador C., Bellomo G., Gaboli M., Pandolfi P.P., Gariglio M. and Landolfo S. (1998) Inhibition of cell proliferation by the interferon-inducible 204 gene, a member of the Ifi 200 cluster. *Oncogene* 16: 1543-1551

Lemke G. (1988) Unwrapping the genes of myelin. *Neuron* 1: 535-543

Lemke G., Lamar E. and Patterson J. (1988) Isolation and analysis of the gene encoding peripheral myelin protein zero. *Neuron* 1: 73-83

Lengyel P. (1982) Biochemistry of interferons and their actions. *Annu. Rev. Biochem.* 51: 251-282

Lengyel P., Choubey D., Li S.-J. and Datta B. (1995) The interferon-activatable gene 200 cluster: from structure toward function. *Sem. Virol.* 6: 203-213

Li W.H. and Graur D. (1991) Fundamentals of molecular evolution. Sinauer, Sunderland, Mass.

Li W.H., Gu Z., Wang H. and Nekrutenko A. (2001) Evolutionary analyses of the human genome. *Nature* 409: 847-849

Li T., Spearow J., Rubin C.M. and Schmid C.W. (1999) Physiological stresses increase mouse short interespersed element (SINE) RNA expression *in vivo*. *Gene* 239: 367-372

Liang Q., He J.S. and Fulco A.J. (1995) The role of Barbie box sequences as cis-acting elements involved in the barbiturate-mediated induction of cytochromes P450BM-1 and P450BM-3 in *Bacillus megaterium*. *J. Biol. Chem.* 270: 4438-4450

Liao D. (1999) Concerted evolution: molecular mechanism and biological implications. *Am. J. Hum. Genet.* 64: 24-30

Liu C.-J., Wang H. and Lengyel P. (1999) The interferon-inducible nucleolar p204 protein binds the ribosomal RNA-specific UBF1 transcription factor and inhibits ribosomal RNA transcription. *EMBO J* 10: 2845-2854

Liu C.-J., Wang H., Zhao Z., Yu S., Lu Y.-B., Meyer J., Chatterjee G., Deschamps S., Roe B.A. and Lengyel P. (2000) MyoD-dependent induction during myoblast differentiation of p204, a protein also inducible by interferon. *Mol. Cell. Biol.* 20: 7024-7036

Liu W.M., Chu W.M., Choudary P.V. and Schmid C.W. (1995) Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res.* 23: 1758-1765

Liu Y. and Janeway C.A. Jr. (1990) Interferon gamma plays a critical role in induced cell death of effector T cell: a possible third mechanism of self-tolerance. *J. Exp. Med.* 172: 1735-1739

Loeffen J., Elpeleg O., Smeitink J., Smeets R., Stockler-Ipsiroglu S., Mandel H., Sengers R., Trijbels F. and van den Heuvel L. (2001) Mutations in the complex I NDUFS2 gene of patients with cardiomyopathy and encephalomyopathy. *Ann. Neurol.* 49: 195-201

Loeffen J., van den Heuvel L., Smeets R., Triepels R., Sengers R., Trijbels F. and Smeitink J. (1998) cDNA sequence and chromosomal localization of the remaining three human nuclear encoded iron sulphur protein (IP) subunits of complex I: the human IP fraction is completed. *Biochem. Biophys. Res. Commun.* 247: 751-758

Lonberg N. and Gilbert W. (1985) Intron/exon structure of the chicken pyruvate kinase gene. *Cell* 40: 81-90

Lotti M., Stoffler-Meilicke M. and Stoffler G. (1987) Localization of ribosomal protein L27 at the peptidyl transferase centre of the 50S subunit, as determined by immuno-electron microscopy. *Mol. Gen. Genet.* 210: 498-503

Mallon A.-M., Platzer M., Bate R., Gloeckner G., Botcherby M.R.M. *et al.* (2000) Comparative genome sequence analysis of the *Bpa/Str* region in mouse and man. *Genome Res.* 10: 758-775

Maraia R.J. (1991) The subset of mouse B1 (Alu-equivalent) sequences expressed as small processed cytoplasmic transcripts. *Nucleic Acids Res.* 19: 5695-5702

Martinon F., Hofmann K. and Tschopp J. (2001) The pyrin domain: a possible member of the death domain-fold family implicated in apoptosis and inflammation. *Curr. Biol.* 10: R118-R120

Martin-Padura I., Lostaglio S., Schneemann M., Williams L., Romano M., Fruscella P., Panzeri C., Stoppacciaro A., Ruco L., Villa A., Simmons D. and Dejana E. (1998) Junctional adhesion molecule, a novel member of the immunoglobulin superfamily that distributes at intercellular junctions and modulates monocyte transmigration. *J. Cell Biol.* 142: 117-127

Matringe M., Camadro J.M., Block M.A., Joyard J., Scalla R., Labbe P. and Douce R. (1992) Localization within chloroplasts of protoporphyrinogen oxidase, the target enzyme for diphenylether-like herbicides. *J. Biol. Chem.* 267: 4646-4651

Matsuo Y. (2000) Molecular evolution of the histone 3 multigene family in the Drosophila melanogaster species subgroup. *Mol. Phylogenet. Evol.* 16: 339-343

Maxam A.M. and Gilbert W. (1977) A new method for sequencing DNA. *Proc. Nat. Acad. Sci.* 74: 560-564

Maxson R., Cohn R. and Kedes L. (1983) Expression and organization of histone genes. *Annu. Rev. Genet.* 17: 239-277

McBride L.J., Koepf S.M., Gibbs R.A., Salser W., Mayrand P.E., Hunkapiller M.W. and Kronick M.N. (1989) Automated DNA-sequencing methods involving polymerase chain reaction. *Clin. Chem.* 35: 2196-2201

McPherson P. *et al.* (2001) A physical map of the human genome. *Nature* 409: 934-941

Martin M., Romero X., de la Fuente M.A., Tovar V., Zapater N., Esplugues E., Pizcueta P., Bosch J. and Engel P. (2001) CD84 functions as a homophilic adhesion molecule and enhances IFN-gamma secretion: adhesion is mediated by Ig-like domain 1. *J. Immunol.* 167: 3668-3676

Masi L., Becherini L., Colli E., Gennari L., Mansani R. *et al.* (1998) Polymorphisms of the calcitonin receptor gene are associated with bone mineral density in postmenopausal Italian women. *Biochem. Biophys. Res. Commun.* 248: 190-195

Mathieson B.J., Sharrow S.O., Bottomly K. anf Fowlkes B.J. (1980) Ly9, an alloantigenic marker of lymphocyte differentiation. *J. Immunol.* 125: 2127-2136

Matsumoto J. *et al.* (1999) ASC, a novel 22-kDa protein, aggregates during apoptosis of human promyelocytic leukemia HL-60 cells. *J. Biol. Chem.* 274: 33835-33838

Medema J.P., Scaffidi C., Kischkel F.C., Shevchenko A., Mann M., Krammer P.H. and Peter M.E. (1997) FLICE is activated by association with the CD95 death-inducing signaling complex (DISC) *EMBO J.* 16: 2794-2804

Meurs E., Chong K., Galabru J., Shaun N., Thomas B., Kerr I., Williams B. and Hovanessian A. (1990) Molecular cloning and characterization of the human double-stranded RNA-activated protein kinase induced by interferon. *Cell* 62: 379-390

Min W., Ghosh S. and Lengyel P. (1996) The interferon-inducible p202 protein as a modulator of transcription: inhibition of NF-kB, c-Fos, and c-Jun activities. *Mol. Cell. Biol.* 16: 359-368

Mizui Y., Yamazaki K., Kuboi Y., Sagane K. and Tanaka I. (2000) Characterization of 5'-flanking region of human aggrecanase-1 (ADAMTS4) gene. *Mol. Bio. Rep.* 27: 167-173

Morrison N.A., Qi J.C., Tokita A., Kelly P.J., Crofts L. *et al.* (1994) Prediction of bone density from vitamin D receptor alleles. *Nature* 364: 284-287

Murdoch J.N., Eddleston J., Leblond-Bourget N., Stanier P. and Copp A.J. (1999) Sequence and expression analysis of Nhlh1: a basic helix-loop-helix gene implicated in neurogenesis. *Dev. Genet.* 24: 165-177

Myers E.W. and Miller W. (1989) Approximate matching of regular expressions. *Bull. Math. Biol.* 51: 5-37

Nadeau J.H. (1989) Maps of linkage and synteny homologies between mouse and man. *Trends Genet.* 5: 82-86

Nagata S. (1997) Apoptosis by death factor. *Cell* 88: 355-365

Nakai K. and Horton P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24: 34-36

Nirenberg M.W. and Leder P. (1964) RNA code words and protein synthesis. *Science* 145: 1399-1407

Nirenberg M.W. and Matthaei J.H. (1961) The dependence of cell-free protein synthesis in *E. coli* upon naturally occuring or synthetic polyribonucleotides. *Proc. Nat. Acad. Sci.* 47: 1588-1602

Nomura T., Takizawa M., Aoki J., Arai H., Inoue E., Wakisaka N., Yoshizuka N., Imokawa G., Dohmae N., Takio K., Hattori M. and Matsuo N. (1998) Purification, cDNA cloning, and expression of the UDP-Gal:glucosylceramide beta-1,4-galactosyltransferase from rat brain. *J. Biol. Chem.* 273: 13570-13577

Oeltjen J.C., Malley T.M., Muzny D.M., Miller W., Gibbs R.A. and Belmont J.W. (1997) Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.* 7: 315-329

O'h Uigin C. and Li W.H. (1992) The molecular clock ticks regularly in muroid rodents and hamsters. *J. Mol. Evol.* 35: 377-384

Olson E.N. and Klein W.H. (1994) bHLH factors in muscle development: dead lines and comitments, what to leave in and what to leave out. *Genes Dev.* 8: 1-8

Opdenakker G., Snoddy J., Choubey D., Toniato E., Pravtcheva D.D., Seldin M.F., Ruddle F.H. and Lengyel P. (1989) Interferons as gene activators. A cluster of six interferon activatable genes is linked to the erythroid alpha-spectrin locus on murine chromosome 1. *Virology* 171: 568-578

Ota N., Hunt S.C., Nakajima T., Suzuki T., Hosoi T., Orimo H., Shirai Y. and Emi M. (1999) Linkage of interleukin 6 locus to human osteopenia by sibling pair analysis. *Hum. Genet.* 105: 253-257

Packham G. and Cleveland J.L. (1995) c-Myc and apoptosis. *Biochim. Biophys. Acta* 1242: 11-28

Palou E., Pirotto F., Sole J., Freed J.H., Peral B., Vilardell C., Vilella R., Vives J. and Gaya A. (2000) genomic characterization of CD84 reveals the existence of five isoforms differing in their cytoplasmic domains. *Tissue Antigens* 55: 118-127

Pardee A.B., Jacob F. and Monod J. (1959) *J. Mol. Biol.* 1: 173

Pawlowski K., Pio F., Chu Z.-L., Reed J.C. and Godzik A. (2001) PAAD-a new protein domain associated with apoptosis, cancer and autoimmune diseases. *Trends Biochem. Sci.* 26: 85-87

Peck S.R. and Ruley H.E. (2000) Ly108: a new member of the mouse CD2 family of cell surface proteins. *Immunogenetics* 52: 63-72

Pekarsky Y., Campiglio M., Siprashvili Z., Druck T., Sedkov Y., Tillib S., Draganescu A., Wermuth P., Rothman J.H., Huebner K., Buchberg A.M., Mazo A., Brenner C. and Croce C.M. (1998) Nitrilase and Fhit homologs are encoded as fusion proteins in Drosophila melanogaster and Caenorhabditis elegans. *Proc. Nat. Acad. Sci.* 95: 8744-8749

Petska S., Langer J.A., Zoon C.K. and Samuel C.E. (1987) Interferons and their actions. *Annu. Rev. Biochem.* 56: 727-777

Pietschmann F., Breslau N.A. and Pak C.Y.C. (1992) Reduced vertebral bone density in hypercalciuric nephrolithiasis. *J. Bone Miner. Res.* 7: 1383-1388

Pollock D.D., Eisen J.A., Doggett N.A. and Cummings M.P. (2000) A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. *Mol. Biol. Evol.* 17: 1776-1788

Polymeropoulos M.H., Ortiz de Luna R.I., Ide S.E., Torres R., Rubenstein J. and Francomano C.A. (1995) The gene for pycnodysostosis maps to human chromosome 1cen-q21. *Nat. Genet.* 10: 238-239

Prober J.M., Trainor G.L., Dam R.J., Hobbs F.W., Robertson C.W., Zagursky R.J., Cocuzza A.J., Jensen M.A. and Baumeister K. (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 238: 336-341

Purtilo D.T. (1985) Epstein-Barr virus-induced diseases in the X-linked lymphoproliferative syndrome and related disorders. *Biomed. Pharmacother.* 39: 52

Putney S.D., Herlihy W.C. and Schimmel P. (1983) A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* 302: 718-721

Quentin Y. (1992) Origin of the Alu family: a family of Alu-like monomers gave birth to the left and the right arms of the Alu elements. *Nucleic Acids Res.* 20: 3397-3401

Ra C., Jouvin M.H. and Kinet J.-P. (1989) Complete structure of the mouse mast cell receptor for IgE (Fc epsilon RI) and surface expression of chimeric receptors (rat-mouse-human) on transfected cells. *J. Biol. Chem.* 264: 15323-15327

Rapaport D., Taylor R.D., Kaser M., Langer T., Neupert W. and Nargang F.E. (2001) Structural requirements of Tom40 for assembly into preexisting TOM complexes of mitochondria. *Mol. Biol. Cell* 12: 1189-1198

Raymond M.H., Schutte B.C., Torner J.C., Burns T.L. and Willing M.C. (1999) Osteocalcin: genetic and physical mapping of the human gene BGLAP and its potential role in postmenopausal osteoporosis. *Genomics* 60: 210-217

Reed B.Y., Heller H.J., Gitomer W.L. and Pak C.Y.C. (1999) Mapping a gene defect in absorptive hypercalciuria to chromosome 1q23.3-q24. *J. Clin. Endocrinol. Metab.* 84: 3907-3913

Rehling P., Wiedemann N., Pfanner N. and Truscott K.N. (2001) The mitochondrial import machinery for preproteins. *Crit. Rev. Biochem. Mol. Biol.* 36: 291-336

Reyes H., Reisz-Porszasz S. and Hankinson O. (1992) Identification of the Ah receptor nuclear translocator protein (Arnt) as a component of the DNA binding form of the Ah receptor. *Science* 256: 1193-1195

Rice A.P., Duncan R., Hershey J.W.B. and Kerr I.M. (1985) Double-stranded RNA-dependent protein kinase and 2-5A system are both activated in interferon-treated, encephalomyocarditis virus-infected HeLa cells. *J. Virol.* 54: 894-898

Riley P., Anson-Cartwright L. and Cross J.C. (1998) The Hand1 bHLH transcription factor is essential for placentation and cardiac morphogenesis. *Nat. Genet.* 18: 271-275

Roomi M.W., Ishaque A., Khann N.R. and Eylar E.H. (1978) The P0-protein . The major glycoprotein of peripheral nerve myelin. *Biochim. Biophys. Acta.* 536: 112-121

Rosen C.J., Kurland E.S., Vereault D., Adler R.A., Rackoff P.J. *et al.* (1998) Association between serum insulin-like growth factor-I (IGF-I) and a simple sequence

repeat in IGF-I gene: implications for genetic studies of bone mineral density. *J. Clin. Endocrinol. Metab.* 83: 2286-2290

Rozzo S.J., Allard J.D., Choubey D., Vyse T.J., Izui S., Peltz G. and Kotzin B.L. (2001) Evidence for an interferon-inducible gene, Ifi202, in the susceptibility to systemic lupus. *Immunity* 15: 435-443

Rubin G.M et al. (2000) Comparative genomics of the eukaryotes. *Science* 287: 2204-2215

Sakahira H., Enari M. and Nagata S. (1998) Cleavage of CAD inhibitor in CAD activation and DNA degradation during apoptosis. *Nature* 391: 96-99

Salamov A.A. and Solovyev V.V. (2000) *Ab initio* gene finding in Drosophila genomic DNA. *Genome Res.* 10: 516-522

Salonpaa P., Kottari S., Pelkonen O. and Raunio H. (1997) Regulation of CYP 2 A 5 induction by porphyrinogenic agents in mouse primary hepatocytes. *Naunyn Schmiedebergs Arch. Pharmacol.* 355: 8-13

Samanta H., Pravtcheva D.D., Ruddle F.H. and Lengyel P. (1984) Chromosomal location of mouse gene 202 which is induced by interferons and specifies a 56.5 kD protein. *J. Interferon Res.* 4: 295-300

Samuel C.E. (1988) Mechanisms of the antiviral action of interferon. *Prog. Nucleic Acid Res. Mol. Biol.* 35: 27-72

Samuel C.E. (1991) Antiviral actions of interferon. Interferon-regulated cellular proteins and their surprisingly selective antiviral activities. *Virology* 183: 1-11

Sandrin M.S., Henning M.M., Lo M.F., Baker E., Sutherland G.R. and McKenzie I.F. (1996) Isolation and characterization of cDNA clones for Humly9: the human homologue of mouse Ly9. *Immunogenetics* 43: 13-19

Sanger F., Coulson A.R., Hong G.F., Hill D.F. and Petersen G.B. (1982) Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.* 162: 729-773

Sanger F., Nicklen S. and Coulson A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Nat. Acad. Sci.* 74: 5463-5467

Sassa S. (1976) Sequential induction of heme pathway enzymes during erythroid differentiation of mouse Friend leukemia virus-infected cells. *J. Exp. Med.* 143: 305-315

Sato T., Furukawa K., Bakker H., Van den Eijnden D.H. and Van Die I. (1998) Molecular cloning of a human cDNA encoding a novel beta-1,4-galactosyltransferase with 37% identity to the mammalian UDP-Gal:GlcNAc beta-1,4-galactosyltransferase. *Proc. Nat. Acad. Sci.* 95: 472-477

Sayos J., Martin M., Chen A., Simarro M., Howie D., Morra M., Engel P. and Terhorst C. (2001) Cell surface receptors Ly-9 and CD84 recruit the X-linked lymphoproliferative disease gene product SAP. *Blood* 97: 3867-3874

Sayos J., Wu C., Morra M., Wang N., Zhang X., Allen D., van Schaik S., Notarangelo L., Geha R., Roncarolo M.G. *et al.* (1998) The X-linked lymphoproliferative disease gene product SAP regulates signals induced through the co-receptor SLAM. *Nature* 395: 462

Scaffidi C., Medema J.P., Krammer P.H. and Peter M.E. (1997) Two FLICE isoforms, caspase 8/a and caspase 8/b, are recruited and activated by the CD95 death-inducing signaling complex (DISC). *J. Biol. Chem.* 272: 26953-26958

Schmid C.W. (1998) Does SINE evolution precludes Alu function? *Nucleic Acids Res.* 26: 4541-4550

Schwartz S., Zhang Z., Frazer K.A., Smit A., Riemer C., Bouck J., Gibbs R., Hardison R. and Miller W. (2000) Pipmaker - A web server for aligning two genomic DNA sequences. *Genome Res.* 10: 577-586

Schwientek T., Almeida R., Levery S.B., Holmes E., Bennett E.P. and Clausen H. (1998) Cloning of a novel member of the UDP-galactose:beta-N-acetylglucosamine

beta-1,4-galactosyltransferase family, beta4Gal-T4, involved in glycosphingolipid biosynthesis. *J. Biol. Chem.* 273: 29295-29305

Seed B. (1987) An LFA-3 cDNA encodes a phospholipid-linked membrane protein homologous to its receptor CD2. *Nature* 329: 840-842

Seeman E. and Hopper J.L. (1997) Genetic and environmental components of the population variance in bone density. *Osteoporosis Int.* 7: s10-s16

Selvaraj P., Plunkett M.L., Dustin M., Sanders M.E., Shaw S. and Springer T.A. (1987) The T lymphocyte glycoprotein CD2 binds the cell surface ligand LFA-3. *Nature* 326: 400-403

Shaper N.L., Shaper J.H., Meuth J.L., Fox J.L., Chang H., Kirsh I.R. and Hollis G.F. (1986) Bovine galactosyltransferase: identification of a clone by direct immunological screening of a cDNA expression library. *Proc. Nat. Acad. Sci.* 83: 1573-1577

Shapiro L., Doyle J.P., Hensley P., Colman D.R. and Hendrickson W.A. (1996) Crystal structure of the extracellular domain from P0, the major structural protein of peripheral nerve myelin. *Neuron* 17: 435-449

Shiraki M., Shiraki Y., Aoki C., Hosoi T., Inoue S. *et al.* (1997) Association of bone mineral density with apolipoprotein E phenotype. *J. Bone Miner. Res.* 12: 1438-1445

Shull G.E. (2000) Gene-knockout studies of Ca^{2+}-transporting ATPases. *Eur. J. Biochem.* 267: 5284-5290

Sicheritz-Ponten T. and Andersson S. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res.* 29: 545-552

Siebenlist U., Franzoso G. and Brown K. (1994) Structure, regulation and function of NF-kB. *Annu. Rev. Cell Biol.* 10: 405-455

Sjöström J. and Bergh J. (2001) How apoptosis is regulated, and what goes wrong in cancer. *B.M.J.* 322: 1538-1539

Skehel J.M., Fearnley I.M. and Walker J.E. (1998) NADH:ubiquinone oxidoreductase from bovine heart mitochondria: sequence of a novel 17.2 kDa subunit. *FEBS Lett.* 438: 301-305

Slemenda C.W., Christian J.C., Williams C.J., Norton J.A. and Johnston Jr. C.C. (1991) Genetic determinants of bone mass in adult women: a reevaluation of the twin model, and the potential importance of gene interaction on heritability estimates. *J. Bone Miner. Res.* 6: 561-567

Smeitink J.A.M., Loeffen J.L.C.M., Triepels R.H., Smeets R.J.P., Trijbels J.M.F. and van der Heuvel L.P. (1998) Nuclear genes of human complex I of the mitochondrial electron transport chain: state of the art. *Hum. Mol. Genet.* 7: 1573-1579

Smith L.M., Sanders J.Z., Kaiser R.J., Hughes P., Dodd C., Connell C.R., Heiner C., Kent S.B. and Hood L.E. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321: 674-679

Smith T.S. and Southan C. (2000) Sequencing, tissue distribution and chromosomal assignment of a novel ubiquitin-specific protease USP23. *Biochim. Biophys. Acta* 1490: 184-188

Sonnhammer E.L.L. and Durbin R. (1995) A dot matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167: GC1-10

Staeheli P. (1990) Interferon-induced proteins and the antiviral state. *Adv. Virus Res.* 38: 147-200

Staub E., Dahl E. and Rosenthal A. (2001) The DAPIN family: a novel domain links apoptotic and interferon response proteins. *Trends Biochem. Sci.* 26: 83-85

Stegh A.H., Schickling O., Ehret A., Scaffidi C., Peterhansel C., Hofmann T.G., Grummt I., Krammer P.H. and Peter M.E. (1998) DEDD, a novel death effector domain-containing protein, targeted to the nucleolus. *EMBO J.* 17: 5974-5986

Steller H. (1995) Mechanisms and genes of cellular suicide. *Science* 267: 1445-1449

Stenesh J. and Roe B.A. (1972) DNA polymerase from mesophilic and thermophilic bacteria. I. Purification and properties of DNA polymerase from *Bacillus licheniformis* and *Bacillus stearothermophilus*. *Biochim. Biophys. Acta* 272: 156-166

Stewart T.L. and Ralston S.H. (2000) Role of genetic factors in the pathogenesis of osteoporosis. *J. Endocrinol.* 166: 235-245

Su Y., Brooks D.G., Li L., Lepercq J., Trofatter J.A., Ravetch J.V. and Lebo R.V. (1993) Myelin protein zero gene mutated in Charcot-Marie-Tooth type 1B patients. *Proc. Nat. Acad. Sci.* 90: 10856-10860

Sueyoshi T., Kawamoto T., Zelko I., Honkakoski P. and Negishi M. (1999) The repressed nuclear receptor CAR responds to phenobarbital in activating the human CYP2B6 gene. *J. Biol. Chem.* 274: 6043-6046

Suzuki H., Okazawa Y., Komiya T., Saeki K., Mekada E., Kitada S., Ito A. and Mihara K. (2000) Characterization of rat TOM40, a central component of the preprotein translocase of the mitochondrial outer membrane. *J. Biol. Chem.* 275: 37930-37936

Swartzman A., Kapoor S., Graham A.F. and Meighen E.A. (1990) A new *Vibrio fischeri* lux gene precedes a bidirectional termination site for the lux operon. *J. Bacteriol.* 172: 6797-6802

Tabor S. and Richardson C.C. (1989) Selective inactivation of the exonuclease activity of bacteriophage T_7 DNA polymerase by *in vitro* mutagenesis. *J. Biol. Chem.* 264: 6447-6458

Tabor S. and Richardson C.C. (1995) A single residue in DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Nat. Acad. Sci.* 92: 6339-6343

Tachibana K., Nakanishi H., Mandai K., Ozaki K., Ikeda W., Yamamoto Y., Nagafuchi A., Tsukita S. and takai Y. (2000) Two cell adhesion molecules, nectin and cadherin, interact through their cytoplasmic domain-associated proteins. *J. Cell Biol.* 150: 1161-1176

Tada Y., Horio Y., Takumi T., Terayama M., Tsuji L., Copeland N.G., Jenkins N.A. and Kurachi Y. (1997) Assignment of the glial inwardly rectifying potassium channel KAB-2/Kir4.1 (Kcnj10) gene to the distal region of mouse chromosome 1. *Genomics* 45: 629-630

Tall A. (1990) Plasma high density lipoproteins. Metabolism and relationship to atherogenesis. *J. Clin. Invest.* 86: 379-384

Tamir I. and Cambier J.C. (1998) Antigen receptor signaling: integration of protein-tyrosine kinase functions. *Oncogene* 17: 1353-1364

Tang B.L. and HongW. (1999) ADAMTS: a novel family of proteases with an ADAM protease domain and thrombospondin 1 repeats. *FEBS Lett.* 445: 223-225

Tannenbaum C.S., Major J., Ohmori Y. and Hamilton T.A. (1993) A lipopolysaccharide-inducible macrophage gene (D3) is a new member of an interferon-inducible gene cluster and is selectively expressed in mononuclear phagocytes. *J. Leuk. Biol.* 53: 563-568

Tatsuo H., Ono N., Tanaka K. and Yanagi Y. (2000) SLAM (CDw150) is a cellular receptor for measles virus. *Nature* 406: 893-897

Teegarden D., Proulx W.R., Martin B.R., Zhao J., McCabe G.P. *et al.* (1995) Peak bone mass in young women. *J. Bone Miner. Res.* 10: 711-715

Theintz G., Buchs B., Rizzoli R., Slosman D., Clavien H. *et al.* (1992) Longitudinal monitoring of bone mass accumulation in healthy adolescents: evidence for a marked reduction after 16 years of age at the levels of lumbar spine and femoral neck in female subjects. *J. Clin. Endocrinol. Metab.* 75: 1060-1065

Thompson-Stewart D., Karpen G.H. and Spradling C. (1994) A transposable element can drive the concerted evolution of tandemly repetitious DNA. *Proc. Nat. Acad. Sci.* 91: 9042-9046

Thornberry N.A. and Lazebnik Y. (1998) Caspases: ennemies within. *Science* 281: 1312-1316

Tiefenthaler G., Hunig T., Dustin M.L., Springer T.A. and Meuer S.C. (1987) Purified lymphocyte function-associated antigen-3 and T11 target structure are active in CD2-mediated T cell stimulation. *Eur. J. Immunol.* 17: 1847

Tortorella M.D., Burn T.C., Pratta M.A., Abbaszade I., Hollis J.M., Liu R., Rosenfeld S.A., Copeland R.A., Decicco C.P., Wynn R., Rockwell A., Yang F., Duke J.L., Solomon K., George H., Bruckner R., Nagase H., Itoh Y., Ellis D.M., Ross H., Wiswall B.H., Murphy K., Hillman Jr. M.C., Hollis G.F., Newton R.C., Magolda R.L., Trzaskos J.M. and Arner E.C. (1999) Purification and cloning of aggrecanase-1: a member of the ADAMTS family of proteins. *Science* 284: 1664-1666

Tortorella M.D., Pratta M., Liu R.-Q., Austin J., Ross O.H., Abbaszade I., Burn T. and Arner E. (2000a) Sites of aggrecan cleavage by recombinant human aggrecanase-1 (ADAMTS-4). *J. Biol. Chem.* 275: 18566-18573

Tortorella M., Pratta M., Liu R.-Q., Abbaszade I., Ross H., Burn T. and Arner E. (2000b) The thrombospondin motif of aggrecanase-1 (ADAMTS-4) is critical for aggrecan substrate recognition and cleavage. *J. Biol. Chem.* 275: 25791-25797

Tovar V., de la Fuente M.A., Pizcueta P., Bosch J. and Engel P. (2000) Gene structure of the mouse leukocyte cell surface molecule Ly9. *Immunogenetics* 51: 788-793

Trapani J.A., Browne K.A., Dawson M.J., Ramsay R.G., Eddy R.L., Shows T.B., White P.C. and Dupont B. (1992) A novel gene constitutively expressed in human lymphoid cells is inducible with interferon-gamma in myeloid cells. *Immunogenetics* 36: 369-376

Trapani J.A., Dawson M.J., Apostolidis V.A. and Browne K.A. (1994) Genomic organization of IFI16, an interferon-inducible gene whose expression is associated with human myeloid cell differentiation: correlation of predicted protein domains with exon organization. *Immunogenetics* 40: 415-424

Trent J.M., Stanbridge E.J., McBride H.L., Meese E.U., Casey G., Araujo D.E., Witkowski C.M. and Nagle R.B. (1990) Tumorigenicity in human melanoma cell lines controlled by introduction of human chromosome 6. *Science* 247: 568-571

Uberbacher E.C. and Mural R.J. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Nat. Acad. Sci.* 88: 11261-11265

Uitterlinden A.G., Burger H., Huang Q., Yue F., McGuigan F.E. *et al.* (1998) Relation of alleles of the collagen type I alpha 1 gene to bone density and the risk of osteoporotic fractures in postmenopausal women. *N. Engl. J. Med.* 338: 1016-1021

Vainberg I.E., Lewis S.A., Rommelaere H., Ampe C., Vandekerckhove J., Klein H.L. and Cowan N.J. (1998) Prefoldin, a chaperone that delivers unfolded proteins to cytosolic chaperonin. *Cell* 93: 863-873

Vaughn J., Wolford J.K., Prochazka M. and Permana P.A. (2000) Genomic structure and expression of human KCNJ9 (Kir3.3/GIRK3). *Biochem. Biophys. Res. Commun.* 274: 302-309

Venter J.C. *et al.* (2001) The sequence of the human genome. *Science* 291: 1304-1351

Wang H., Chatterjee G., Meyer J.J., Liu C.-J., Manjunath N.A., Bray-Ward P. and Lengyel P. (1999) Characteristics of three homologous 202 genes (*Ifi202a*, *Ifi202b* and *Ifi202c*) from the murine interferon-activatable gene 200 cluster. *Genomics* 60: 281-294

Wang H., Liu C.-J., Lu Y., Chatterjee G., Ma X.Y., Eisenman R.N. and Lengyel P. (2000) The interferon- and differentiation-inducible p202a protein inhibits the transcriptional activity of c-Myc by blocking its association with Max. *J. Biol. Chem.* 275: 27377-27385

Wang N., Morra M., Wu C., Gullo C., Howie D., Coyle T., Engel P. and Terhorst C. (2001) CD150 is a member of a family of genes that encode glycoproteins on the surface of hematopoietic cells. *Immunogenetics* 53: 382-394

Warden C.H., Hedrick C.C., Qiao J.-H., Castellani L.W. and Lusis A.J. (1993) Atherosclerosis in transgenic mice overexpressing apolipoprotein AII. *Science* 261: 469-472

Watson J.D. and Crick F.H.C. (1953) Molecular structure of nucleic acids. *Nature* 171: 737-738

Waxman D.J. (1999) P450 gene induction by structurally diverse xenochemicals: central role of nuclear receptor CAR, PXR and PPAR. *Arch. Biochem. Biophys.* 369: 11-23

Wei P., Zhang J., Egan-Hafley M., Liang S. and Moore D.D. The nuclear receptor CAR mediates specific xenobiotic induction of drug metabolism. *Nature* 407: 920-923

Weidner U., Geier S., Ptock A., Friedrich T., Leif H. and Weiss H. (1993) The gene locus of the proton-translocating NADH:ubiquinone oxidoreductase in Escherichia coli: organization of the 14 genes and relationship between the derived proteins and subunits of mitochondrial complex I. *J. Mol. Biol.* 233: 109-122

Weiler S.R., Gooya J.M., Ortiz M., Tsai S., Collins S.J. and Keller J. R. (1999) D3: a gene induced during myeloid cell differentiation of Lin^{lo} c-Kit^{+} Sca-1^{+} progenitor cells. *Blood* 93: 527-536

Weissmann C. and Weber H. (1986) The interferon genes. *Prog. Nucleic Acid Res. Mol. Biol.* 33: 251-302

Wen Y., Yan D.H., Spohn B., Deng J., Lin S.Y. and Hung M.C. (2000) Tumor suppression and sensitization to tumor necrosis factor alpha-induced apoptosis by an interferon-inducible protein, p202, in breast cancer cells. *Cancer Res.* 60: 42-46

Wen Y., Yan D.H., Wang B., Spohn B., Ding Y., Shao R., Zou Y., Xie K. and Hung M.C. (2001) p202, an interferon-inducible protein, mediates multiple antitumor activities in human pancreatic cancer xenograft models. *Cancer Res.* 61: 7142-7147

Wilkins M.H.F., Gosling R.G. and Seeds W.E. (1951) Physical studies of nucleic acids. *Nature* 167: 759-760

Wilkinson K.D. (1997) Regulation of ubiquitin-dependent processes by deubiquitinating enzymes. *FASEB J.* 11: 1245-1255

Williams R.C. (1977) *Proc. Nat. Acad. Sci.* 74: 2313

Wilson R. *et al.* (1994) 2.2 Mb of continuous nucleotide sequence from chromosome III of *C. elegans. Nature* 368: 32-38

Woese C.R. (1987) Bacterial evolution. *Microbiol. Rev.* 51: 221-271

Wong Y.W., Williams A.F., Kingsmore S.F. and Seldin M.F. (1990) Structure, expression, and genetic linkage of the mouse BCM1 (OX45 or Blast-1) antigen. *J. Exp. Med.* 171: 2115-2130

Woolhiser M.R., Okayama Y., Gilfillan A.M. and Metcalfe D.D. (2001) IgG-dependent activation of human mast cells following up-regulation of FcgammaRI by IFN-gamma. *Eur. J. Immunol.* 31: 3298-3307

Wower I.K., Wower J. and Zimmermann R.A. (1998) Ribosomal protein L27 participates in both 50S subunit assembly and the peptidyl transferase reaction. *J. Biol. Chem.* 273: 19847-19852

Wrabetz L., Feltri M.L., Quattrini A., Imperiale D., Previtali S., D'AntonioM., Martini R., Yin X., Trapp B.D., Zhou L., Chiu S.-Y. and Messing A. (2000) P0 glycoprotein overexpression causes congenital hypomyelination of peripheral nerves. *J. Cell Biol.* 148: 1021-1033

Xie J., Briggs J.A., Olson M.O.J., Sipos K. and Briggs R.C. (1995) Human myeloid cell nuclear differentiation antigen binds specifically to nucleolin. *J. Cell. Biochem.* 59: 529-536

Xie J., Briggs J.A., Morris S.W., Olson M.O.J., Kinney M.C. and Briggs R.C. (1997) MNDA binds NPM/B23 and the NPM-MLF1 chimera generated by the t(3;5) associated with myelodysplastic syndrome and acute myeloid leukemia. *Exp. Hemat.* 25: 1111-1117

Xie J., Briggs J.A. and Briggs R.C. (1998) Human hematopoietic cell specific nuclear protein MNDA interacts with the multifunctional transcription factor YY1 and stimulates YY1 DNA binding. *J. Cell. Biochem.* 70: 489-506

Xie T. and Ding D. (2000) Investigating 42 candidate orthologous protein groups by molecular evolutionary analysis on genome scale. *Gene* 261: 305-310

Yan D.-H., Wen Y., Spohn B., Choubey D., Gutterman J.U. and Hung M.-C. (1999) Reduced growth rate and transformation phenotype of the prostate cancer cells by an interferon-inducible protein, p202. *Oncogene* 18: 807-811

Yano K. and Zarain-Herzberg A. (1994) Sarcoplasmic reticulum calsequestrins: structural and functional properties. *Mol. Cell. Biochem.* 135: 61-70

Yokoyama C., Wang X., Briggs M.R., Admon A., Wu J., Hua X., Goldstein J.L. and Brown M.S. (1993) SREBP-1, a basic-helix-loop-helix-leucine zipper protein that controls transcription of the low density lipoprotein receptor gene. *Cell* 75: 187-197

Yoon J.C., Puigserver P., Chen G., Donovan J., Wu Z., Rhee J., Adelmant G., Stafford J., Kahn C.R., Granner D.K., Newgard C.B. and Spiegelman B.M. (2001) Control of hepatic gluconeogenesis through the transcriptional coactivator PGC-1. *Nature* 413: 131-138

Zelko I. and Negishi M. (2000) Phenobarbital-elicited activation of nuclear receptor CAR in induction of cytochrome P450 genes. *Biochem. Biophys. Res. Commun.* 277: 1-6

Zelko I., Sueoshi T., Kawamoto T., Moore R. and Negishi M. (2001) The peptide near the C-terminus regulates receptor CAR nuclear translocation induced by xenochemicals in mouse liver. *Mol. Cell. Biol.* 21: 2838-2846

Zhang K., Merazga Y. and Filbin M.T. (1996) Mapping the adhesive domains of the myelin P0 protein. *J. Neurosci. Res.* 45: 525-533

Zhang M.Q. (1997) Identification of protein coding regions in the human genome based on quadratic discriminant analysis. *Proc. Nat. Acad. Sci.* 94: 565-568

Zhong S., Goldberg I.J., Bruce C., Rubin E.M., Breslow J.L. and Tall A. (1994) Human AII inhibits hydrolysis of HDL triglyceride and the decrease of HDL size induced by hypertriglyceridemia and CETP in transgenic mice. *J. Clin. Invest.* 94: 2457-2467

Zmuda J.M., Eichner J.E., Ferrell R.E., Bauer D.C., Kuller L.H. *et al.* (1998) Genetic variation in alpha 2HS-glycoprotein is related to calcaneal broadband ultrasound attenuation in older women. *Calcif. Tissue Int.* 63: 5-8