INFORMATION TO USERS

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

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

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

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

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

Shen, Li-Ming, Ph.D.
The Ohio State University, 1994
STRUCTURAL AND MOLECULAR GENETIC STUDIES OF
THE MHC NOVEL GENE RP

DISSERTATION

Presented in Partial Fulfillment of the Requirement for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By

Li-Ming Shen, B.S.

The Ohio State University
1994

Dissertation Committee:
C. Yung Yu,  D.Phil.
Arthur H. M. Burghes,  Ph.D.
Philip R. Johnson,  M.D.
M. Sue O’Dorisio,  M.D., Ph.D.

Approved by:

Advisor
The Ohio State University
Biochemistry Program
To My Parents
ACKNOWLEDGEMENTS

My greatest thanks go to my advisor, Dr. C. Yung Yu for his unending support and guidance throughout my research. I would like to thank Dr. Ross Dalbey for his teaching in the early years of my graduate studies. Many heartfelt thanks are given to members of my dissertation committee, Dr. Sue O'Dorisio, Dr. Philip Johnson and Dr. Arthur Burghes for their encouragement and constructive advice.

I am grateful for the technical assistance I have received from past and present members of the laboratory: Anna Mendoza, Rujju Chen, Salih Sanlioglu, Dr. Andrew Dangel and Brad Baker. I am indebted to Dr. Lai-chu Wu whose work laid the foundation of the RP project.

Finally, I am extremely grateful to my family. This work would not be possible without their support and encouragements.
VITA

August 27, 1963  Born - Taiwan, Republic of China
1981-1985  B.S., Department of Chemistry, National Taiwan University
            Taipei, Taiwan
1985-1987  Second Lieutenant, Marine Corps, Republic of China
1987-1990  Graduate Teaching Assistant, Department of Chemistry
            The Ohio State Biochemistry Program
1990-present  Graduate Research Assistant, Children's Hospital Research
              Foundation, The Ohio State Biochemistry Program

PUBLICATIONS


FIELD OF STUDY

Major Field:  Biochemistry
            Studies in Molecular Genetics
TABLE OF CONTENTS

DEDICATION .................................................................................................................... ii
ACKNOWLEDGEMENTS ............................................................................................ iii
VITA .............................................................................................................................. iv
LIST OF TABLES ............................................................................................................ ix
LIST OF FIGURES ........................................................................................................ x
ABBREVIATIONS ......................................................................................................... xiv

CHAPTER I INTRODUCTION

1.1 Structure and function of the MHC ................................................................. 1
1.2 MHC-associated genetic diseases and immunological disorders .................. 3
1.3 Genes at the MHC class III region .................................................................. 8
1.4 Duplication of the complement C4-CYP21 gene cluster ............................... 10
1.5 Evolution of C4-CYP21 locus and its multiplication ..................................... 13
1.6 Research goals of this thesis ........................................................................... 14

CHAPTER II STRUCTURE AND GENETICS OF THE PARTIALLY DUPLICATED GENE RP IN THE HLA CLASS III REGION

Introduction ................................................................................................................. 23
Materials and methods ............................................................................................... 25
Results
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Isolation of cDNA clones for the novel gene RP upstream of C4</td>
<td>30</td>
</tr>
<tr>
<td>2.2</td>
<td>Gene structure of RP1</td>
<td>32</td>
</tr>
<tr>
<td>2.3</td>
<td>A composite retroposon is present in the RP1 gene</td>
<td>33</td>
</tr>
<tr>
<td>2.4</td>
<td>RP gene is duplicated in most human and primate genomes</td>
<td>35</td>
</tr>
<tr>
<td>2.5</td>
<td>Partial gene duplication of RP and Gene X</td>
<td>36</td>
</tr>
<tr>
<td>2.6</td>
<td>A common breakpoint region for duplication of the RCCX modules in great apes</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>Discussion</td>
<td>40</td>
</tr>
</tbody>
</table>

**CHAPTER III**

**GENERATION OF POLYCLONAL ANTIBODIES AND BIOCHEMICAL CHARACTERIZATION OF THE ENDOGENOUS RP PROTEINS**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Strategies for construction of MBP-RP fusion protein clones</td>
<td>82</td>
</tr>
<tr>
<td>3.2</td>
<td>Induction of fusion protein synthesis</td>
<td>83</td>
</tr>
<tr>
<td>3.3</td>
<td>Optimization of the synthesis and purification of fusion proteins</td>
<td>84</td>
</tr>
<tr>
<td>3.4</td>
<td>Immunization and antibody purification</td>
<td>86</td>
</tr>
<tr>
<td>3.5</td>
<td>ELISA assay</td>
<td>87</td>
</tr>
<tr>
<td>3.6</td>
<td>Western blot analyses</td>
<td>88</td>
</tr>
<tr>
<td>3.7</td>
<td>Biosynthetic labeling and immunoprecipitation</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>Discussion</td>
<td>90</td>
</tr>
</tbody>
</table>
CHAPTER IV
CHARACTERIZATION OF THE RP
TRANSCRIPTS AT THEIR UPSTREAM
REGIONS AND DETERMINATION OF
THE RP GENE EXPRESSION

Introduction ............................................................................................................. 107
Materials and methods ..................................................................................... 109
Results
4.1 RT-PCR for the 5'-end of human RP cDNA ........................................ 113
4.2 Primer extension analyses ...................................................................... 115
4.3 Ribonuclease protection assays ........................................................... 116
4.4 Use of a hybrid system for the transcription factor activity
   of RP ............................................................................................................ 117
Discussion ......................................................................................................... 118

CHAPTER V
MOUSE RP1-SLP AND RP2-C4 GENE
STRUCTURES IN THE H-2 S REGION

Introduction ............................................................................................................. 138
Materials and methods ..................................................................................... 140
Results
5.1 Cloning and characterization of mouse cDNA ................................. 142
5.2 RP genes are also present in mouse and other mammalian
   genomes ........................................................................................................ 143
5.3 Isolation and identification of mouse RP genes ................................. 144
5.4 Duplication of mouse RP1-Slp and RP2-C4 gene loci ......................... 147
5.5 The exon-intron structure and the sequence variations of
   mouse RP1 and RP2 genes ......................................................................... 149
5.6 The expression of mouse RP .................................................................... 150
Discussion ........................................................................................................... 151
CHAPTER VI POLYMORPHISM AND GENOMIC REARRANGEMENT OF THE RP GENE

Introduction ......................................................................................................... 173
Materials and methods .................................................................................... 176
Results
6.1 A sequence comparison between the 3'-ends of mouse RP1 and RP2 genes .................................................................................................... 178
6.2 A sequence alignment among the 3'-ends of a RP1 and three RP2 genes: evidence for a gene conversion event in the human RP locus .............................................................................. 179
6.3 Genomic rearrangement of RP in a NPC cell line ........................................ 180
6.4 Cloning and determination of rearrangement breakpoint in NPC 181
Discussion ............................................................................................................ 182

CHAPTER VII GENERAL DISCUSSION ................................................... 197

REFERENCES ..................................................................................................................203
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. HLA disease associations</td>
<td>19</td>
</tr>
<tr>
<td>2. Diseases association with null alleles of the complement C4 gene</td>
<td>22</td>
</tr>
<tr>
<td>3. A summary of the exon-intron structure of the human RP1 gene</td>
<td>69</td>
</tr>
<tr>
<td>4. ELISA of RP antibodies</td>
<td>100</td>
</tr>
<tr>
<td>5. Exons and introns of the mouse RP genes</td>
<td>172</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Organization of the human major histocompatibility complex on the short arm of the chromosome 6</td>
<td>2</td>
</tr>
<tr>
<td>2.</td>
<td>A graphic representation of linkage disequilibrium between HLA-A and HLA-DR</td>
<td>16</td>
</tr>
<tr>
<td>3.</td>
<td>Haplotype analysis of unrelated patients with type I diabetes</td>
<td>18</td>
</tr>
<tr>
<td>4.</td>
<td>A diagram for adrenal steroid biosynthesis</td>
<td>20</td>
</tr>
<tr>
<td>5.</td>
<td>Northern blot analysis of human RP</td>
<td>45</td>
</tr>
<tr>
<td>6.</td>
<td>cDNA and predicted amino acid sequences of human RP</td>
<td>46</td>
</tr>
<tr>
<td>7.</td>
<td>Structural properties of the RP protein</td>
<td>48</td>
</tr>
<tr>
<td>8.</td>
<td>A comparison of human RP cDNA sequences from RP1.1, RP410, and J6 cDNA libraries</td>
<td>49</td>
</tr>
<tr>
<td>9.</td>
<td>A restriction map of the human RP1 gene</td>
<td>50</td>
</tr>
<tr>
<td>10.</td>
<td>The complete DNA sequence of the human RP1 gene</td>
<td>51</td>
</tr>
<tr>
<td>11.</td>
<td>The organization of human RP1 gene</td>
<td>60</td>
</tr>
<tr>
<td>12.</td>
<td>The VNTRs in the human RP1 gene</td>
<td>61</td>
</tr>
<tr>
<td>13.</td>
<td>A comparison of the structures of the composite SVA retroposons in RP1, C2 (B allele), and CYP1A1 genes</td>
<td>62</td>
</tr>
<tr>
<td>14.</td>
<td>An alignment of the Alu sequences from SVA-RP (Alu-RP), SVA-C2 (Alu-C2), and SVA-1A1 (Alu-1A1) with the Alu consensus</td>
<td>63</td>
</tr>
<tr>
<td>15.</td>
<td>Southern blot analysis of RP genes in humans and monkeys</td>
<td>64</td>
</tr>
</tbody>
</table>
16. Partial gene duplication of Gene X and RP ......................................................... 65
17. The breakpoint for the partial gene duplication of RP and Gene X .............. 66
18. A molecular map of the RCCX gene modules in the HLA class III region 67
19. A common breakpoint region for gene duplication of Gene XA-RP2 for human and great apes ................................................................. 68
20. Constructs of the RP-fusion protein clones ........................................................... 94
22. Determination of the optimal conditions for fusion protein production and purification ............................................................................................................. 97
23. Large scale MBP-RP/a fusion protein production and purification .......... 98
24. Large scale MBP-RP/b fusion protein production and purification .......... 99
25. ELISA analysis of six antisera against the RP proteins ............................... 101
26. Western blot analysis with unpurified anti-MBP-RP sera ....................... 102
27. Western blot analysis of HT29 and NIH3T3 cells .......................................... 103
28. Western blot analysis by using both unpurified and purified antisera for HT29 and CNE cells ......................................................................................... 104
29. Analysis of six antisera by immunoprecipitation and the use of Ab6 on TEJES and MOLT4 cell lysates .............................................................................. 105
30. Detection of a 55 kDa polypeptide for RP by immunoprecipitation ........ 106
31. RT-PCR analysis and subcloning from BJAB cells ....................................... 123
32. Subcloning of RT-PCR products for 5'-end of RP cDNA ......................... 124
33. A sequence comparison among RT-PCR products and genomic DNA on their splice junction ................................................................. 125
34. Construction of a full-length RP cDNA by combining R1.1 and a RT-PCR product ................................................................. 127

35. A primer extension analysis using primer RP53 and 0.8K-R1 .......... 129

36. Mapping of transcription start sites from primer extension experiments 131

37. A RNAse protection assay to determine the RP transcript .......... 132

38. Nucleotide sequence of the 5' regulatory region of the human RP gene 134

39. Reporter and effector constructs for the GAL4 hybrid system ........ 136

40. Cloning strategy for the construction of two GAL4-RP hybrid proteins 137

41. Partial sequence of the mouse RP cDNA and its deduced amino acid sequences ................................................................. 155

42. Genomic Southern blot analysis of the mouse RP genes ............ 157

43. Molecular organization of the mouse RP gene loci and the BamHI restriction map of the mouse RP1 gene 158 ................................. 158

44. The genomic organization of the mouse RP2 gene .................... 160

45. Analysis of RP1 genomic DNA clones from the 129SVJ Lambda library 162

46. Analysis of RP2 genomic DNA clones from the 129SVJ Lambda library 163

47. A comparison of nucleotide sequences of mouse RP1 and RP2 from exon 7 to exon 9 region ............................................. 164

48. A schematic diagram for the genomic organization of mouse RP1 and RP2 genes at their 3' regions .............................................. 166

49. Variations of the coding DNA sequences and its associated alternations in amino acid residues between RP1 and RP2 .................. 168

50. RT-PCR analysis of various mouse strains and tissues .............. 169
51. The gene-targeting strategy for mouse RP gene ........................................... 170

52. A summary of sequence variations for six mouse RP gene at their exon 7 to exon 9 regions .......................................................... 187

53. A sequence alignment of human RP1 and RP2 genes at the exon 8 flanking region ................................................................. 189

54. Cloning of the rearrangement partner fragment by RAGE (Rapid Amplification of Genomic End) technique ........................................ 191

55. The PCR products for the cloning of the CNE rearranged fragment ... 193

56. Sequence analysis of the vector-anchored PCR product for the NPC rearrangement breakpoint .................................................... 194

57. The sequence of the PCR product for NPC rearrangement study ...... 195

58. A schematic representation of the putative mechanism of the slippage synthesis for the rearrangement of the CNE cells ...................... 196


LIST OF ABBREVIATIONS

Ab                  antibody
bp                  base pair
cDNA                complementary DNA
Ci                  Curie
cpm                 counts per minute
dATP                2'-deoxyadenosine 5'-triphosphate
dNTP                deoxynucleotide triphosphate
DNA                 deoxyribonucleic acid
EDTA                ethylene diamine tetraacetic acid
ELISA               enzyme-linked immunosorbent assay
kb                  kilobase
kDa                 kilo-Dalton
mg                  milligram
ml                  milliliter
μg                  microgram
μl                  microliter
mRNA                messenger ribonucleic acid
O.D.                optical density
PCR                 polymerase chain reaction
rpm                 rotation per minute
RT-PCR              reverse transcription polymerase chain reaction
SINE                short interspersed repeats
SDS                 sodium dodecyl sulfate
SSC                 salt-sodium citrate buffer
ss DNA              single-stranded DNA
SSPE                salt-sodium phosphate EDTA buffer
TBE                 Tris-borate EDTA buffer
Tris                tris(hydroxymethyl)aminomethane
VNTR                variable number tandem repeats
CHAPTER I
INTRODUCTION

1.1 Genomic structure and function in the MHC region

The human major histocompatibility complex (MHC) is one of the most fascinating and the most intensely studied regions of the human genome. More is known about the 4 Mb of DNA (0.1% of the genome) on the short arm of chromosome 6 (6p21.3) than any other region of similar size. In addition to a crucial role in the immune defense system, the MHC is one of the most polymorphic regions in the mammalian genomes. It is therefore a useful model for the study of gene organization, polymorphism, linkage disequilibrium and recombination. The human MHC (also known as the HLA, Human Leukocyte Antigen) comprises a chromosomal segment of 3500 kb. A molecular linkage map of the entire MHC has been determined by pulsed-field gel electrophoresis (Carroll et al., 1984; Dunham et al., 1987) and it is divided into class I, II and III regions. The class III region (about 1000 kb in size) is flanked by the class I loci (about 1800 kb) at the telomeric end and the class II loci (about 1000 kb) at the centromeric end (Figure 1).

The protein products encoded by class I (HLA-A, -B, -C) and class II (HLA-DR, -DQ, -DP) genes are involved in the presentation of processed antigens to T cells, related to the cell-mediated immune response (reviewed in Monaco, 1992). While the class I molecules are usually associated with presentation of endogenous-synthesis proteins (such as those from virus), class II molecules are thought to bind
Figure 1

Organization of human major histocompatibility complex on the short arm of chromosome 6 (6p21:31-33) (Adapted from Srivastava et al., 1991).
extracellular peptides in a specialized endosome-lysosome compartment. In addition to these polymorphic surface glycoproteins involved in antigen recognition, several structurally unrelated genes in class II region encode protein with the function of antigen processing and transportation. For instances, the products of TAP1 and TAP2 genes are proposed to form an ATP-binding cassette and an endoplasmic reticulum (ER) peptide transporter (Parham, 1990); LMP2 and LMP7 genes encode inducible components of the proteasome complex (Monaco, 1992). The protein products encoded by class III region are of heterogeneous function, unrelated to each other as well as to both class I and II histocompatibility antigens. These include several known genes encoding the complement components (C4, C2 and FB), cytochrome P450 steroid 21-hydroxylase (CYP21), tumor necrosis factor (TNF), and three members of the major heat shock protein HSP70 family (see below).

1.2 MHC-associated genetic diseases and immunodeficiency disorders

The alleles of the polymorphic HLA class I, II and III genes undergo an unequal combination or linkage disequilibrium in the population, in which specific alleles at the closely linked loci are found together in individuals more often than one would predict from the frequency of the individual allele. The combination of these alleles in a haploid genome is known as a haplotype (Figure 2). Certain MHC haplotypes have been associated with more than 40 genetic or autoimmune diseases (reviewed by Sinha et al., 1990) including insulin dependent diabetes mellitus (IDDM) (Figure 3), rheumatoid arthritis, multiple sclerosis, etc. The high level of polymorphism of the MHC system has provided excellent genetic markers for understanding the genetic components of complex diseases. A list of diseases with strong HLA associations is given in Table 1. For instance, nearly 90% of individuals with ankylosing spondylitis have B27, while only 9% of general
population with the B27 antigen.

Basically two possibilities exist to explain the MHC disease associations. Firstly, the disease susceptibility is the direct result of the presence of a particular HLA antigen with the involvement of immune defense. The second is that the association is a result of linkage disequilibrium (nonrandom association) between the HLA marker allele(s) and a nearby locus which confers susceptibility to disease. However, the genetics and the determination of the modes of inheritance are rather complex for these diseases due to multiple disease loci, heterogeneity in the mechanisms of the disease predisposition, and incomplete penetrance (reviewed by Thomson, 1990).

Individuals with haplotypes containing null and/or rare alleles of C4, such as C4AQ0 and C4B3 are more susceptible to IgA deficiency (IgA-D), common variable immunodeficiency (CVID) (Wilton et al., 1985; Schaffer et al., 1989; French and Dawkins, 1990), IDDM (Partenen et al., 1986), myasthenia gravis (Skanes, 1985) and juvenile rheumatoid arthritis (Arnaiz-Villena et al., 1989). Those patients with diseases such as systemic lupus erythematosus (SLE), scleroderma, autism, schizophrenia, and many others (see Table 2 for listing) have also higher frequencies of C4-deficiency (null alleles at the C4 locus).

IgA deficiency is a common immunological disorder with a prevalence of about 1 in 800 Caucasians (Eibl et al., 1989). The syndromes associated with the IgA deficiency are highly variable. Many affected individuals have no obvious health problem, at least for many decades, while others may have recurrent infections, gastrointestinal disorders, autoimmune diseases, allergies, or malignancies (reviewed by Schaffer et al., 1991). Current evidence indicates that pathogenesis of IgA deficiency is a defect of B cell differentiation. IgA-D individuals have normal numbers of IgA-bearing B cells, but fail to undergo terminal differentiation into IgA-secreting plasma cells (Conley et al., 1981). It has
been proposed that an unknown genomic element is involved in the isotype switching and immune-regulation (Schaffer et al., 1989).

Common variable immunodeficiency (CVID), another immuno-deficiency syndrome, exhibits decreased level of multiple immunoglobulin classes. Similar to the IgA deficiency, the majority of the individuals with CVID have normal numbers of immunoglobulin-bearing B cell precursors but their B cells are unable to differentiate into antibody-secreting plasma cells (Copper et al., 1971). Schaffer et al. (1989) proposed that both IgA-D and a subset of CVID may represent polar ends of spectrum of a B cell differentiation, suggesting that these two diseases may have a genetic defect of common origin. In addition, an autosomal recessive mode of inheritance was observed in family studies (Wilton et al., 1985), and these two disorders occurred in immediate relatives of the same family (Wollheim et al., 1965). Previous studies have demonstrated that particular MHC haplotypes were associated with both IgA-D and CVID (Schaffer et al., 1989), and the MHC susceptibility gene (or genes) was localized to the locus of MHC class III region. However, several reports also indicated that a number of class I and/or class II HLA antigens, including A1, A28, B8, B14, B40, DR3, and DR7, were associated with IgA-D and other autoimmune diseases (Ambrus et al., 1977). Particularly, variation of amino acid residue at position 57 of HLA-DQ β-chain was shown to associate with susceptibility and resistance to IgA deficiency (Olerup et al., 1990). However, Volanakis et al. (1992) reported no statistically significant difference of the amino acid at DQ β-chain allele after analyzing 31 immunodeficient individuals, when compared with their healthy relatives and other unrelated controls. Hence, it seems possible that there are predisposing genetic factors for the development of IgA-D and CVID and these factors are in linkage disequilibrium with class I and/or class II alleles. According to allelic polymorphism data (Schaffer et al., 1989), a small number of MHC haplotypes are shared by most
individuals with IgA-D and CVID. It has been speculated that the susceptibility gene or genes for a large subset of both immunodeficiencies is located in the MHC class III region, perhaps between C4B and C2 genes (Volanakis et al., 1992).

At least two human malignant neoplasias were associated with MHC: nasopharyngeal carcinoma (NPC) (Lu et al., 1991) and squamous cell carcinoma of the cervix in humans (Wank et al., 1991). Both were shown with viral-MHC associations and probably contained susceptibility loci closely linked to genetic variation in the HLA region. NPC was linked to the Epstein-Barr virus (EBV), while cervical carcinoma was believed to be human papilloma virus (HPV) induced. The work of Wank et al. (1991) suggested that genes of the MHC may influence the effectiveness of the immune response against tumor cells. Moreover, a linkage of tumor regression and malignant conversion of rabbit viral papilloma was found to be associated with the MHC genes (Han et al., 1992). Recently, a report demonstrated that specific MHC haplotypes were significantly associated with either susceptibility or with resistance to cervical carcinoma (Apple et al., 1994). Hodgkin's disease, another cancer associated with EBV, was also linked to a predisposition loci at the HLA (Klitz et al., 1994). Allelic variation or the polymorphic genes of the host immune system may play a key role in the host response. The genes on MHC alleles therefore could have both protective and susceptible effects on the viral infection (Klitz, 1992).

The search for the alteration or the loss of function through gene mutation is the "Holy Grail" of disease genetics. Locating the gene within the chromosome and pinpointing the specific mutations can lead to understanding the disease pathology, early diagnosis and possible treatment. A wide range of disease susceptibility loci have been mapped to MHC. However, the association with diseases in the HLA region involving genetic variation is quite different from many monogenic hereditary genetic diseases such as cystic fibrosis and Duchenne
muscular dystrophy. Reported allelic or haplotype association to certain disease is never 100%. In other words, only a fraction of the individuals carrying a susceptibility allele will develop the disease, while most individuals will be normal and symptom-free. As a result, the debate about the HLA-disease association is not uncommon and some further studies could not confirm previous reports (Glew et al, 1992). Two explanations may address the problem: Firstly, the disease heterogeneity may be responsible for the diversity of the HLA association while the developed phenotypes are indistinguishable. For instance, in two forms of diabetes only insulin-dependent diabetes has strong HLA association. Therefore there are apparently two types of mechanisms leading to diabetes. Secondly, the differences in the HLA-bearing among various ethnic populations adds to the difficulty of confirming HLA-association. Due to the HLA linkage disequilibrium, one could show the association of HLA-disease in one ethnic population but not in another. That might explain some inconsistencies among HLA studies in cervical cancer (Glew et al., 1992). Therefore the identification of the disease susceptibility gene locus from the MHC region will be a challenge of human genetics. That is also the very reason to study the newly identified genes in the region. The characterization of the genetics and structural properties of these novel genes may shed light on the disease pathology.

More evidence from the genetic studies on congenic mice and rats indicated that some disease susceptibility gene loci were linked to MHC class III region: 1) There is a tumor suppressor gene in the rcc locus in the rat grc region, where the grc region was linked to growth and reproduction defects and might coincide with the class III region of human MHC (Melhem et al., 1993). 2) Others have shown that strains of inbred mice with different MHC background differ greatly in their susceptibility to the development of tumors. This implied the presence of certain MHC genes of non-immunological function in susceptibility to tumorigenesis
3) A genetic locus responsible for a predisposition to skin cancer in mice has been studied and the equivalent locus in the human genome resides between C4 and DR genes in the MHC class III region (Molvig et al., 1988). Nonetheless, the precise location and the putative function of these tumor susceptibility gene/genes have not been identified.

1.3 Genes in the MHC class III region

The class III region of human leukocyte antigen (HLA) locus on chromosome 6p21 is comprised of about 1,200 kb, containing many duplicated genes with very diverse functions. These include tumor necrosis factor α and β, heat shock proteins HSP70, complement cascade factors B, C2, C4A, and C4B, adrenal steroid 21-hydroxylase gene 21A and 21B. There are many other newly identified genes in this region with unknown function. Their characterization may shed light on clinical and genetic studies for a number of MHC-associated diseases, and on the genetics and evolution of the MHC complex.

C4 is the fourth component of the serum complement system. It is an essential subunit of the C3 and C5 convertases of the classical activation pathway (review by Porter, 1984). In humans there are two tandem C4 loci, locus I and locus II about 12 kb apart encoding two isotypes, C4A and C4B, respectively (review by Porter, 1985; Carroll et al., 1985). Despite the 99% sequence identities, C4A and C4B are highly polymorphic at the protein level and show remarkable difference in class-specific, chemical reactivities (Isenman et al., 1984; Law et al., 1984; Dodds et al., 1986). By the conventional typing method thirty-four allotypes of C4 in the two classes have been detected. Some particular haplotypes are believed to account for some HLA-linked autoimmune diseases (reviewed by Sinha et al., 1990). For example, the C4 null alleles, particularly C4A Q0, had a significantly higher relative risk for systemic lupus erythematosus (SLE) (Awdeh
et al., 1983 and Howard et al., 1986). Many autoimmune diseases and other disorders as shown in Table 2, where the patients possess C4 null alleles were compared with the normal individuals (control).

Steroid 21-hydroxylase (CYP21) is an adrenal-specific microsomal cytochrome P450 required for the synthesis of both glucocorticoids and mineralocorticoids (Figure 4). It is an enzyme involved in the conversion of cholesterol to aldosterone or cortisol in the cortex of the vertebrate adrenal gland (Miller, 1988). The human 21-hydroxylase B (CYP21B) is a functional gene, while the CYP21A gene has various mutations, resulting in a frameshift and premature stop codons. Hence CYP21A is considered to be a nonfunctional pseudogene. Gene lesions of CYP21B cause 21-hydroxylase deficiency, a disease affecting about 1 in 12,000 newborns. Congenital adrenal hyperplasia (CAH) is a family of autosomal recessive disorders resulting from malfunction in genes encoding the enzymes for steroidogenesis pathways. The most common form of CAH involves CYP21, and is characterized by a wide range of phenotypes such as salt-wasting and/or virilization phenomena (reviewed by Miller et al., 1989). Gene conversion events between CYP21A and CYP21B genes through frequent genetic recombination were proposed to be the cause of most cases of CYP21B mutations (Collier et al., 1993).

At least three new genes have been located to an approximately 120 kb region, between the complement C2 and CYP21B. A single copy gene RD, which is located 6 kb downstream from the BF gene, encodes a polypeptide of 42 kDa with a dipeptide motif of alternating basic and acidic residues Arg (R) and Asp (D) (Levi-Strauss et al., 1988). A pair of duplicated genes, Tenascin Tn-XA and Tn-XB, are located to the 3' end of the CYP21A and CYP21B, respectively (Morel et al., 1989). The biological functions of these new genes have not been established.
To date 39 genes have been localized in the MHC class III region. Two major approaches used to detect the putative coding sequences from new genes are to search for sequences conserved in evolution (for instance, among the species of mammals) and to define the position of CpG islands. These under-methylated CpG-rich sequences are frequently associated with the 5' end of housekeeping, and also some tissue specific genes (Bird, 1987; Gardiner et al., 1987). The CpG-islands can be detected in chromosomal DNA as clustered sites for certain infrequently cutting restriction enzymes with CpG dinucleotide as part of their recognition sequences. There are at least 9 different transcripts, BAT1-BAT9 (BAT represents B-Associated Transcripts), encoded by the genomic region between the C2 and HLA-B interval of the MHC class III region (Sargent et al., 1989; Spies et al., 1989). Furthermore, 12 novel genes (G1-G12) have been identified in the segment of 541 kb of DNA separating the C2 and the tumor necrosis factor (Sargent et al., 1989). Seven other genes (G12-G18) are localized within a 160 kb DNA region extending from the complement cluster toward HLA-DR (Kendall et al., 1990). However, many of the G transcripts are identical to BAT-transcripts discovered by Spies and colleagues. The biological functions of these new genes remain to be determined.

1.4 Gene duplication in complement C4-CYP21 gene cluster

It is noteworthy that many genes in the MHC Class III region are duplicated (C4, CYP21, TNF, HSP70 and others). Duplications are also common in MHC class I and Class II regions. Gene duplication plays an important role in genomic evolution of various organisms. The most striking examples are the multigene or supergene families, such as genes for ribosomal RNA, immunoglobulin genes, and haptoglobin genes (Hu et al., 1991). The sequence similarities among different proteins provide a clue that these proteins may be evolutionarily related. There are more than 30 unrelated partial gene duplications that have been implicated to
contribute to a mutation spectrum for several human genetic diseases (review see Hu and Worton, 1992). Hence the study of the tandemly duplicated C4/CYP21 gene cluster of the MHC class III region became much more interesting because of the relationship to human diseases. Gene duplication, deletion, and rearrangement occur at a high frequency in this gene cluster. It has been estimated that about 85% of 21-hydroxylase deficiencies were due to frequent genetic recombination (or gene conversion) in this locus (reviewed by Morel et al., 1991).

Gene duplication was considered to be an evolutionary strategy, where one copy of the duplicated gene continues to perform the original function, and the other copy is free to mutate and to assume a new function (Ohno et al., 1970). Gene duplications in the MHC region are frequent events. In addition to the polymorphic locus for Class I and Class II gene clusters, other MHC genes have undergone duplication, including HSP70, TNF, C4, C2, factor B, TPA (an ATP-binding cassette transporter superfamily, Parham, 1990) and LMP (a subunit of proteasome, reviewed by Monaco, 1992). In haplotypes carrying multiple copies of C4 and CYP21 genes, the genomic arrangements always alternate on the linkage map in the order: C4...CYP21...C4...CYP21. This observation suggested that the basic unit for duplication is a C4-CYP21 module and multi-modular haplotypes arise from the multiplication of this unit (Kawaguchi et al., 1991). One pair of newly discovered genes, termed XA and XB, is also tandemly duplicated in the C4/CYP21 locus (Morel et al., 1989). Hence, the third member of this modular structure was added. This study shows that this module also includes the RP gene located upstream of the C4 gene. Together with RP1 and RP2 genes, the complete four-gene cluster structure was named RCCX module (RP-C4-CYP21-Gene X). Gene XA and XB overlap the last two exons of CYP21A and CYP21B genes, respectively, and are transcribed in the opposite orientation relative to the remaining three genes. Gene XB encodes a protein that is closely similar to
Tenascin is an extracellular matrix protein with five distinct domains: a signal peptide, a hydrophobic domain containing three heptad repeats, a series of EGF-like repeats, fibronectin type III repeats, and a carboxyl-terminal fibrinogen-like domain (Matsumoto et al., 1992; Gitelman et al., 1992). The XB gene probably spans over 65 kb with a transcript of 12 kb and encodes a protein about 400 kDa in molecular weight. Gene X is expressed ubiquitously in human fetal tissues. It has been suggested to be essential for development and that its deletion might be fatal (Bristow et al., 1993).

Many disease-causing duplications or deletions might be attributed to illegitimate recombination. For instance, certain repetitive DNA elements such as the Alu-elements may be involved in genetic alteration. The breakpoints of the duplicated LDL receptor gene and the C1 inhibitor gene are generated by Alu-Alu recombination (Ariga et al., 1990; Stoppa-Lyonnet et al., 1990). The presence of a similar sequence at the proximity of the duplication boundary is a hallmark for homologous recombination. It is thus of interest to determine whether the rearrangement of the RCCX locus involves the repetitive DNA elements at the breakpoint.

Nonetheless, nonhomologous recombination has been suggested to be responsible for the majority of genomic rearrangement in mammalian cells (Roth et al., 1986). Due to the consequence of DNA breakage and rejoining for the process of nonhomologous recombination, it was conceivable that multiple pathways might be required in mammalian cells. The type I and type II DNA topoisomerases were suggested to mediate nonhomologous recombination in vitro (Bullock et al., 1985, and Bae et al., 1988). Other enzymes involving DNA scission and ligation such as various DNA repair enzymes and DNA ligases are the putative factors responsible for nonhomologous recombination. However, no evidence so far could support this hypothesis. A systematic analysis of 110 junction sequences
(Roth et al., 1985) generated by nonhomologous recombination indicated that about 40% of the junction breakpoints were associated with very short regions of homology (usually 2 to 6 nucleotides). Therefore, it is interesting to study the duplication breakpoint of the RCCX gene cluster in order to understand the possible mechanism of frequent genomic rearrangements at this MHC region.

1.5 Evolution of C4-CYP21 haplotypes

In mammals, the number of complement C4 and CYP21 gene copies varies from species to species and also among individuals of certain species. Single C4-CYP21 haplotypes are present in humans (Mclean et al., 1988), hamsters (Levi-Strauss et al., 1985), dogs (Kay et al., 1984), cats (Kroon et al., 1986), guinea pigs (Bitter et al., 1977), and several whale species (Spilliaert et al., 1990).

Multiplications of C4-CYP21 haplotype are commonly seen in different mammalian orders. The haplotypes with two C4 and two CYP21 genes are present in humans, chimpanzees (Kawaguchi et al., 1990), gorillas (Kawaguchi et al., 1992), macaques (Mevag et al., 1983), mice (Roos et al., 1978), rats (Tosi et al., 1985), cattles (Yoshioka et al., 1986), and pigs (Kirszenbaum et al., 1985). Haplotypes with three or four copies of C4-CYP21 have been reported in humans (Mclean et al., 1988; Zhang et al., 1993), and in orangutans (Kawaguchi et al., 1992).

Previous studies indicated that primates, ungulates (cattle), and rodents share similar duplicated C4-CYP21 loci (reviewed by Miller et al., 1992). This observation might imply that the duplication of the C4-CYP21 locus predated mammalian speciation, which probably occurred about 85 million years ago (reviewed by Klein et al., 1993). However, this hypothesis has been challenged by the following two lines of evidence. Firstly, the duplication boundary of mouse is different from that of primate (Nonaka et al., 1986 and our unpublished
observation). The intergenic region between CYP21A and C4B of mouse genome is about 60 kb instead of 7 kb in the human genome. Secondly, the sequences of mouse C4 or CYP21 genes are more similar to the duplicated counterparts than they are to any of the human C4 or CYP21 genes (Tosi et al., 1985). These observations suggest that the duplication of the RCCX locus of these mammalian species occurred independently. These post-speciation events and unrelated duplications at different locations of the analogous locus suggest a dynamic and plastic nature of the genomic structure of the HLA locus.

1.6 The research goals of this thesis

A. **Determine the complete cDNA structure of RP to deduce the protein primary structure**

Several complementary (cDNA) sequences were isolated from cDNA libraries and completely sequenced. The 5' end region of the cDNA was obtained by RT-PCR technique. The RP cDNA sequence and its derived amino acid sequence were used to screen national databases in order to compare RP with other known genes and proteins.

B. **Determine the genomic sequences of the RP1 and RP2 genes**

Several DNA subclones derived from cosmid and λ-clones corresponding to RP1 and RP2 genes were characterized. The DNA sequences were determined to elucidate the exon-intron structures, to define the functional/transcribed gene or pseudogene, and to complete the C4-CYP21 duplicated modular structure. The entire DNA sequence in this locus was analyzed to reveal unusual genomic elements.
C. Study of the RP gene expression

The RP expression spectrum was determined by using RNA isolated from cell lines of various origins. The transcriptional initiation site of the RP gene was investigated by RNase protection assays and primer extension analysis. The regulatory region of the RP gene was analyzed by using a transcriptional reporter construct.

D. Study of the biochemical properties of RP protein

Fusion proteins of RP were produced, purified, and used to immunize rabbits in order to generate RP specific antibody. The endogenous RP protein was characterized biochemically with immunochemical techniques.

E. Characterize the mouse RP gene

Mouse RP cDNA and genomic DNA clones were screened from cDNA libraries and a genomic libraries. The mouse RP cDNAs were completely sequenced and the amino acid sequence was deduced and compared with the human RP. The evolutionary implication was elucidated from the genomic structure. DNA constructs for gene-targeting animal models were prepared for investigating the biological function of the RP gene.

F. Study the genetics of RP gene and its correlation to HLA-linked diseases

Sequences of human and mouse RP genes were analyzed to elucidate the genomic diversity and plasticity in this MHC locus and to propose putative genomic exchange events. Subcloning and sequencing of a unique RP rearranged fragment from a HLA-associated tumor cell line were performed.
Figure 2
A graphic representation of linkage disequilibrium between HLA-B, HLA-DR. Each HLA-B and DR specificity is represented by a row and column, the height or width of which is proportional to allele frequency in the population of normal chromosomes. The extent to which each HLA-B-DR box is filled with a specific symbol (complotype) reflects the three-point linkage disequilibrium. Each complotype (BF, C2, C4A, C4B) is represented by a different symbol as shown at the right of the figure (Adapted from Awdeh et al., 1983). For instance, majority of the individuals bearing SCO1 (HLA class III, shown as triangle symbols) have B8 DR3 haplotype. F and S alleles are the subtypes of the BF gene, while B and C alleles are two varieties for the C2 gene. Different typings for C4A and C4B are indicated as numbers.
Figure 2
Haplotype analysis of unrelated patients with type I diabetes.

The haplotypes are arranged by HLA-DR types. For those haplotypes (labeled as A to J) which link with HLA class III are Extended Haplotypes. Other haplotypes (other than extended haplotypes) are shown as cross-hatched bars. Specific haplotypes are shown as open area and letters are referred to the upper right portion of the figure. The HLA class III designation (i.e., SCO1) refers to Figure 2. The left-hand bars in each set represents patient haplotypes. The middle bar are family normal controls and the right-hand bars represent overall caucasian control chromosomes. Noted that the increase among patients of DR3 and DR4-bearing and the decrease of DR2-bearing haplotypes is evident (adapted from Raum et al., 1984).
<table>
<thead>
<tr>
<th>Antigen</th>
<th>Patients %</th>
<th>Controls %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ankylosing spondylitis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B27</td>
<td>89</td>
<td>9</td>
</tr>
<tr>
<td>Idiopathic hemochromatosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>72</td>
<td>28</td>
</tr>
<tr>
<td>Insulin dependent diabetes mellitus (IDDM)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B8</td>
<td>40</td>
<td>21</td>
</tr>
<tr>
<td>B15</td>
<td>22</td>
<td>14</td>
</tr>
<tr>
<td>DR3</td>
<td>52</td>
<td>23</td>
</tr>
<tr>
<td>DR4</td>
<td>74</td>
<td>24</td>
</tr>
<tr>
<td>DR3 or DR4</td>
<td>93</td>
<td>43</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR4</td>
<td>68</td>
<td>25</td>
</tr>
<tr>
<td>Celiac disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR3</td>
<td>79</td>
<td>22</td>
</tr>
<tr>
<td>DR7</td>
<td>60</td>
<td>15</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR2</td>
<td>51</td>
<td>27</td>
</tr>
<tr>
<td>Narcolepsy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR2</td>
<td>100</td>
<td>22</td>
</tr>
<tr>
<td>DR2*</td>
<td>100</td>
<td>34</td>
</tr>
</tbody>
</table>
Figure 4

Diagram of biosynthetic pathways for steroids in adrenal cortex

The name of each steroid is shown below a structural diagram of that molecule. The enzyme that is responsible for each conversion is accompanied by an arrow (adapted from New et al., 1983). DHEA, dihydro-epiandrosterone; 3β-HSD, 3β-hydroxysteroid dehydrogenase
Figure 4
Table 2
The diseases association with Complement C4 null alleles

<table>
<thead>
<tr>
<th>Diseases</th>
<th>Frequency(%)</th>
<th>Control(%)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autism</td>
<td>58</td>
<td>27</td>
<td>Warren et al., 1991</td>
</tr>
<tr>
<td>Bacterial Meningitis</td>
<td>10.9</td>
<td>3.1</td>
<td>Rowe et al., 1989</td>
</tr>
<tr>
<td>Biliary Cirrhosis</td>
<td>72</td>
<td>34.5</td>
<td>Manns et al., 1991</td>
</tr>
<tr>
<td>Brazilian paracoccidioidomycosis</td>
<td>3/69</td>
<td>2/223</td>
<td>Messias et al., 1991</td>
</tr>
<tr>
<td>Early-onset autoimmune hepatitis</td>
<td>13/24</td>
<td>12/90</td>
<td>Scully et al., 1993</td>
</tr>
<tr>
<td>Felty's syndrome</td>
<td>50</td>
<td>20</td>
<td>Clarkson et al., 1990</td>
</tr>
<tr>
<td>Herpes gestationis</td>
<td>38/42</td>
<td></td>
<td>Shornick et al., 1993</td>
</tr>
<tr>
<td>Juvenile dermatomyositis</td>
<td>13/20</td>
<td>SCO1</td>
<td>Robb et al., 1988</td>
</tr>
<tr>
<td>Myositis</td>
<td>47</td>
<td>25</td>
<td>Moulds et al., 1990</td>
</tr>
<tr>
<td>Prolonged asymptomatic after HIV-infection</td>
<td>58</td>
<td>33</td>
<td>Cameron et al., 1990</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>7.89</td>
<td>1.21</td>
<td>Rudduck et al., 1985</td>
</tr>
<tr>
<td>Sjogren's syndrome</td>
<td>10/28</td>
<td>1/63</td>
<td>Moriuchi et al., 1991</td>
</tr>
<tr>
<td>Vitiligo</td>
<td>17</td>
<td>2</td>
<td>Venneker et al., 1992</td>
</tr>
</tbody>
</table>
CHAPTER II
STRUCTURE AND GENETICS OF THE PARTIALLY DUPLICATED GENE RP LOCATED IN THE HLA CLASS III REGION

INTRODUCTION

Autoimmune, genetic and malignant diseases are associated with the major histocompatibility complex (MHC) in humans (also known as the HLA) (Reviewed by Porter, 1983; Alper, 1991). This may be attributed to (i) the presence of disease susceptibility genes, oncogenes or tumor suppressor genes in the MHC; or (ii) the variable efficiencies of the MHC class I, class II and possibly class III molecules for antigen presentation. Indeed, more than 36 new genes have been identified in the MHC and many of those gene products have been found to be involved in important cellular processes (Trowsdale et al., 1991). Characterization of these novel genes and identification of their biological functions are essential to understand the molecular basis of MHC-linked diseases.

When the promoter regions of the C4A and the C4B genes were characterized, a polyadenylation (poly-A) signal, AATAAA, was found located 633 bp upstream of the transcriptional initiation sites of each C4 gene (Wu, L.C. and colleagues, manuscript in preparation). Each of these poly-A sites is followed by a stretch of GT-rich sequence that is a characteristic feature for the 3' end of a mammalian gene (Proudfoot, 1991). A 655 bp BstEI-AccI fragment, which is 274 bp upstream of the major transcriptional start site of the C4A gene, was used to screen cDNA libraries. This DNA fragment was also used as a probe for Northern
blot analysis to investigate the expression of RP transcripts isolated from cell lines of different origins (Figure 5). This probe hybridized to a message of 16-1.8 kb in size from RNA samples isolated from the liver, hepatoma cell line HepG2, lymphocytic cell lines MOLT4 and RPMI 8402, and monocytic cell line U937. The newly identified gene was named RP to memorialize the late Professor Rodney Porter.

A molecular map of the MHC class III region has been generated by linkage studies using pulse field gel electrophoresis and cosmid cloning (Carroll et al., 1984 and Kendall et al., 1990). Genomic cosmid clones containing RP and C4A genes and partial cDNA clones for RP had been isolated in our laboratory. The purpose of this study was (1) to determine the complete sequence of the RP cDNA in order to deduce its amino acid sequence, (2) to determine the genomic sequence of human RP in order to deduce its exon-intron structure and (3) to investigate the molecular genetics of human RP.
MATERIALS AND METHODS

Isolation of RP cDNA clones

A 655 bp BstEII - AccI human genomic DNA fragment was obtained from a 1.3 kb BstE II subclone of cos 3A3 (Carroll et al., 1984 and Yu, 1991) and labeled by multiprime labeling method (Feinberg et al., 1984) with a USB kit and α32P-dCTP (Amersham). A λ gt10 cDNA library made from the human T-cell line RPMI 8402 was a gift from Dr. Terry Rabbitts (MRC Laboratory of Molecular Biology, Cambridge). Two other cDNA library screening were performed: U937 λ gt10 library made from human monocyte cell line (Clontech), and J6 λ gt10 library (T-cell line, kindly provided by Dr. Lai-chu Wu).

The screening of cDNA libraries were performed by using standard protocols (Huynh et al., 1985). The λ cDNA library was plated out in ten 150 mm petri dishes, each containing 10^5 phages. E. coli strain C600-Hfl was used as bacterial host. After overnight incubation at 37°C, the phage plaques were lifted onto the master and replica Hybond N membranes (Amersham). Twenty filters were denatured in 1.5 M NaCl/0.4N NaOH solution, and neutralized with 1.5 M NaCl/0.5 M Tris pH 7.4. To fix the DNA to the filters, membranes were irradiated with UV by crosslinker, following manufacturer's recommended conditions (Stratagene). The filters were incubated with Prehyb solution (50% formamide, 5X SSPE, 5X Denhardt's solution, 0.1% SDS, 100 μg/ml denatured salmon sperm DNA) at 42°C for 2 hours with rotation. Appropriated DNA probes were denatured by boiling at 100°C for 10 minutes, chilled quickly on ice. The denatured probe was added to the hybridization solution. The filters were incubated in hybridization solution at 42°C for 16-20 hours. The filters were first washed in 2X SSC, 0.1% SDS solution at room temperature for 2 x 20 min, followed by 0.1X SSC, 0.5% SDS solution at 65°C for 2 x 20 min. Autoradiography was performed and the positive clones were rescreened for three times.
Determination of RP cDNA sequences.

Determination of RP cDNA sequence from clone R1.1 was achieved by shot-gun cloning of randomly sonicated DNA fragments into M13 Smal cut and phosphatased vector, and ssDNA dideoxy sequencing (Bankier et al., 1989 and Messing et al., 1989). Gel readings were assembled with Staden's DNA Analysis software (Staden, 1982).

Synthetic PCR primers

Oligonucleotides were synthesized by Applied Biosystem Model 390B DNA Synthesis machine. Sequences of PCR primers used in this study are listed below. The relative positions of primers 1 to 5 in the RP1 gene (Figure 10) are numbered in parenthesis (with reverse orientation primers abbreviated as r.o.). Sequences added to the primers to facilitate cloning are in lower cases.

1. **RP5**  
   5' AAG AGG ATC CGA CTC CAC AGG CCC 3' (2,423-2,400, r.o)

2. **1.6K-F1**  
   5' CTC TGG GCC CGA GCG TTC 3' (908-925)

3. **1.6K-F2**  
   5' CGT CAG CAG TTT TGT CAG GTG CCC 3' (1,009-1,032)

4. **0.8K-F1**  
   5' TCC TCC AAA TGC AGT GAG GT 3' (1,795-1,815)

5. **HRP3**  
   5' tcc gaa TTC ATG TCT CTG GCA GGC G 3' (10,964-10,946, r.o.)

6. **YMRi**  
   5' agg gaa TTC AGG GGC CTC TGG GGC TAA CTC 3'

Determination of the 5' sequence of RP cDNA

Total RNAs were isolated from cultured cells MOLT4 (T leukemia cell line) and from HT29 (colon carcinoma cell line, kindly provided from Dr. Sue O'Dorisio) by guanidine isothiocyanate lysis and CsCl ultracentrifugation (Ausubel et al., 1987). The purity and quantity of RNAs were determined by the spectroscopy analysis. Reverse transcription (RT) was carried out using oligo-dT as the first primer and 1-5 μg of RNA according to the Perkin Elmer Cetus
RNA PCR protocol (Kawasaki et al., 1990). Amplification of RP cDNA was achieved by two rounds of PCR. The first PCR was performed using primers corresponding to the 3' end of the RP cDNA (HRP3) and 5' end of RPI gene (1.6K-F1 or 1.6K-F2). Approximately 10% of product from the first PCR was used for the second PCR using "nested" RP primers, 1.6K-F2 and RPR5, and 0.8K-F1 and RPR5, corresponding to the 5' end of the RPI gene (1.6K-F2 and 0.8K-F1) and to RPI Exon 3 (RPR5), respectively. The 1.6K-F1 and 1.6K-F2 sequences were assumed to be present in the RPI transcript by the presence the Kozak consensus for an initiation codon (Kozak, 1987). PCR conditions were: 1 cycle at 94 °C for 5 min; 30 cycles at 94 °C for 1 min, 54 °C for 1 min and 72 °C for 1 min; and 1 cycle at 72 °C for 10 min. The PCR products were cloned into TA cloning vector (Invitrogen) and sequenced.

Sequence determination of RPI and RP2 genes

A) RPI gene

Restriction fragments containing the entire RPI gene were subcloned from cos 3A3 and designated pSH13 and pBH4. The majority of DNA sequences for the RPI gene were obtained by shotgun cloning of sonicated DNA fragments into M13mp18 or into Bluescript KS vectors (Stratagene, LaJolla, CA) and ssDNA dideoxy DNA sequencing (Bankier et al., 1989). Sequenase kit (USB, Cleveland OH) and 35S-dATP were employed for sequencing reactions. Gel readings were compiled with Staden's DNA sequence analysis softwares (Staden, 1982). Gaps in sequence contigs were filled by further subcloning of the appropriate restriction fragments and primer walkings using new primers based on known DNA sequences for sequencing reactions. Sequence contigs were joined together through sequence determination of PCR-amplified DNA fragments overlapping the junctions. Overall, each nucleotide was determined more than 3 times and confirmed by sequences
from both strands.

B) RP2 sequence from λ JM-2a

A 5.3 kb Taq I fragment corresponding to the upstream region of a C4B5 gene from λ-JM2a (haplotype C4A4 C4B5, Yu et al., 1986) was subcloned into pUC18 vector. From this plasmid, a 2.1 kb BamHI - Taq I fragment was sequence-determined after shot-gun cloning into M13mp18 SmaI cut vector. Gel readings were assembled by Intelligenetics PC Gene softwares.

Southern blot analysis

About 5-10 μg of genomic DNA was digested to completion with appropriate restriction enzymes, resolved in a 0.8% agarose gel. The DNAs in the gel were denatured in 0.5M NaOH/1M NaCl for 30 min. Following the neutralization washing with 1M Tris-Cl/0.6M NaCl for 40 min, the gel was soaked in 10X SSC for 20 min and blotted to Hybond N membrane using 10X SSC (Southern, 1975). The membrane was hybridized with α32P-dCTP labeled R1.1 probe (Random Primed Labeling Kit, USB) and autoradiographed.

PCR of genomic DNAs

Locations of possible breakpoints for Gene XA and RP2 hybrid regions in primate and mouse genomes were determined by PCR (Saiki, 1990) with primers YMR1 and HRP3. For each reaction, 500-1,000 ng of genomic DNA and ~250 ng of each primer were used. Conditions for PCR were: 1 cycle at 94°C for 5 minutes; 30 cycles at 94°C for 1 minute, 54°C for 1 minute, and 72°C for 1 minute and 15 seconds; and 1 cycle at 72°C for 10 minutes.
Protein and DNA sequence analyses

Comparison of the RP amino acid and DNA sequences with national databases were performed by GCG FASTA program from Pittsburgh Supercomputer Center. Hydrophobicity and Acid-Base plots of RP protein sequence were achieved with DNA Strider. A dendrogram of the RP variable number tandem repeats (VNTR) were performed by PC Gene Program Clustal. Other sequence analysis programs used included DOTPLOT, BESTFIT, PILEUP PRETTY, and PUBLISH of the GCG package (Genetics Computer Group, Inc. 1990).
RESULTS

2.1 Isolation of cDNA clones for the novel gene RP upstream of C4

A cDNA clone R1.1 was screened and isolated from the RPMI 8402 library. Determination of the 1.1 kb DNA sequence of R1.1 revealed a 26 bp poly-A tail which is 16 bp downstream of the polyadenylation signal (Figure 6). The 3' sequence of the R1.1 cDNA clone is identical to the 5' regulatory sequences for the C4A and the C4B genes. This implies that there may be a pair of duplicated genes, designated RP1 and RP2, present immediately upstream of the C4A and the C4B genes, respectively.

The R1.1 cDNA was used as a probe for Northern blot analysis to investigate the expression of RP transcripts isolated from cell lines of different origins. This probe hybridized to a message of 1.6-1.8 kb in size from RNA samples isolated from the liver, hepatoma cell line HepG2, lymphocytic cell lines MOLT4 and RPMI 8402, and monocytic cell line U937 (see Introduction). RP transcripts of similar size were also detected from Northern blot analysis of RNA samples from colon carcinoma cell line HT29 and neuroblastoma cell line IMR32. Thus, RP appears to be ubiquitously expressed with a major transcript size about 1.6-1.8 kb.

Since R1.1 contains the poly-A tail and is only 1.1 kb in size, the 5' region of the RP full length cDNA is missing in this clone. Determination of the RP 5' cDNA sequence was achieved by RT-PCR (Kawasaki, 1990) with RNAs isolated from HT29 and MOLT4, using PCR primers derived from the upstream sequence of cDNA clone R1.1, and from the 5' sequence of the RP1 gene. The coding sequence for RP1 is shown in Figure 6. There is an open reading frame coding for 364 amino acid residues. An in-frame stop codon is located 33 nt 5' to the putative initiation codon. Thus, it is likely that the predicted amino acid sequence for RP is complete.
The RP protein is extremely hydrophilic in its N-terminal half. There are several alternate hydrophilic and hydrophobic regions at the carboxyl half of the protein (Figure 7). Overall, there are 53 positively charged residues (Arg or Lys) and 37 negatively charged residues (Asp or Glu) in the RP protein. Thus, there is a net positive charge of 16 in the RP protein. In addition, there are eight histidine residues which may also be positively charged.

Located between residues 114-131 is a bipartite, positively charged nuclear localization signal (Robbins et al., 1991), KRHLIPPETFGVKKRRKR. Hence, the RP protein may be targeted to the nucleus. The RP protein is rich in Gly (9.6%), Pro (6.9%), and Leu/Ile (12.6%) residues. While most of the Pro and Gly residues cluster at the N-terminal portion, the majority of Leu/Ile residues are present at the C-terminal region.

There are no N-linked glycosylation sites in the predicted RP sequence. Search of RP for protein modification sites with the PC/GENE PROSITE program revealed that there is a single potential tyrosine kinase phosphorylation site at residue 231; five potential kinase C phosphorylation sites at residues 80, 105, 112, 145 and 277; five potential casein kinase II phosphorylation sites at residues 6, 76, 85, 183 and 277; two amidation sites at residues 61 and 323; and nine potential N-myristoylation sites at residues 25, 27, 65, 69, 70, 74, 104, 144, and 211. Whether the endogenous RP protein is post-translationally modified as revealed by these potential sites remained to be determined.

Two more cDNA libraries, monocyte U937 and T-Lymphocyte J6, were screened by using R1.1 as a probe. A clone with a ~800 bp insert was obtained from the U937 library, while a clone with a ~700 bp insert was derived from the J6 library. Partial sequences were determined for both clones and revealed identity with R1.1 sequence. However, the sequence derived from the J6 clone exhibited aberrant splicing patterns (see also the comparison with another RP cDNA clone,
RP410 in Figure 8). RP410 clone acquired portion of intron 3 sequence at the splicing junction of exon 4; and 10 bp of intron 5 sequence at exon 5 junction. The J6 clone, however, contained partial intron 4 sequence without the splicing at the exon 4 region (Figure 8).

2.2 Gene structure of RP1

The human RP1 gene located upstream of a C4A3 gene in a cosmid clone was subcloned into plasmids (Figure 9). The DNA sequence for the entire RP1 gene and the intergenic region between C4A and RP have been completely determined (a ~2 kb fragment between HindIII site to C4 in clone pBH4 was previously sequenced by Dr. L-C Wu). A sequence of 12,121 bp is shown in Figure 10. The RP1 gene consists of 9 exons (Figure 11). The 5' boundary for Exon 1 has not been defined precisely partly because of the multiple initiation sites of transcription (also see Chapter 4). The 3' end of Exon 1 was located because a continuous cDNA sequence spanning 191 bp for the 3' region of Exon 1 and about 70 bp for the 5' region of Exon 2 has been obtained. The putative initiation codon is 128 bp downstream of the splice junction of Exon 2. The coding capacity of the exons ranges from 14 amino acid residues (for Exon 9) to 74 amino acid residues (for Exon 2). Exon 9 also contains a 3'-UT region of 397 bp. The size of the introns ranges from 85 bp for Intron 2 to 6,136 bp for Intron 4 (Table 3).

There are several notable features of the RP1 gene. First, the 5' region of the RP1 gene is rich in CpG sequences. For example, there are 130 CpG dinucleotides from nucleotide 500-3,000, compared with 30 from nucleotide 8,500-11,000 (Figure 10). The CpG sequences are generally under-represented in mammalian DNA and their presence (at the 5' region of a gene) is strongly correlated with housekeeping genes (Bird, 1986). This would infer that RP is involved in an essential function.
Second, there are eight complete copies of Alu elements present in Intron 4 and one of these elements (copy number 6) is actually located within another Alu element (copy number 5). Among the eight Alu elements, copy number 1, 2 and 8 belong to the Alu-J subfamily that is similar to the 7SL DNA (Jurka et al., 1988). Copy number 4 appears to be an integral component of a composite retroelement and will be discussed below. The other four Alu elements belong to the Alu-S subfamily with copy number 6 categorized to the b branch and copy number 3 categorized to the c branch (Jurka et al., 1988). Alu-S/b is considered to be a young branch of the Alu family, which is consistent with our data as the incorporation of copy number 6 into the RP gene has to be after the presence of copy number 5. Copy number 3 has an atypical trimeric structure in contrast to the dimeric structure for most Alu elements. The additional structure was labeled Alu-3.1 in Figure 10.

Third, there is a stretch of highly repetitive DNA sequences located between nucleotide 5,717 to 6,579 (Figure 10). These repeats can be illustrated by multiple diagonal lines in a dotplot analysis (Figure 12C). There are 21 copies of very similar but non-identical, tandem repeats, each of which has a GC content of 72-84% and a size of 35-45 bp. A dendrogram showing the relatedness of these 21 tandem repeats is presented in Figure 12A. These 21 copies of tandem repeats together are flanked by hexameric sequences, TGGGCA (boxed in Figure 10). The presence of these hexamers precisely at both ends of the 35-45 bp tandem repeats suggests that these tandem repeats as a whole might exist as a structural unit in a transposition/retroposition event resulting in the generation of the hallmark signal, direct repeats (Li et al., 1991).

2.3 A composite retroposon is present in the RP1 gene

Comparison of the tandem repeat sequences in the RP1 gene with the
GenBank database through the GCG Fasta program revealed striking similarities with a group of nonviral retroelements, SINE-R11, -R14 and -R19 (Ono et al., 1987), but arranged in the reverse orientation. Members of the SINE-R elements contain two basic components: (a) 3-6 copies of variable number tandem repeats (VNTR) which are highly similar to those 35-45 bp tandem repeats present in the RP1 gene, and (b) a short interspersed repeat element (SINE) of ~490 bp that is homologous to the genomic region between the env gene and the 3' long terminal repeat (LTR) of an endogenous retrovirus HERV-K10 (Ono et al., 1986). The SINE sequence in RP1 is only 135 bp in size. Most of the sequence corresponding to the HERV-K10 LTR region in the SINE-R retroposons is absent in the RP1 gene. Immediately preceding the SINE element in RP is a stretch of T residues of 16 nucleotides.

In addition to the RP1 gene, the SINE-R related sequences have also been detected in introns of two other human genes, the complement C2 (also known as SINE.R-C2) (Zue et al., 1992) and the cytochrome P450 CYP1A1 (Kawajiri et al., 1986) genes. The copy number for the VNTRs varies from sixteen in CYP1A1, seventeen in complement C2 (B allele), twenty-one in RP1, to twenty-three in another complement C2 gene (A allele). Analysis of the RP, C2 and CYP1A1 genomic sequences reveals a more complex organization of the reiterative sequences than those of SINE-Rs. Located immediately downstream of the VNTRs in each gene is a highly conserved region of 370-372 bp (Figure 13) with sequence identities of about 95%. Present in each of these sequences are three stretches of Alu-related sequences of 25 bp, 54 bp and 246 bp (Figure 13) and a less well-defined sequence of 32 bp. One of the Alu-related sequences (i.e. 246 bp; Alu copy number 4 in Figure 10) appears to be an entire Alu element and is flanked by a pair of target site repeats. This particular Alu element in RP1, C2 or CYP1A1 genes is unusual in that there are many conserved changes such as three deletions
at nucleotides 1-22, 200-210, and 281-297; two mini-insertions at nucleotides 121-122 and 258-262; and 34 scattered point mutations, when compared with the consensus Alu sequence (Jurka et al., 1988) (Figure 14).

For the case in RP1, the poly-T track, SINE-R, 21 copies of VNTRs, and the described ~370 bp Alu-related sequences are enclosed by a direct repeat sequence of 13 bp, GATAATTCCACTA (boxed in Figure 10). Homologous organization enclosed by a pair of direct repeats of 18 bp is present in the complement C2 gene. Hence, these retroelements appear to form a family of retroposons with discrete, composite units (i.e. SINE, VNTRs and Alu) proliferated in the human genome. This composite retroposon is named SVA in light of its composition. Part of the DNA sequence for the SVA retroposon in the CYP1A1 gene is not available but the existing data reveal a structure very similar to the SVA-RP and SVA-C2 (Figure 13).

2.4 RP gene is duplicated in most human and primate genomes

Since RP sequences are present upstream of the complement C4A and the C4B genes, it infers that there may be two copies of RP genes in a haploid genome. A Southern blot analysis of BamHI digested genomic DNAs, which were isolated from human peripheral blood lymphocytes from CAH patients (lanes 1-3), Prader Willi patients (lanes 4 and 5), a nasopharyngeal patient (lane 6), normal individuals (lanes 7, 8 and 10), human tumor cell lines (lanes 9 and 11), an African green monkey (AGM) cell line (COS7) and a cotton top tamarin cell line (NPC/LC), using a RP cDNA probe (R1.1) is shown in Figure 15. Two distinct RP specific, BamHI fragments of 9.6 kb and 5.0 kb in size were detected in most human samples, but only a single 9.6 kb fragment was detected in the genomes of CAH-E1 (lane 1) and of SC01 (lane 10). CAH-E1 was a congenital adrenal hyperplasia patient with homozygous deletion of the CYP21B genes (manuscript in
preparation). SC01 was a normal individual who was typed as HLA B8 DR3 C4AQ0 C4B1 (there is a homozygous deletion of C4A genes in this individual). Subsequent restriction mapping and DNA sequencing data revealed that the 9.6 kb \textit{BamHI} fragment corresponds to the RP1 gene, while the 5.0 kb \textit{BamHI} fragment corresponds to the RP2 gene. Two RP specific \textit{BamHI} fragments of 9.3 kb and 4.7 kb were detected in AGM (lane 12), but a single fragment of 10 kb was detected in cotton top tamarin (lane 13). These results suggest that the RP genes are duplicated in the majority of the human population, but in some individuals only the RP1 gene is present. They also infer that there may be two RP genes in an Old World monkey AGM but a single RP gene in a New World monkey cotton top tamarin. This same conclusion was obtained from a genomic Southern blot analysis of \textit{TagI} digested DNAs with the R1.1 probe that two copies of RP genes were present.

\section*{2.5 Partial gene duplications of RP and Gene X}

Located in the approximately 12 kb intergenic region between C4A and C4B are the CYP21A pseudogene, RP sequence, and Gene XA which overlaps CYP21A at the 3' ends. CYP21A is about 3.2 kb in size and located 3.0 kb downstream of the C4A gene; thus, Gene XA and the RP2 gene are localized in a region of 6 kb. This observation appeared paradoxical as the size of the RP1 gene is about 11.5 kb (Figure 10), while that of the Gene XB may be as large as 70 kb (Matsumoto et al., 1992a and 1992b).

In order to solve this puzzle, a 12 kb \textit{BgIII} fragment was subcloned from \textit{cas} 4A3, which corresponds to the intergenic region between two C4B genes in an unusual haplotype C4A2 C4B1 C4B2 (Carroll et al., 1984) (Figure 16A). The RP2-specific 5.0 kb \textit{BamHI} restriction fragment (Figure 15) is located in this subclone, which was completely sequenced by shot-gun cloning and dideoxy
sequencing. A comparison of this sequence with those for RP and Gene X cDNAs and the RP1 gene reveals a hybrid sequence derived from RP and Gene X (Figures 16B and 17). Specifically, this 4,971 bp sequence contains a 1,566 bp fragment corresponding to part of the 5' UT region of a C4 gene (42 bp), the RP-C4 intergenic region (611 bp) and Exon 7- Exon 9 of the RP1 gene (913 bp) which is fused to a 3,405 bp fragment corresponding to the 3' region of Gene X (Figure 16B). The 5' ends for both RP and Gene X in this hybrid region are truncated and therefore duplications for these two genes are incomplete. With respect to the RP1 gene, the breakpoint of gene duplication for the RP2 sequence is located in Exon 7 (Figure 17) and is 2,093 bp downstream of the Alu clusters and SVA element. A DNA sequence corresponding to Gene X cDNA is found 795 bp upstream of RP2. Nine hypothetical Gene X exons, Exons a to i can be deduced based on this genomic sequence and the published cDNA sequence (Figure 16B) (Morel et al., 1989). Gene XA is arranged in the opposite orientation with respect to RP, C4 and CYP21 genes. The first 332 bp of the published Gene X (partial) cDNA sequence is absent in the Gene XA sequence. In addition, there is an internal deletion of 91 bp in the hypothetical Exon e. The truncation of the 5' region and the internal deletion may change the reading frame and result in premature termination with respect to Gene XB cDNA (Morel et al., 1989). It remains to be determined whether the changes in Gene XA would lead to the generation of a new gene product. To date the exon-intron structure of the presumed intact Gene XB has not been fully elucidated and therefore it was not possible to name the hypothetical Gene XA exons with respect to those of Gene XB. An independent study by Gitelman and colleagues has shown that the 795 bp DNA sequence upstream of the first hypothetical exon of Gene XA (referred to here as Exon a) corresponds to intronic sequence of Gene XB (Gitelman et al., 1992). Thus, the Gene XA-RP2 hybrid was formed by a recombination at a Gene X intron and Exon 7 of RP.
To determine if there is a common breakpoint for gene duplication of Gene X-RP-C4-(CYP21), a 2.1 kb \textit{TaqI-BamHI} restriction fragment from \textit{AJM-2a} spanning the RP2 sequence and the 5' region of a C4B5 gene of the C4A4 C4B5 haplotype (Yu et al., 1986) was determined (Figure 16C). This fragment covers the entire 913 bp RP2 sequence and also 573 bp of the Gene X sequence and its sequence is identical to the corresponding region obtained from \textit{cos 4A3}, except for the presence of two point mutations. In other words, the breakpoint of the hybrid Gene XA-RP2 in a C4A4 C4B5 haplotype is identical to that of the C4A2 C4B1 C4B2 haplotype. (Further analysis of the polymorphism of Gene XA-RP2 sequences will be described in Chapter 6).

Thus, the genomic region between C4A and C4B genes contains pseudogenes or truncated sequences for three different genes, i.e., CYP21A, Gene XA and RP2. The tandem genes for RP, C4, CYP21 and Gene X appear to form a four-gene module RCCX that may be duplicated together in the MHC class III region (Figure 18). However, duplications for the flanking RP and Gene X are incomplete. Homozygous deletions of RP2 in individuals CAH-E1 and SC01 were concurrent with Gene XA. This is because the 5.0 kb \textit{BamHI} restriction fragment containing Gene XA-RP2 sequences were absent in these individuals (Figure 15). We have also found homozygous deletions of a C4 gene and a CYP21 gene in the genomes of CAH-E1 and SC01. In other words, these individuals have single RCCX modular structures (manuscript in preparation).
2.6 A common breakpoint region for duplication of the RCCX modules in great apes

To determine if the RP-C4-CYP21-Gene X modules are duplicated with a Gene XA-RP2 hybrid region in humans and apes, PCR was performed with a set of primers corresponding to the RP (Exon 9) at one end and to Gene X at the other end (Figure 16D). As shown in Figure 19 (Panel A), a 1.36 kb fragment was amplified from cosmid DNA (cos 5) that spans a long C4A and a short C4B gene (lane 1), from human genomic DNAs with RCCX bimodular structures, e.g., Raji (lane 3) and HepG2 (lane 4), and from chimpanzee (lane 5) and orangutan genomic DNAs (lane 6). Southern blot analysis of the samples shown in Panel A using a R1.1 probe confirmed that the 1.36 kb fragment contained RP-specific sequence (Panel B). Thus, there is a common breakpoint region for gene duplication of the RCCX modules in the great apes. On the other hand, no amplified products were detected from human genomic DNA, CAH-E1, with a single RCCX modular structure (lane 2), or from mouse genomic DNA (lane 7 and see Chapter 5). The former was expected because the corresponding PCR primers in a single modular haplotype are oriented in a head-to-head configuration, located 20-30 kb apart and therefore could not be amplified by PCR. The breakpoint of gene duplication for mouse RP is undetermined but available sequence data corresponding to the 5' regions of the C4 and the Slp genes exclude the possibility for an identical breakpoint as in humans and apes.
DISCUSSION

Here we report the identification, cloning and characterization of the novel gene RP located 611 bp upstream of the human complement component C4A and the C4B genes in the class III region of the HLA. The unusual modular duplication (and deletion) of RP together with its neighboring genes Gene X, complement C4, and steroid 21-hydroxylase CYP21, and the association of the HLA with autoimmune and genetic diseases motivates an intensive investigation on the structure, genetics and function of RP.

Although the deduced amino acid sequence of RP does not reveal striking similarities to any known proteins, it sheds light on the properties and possible function of this ubiquitously expressed molecule. The presence of a bipartite nuclear localization signal suggests RP may be a nuclear protein. The highly hydrophilic and basic nature of the protein infers that the protein might interact with negatively charged molecules such as DNA or acidic domains of transcriptional factors. A comparison of the RP protein sequence with other protein sequences in national databases revealed that the N-terminal portion of 157 residues in RP is 22.9% identical to the hypothetical 119.5 kDa uvr-A protein in bacteria Micrococcus (Shiota et al., 1989), while the carboxyl portion of RP is about 20% identical to the yeast RAD7 protein (Perozzi et al., 1986). Both uvr-A and RAD7 are involved in the DNA repair mechanism. Similar to RP, a bipartite nuclear localization signal and a leucine-rich region are present in the RAD7 sequence (Schneider et al., 1991). Mutation of the RAD7 gene in yeast resulted in decreased proficiency of excision repair of DNA damaged by UV light (Perozzi et al., 1986). While analogs for many of the components involved in the DNA repair mechanism of yeast have been isolated, the human analog for the yeast RAD7 has
not been cloned.

Immunological disorders such as systemic lupus erythematosus (SLE) (Atkinson, 1986 and Fielder et al., 1983), immunoglobulin IgA deficiency and common variable immunodeficiencies (Schaffer et al., 1989 and Wilton et al., 1985), and malfunctions in reproduction such as recurrent spontaneous abortions (Laitinen et al., 1991) have been related to the null alleles of C4A. In this case null alleles of C4A imply a gross deletion of a C4A gene together with other genes in the RCCX module, or mutations of the C4A gene that may also be concurrent to RP, CYP21 or Gene X. A typical example for the latter can be found in a HLA B44 haplotype where the conversion of a C4B gene to C4A in the second C4 locus was concurrent with mutations of the CYP21B gene (Yu et al., 1987 and Rodrigues et al., 1987). The diversities of many disorders correlated with the C4A null alleles infer deficiencies of different genes in the close proximity of C4A, and/or the presence of a malfunctioning gene with widespread functional properties.

The mouse also contains RP genes upstream of the C4 and the Sip genes, although the breakpoint of gene duplication or deletion for RP appears to be different from that in humans (Shen and colleagues, manuscript in preparation). Whether both RP genes in the mouse are functional is yet to be determined. It was shown that a crossover at the C4-CYP21 region led to the lethality of homozygous embryos (Shiroishi et al., 1987), which suggests the presence of an essential gene at the region of crossover. On the other hand, breeding experiments for congenic rats revealed the existence of a growth and reproduction complex (grc) with several genes closely linked to, if not present in, the MHC. One of the genes in the grc has been inferred to be a tumor suppressor gene (Melhem et al., 1993 and Kunz et al., 1980). Our zoo-blot experiment suggested that there are two copies RP genes in the haploid genome of rat (unpublished). The physical location and the structural information together suggest that genes of the RCCX modules could be related to
The concept for the modular organization of the C4 and CYP21 genes was first suggested by Klein and colleagues (Kawaguchi et al., 1991). This study extends the concept of the modular gene duplication to include the genes flanking C4 and CYP21, RP and Gene X. In a RCCX bimodular (or trimodular) structure, the breakpoint of the four-gene duplication is present at Exon 7 of the RP1 gene and an intron of the Gene XB. This resulted in the complete duplication of a C4 gene and a CYP21 gene, but only partial duplications of RP and Gene X. The truncated sequences, RP2 and Gene XA, form a chimeric hybrid at the intergenic region of the two C4 genes. This modular duplication pattern involving four structurally and functionally unrelated genes is unusual. Partial gene duplication has been suggested to be one of the major mechanisms leading to genetic diseases (Hu et al., 1992). This is because a partially duplicated DNA sequence may mutate at a faster rate and the deleterious mutations can be incorporated into the functional gene through recombinations or gene conversions, as observed in the CYP21B genes (Miller et al., 1989). Thus, the Gene XA-RP2 hybrid sequences could play a role in disrupting the gene function and in the genetic instabilities of the RCCX modules in the population.

Although the bimodular structures of RCCX are prevalent in the population, the single modular structures account for 10-30% of the human genomes (Schneider et al., 1986). Invariably, the single modular structures consist of the intact RP1 and Gene XB loci. In genomes with single RCCX modules, the absence of CYP21B leads to CAH, while the absence of C4A is a predisposing factor for SLE. Thus, it is important to understand the mechanism leading to the deletion of genes of the RCCX modules.

In many situations a repetitive element such as an Alu element or an endogenous retrovirus was found at or proximal to the breakpoint of DNA
rearrangements (Hu et al., 1992). In the RCCX modules, a cluster of Alu elements and a composite retroposon SVA with 21 copies of VNTRs are present at Intron 4 of the RP1 gene, which is 2,093 bp upstream of the corresponding breakpoint of RP gene duplication. Notably dimeric sequences for Alu elements have not been reported in the C4A, CYP21A, Gene XA, RP2, C4B and CYP21B, a genomic region more than 50 kb in size.

Elucidation of the composite retroposon SVA was the result of a deliberate sequence comparison with DNA sequences in the GenBank database. In contrast to a simpler retroposon SINE-R, the composite retroposon SVA in the RP1, C2 or CYP1A1 genes contains 16-23 copies of unusually GC-rich VNTRs and also additional sequences with an Alu element characterized by distinct deletions and mutations among SVA retroposons. The SINE and Alu elements are arranged in the opposite, head-to-head configurations. Since possible gene products of the SVA have not been defined at this stage, the sense DNA strand of SVA cannot be specified with confidence. However, the presence of multiple T residues at one end of the SVA could reflect the presence a poly-A structure similar to an mRNA that was reversely transcribed and subsequently incorporated into the human genome. If this were the case, the SVA would be orientated in the opposite direction with respect to the resident genes RP, C2 and CYP1A1. The striking similarities in the organization and high sequence identities of SVA among the three genes suggest that SVA may be a recently evolved retroposon. All three SVA elements described above are located in an Alu-rich region. For example, the SVA-RP is present within an Alu cluster and the entire repetitive DNA region spans 4.4 kb in size. The SVA-C2 and SVA-CYP1A1 have acquired an additional structure with 7 copies of hexameric sequences located immediately after the SVA-specific Alu element (Figure 13). The SVA-RP contains only 135 bp of the 490 bp SINE element present in SINE-Rs. The missing region consists of a responsive element
for glucocorticoids (Ono et al., 1986 and 1987) and therefore the gene activity of RP may not be induced by these steroids. Whether the SVA element and its GC-rich VNTRs play a role in the function of RP, or in the unusually frequent RCCX modular variations such as gene duplications and/or deletions and polymorphisms, remains to be determined. Only three SVA retroposons have been elucidated so far but two of them (i.e. SVA-RP and SVA-C2) are localized about 20 kb apart in the polymorphic HLA class III region. It is also of considerable interest to note that about seven copies of the SVA-related VNTR sequences are found close to the meiotic recombinational breakpoint of the HLA DQB1 gene in the DR7 DQw2 haplotype (Satyanarayana et al., 1992).
Figure 5
Northern blot analysis of human RP.
Total RNAs isolated from human liver tissue (lane 1) or from cultured cell lines HepG2 (lane 2), U937 (lane 3), U937 treated with phorbol ester PMA (lane 4), RPMI 8402 (lane 6) and poly-A⁺ RNA from MOLT4 (lane 5) were resolved by formaldehyde agarose gel electrophoresis, blotted to Hybond N membrane and hybridized with R1.1 probe.
Figure 6

cDNA and predicted amino acid sequences of human RP.
The first 521 bp was cloned by RT-PCR technique. The rest of the sequence was
derived from clone R1.1. The putative nuclear localization signal is *highlighted.*
The basic residues (K and R) are in **bold** and the acidic residues (D and E) are
*italicized.* The stop codon is asterisked.
Figure 7
Structural properties of the RP protein
(A) The hydropathy profile based on the method of Kyte-Doolittle method. Those regions with more hydrophobic residues are placed above the baseline at zero, while these regions are hydrophilic are indicated with negative values.
(B) The amino acid distribution of acidic and basic residues of the RP protein. Acidic residues are presented on top, Glu and Asp residues are indicated with full and partial bars, respectively. Basic residues are on the bottom. Arg, Lys, and His are shown as three bars with different sizes, from longest ones to shorter ones, respectively.
Figure 8
A comparison of human RP cDNA sequences from RP1.1, RP410, and J6 cDNA libraries
The deviations of the cDNA sequences between RP1.1 and other cDNA clones from J6 and RP410 libraries. The positions of aberrant splicing occurring at the exon-intron junction are indicated, and the partial sequences of RP1 introns are shown. The relative position is derived from Figure 6.
Figure 9
A restriction map of the human RP1 gene.
Plasmid subclones derived from cos 3A3 corresponding to the RP1 and the 5' region of the complement C4A gene are shown under the restriction map. DNA sequence for RP1 was obtained by sequencing to completion the consecutive 1) 1.6 kb BamHI, 2) 0.8 kb BamHI, 3) 7.5 kb BamHI-HindIII, and 4) 3.8 kb HindIII-BglII restriction fragments. DNA sequences obtained from shot-gun clones of randomly sonicated fragments are indicated as (sonicated). Overlapping of sequence contigs were achieved by sequencing PCR amplified DNA fragments across the junctions (shown by solid bars). Horizontal arrows represent the direction and location of transcriptional initiation sites of RP1 or C4A; a vertical arrow indicates the location of the RP1 poly-A site.
Figure 10
The complete DNA sequence of the human RP1 gene.
Amino acid sequences for each exon are translated under nucleotide sequences. Donor and acceptor sites of introns are underlined. Orientations of Alu and SINE are indicated by arrows. Putative transposition target sites (if identified) of Alu elements are in bold and underlined. These 21 copies of VNTRs are numbered V1 to V21 with the first nucleotide of each repeat high-lighted. Direct repeats flanking the VNTRs and also those flanking the SVA composite retroposon (SINE-VNTR-Alu) are boxed. An asterisk indicates the precise location of the feature described in the figure.
Figure 10-continued

GGTACAGAGTTTCATCTAGACAGGGGCTGAGAAGGATGACTGGTTTAGGACTAAC 1680
CCAGGCCCACCTGATCGCCGGAGCTCTGCTGCTTGAACATTGACCAGGGGCCAGGCCCC 1740
Exon 2.
GGCCAAAGCTCTCATCTGCCCTCCTTTTCTTTCCTACGACCACCTCCCTCAAGGTCTCTCC 1800
AAATGCAAGTGAGTTGAAAGACGCTCTGGCTCAGATGAAAATCAGTTACTCAAAAG 1860
CTCCCCAAACTTCACCTGCAGACTATGACATGGCAAGGACGCCAGAGGGCGGAAG 1920
Taq I
GATGCAAAAGTGGTTTCTCTGCTTGCAGTACATTCACGGACAGTGGCAGCGGCGGGAAG 1980
MQKWFSAFDDAIRIQQRWRAN
CCCCTCCGGGGCGGGGAGGTCTAGCTCAGAAGGATGTCAGGGGAGCATAGGGGGCCAC 2040
GAPPRRQRPVGRACPWRPI
CCGGGGCCCTCCAGCGCGCCAAAGCAGTCCCGCGCGGTGGCTGCTCCTGGGAAGGGAATAT 2100
RGRGARPGCGDAG(74)
CCCTCCGCACTGAGGGCTGAGACGAAAACTTGACTCTCTCTGCTCATACCTGG 2221
Exon 3.
TCTAGGAGGACGCCCGGGGACGCTATGCTCCTCGGGCCAGAAGACCCTATTTT 2281
(75)GTPGETVRHKCSAPEDPIF
CACGGTTCTTTTCTCCATTTACCTTCCCTCGCTCACGCTACTGGCGTACGATATAG 2341
RFSSLHSPFPGTIKSRDMS
CTGGAGAGGAGATACGTGACCTGGAGGCTTCTTGGACTTAAAGGGCCGGAGGGAAGGGG 2401
WKHRHPLPETFGVKRRRRKRG
BamHI I
GCCTGTGGAATGCGATCTCCCTCTGCGGGTGAGCCAGTACATGCAACCCCGGGGGGTGG 2461
PVESDPLREPG(144)
GTCTCTCGTCCGCCGGTAAGCCAGTGTGGTAAGGAACGCTGCTGCCTCTCGGT 2521
Exon 4
CCGGTGAGCCATTICCGGCAGGGCTGCTGCTCAGAGCATGCAACTGACGCC (145)SARAAVSEMLQL 2581
TGTTCCGCGGAGCCCTGTGGAGGAAGCGCTGCTGCCCACATGTCTGAGGACCCAGGTTGT 2641
FPRGLFEDALPPIVFLRSQVY
ACAGCCTGTCGCGGACAGCCGGGGCCGACGGGCTGCTGGTAC 2701
SLVPDRTVADRLQL(189)
EcoRI I
CCGGGCAAGCCCATTTCCTGAAATCTCCTCTCCTCCTGCTCACTGCTCTCCGCGACCC 2761
GTTGAGCAAGCATTAGGCTTTTCAGGGGAGGGCACTGCTCCAGGGCACAGGGGCAC 2821
AGACGTCTTAAAAAGCCCTCCACGCTAGTACAGGGGCAATTCAATTGAGGATAGAAGAA 2881
Taq I
CGAGGCTTTGCGGAAAGCGTCATGGGGGCATGCTGGAACATACCCCGGCTGGGAGGTGC 2941
AGTCAACGTCAGATCAGACGCCCTGTGACATGATGTTTTCCATCTTACACTGAGG 3001
Figure 10-continued

Alu (7) -

| Alu (8) |

GGCTTGAGCCATCTGCCCGGCCTACTTAGTTATATTAGTGTAATTCCCTGTTATCC 8041
TGTGAGCTTTTGTCTCTTAAGAGTTTTTTTTAAGAGATGGGCTCTAGCTGTGGCC 8101
AGTTGCAATCATATCTTACCTGAGCCCTCAAACTCTGGCGTCAAGTACTGTCTTGGCC 8161
GTTCCTCAATGAGCTTAGGACCATAGGTGTCTTGCGCCACGCTGGCTGTTTTTACATTTT 8221
TGGTAGAGATGTTCCCTGGGCTCTACCATGTTTGGGCACTGTCCTGAACTCC 8281
TTCTCAATTGATCTGACTCGCTCCCTCCAAAATGCTGAATTACAGGCTAGGCA 8341

Exon 5

AGCATCAACAGCAGGCGAGGGGAGATCGAGAATCGTCCAGCTGGGCTTCGACTTGGATGCCCAT 8881

GI I F T E D Y R T R (221)

GGGCGGACGAGGAAACGCGCCGCCTGGGGGAGGCCATACAGGACTCAGAAATCACGAGACTTGT 9001

QDQMTQTFGFRDSEIT (268)
Figure 10-continued

```
GGAGCCCAACAAAGTCAGGCATCTGCTCTCCCTGCTCCCCCTCTCCAGTTCCATCCAGCCT
9540
GTCCTCTGTTTTTGGTGAAAGCTGGAAGAAAAAGCTCCTTTTTTTTTTTTT
9600
TAATAAAATGACCACATGTTATTTCCCGAACAAGAGAGGCGCTATCTTTACTGGGG
EcoR I
9660
CCTAGGAAGAGCCCTGAATAGAATTCGAGTCTTGCTGCTATCAAAGACACAGGG
Hind III
9720
GGCTGGCAGGAAATCTGTAAAAAGCTGAATTTAAATCTCAATCAGTACGATCTCA
9780
GTTAGGAGGGATGATCCCTAAAGATCATCAGTTAAACAAAAAAGCAGACTAAA
9840
AAACACTGCGGACCATTTAATAAAAATCTCAAATTTTGAAGGCCTATCTGGCTCAAGG
9900
TCCTACCTTCTAACCAGGAAATGAAATGTAAAGAGATGGTGACATGGGCTACCTAG
9960
CATGGGACAGCAACAATCTCTCTGAGATCTTGTTGGGAGCAGTGGCCAGCTAGCA
10020
GTGGCAGAGCCGACAGGACTGCTGATCTCCTCCTCGGTCTCTAAAGCAGATGTGGCTCA
10080
GGGCAGCTGCTGGGAGAAGAAGCAGCTTTATCTGAGCTTACGAGGAACATGAGAGCA
10140
AAAGTGAGGGAGGGGACGTGACCTAGGAGGCGCTTTGAAAGACATGGCAGAAGAGG
10200
CATTGGTTGGGACAGCAAAACGACCGAGTGGTTATTAGTTATCCACGTCCTCTTGGC
10260
TGCTGCCTGGTTCTGGGACAGTTTGTGTAACATTAATGAGTACGAGGAGAGATATGGTC
10320
TCAGAGCCGCAGGCTCTGATCTCTGGCTGCTGCTTCCGACTGGCGCTGACATCTGGTGAATGCT
10380
(269) HLVNA
10440
GGAGTCCTCAGGGCCTCCGGAGCTCTGGGCTAGCTGCTGCTGAGCTGAGCCCATGGGAG
GVLTVRDAGSWSWLAVPGAGR
(* Breakpoint of RP2)
10500
TTGATCAAGCTTTTTTTAAAAGTCATCAGATCTGCTGAAGGCTAAGGCTACCCCTCAGCT
FITYFVKG(301)
10560
TTGTTGGCTCTCAGGGCGTCTAGGGCTTTATCAGCCTTTTTCCCCTCTCTGGCCAGTTCT
Exon 8
10620
CGTCTAGGGCGCCACGGTCTCCTAGGCTAGGAGCTGAGTTGCTTTCCCTGGCTCTGCC
(302) RQAVLSMVRKAKYRELL
10680
CTATGACGAGCTCTGGGCGGCGGCGGCGCTTCTGTGCTGGCTGCTGCTGGCCTACCTACAT
LSELLGRRAAPVVVVRGLGLYH
10740
GTGCAGACATCAGTCATGTTGGGCGGCGGCGGCGCTTCTGTGCTGGCTGCTGCTGGCCA
VHDLIGAQVLVDC(350)
10800
GATTATGAACATGAGCGCACAAGTGGATTTTCTCCAGCTCATGGATTTTCTTGTCTATCG
10860
AAGGCCAGGCTGCTCAATTTTTCATCTCTGGAACCCCGGGGACACCCTTCCCAAGGTCTAACT
Exon 9
10920
TCCTCTGAGGCTAGGGCTCTGGCTGCTCCACGCTTGTTCTTCTGGCTGAGAGCAGCAT
(351) I
```
CTCTACCTTGAGAAACCTCCTCGCCACTGACAGACATGAGAATCTGCTCATGATT
S T T S G T L L R R L P E T * (364)

GCTCAGGTCTGAGACTGGCCGCAGGGAAGGACTGAGAGCTGCATGACTCCCTGAGA

CAGGCTCAGCTTTGGAAAGCTTTGGAGCAATGAGTGTGGGCTTGGCTGCAAA

AGGTCAGAGCTGACCTGGCGTCGGAAGGGGATCTGGGCTTGGAGAA

TGGTTCTGCTGGCGCATGAGGAGCTTGAGTGGGCTTGGCTGCAAA

CTTTTCGGTAATAATGCTGCTTTATTTACTCTGCTGGATAGCTGCTGTGGATTTCTG

* RPI poly-A signal

GGCAGGGCTGGAGAGGGGCCAGGCTTCTCTCTCTCTACGTCATGTTAAAATGGCCAA

* Poly-A site

TAAGTACCTCGCTGCTGATATTTTCTGC

GGGAGGCTGGTGAAAGGAGGGCTTTTTCTCTACGTGTCATGTTAAAATGGCCAA

Poly-A size

AGTACCTCTGCCTGTGATATTTTCTGCAGGCTGCTCTGCGGGCTGATCTGGGCATCCAGCT

Leader peptide of C4

GAGTCAACTCTGGCCCAGGCTACTGCGTGGCCGAAAGGAACGACAGACAGGATCT

† BamHI

AACCTCTTGGATTCTCCAGGCCATGAGCTTGCTGGGCTTGGGATCTGGGCCATCGCCT

MRLLWGLIWASSF

Leader peptide of C4

Fig. 10-continued

FTLSLQKPR (C4 Intron I)

* N-terminus of C4
Figure 11

The organization of human RP1 gene.

The coding exons are in solid, black boxes and non-coding sequences in exons are in *stripped boxes*. Categorization of *Alu* elements [i.e. J and S classes, ref: Jurka and Smith, 1988] are marked. An arrow shows the orientation of transcription.
Figure 12
The VNTRs in the human RP1 gene.
(A) A dendrogram showing the relationship of the 21 copies of VNTRs (RPV1 to RPV21); (B) an alignment of the VNTRs, gaps are shown by *dash lines*; (C) a dotplot of the VNTRs, *diagonal lines* represent internal repeats in the sequence.
Figure 13

A comparison of the structures of the composite SVA retroposons in RP1, C2 (B allele) and CYP1A1 genes.

Numberings represent sizes in nucleotides. **Shaded circles and boxes** stand for **Alu**-related sequences. **Solid triangles** represent terminal, direct repeats flanking the SVA retroposons. **Hatched triangles** represent direct repeats flanking **Alu** elements. **Solid, vertical stripes** stand for tandem hexameric repeats (5' AGAGGG 3') present in SVA-C2 and in SVA-1A1.
**Figure 14**

An alignment of the *Alu* sequences from SVA-RP (*Alu-RP*), SVA-C2 (*Alu-C2*), SVA-1A1 (*Alu-1A1*) with the *Alu* consensus. Deletions are shown by dots; DNA sequences identical to that of *Alu*-RP are represented by dash lines, dissimilar sequences are shown at the corresponding positions.
Figure 15
Southern blot analysis of RP genes in humans and monkeys. Human genomic DNAs from CAH-E1 (lane 1), CAH-2 (lane 2), CAH-3 (lane 3), PW-1 (lane 4), PW-2 (lane 5), NPC-A (lane 6), B1 (lane 7), L1 (lane 8), HepG2 (lane 9), SC01 (lane 10), MOLT4 (lane 11), and African green monkey (COS 7; lane 12) and cotton top tamarin (NPC-LC; lane 13) genomic DNAs were digested with BamHI restriction enzyme, resolved by 0.8 % agarose gel electrophoresis, blotted to Hybond N membrane and hybridized with the R1.1 probe.
Figure 16
Partial gene duplications of Gene X and RP.
(A) A subclone of a 12 kb BglII restriction fragment from cos 4A3 spanning the intergenic region between two C4B genes; (B) Hypothetical exon-intron structures for Gene XA and RP2 in a 5.0 kb BamHI restriction fragment that has been completely sequenced (Database Accession number L26263). The breakpoint for gene duplication for RP and Gene X at the chimeric region is marked by an arrow. (C) The relative position of a 2.14 kb Taq I-BamHI restriction fragment from clone λ JM-2a corresponding to the intergenic region between the C4A4 and C4B5 genes. This fragment has been completely sequenced (Database accession number L26262). (D) The relative location of the PCR primers to determine the breakpoint of gene duplication for the Gene XA-RP2 (please refer to Figure 19).
Figure 17

The breakpoint for the partial gene duplication of RP and Gene X. The RP cDNA, RP1 gene and the Gene XA-RP2 hybrid sequences are aligned and identical sequences are marked by vertical lines. Intrinsic sequences corresponding to RP1 gene and to Gene XA are in lower cases. The exon-intron boundaries for RP1 gene and the hybrid are underlined. The breakpoint for the partial gene duplication of RP and Gene X is marked by a ^.
Figure 18
A molecular map of the RCCX gene modules in the HLA class III region. *Arrows* represent the transcriptional orientation of functional genes. 21A, CYP21A; 21B, CYP21B. Data taken from references (Trowsdale et al., 1991; Carroll et al., 1984; Gilelman et al., 1992; Yu, 1991; Matsumoto et al., 1992; Levi-Strauss et al., 1988; Speiser & White, 1989) and this work.
Figure 19
A common breakpoint region for gene duplication of Gene XA-RP2 for human and great apes.

(A) PCR amplification of the Gene XA-RP2 breakpoint region using primers YMR1 and HRP3 (Figure 16D) and genomic DNAs. DNA templates for PCR are from cos 5 (lane 1), CAH-E1 (with RCCX single modular structure; lane 2), Raji (with RCCX double modular structure; lane 3), HepG2 (with RCCX double modular structure; lane 4), WES (a chimpanzee cell line with RCCX bimodular structure; lane 5) and PUT (an orangutan cell line with RCCX bimodular structure; lane 6). Mol. wt. markers were 1 kb ladder (Gibco-BRL). (B) Autoradiograph of the Southern blot from (A) hybridized with R1.1 probe.
TABLE 3
A Summary of the Exon-Intron Structure of the Human RP1 Gene

<table>
<thead>
<tr>
<th>Exons</th>
<th>Intron</th>
</tr>
</thead>
<tbody>
<tr>
<td>nucleotide number*</td>
<td>size (bp)</td>
</tr>
<tr>
<td>1) - 1,199*</td>
<td>?</td>
</tr>
<tr>
<td>2) 1,794 - 2,141</td>
<td>348</td>
</tr>
<tr>
<td>3) 2,227 - 2,436</td>
<td>210</td>
</tr>
<tr>
<td>4) 2,546 - 2,682</td>
<td>137</td>
</tr>
<tr>
<td>5) 8,819 - 8,914</td>
<td>96</td>
</tr>
<tr>
<td>6) 9,330 - 9,469</td>
<td>140</td>
</tr>
<tr>
<td>7) 10,365 - 10,462</td>
<td>98</td>
</tr>
<tr>
<td>8) 10,568 - 10,715</td>
<td>148</td>
</tr>
<tr>
<td>9) 10,918 - 11,360</td>
<td>442</td>
</tr>
</tbody>
</table>

* nucleotide numberings correspond to those in Figure 10.
+ transcriptional start site(s) of RP have not been mapped, therefore, the size of exon 1 undefined.
+ after Patty, 1987
CHAPTER III
GENERATION OF POLYCLONAL ANTIBODIES AND THE BIOCHEMICAL CHARACTERIZATION OF THE RP PROTEIN

INTRODUCTION

The RP gene structure and its cDNA sequence have been completely determined. However, the presence of RP remains hypothetical unless the endogenous products is demonstrated. Despite of a limited degree of amino acid sequence similarities to two DNA repair proteins of lower organisms, the primary structure of RP did not reveal any significant functional domain (see discussion chapter). Several immediate questions regarding biological and biochemical properties can be addressed. They are: the molecular weight of the RP protein; the expression of the RP protein in various tissue types or cells; the subcellular localization of the RP protein such as the nucleus; the putative post-translation modifications of the RP protein such as potential phosphorylation and myristoylation; and whether the RP protein may form an oligomer structure among themselves or may associate with other protein molecules. The investigation of these questions requires a powerful biochemical tool: a specific antibody against the RP protein.

To generate a RP-specific antibody requires immunization with sufficient quantities of the RP polypeptide. This could be achieved by the production of the RP cDNA encoded protein through an efficient bacteria expression system. Thus we applied protein fusion and purification system from New England Biolabs (Guan et al., 1987 and Maina et al., 1988) to express and purify the maltose
binding protein (MBP) with the RP fusion. The pMAL vector contains a strong Ptac promoter and a repressor, encoded by lacI gene, which keeps expression of the fusion protein low in the absence of IPTG induction. Under the regulation of the Ptac promoter, a gene fusion of malE-lacZα or the malE-RP fusion protein could be induced to express at a high-level. A polylinker site between malE and lacZα is available for inserting the coding sequence of RP gene, which will interrupt the malE-lacZα fusion production (and allow a blue and white selection via α-complementing host cell such as TB1 cell strain). This system provides a method to express the fusion protein at high yield, that could account for 20-40% of the total cellular proteins.

One of the most important factors in producing a highly specific polyclonal antiserum is the purity of the antigen preparation used for immunization. Another consideration is the structure of the antigen. For functional studies of RP, it is preferable to ensure that the antigen for injection is in native conformation so that the antibodies generated can recognize the native proteins. However, the antibodies specific for denatured forms of the antigen can be useful for techniques such as Western blot analysis. The over-expressed fusion protein can be purified by use of the affinity property between maltose-binding-protein and amylose resin or maltose molecules (Kellerman et al., 1982). The fusion protein was then used to immunize rabbit to obtain RP-specific polyclonal antibodies. To enhance the titer of specific antibodies and to minimize animal trauma, an improved adjuvant (TierMax) was used to allow the antigen to be released steadily, ensuring the continual stimulation of the immune system. In this study, six rabbits were immunized with two different MBP-RP fusion proteins. These anti-MBP-RP antisera were characterized by ELISA, partially purified and used for immunochemical studies such as Western blot analysis and immunoprecipitation experiments.
MATERIALS AND METHODS

A) Construction and Purification of RP Fusion Protein

1. RP-fusion protein with a truncated structure of the RP gene

A 1.1 kb *BamHI* - *EcoRI* fragment from cDNA clone R1.1 (Shen et al., 1994) was inserted into the *StuI* site of pMAL-c vector (New England Biolabs, Beverly, MA) by blunt end ligation. The 1.1 kb fragment was end-filled by using Klenow DNA polymerase (3 U) in 0.5 mM dNTP buffer at 37°C for one hour. The *StuI* cut vector was treated by phosphatase reaction (5 U) at phosphatase buffer (50 mM Tris pH 8.0, 0.1 mM EDTA) at 37°C for 30 min. The translation reading frame of the RP insert was maintained as the same phase as that of the vector's *malE* gene. A ~60 kDa MBP-RP fusion product was produced with 228 amino acid residues of the RP. This fusion protein construct was designated MBP-RP/a.

2. RP-fusion protein containing Nuclear Localization Signal

A second fusion protein construct, MBP-RP/b, was generated by incorporating 21 more amino acids residues N-terminal to MBP-RP/a which contains a putative nuclear localization signal. This was achieved as follows: A 0.8 kb *BamHI* genomic DNA, which is the 5'-end of the *BamHI* site of cDNA clone, was ligated with cDNA R1.1 to generate a clone contained 1.9 Kb insert (designated as RP1.1+0.8K). The PCR was then performed using primers and both primers contained *EcoRI* sequences.

HRP5 (5'-ccgGAATTTCATGAGCTGGAAGAGGC-3')
HRP3 (5'-tccGAATTTCATGTCTCTGGCAGGCG-3'),

The reaction conditions were one cycle at 94°C for 5 min; 30 cycles at 94°C 1 min, 54°C for 1 min, 72°C for 1 min; and one cycle at 72°C for 10 min.
The amplified PCR products of 700 bp fragment were digested with EcoRI and purified as described previously. The fragment was ligated into EcoRI-restricted and phosphatase-treated vector pMAM-cRI.

3. Analytical characterization of the fusion protein production

The ligated mixture of fusion protein constructs were transformed to competent TB1 cells araΔ(lac proAB) rpsL (φ80 lacZ ΔM15) hsdR, which are E.coli host cell with α-complementation activity. The vector express the malE gene fused to the lacZα gene under the regulation of a strong promoter Ptac and a repressor of lacI gene (Figure 20). The transformants could be selected by blue/white colonies in the presence of 80 μg/ml X-gal and 0.1 mM IPTG. However, the low efficiency of transformation was probably due to the over-expression of MBP-RP fusion protein that might be toxic to E.coli (even with as low as 10 μM IPTG). We then performed the screening of positive clones for fusion protein constructs by colony hybridization.

Overnight cultures of cells transformed with fusion plasmids were inoculated in an rich medium (per liter: 10 g tryptone, 5 g yeast extract, 5 g NaCl, 2 g glucose, supplemented with 100 μg/ml ampicillin after autoclave). The cells were cultured at 37°C with vigorous shaking until 2 x 10⁸ cells/ml were obtained (A₆₀₀ of 0.4). A 1 ml sample was saved as uninduced control. IPTG was added into the remaining culture to a final concentration of 0.3 mM. At different time points of induction for 1-3 hours, samples were collected, centrifuged and resuspended in SDS-PAGE sample buffer (4% SDS, 100 mM Tris-Cl pH 6.8, 20% glycerol, 1.5% 2-mercaptoethanol, 0.02% bromophenol blue). All samples were boiled for 5 min, spun for 1 min and loaded onto 10% SDS-PAGE gel to determine the optimal time point for the harvest of the fusion protein for large scale preparation.
To identify whether the expressed fusion proteins were present in the soluble or in insoluble fraction such as inclusion bodies and to determine the ability of binding to amylose resin. The IPTG induced cell lysates were subjected to freeze/thaw conditions, sonicated, and centrifuged at 9,000 x g for 20 min. The supernatant (Extract 1, soluble proteins) and the inclusion body pellet (Extract 2) were resuspended in lysis buffer (10 mM sodium phosphate, 30 mM NaCl, 0.25% Tween 20, 10 mM β-mercaptoethanol, 10 mM EDTA, 10 mM EGTA, adjust to pH 7.0 with NaOH) and saved for loading onto SDS-PAGE gel. Around 50 μl Extract 1 was mixed with equal volume of a slurry of amylose resin in column buffer (10 mM sodium phosphate buffer, pH 7.2, 0.5 M NaCl, 1 mM sodium azide, 10 mM β-mercaptoethanol). The mixture was incubated on ice for 15 min. After centrifuging for 1 min, the supernatant was discarded, and the pellet was washed with 1 ml column buffer. The washed pellet was resuspended in the sample buffer and resolved in a SDS-PAGE gel.

4. Affinity chromatography for large scale purification

One liter rich media was inoculated with 10 ml of overnight culture and cultured until the OD reach A600~0.4. IPTG was added to a final concentration of 0.3 mM, the culture continued to grow at 37°C for 3 hours. The cells were harvested by centrifugation at 4000 x g for 20 min and resuspended in 50 ml lysis buffer, supplemented with 5 mM of PMSF to inhibit serine proteases. The mixture was sonicated for ~2 min to ensure the maximum release of nucleic acids with complete cell lysis (which could be monitored by determining the A260). NaCl was added to 0.5 M final concentration. The cells were centrifuged at 9,000 x g for 30 min and diluted with 5 bed volumes of column buffer with 0.25% Tween 20. The MRP-RP fusion proteins in the cell extracts were purified with an affinity chromatography column.
The affinity column was prepared as follows. About 1.5 g of amylose resin was swollen in 50 ml of column buffer for 30 min, poured into a 60 ml syringe (a substitution for 2.5 x 10 cm column) plugged with silanized glass wool. The column was washed with 3 bed volumes of column buffer + 0.25% Tween 20. The diluted cell extracts (1:5 dilution with the same buffer) were loaded onto the column at a flow rate of 1 ml/min. Washing of the nonspecific proteins from the column was achieved by passing 3 bed volumes of column buffer + 0.25% Tween 20 and 5 bed volumes of column buffer without Tween 20. The elution of resin-bound fusion protein was accomplished by running column buffer + 10 mM maltose. About 1 ml fractions were collected and monitored by UV A absorbance for the presence of the fusion protein. The purity of the protein was determined by loading onto a SDS-PAGE gel and stained with Coomassie Brilliant Blue (CBB). The concentration of the purified fusion protein was determined by the procedure provided by the manufacturer of a protein assay kit (BioRad, Richmond, CA).

5. Fusion protein cleavage by protease factor Xa

The protein-containing fractions were pooled and dialyzed with buffer (4 x 100 volumes 10 mM Tris-Cl, 100 mM NaCl, pH 8.0) in order to remove the maltose from the fusion protein. The proteins were then concentrated to 1 mg/ml by using Amicon Centricon. To determine the cleavage activity of Factor Xa, 20 μl of fusion protein at 1 mg/ml were added to 1 μl factor Xa at 200 μg/ml. After incubation at room temperature for 2, 4, 8, and 24 hours, 5 μl samples were resolved in a SDS-PAGE gel to identify the optimal reaction time.

6. Immunization of rabbits and collection of antisera

Two RP fusion proteins MBP-RP/a and MBP-RP/b were emulsified with two different adjuvants: complete Freunds adjuvant (Sigma, St. Louis, MO) and
Hunter's TiterMax #R1 (CytRx Corp. Norcross, GA). For the complete Freunds adjuvant (CFA), 1 ml of ~1 mg/ml fusion protein in PBS was mixed with 1 ml of CFA. For those rabbits injected with TiterMax adjuvant, 500 µg of fusion protein of 1 mg/ml was mixed with equal volume of adjuvant. Emulsification of antigen with adjuvant was achieved by connecting the two syringes via a 3-way stopcock, pushing half volume of the antigen into the adjuvant syringe first, so that the aqueous phase would enter into the oil phase rather than vice versa. The other half of antigen was mixed in the same way by forcing the mixture back and forth from one syringe to the other repeatedly. When the mixture turned homogeneous and milky, a simple test was performed to ensure that the emulsion was in good water-in-oil phase: A tiny drop of emulsion was extruded onto the surface of water, it should hold together for a good emulsion. In the event that the drop dispersed on the surface of water, the syringes were reconnected and emulsify for another 2 minutes.

Six New Zealand White rabbits (4-6 months old) were immunized with RP fusion proteins. Prior to immunization, the rabbits were bled and the blood samples were collected for pre-immune antisera controls. The emulsion mixture was injected into two intramuscular or subcutaneous sites with two divided doses. Two weeks following the priming immunization, rabbits were bled and booster immunized with the same antigen emulsified with either Titer Max or Incomplete Freund's Adjuvant. Further booster immunization at 2- to 3-week intervals was administered, and the animals were bled about one week after each boost.

Antisera were allowed to stand for 4 hours at room temperature and 4°C overnight in order to clot. The clot was loosen from the sides of the tube with an applicator stick. The sera were obtained by centrifuge and pelleting any remaining blood cells and debris at 2700 x g, 4°C. Sera were stored in aliquots at -20°C.
B) Immunoaffinity Purification of Antisera

1. Coupling MBP-RP fusion proteins to activated beads

Reacti-Gel 6X (PIERCE, Rockford, Il), an immobilized carbodiimide agarose was used as a coupling solid support for the RP-fusion protein. One ml of this activated agarose slurry was filtered and washed with three bed volumes of ice-cold water to remove excess acetone. One mg of RP-fusion protein was diluted in chilled coupling buffer (0.1 M Sodium Carbonate pH 9.16) and added to the washed gel. The tube was incubated at 4°C at pH 9.4 with gentle shaking for 30 hours which allowed the coupling reaction to reach completion and to remove active imidazolyl-carbamate group. The coupling buffer was removed and the blocking buffer (1 M Tris pH 8.8) was added to react with the unreacted activated-beads. The beads coupled with the fusion protein were then transferred to a column and washed with 10 bed-volumes of 10 mM Tris (pH 7.5), followed by 10 bed-volumes of 100 mM glycine (pH 2.5), then 10 mM Tris (pH 8.8). After the pH of last drop from the column turned 8.8, another 10 bed-volumes of 10 mM Tris (pH 7.5) was passed through the column until the pH decreased to 7.5.

2. Binding and eluting of antisera from column

The polyclonal serum was diluted at 1:2 of 10 mM Tris (pH 7.5), and was spun to remove any debris prior to loading on the column. The serum was applied through the column at a slow rate (drop by drop) for three times to ensure complete binding with the coupled antigen. About 20 bed-volumes of 10 mM Tris (pH 7.5) was used to wash the column, and then with another 20 bed-volumes of 500 mM NaCl, 10 mM Tris (pH 7.5). The specific-binding antibodies were eluted by passing 10 ml of 100 mM glycine (pH 2.5) through the column. The eluate of 0.5 ml fractions were collected and neutralized with 0.5 ml of 1M Tris (pH 8.0).

For the elution of base-sensitive interacted antibodies, 10 ml of 100 mM triethylamine (pH 11.5) were passed through the column and 0.5 ml fractions were
collected. Each fraction was neutralized with 0.5 ml of 1M Tris (pH 8.0). The column was regenerated by washing with 10 mM Tris (pH 7.5) and stored in buffer with 0.02% sodium azide at 4°C.

C) Biosynthetic Labeling of Protein and Immunoprecipitation

1. Cell lines

Human skin fibroblasts with xeroderma pigmentosum symptom XP14BE (TeJes) and XP2RO (XP-E) cell lines were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% heat-inactivated fetal bovine serum (FBS) and grown in a humidified 5% CO₂ atmosphere.

MOLT4 and BJAB lymphocytes cell lines were maintained in RPMI 1640 medium supplemented with 10% FBS, 1% glutamine, 6 ml/L S/P and grown at 37°C in CO₂ independent atmosphere.

2. Labeling of suspension cell with [³⁵S]-Methionine

Culture cells were maintained in suspension until they were growing in exponential phase. About 10⁷ to 10⁸ cells were harvested by centrifuging 5 min at 300 x g, at room temperature. 10 ml of methionine-free medium (RPMI1640, Gibco) was used to wash 10⁷ cells in conical tubes. The suspended cells then were centrifuged for 5 min at 300 x g. Supernatant was discarded and the cell pellets were resuspend by gently tapping on the bottom of the tubes and a repeat wash as above was done.

Cells were resuspended at 2 x 10⁶ cells/ml in Met-free medium supplemented with 2% dialyzed FCS (against 100 X volume of PBS at 4°C overnight) and incubated for 45 min at 37°C to deplete intracellular pools of methionine. The tube was swirled periodically to suspend the aggregated cells at the bottom.
The cells were centrifuged at room temperature for 5 min at 300 x g, and the supernatant was discarded. Cells were then resuspended ~4 ml total to a final concentration of 5 x 10^6 cells/ml in Met-free medium supplemented with 10% dialyzed FCS. Cells were radiolabeled with [35S]-methionine (Trans35S label [ICN, Irvine, CA] or Expre35S35S label [Dupont, Wilmington, DE]) at 50 µCi/ml concentration, and then were incubated at 37°C for 4 hours or other time periods.

3. Cell lysis

The radiolabeled cells were washed twice with ice-cold PBS, pelleted by centrifugation and resuspended with lysis buffer (10 mM Tris-HCl pH 7.4, 1% NP-40, 150 mM NaCl, 1 mM EDTA) in the presence of 1 mg/ml BSA and 1 µl of 1 mM PMSF per 2 x 10^7 cells. The cells were then incubated on ice for 30 min. Cell lysate was centrifuged at 10,000 g for 5 min at 4°C to pellet the nuclei. The lysate was stored at -70°C and used for immunoprecipitation experiments.

4. Pre-clearing lysate and Immunoprecipitation

Staph A or Protein A-sepharose was washed twice in lysis buffer and resuspend as a 10% solution in lysis buffer plus BSA. The NaCl concentration was adjusted to 0.5 M. 10 µl of pre-immune sera was added into 0.5 ml lysate (~10^7 cells) and the reaction mixture in a tube was rotated for 1 hour at 4°C. 50 µl of Staph A or Protein A-sepharose was added and rotated at 4°C for 1 hour. The tube was then spun at 10,000 x g for 5 min at 4°C in a microcentrifuge to pellet the Staph A. Lysate was transferred to a new tube, followed by adding 30 µl antiserum (~10 µg purified antibody) or other control antisera to each tube. The mixtures of lysates and antisera were incubated for 1 hour at 4°C with constant shaking. 100 µl of Staph A or protein A-sepharose was added and rotated for 1 hour at 4°C.

5. Washing

The immunoprecipitated complex was spun at 10,000 x g for 5 min at 4°C. The remaining lysate could be saved in -20°C for other immunoprecipitation with
other antisera. 1 ml of washing buffer I (lysis buffer as mentioned at previous step, plus 0.5 M NaCl) was added to the Staph A pellet to suspend them well. The solution was centrifuged at 10,000 x g for 5 min at 4°C in a microcentrifuge. The supernatant was discarded and the staph A pellet was consequently washed with wash buffer II (lysis buffer + 0.1% SDS) and wash buffer III (10 mM Tris-HCl, pH 7.4, 0.1% NP-40) as the washing procedure of wash buffer I. 50 µl protein sample buffer was added to precipitate the pellet after final spin and the sample was then loaded onto an acrylamide gel.

D) Western blot analysis

Method 1

Purified MBP-RP fusion proteins, bacterial lysate of MBP-RP protein constructs, human BJAB cell lysate, and human MOLT4 cell lysate were resolved in SDS-PAGE gel and electroblotted to Hybon C membrane using semi-dry machine (BioRad). The membrane was blocked in 1% bovine serum albumin (Blot-Qualified BSA, Promaga) in Tris buffer saline (TBS, 10 mM Tris-HCl, pH 8.0, 150 mM NaCl) for 30 min to saturate nonspecific protein binding sites. Blocking solution was replaced with TBST (TBS + 0.05% Tween 20) The membrane was then incubated with anti-RP sera with 1:100 to 1:1000 dilution for 1 hr. at room temperature with rocking and washed three times each with TBS + 0.1% Tween 20 (TBST) for 10 minutes. The membrane was transferred to TBST containing 1:7500 dilution of IgG-AP (alkaline phosphatase) conjugate and incubated for 30 min. NBT (nitroblue tetrazolium) and BCIP (5-bromo-4-chloro-3-indolyl phosphate) substrates were mixed and added into the membrane to develop the color. The color development was stopped by addition of deionized water.
Method 2

Protein samples were loaded onto a SDS-PAGE gel. The transfer buffer (57.6 g Glycine, 12.1 g Tris, 800 ml Methanol, adjusted to 4 L with water) was used for electroblotting of proteins onto nitrocellulose membrane. The transfer was performed by a semi-dry machine (BioRad) at 0.6 Amps for 2 hours. The membrane was then incubated for one hour in blocking buffer (50 ml 3 M NaCl, 10 ml 2 M Tris pH 7.4, 50 ml milk, 0.5 ml Tween 20, and adjusted to 1 L). First antibodies were diluted at desired concentration in blocking buffer, added to membrane and the mixture was incubated for 45 min. at room temperature. The membrane was then washed with blocking buffer three times for 5 min. each. The second antibody (goat antibody against rabbit with alkaline phosphatase conjugate, Sigma) was diluted at 1:2000 in blocking buffer and incubated with membrane for 45 min. at room temperature. The membrane was washed three times for 5 min. The substrate (1 mg/ml Naphthol AS-MX phosphate, 1 mg/ml Fast Blue BB salt, from Sigma) was dissolved in substrate buffer (100 mM Tris pH 9.5) and added to the membrane for color development.

E) ELISA analysis

Purified MBP-RP fusion proteins were diluted to 10 μg/ml in carbonate-bicarbonate buffer, coated onto a micro-ELISA plate, and incubated at 4°C for 3 hr. in a humid chamber. The coated antigen was washed by filling the well completely with PBS-Tween 20, incubating at room temperature for 3 min and the buffer was then emptied by inverting and shaking. The washing step was repeated twice manually or by using a washing devise (Bio-Tek instruments).

The test antibodies including the preimmune serum were diluted in PBS-Tween 20 at concentrations varying from 1:500 to 1:10,000. To each well was added 200 μl diluted serum with serial titer and incubated for 2 hr at room
temperature in a humid chamber. Then the plate was washed as described previously.

200 μl of goat-anti-rabbit antibody with alkaline phosphatase conjugate was diluted in PBS-Tween (1:7500). The plate was then incubated for 30 min. at room temperature in a humid chamber. While the plate was incubating, the substrate was prepared by mixing 1 ml of 5x diethanolamine buffer and 4 ml of water for each tablet of P-nitrophenylphosphate. After the plate was washed by PBS-Tween 20, 100 μl of substrate solution was then added to each well. The reacted product was observed by bright yellow color.

The reaction was stopped after 15 min. incubation by adding 0.4 M NaOH at 100 μ/well. The absorbance of the reaction products at 405-420 nm could be detected and recorded by using spectrophotometer reader (Bio-Tek instruments). Statistic package (Qutra program) was used to analyze and plot the absorbance values.

RESULT

3.1 Strategies for the production of MBP-RP fusion protein clones

In order to obtain large quantities of the RP polypeptide, a bacterial protein expression system was utilized. The MBP-fusion protein system (New England Biolabs) was employed for the production of maltose binding protein-RP fusion. The RP cDNA clones were inserted into a cloning linker region of an expression vector pMAL (Figure 20A) The expression of MBP-RP fusion was regulated under a strong promoter, Ptac. Two different RP fusion protein constructs were made. First, an end-filled 1.1 kb BamHI - EcoRI cDNA fragment was ligated into the SstI site of the pMAL-c vector (Figure 20A), and transformed into TB1 cells. Ampicillin resistant colonies resulting from the transformation were screened by
colony hybridization using cDNA R1.1 as a probe. With IPTG induction of the
fusion protein synthesis, we expected a 65 kDa (41+24 kDa from the MBP-RP)
fusion protein to be produced from the MBP-RP/a construct. The R1.1 cDNA
contains codon #137 to #364 and encodes a protein containing 228 amino acid
residues (Figure 20B). Although the N-terminus of the RP protein is lacking in this
cDNA clone, this 228 amino acids should be sufficient to stimulate a strong
immune response for the production of RP antisera.

The second construct included the putative nuclear localization signal (NLS).
This was achieved by the ligation of RP1.1 cDNA clone with 0.8 kb BamHI
genomic DNA corresponding to Exon 2 to Exon 3 of RP. The resulting product
was designated RP1.1K+0.8K (Figure 20C). We then used the PCR method to
amplify the region containing the NLS (primers: HRP5 and HRP3). This 700 bp
fragment was used to ligate with the fusion protein production vector pMAL-cR1,
which provides a modified recognition site for proteinase Xa digestion. Colony
hybridization was performed to screen transformants grown on LB-Amp plates. The
correct configuration of the RP insert with respect to the MBP gene was diagnosed
by BamHI and EcoRI enzymes digestion on those positive plasmid DNAs. The RP
portion of the fusion product encoded 254 amino acids residues from codon #111
to #364. The obtained fusion proteins MBP-RP/b, was predicted as about 68 kDa
(41+27 kDa) in molecular weight (Figure 20C). Both DNA clones for fusion
protein constructs were analyzed by dideoxy sequencing to ensure that no
nucleotide changes occurred during the cloning procedure and thus the appropriate
reading frame had been maintained.

3.2 Induction of the fusion protein synthesis

Since there were no positive transformants when IPTG was included in the
LB-Amp plates for propagation of transformants, it was inferred that the IPTG-
induced synthesis of the MBP-RP fusion protein was toxic to the host cells. Therefore, colony hybridization was adopted to obtain the positive transformants instead of using the conventional blue-white selection with the presence of IPTG and X-gal. Thus, it was important that the fusion protein synthesis was under tight control of an inducible promoter. In the pMAL vector system, the malE coding region is preceded by the Ptac promoter allowing the regulated expression of the MBP-RP fusion protein with the repressor lacI and the repressor inhibitor, IPTG. Cells were grown until high cell density with limited synthesis of the hybrid proteins, and then grown for a certain period of time for over-expression of the hybrid proteins in the presence of the inducer, IPTG.

Three MBP-RP/a and four MBP-RP/b positive clones with correct orientation and without detectable mutations around the cloning junction were tested for their protein synthesis, along with one control clone encoding a MBP-lacZ hybrid (clone 1 of Figure 21). Cell lysates from induced or non-induced bacteria harboring the appropriate plasmids were resolved in a 15% SDS-PAGE gel (Figure 21). In the absence of IPTG induction no maltose binding protein or any hybrid protein was synthesized. In the presence of 0.3 mM IPTG a major polypeptide of approximately 65 kDa corresponding to the expected size of MBP-RP/a (clones 6 and 7) were produced. One MBP-RP/a clone (#8) appeared to be un-inducible by IPTG. The expression of MBP-RP/b clones (clones 2-5) gave rise to unexpected patterns. Clones 3 and 4 expressed truncated polypeptides, while the underlining mutations are unknown. Clones 2 and 5 produced the expected 68 kDa fusion protein, but the levels of expression were low.

3.3 Optimization of the synthesis conditions and fusion protein purification

Serial analytical experiments were carried out to determine the optimal expression time periods for the fusion protein production, to test whether the
proteins are in soluble form and their ability of binding to amylose resins. All cell lysates were isolated at various induction time points from MBP-RP/a and MBP-RP/b in the presence of IPTG and analyzed on a 15% SDS-PAGE gel. The MBP-RP/a fusion protein (65 kDa) could be continuously induced up to 3 hours without substantial degradation (lanes 13-15, Figure 22), while the MBP-RP/b (68 kDa) started to break down after 2-hour of IPTG induction even in the presence of a proteinase inhibitor cocktail (Boehringer Mannheim). The majority of the proteins were degraded (lanes 10-12). This is probably due to the highly positive charged residues in NLS sequence where the positive charged feature may be a target site for serine proteinases. SDS-PAGE analysis showed multiple Coomassie-Blue stained polypeptides at about 30-40 kDa, which were probably degradation products for the 68 kDa MBP-RP/b fusion protein (lane 16, compared with the stable MBP-RP/a protein at lane 17, Figure 22).

To determine the solubility of fusion proteins and their ability to associate with the amylose resin, the following tests for two independent clones were performed. After centrifugation of the bacterial lysates, SDS-PAGE analysis of the resulting supernatant and cell pellets revealed that most of the fusion proteins were present in the soluble form as shown in Extract 1 fraction (lanes 4 and 8). Relatively less amount was retained within the cell debris, or the so-called inclusion body fraction (Extract 2 of lanes 5 and 9). The fusion proteins could successfully associate with amylose resins with high affinity (lanes 2, 3 and 6, 7). Hence both fusion proteins, MBP-RP/a and MBP-RP/b, were purified by affinity chromatography with amylose resin. The resulting fractions of each MBP-RP hybrids were analyzed by loading the samples onto SDS-PAGE gel and stained with Coomassie Blue. The MBP-RP/a proteins were purified at high purity (Figure 23), while the MBP-RP/b proteins exhibited degradation products at about 30-40 kDa in addition to the 68 kDa intact fusion proteins (Figure 24).
To obtain the RP polypeptide without the component of maltose binding protein, the purified fusion protein was dialyzed to remove the remaining maltose and subjected to the digestion of proteinase factor Xa. After several trials of lengthy proteinase Xa incubation for up to 3 days on the denatured fusion proteins (with 6 M guanidine hydrochloride to render the cleavage site accessible to Xa), we still could not cleave the MBP portion from the fusion protein. Hence the purified MBP-RP fusion protein was used directly as an immunogen to raise antibody against the RP protein.

3.4 Immunization of rabbits and antibody purification

Six New Zealand white rabbits were immunized with MBP-RP to generate RP-specific antibodies. The concentration of the fusion protein was determined by using protein assay kit (BioRad). Approximately 50 mg of protein was obtained from the purification of one liter culture. Each rabbit was immunized with 100-500 μg of fusion protein with TiterMax or Freund's complete adjuvant. These rabbits were boosted with a second dose of fusion protein with TiterMax or Incomplete Freund's adjuvant. Each rabbit was bled for 5-10 ml of blood before immunization (preimmune sera) and also bled at the period of two weeks after each boost.

Selected polyclonal antisera from rabbit #4 were pooled and purified through affinity chromatography. We coupled MBP-RP fusion proteins with carbonyldiimidazole-activated beads (Reacti-Gel from PIERCE) to form a stable complex. The antisera were then passed through the antigen-coupled column three times to ensure complete binding. The bound antisera were eluted by either acidic or basic elution buffer. Ten fractions of 0.5 ml each elution were collected and the concentration of each fraction was determined by using protein assay kit. The antibody elution peaked at the second and third fractions with a concentration around 10 μg/ml. The purified and the original antisera were then used for Western
blot analysis, ELISA analysis, and immunoprecipitation for the characterization of the RP protein.

3.5 Enzyme Linked Immunosorbent Assay (ELISA)

In order to monitor the immune response of each rabbit at various time periods and to identify which antiserum of each bleeding gave rise to the highest titer, the sensitive and convenient ELISA technique was employed.

About 1 μg of purified MBP-RP/a fusion protein were coated on micro-ELISA plates and the testing antisera were diluted varying from 1:500 to 1:10,000. The antibody activity against the RP fusion protein was then determined by the enzymatic activity of an alkaline phosphatase (Table 4). The resulting absorbance of the substrate reflecting the enzymatic activity from the titer plate was recorded and analyzed. All antisera revealed significant increase in affinity to coated MBP-RP fusion proteins when compared to the background absorbance of pre-immune antisera control (Table 4 and Figure 25). The No.5 antiserum showed the strongest response to MBP-RP fusion protein consistently, however, the pre-immune sera control from the same rabbit also gave rise to the highest background absorbance (Figure 25 B). The No.4 antiserum exhibited a relatively lower degree of antigenic activity among all six different antisera (No. 1 to No. 6). The antisera obtained from the third bleeding of every rabbit started to show a higher response (data as shown in the Table 4) and the activity was constant up to the final bleeding of these rabbits. The reactivity of each antibody reached a plateau at a dilution about 1/500 to 1/1000 (The reactivity of rabbit No.2 antisera is illustrated in Figure 25A).
3.6 Western blot analysis

To determine the reactivity of the RP antisera and to identify the presence of endogenous RP protein in mammalian cell lysates, serial Western blot analyses were performed. All antisera reacted strongly with the MBP and MBP-RP fusion protein from IPTG induced bacteria lysates (Figure 26 showed the reactivity of Antisera 1). These antisera recognized both fusion protein containing 228 and 254 amino acid residues of RP protein, MBP-RP/a and MBP-RP/b, respectively. Although the MBP-RP/b fusion protein was extremely unstable (as shown in Coomassie blue staining in Figure 24 and description above), an intense protein band with the expected 68 kDa molecular weight was detected in the Western blot of the total bacterial cell lysates (lane 5 of Figure 26).

Due to the fact that the antibodies were raised against the fusion protein with the maltose binding protein, the antisera are also expected to react with the maltose binding protein. The anti-MBP activities were demonstrated by the 51 kDa polypeptides from the bacteria lysate expressing the MBP-LacZ fusion protein (lane 4 of Figure 26). The presence of anti-RP activities in the antisera can be demonstrated by employing an unrelated RP fusion protein (eg. a β-Gal-RP fusion protein) for Western blot analysis, or by the direct identification of the endogenous RP protein from mammalian cell lysates.

The antisera could detect polypeptides of 53 kDa, 55 kDa, and 80 kDa molecular weights from total cell lysate of the human Raji cell line (Figure 26). In the marmoset B cell lysate, a 55 kDa and an 80 kDa proteins were observed. Only the 55 kDa protein was detected in the cell lysate of the mouse EL4 cells. These polypeptide could be candidates of the endogenous RP gene products. The divergence in protein sizes may be a result from the variable post-translation modifications or differential transcripts of the RP gene in a tissue specific manner. Nonetheless we could not exclude the possibility that some of the polypeptides
were the result of cross-reactions of the polyclonal antibodies with some unknown proteins.

In order to circumvent the possible cross-reactivity of the polyclonal antibodies against the unknown polypeptides, immunoaffinity purified antisera were used in immunoblot analysis. Three different cell lines: human colon carcinoma HT29, human nasopharyngeal carcinoma CNE, and mouse fibroblast NIH-3T3 were tested, in which a 55 kDa protein was consistently detected from all three cell lines (Figure 27 and lane 1, 4 of Figure 28). It was noted that the molecular weight markers were purchased from Bio-Rad instead of Gibco-BRL (as been used elsewhere in this chapter, while the discrepancies of these two markers was described in Figure 27). In contrast, the unpurified antisera could react with several other polypeptides in addition to the 55 kDa protein (lane 2 and 5 of Figure 28). The pre-immune control sera could not detected the same 55 kDa polypeptides (lane 3 of Figure 27A and lane 3, 6 of Figure 28). The chance of the mammalian cell lines expressing the bacteria maltose binding protein analog (or its cross-reaction with anti-RP antisera) could be ruled out by the absence of the 55 kDa polypeptide when the same cell lysate was tested negative with the control antibody, anti-MBP antisera (lane 2 of Figure 27A).

3.7 The RP protein detection by immunoprecipitation

Human fibroblast cell line TEJES and human lymphocyte cell line MOLT4, were harvested and metabolically labeled with $^{35}$S-[Met] for 4 hours. The cell lysates were then tested by immunoprecipitation with six different antisera; preimmune control sera; and a polyclonal antisera against CD1. Antibody 6 and the preimmune sera from the same rabbit (No.6) were used as shown in Figure 29A. Besides the common precipitated polypeptides that may cross-react with Ab6 and the preimmune-serum, more nonspecific association on the TEJES cell lysate was
observed (lane 1 of Figure 29A). The polypeptides at about 55 kDa molecular weight was specifically detected by Ab6 but not by the preimmune-serum (lane 2 and 3 of Figure 29A). All sera derived from the six different rabbits as well as the control serum could precipitate several polypeptides. Two major bands were shown at about 43 kDa and 40 kDa for Ab1 and Ab2. Presumably, they were proteins of non-specific association with either antisera or the protein A-sepharose, since they were commonly detected by all sera tested. However, only Ab6 clearly detected a protein at about 55 kDa (lane 6 of Figure 29B).

To test the specificity of Ab6 and its consistency on the detection of the RP protein, three independent experiments were performed using $^{35}$S-[Met] labeled MOLT4 cell lysates. The precipitated samples by using Ab6 (lane 1, 2 and 3 of Figure 30), along with other unrelated antisera (Ab4 and Ab2 for lane 4 and 5 respectively) were resolved in a SDS-PAGE gel (Figure 30). The antisera (Ab6) raised by the RP fusion protein could consistently react with a 55 kDa polypeptide in the total cell lysates of the human MOLT4 cell line but not the other antisera. Although we were aware that the antibody recognition could be the consequence of immune response to an unknown antigen from which the No.6 rabbit was bled. Moreover, a polypeptide at the same 55 kDa molecular weight was detected from Western blot analysis by using Ab4 antisera. It suggested that the Ab6 sera probably recognized the native form of the RP epitope (immunoprecipitation experiments), while the Ab4 recognized a denatured RP epitope in Western blot experiments.

**DISCUSSION**

Two different MBP-RP fusion proteins were purified at large quantity and used to raise anti-RP antibodies. The questions we addressed in this study are how the rabbits respond to the fusion protein; the antigenic activity of the induced
antibody; the detection of the RP protein in various tissue and cell lines, and to identify the protein localization within the cell. These could be achieved by several immunochemistry methods. Those experiments could provide us with important information about the biochemical properties of the endogenous RP protein.

From the ELISA assay, certain portions of antisera collection demonstrated strong antigenic activity toward the MBP-RP fusion protein. The immunization of the rabbits with a fusion protein results in the production of antibodies against both portions of the fusion protein. Since the Maltose-binding-protein fragment of the fusion protein was included for the rabbit immunization, the antibody will cross-react with the MBP polypeptide. Sometimes the dominant immune response was indeed against the bacterial protein with little or no antigenic activity against the specific insert protein. Hence antibody screening using ELISA to detect the immune response against purified fusion protein did not prove definitively whether these antisera were specifically against the RP protein. Therefore the use of antibody against the maltose binding protein is important to be included as an appropriate control sera in any immune-techniques. Alternately, an assay to detect insert specific antibodies was to construct a second fusion protein containing the same RP cDNA insert cloned into β-galactosidase fusion vector (data not shown). Positive interaction to the second fusion protein would suggest the presence of antibodies against the common insert of the RP polypeptide.

A third construct has been made by fusion of a full-length cDNA fragment with pMAL-cR1 EcoRI restricted vector. This 1.0 kb full-length RP cDNA was the PCR product by using primers, RP-met, HRP3 (both contain EcoRI site), and the template BS-1.5K. This fusion protein construct was expected to encode a hybrid product of a maltose binding protein with full-length (364 amino acid residues) RP protein. Expression of this full-length RP protein may provide a better tool for the
studies of the biochemical properties of the RP protein. However, this construction has not been fully characterized and not yet been used in this study.

Both Western blot and immunoprecipitation techniques were employed to detect the expression of the RP protein. As a result, the putative RP polypeptide was detected consistently at 55 kDa molecular weight in cell lysates from various human cell lines, MOLT4, HT29, BJAB, and also from a mouse NIH-3T3 cell line. Protein purification and subsequent polypeptide micro-sequencing could confirm if the 55 kDa precipitated products are clearly the endogenous RP-encoded protein.

Since all Western blot analysis described above were based on the transformed or tumor cell lines, lymphocytes were isolated from peripheral blood of one normal individuals (PW) was then tested with the same antisera. With the same amount of cell lysate (prepared from $10^6$ cells) loading onto the SDS-PAGE gel, a 55 kDa polypeptide was consistently detected in HT29 cell lysate. However, the same 55 kDa protein could not be detected in the PW cell lysate by using purified antisera. A faint smear at about 55 kDa along with other non-specific polypeptides were observed by unpurified antibody. Thus expression of RP protein at a lower extent at least in the lymphocytic cell of this individual with respect to the tumor cells might imply the tissue-specific expression of the RP protein. Another possibility is the involvement of tumorigenesis reflecting on the variant abundance of RP protein between the tumor cells and normal cells. However, more studies needs to be done on the thorough screening of the RP protein expression for other tissues from normal individual.

By use of polyclonal antibodies which were raised against the RP protein could be employed for various studies on biochemical properties of the RP protein. For instance, the proof of putative phosphorylation of the RP protein and the determination of any tightly associated macromolecule can be achieved by
immunoprecipitation techniques; the distribution of the RP expression in various tissues and the localization of the RP protein in a cell can be accomplished by several immunocytochemical methods. Many trials of immunofluorescence and immunohistochemistry assay have been performed on various culture cell lines and frozen tissues. The high background made the result ambiguous. Therefore, we could not confirm the presence of nuclear localization signal of the RP protein in the chosen cell lines. It was probably that these antibodies were not suitable to recognize the RP protein \textit{in situ} under the fixed condition of the cells. Moreover, several factors may be taken into consideration such as the turnover rate of RP proteins; the abundance and the expressing regulation manner in specific cell types. For example, the stress-inducible responses may be involved in the expression and stability of the RP proteins which was proved to be the case for a hypothetical yeast homolog of RP, RAD7, a DNA-damage responding protein (Perozzi et al., 1986). However, other experimental designs could be applied such as the construction of RP protein by using eukaryotic expression system allowing the use of a tag-specific (or a reporter protein fusion) antibody to identify the RP protein translocation. This technique has been commonly utilized for the studies on putative nuclear targeting sequences. For instance, nucleoplasm-pyruvate kinase fusion protein was detected in cellular nucleus with antiserum raised against pyruvate kinase (Robbins et al., 1991); and rabbit anti-β-Gal antibody was applied to immune-stain the expression of protein with Interleukin-1α and β-Galactosidase (β-Gal) fusion within nucleus (Wessendorf et al., 1993).
Figure 20

(A) Fusion protein cloning vector pMAL-c and pMAL-cR1 (NEB)
These vectors contain the inducible Ptac promoter, which can be induced to transcribe a *malE-lacZα* gene fusion. The *lacI* gene encodes the Lac repressor, which keeps the expression of the fusion protein at low extent in the absence of IPTG.
Polylinker sequences of both vectors are shown under the cloning vectors. MBP-RP/a fusion protein construct was ligated into *StuI* site on pMAL-c vector. *EcoRI* site on pMAL-cR1 vector was used for the subcloning of MBP-RP/b fusion protein.

(B) Strategy for the expression of fusion protein MBP-RP/a
MBP-RP/a construct was derived from the cDNA clone R1.1, by which the RP cDNA fragment was subcloned to the pMAL-c vector by blunt-end ligation.

(C) Strategy for the expression of fusion protein MBP-RP/b
PCR reaction was performed to obtain 0.7 kb fragment by using: BS-RP(1.1+0.8) as DNA template, and HRP5, HRP3 as primers. After the *EcoRI* digestion, the PCR fragment was ligated to *EcoRI* site of the pMAL-cR1 vector.
Figure 20
Figure 21
Induction of MBP-RP fusion protein, MBP-RP/a and MBP-RP/b

The protein expression of eight clones (#1 to #8) were tested in the presence (+) and absence (-) of IPTG. Clone #1 corresponded to MBP-β-gal-α fusion with 51 kDa molecular weight. Clones #6, 7, and 8 were the MBP-RP/a constructs. #6 and #7 clones could be induced to express 65 kDa MBP-RP/a fusion protein, while #8 was un-inducible. Clones #2, 3, 4, 5 were the MBP-RP/b constructs. About 68 kDa expected products in clones #2 and #5 were barely detectable after IPTG induction, while degraded polypeptides accumulation was noted at lower molecular weights. Both products from clones #3 and 34 were truncated fusion proteins with molecular weight less than 40 kDa.
Figure 22
Determination of optimal conditions for fusion protein production and purification

Lane 1 shows the induced fusion protein as clone #1 in Figure 21. Two independent cell lysates of MBP-RP/a fusion proteins (clones #6 and #7 of Figure 21) were associated with 200 µl amylose resin (lanes 2 and 6); 50 µl resin (lanes 3 and 7). Extract 1 represented the soluble protein of total lysate (lane 4, 8), while the Extract 2 were insoluble inclusion bodies (lanes 5, 9) of both fusion clones. The expression time points: 1, 2, and 3 hours induction for MBP-RP/b (lanes 10-12), and MBP-RP/a (lanes 13-15). The molecular weights of MBP-RP/b and MBP-RP/a are 68 and 65 kDa, respectively (lanes 16 and 17). Noted the degradation polypeptides at about 30-40 kDa for the total cell lysates sample of MBP-RP/b fusion protein.
Figure 23
Large scale MBP-RP/a fusion protein production and the purification
The flow through of washing fluid before eluting was represented with W. #1-#11 elution fractions were collected. 12 μl of each fraction was resolved a 15% SDS-PAGE gel.
### Figure 24

**Large scale MBP-RP/b fusion protein production and the purification**

W represented the flow through of washing fluid before eluting. #1-#10 elution fractions were collected, 12 μl of each fraction was loaded onto a 15% SDS-PAGE gel. Multiple fragments of polypeptides were noted at about 30-40 kDa molecular weight as described in the Figure 22.
### Table 4

ELISA analysis

<table>
<thead>
<tr>
<th>Rabbit #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dilution</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/500</td>
<td>.221</td>
<td>.231</td>
<td>.300</td>
<td>.181</td>
<td>.334</td>
<td>.311</td>
</tr>
<tr>
<td>1/1000</td>
<td>.295</td>
<td>.291</td>
<td>.448</td>
<td>.135</td>
<td>.357</td>
<td>.319</td>
</tr>
<tr>
<td>1/2000</td>
<td>.189</td>
<td>.232</td>
<td>.324</td>
<td>.130</td>
<td>.237</td>
<td>.238</td>
</tr>
<tr>
<td>1/4000</td>
<td>.224</td>
<td>.170</td>
<td>.271</td>
<td>.119</td>
<td>.184</td>
<td>.214</td>
</tr>
<tr>
<td>1/8000</td>
<td>.177</td>
<td>.188</td>
<td>.262</td>
<td>.138</td>
<td>.184</td>
<td>.174</td>
</tr>
<tr>
<td>1/16000</td>
<td>.143</td>
<td>.158</td>
<td>.194</td>
<td>.130</td>
<td>.152</td>
<td>.136</td>
</tr>
<tr>
<td>1/32000</td>
<td>.137</td>
<td>.132</td>
<td>.139</td>
<td>.126</td>
<td>.131</td>
<td>.127</td>
</tr>
<tr>
<td><strong>pre-immune</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/500</td>
<td>.144</td>
<td>.131</td>
<td>.125</td>
<td>.158</td>
<td>.159</td>
<td>.124</td>
</tr>
<tr>
<td>1/1000</td>
<td>.122</td>
<td>.123</td>
<td>.128</td>
<td>.135</td>
<td>.132</td>
<td>.120</td>
</tr>
<tr>
<td>1/2000</td>
<td>.124</td>
<td>.121</td>
<td>.118</td>
<td>.129</td>
<td>.129</td>
<td>.119</td>
</tr>
<tr>
<td>1/4000</td>
<td>.115</td>
<td>.118</td>
<td>.118</td>
<td>.128</td>
<td>.125</td>
<td>.116</td>
</tr>
<tr>
<td>1/8000</td>
<td>.116</td>
<td>.118</td>
<td>.112</td>
<td>.122</td>
<td>.122</td>
<td>.121</td>
</tr>
</tbody>
</table>
Figure 25
ELISA analysis of six antisera against RP proteins
(A) Schematic presentation of the reactivity from the ELISA of antisera 2. The absorbance of antisera against the RP fusion protein and the pre-immune controls was as indicated at serial dilutions.
(B) Diagram for the immune activity against the fusion protein. Six antisera (rab1-rab6, as indicated) demonstrate the positive responses comparing to the pre-immune sera (the bottom line) at serial dilutions.
Figure 26
Western blot analysis with unpurified anti-MBP-RP sera
Total cell lysates from Raji (lane 1), marmoset B cell line NPC-LC (lane 2), mouse T cell EL4 (lane 3); cell lysates from *E. coli* with the expression of MBP-lacZ (lane 4); expression of MBP-RP/b (lane 5), expression of MBP-RP/a (lane 6). The antisera used were diluted at 1:500 with PBS solution. Molecular weight markers were from Gibco-BRL.
### Figure 27

**Western blot analysis of HT29 and NIH3T3 cells**

(A) About $5 \times 10^3$ HT29 cells were lysed directly in sample buffer, resolved in a 10% PAGE-SDS gel, and electroblotted to Hybond C membrane. The membrane was cut into strips and tested with the following antisera:

- **Lane 1**: purified antibody 4 at 1:500 dilution.
- **Lane 2**: anti-MBP antisera (New England Biolabs) at 1:5,000 dilution.
- **Lane 3**: pre-immune antisera at 1:500 dilution (negative control).

(B) The sample preparation was the same as in (A), lane 1 was HT29 cell lysate; lane 2 was mouse NIH3T3 cell lysate. A 55 kDa polypeptide was detected from both cell lines. Molecular weight markers were purchased from Bio-Rad. However, the same markers from Gibco-BRL were labeled with different molecular weights: Phosphorylase B (106 kDa), bovine serum albumin (80 kDa), ovalbumin (49.5 kDa) and carbonic anhydrase (32.5 kDa) were labeled as 97.4 kDa, 68 kDa, 43 kDa, and 29 kDa, respectively. For consistency, the molecular weight standard from Gibco-BRL were used throughout this chapter.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>kDa</strong></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>106-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>49.5-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32.5-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27.5-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18.5-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 28
Western blot analysis by using unpurified and purified RP antisera with HT29 and CNE cell lysates
HT29 (lanes 1-3) and CNE (lanes 4-6) cell lysates were resolved in a 10% SDS-PAGE gel and electroblotted to Hybond C membrane. The membrane was cut into strips and tested with the following antisera:
Purified antisera at 1:500 dilution (lanes 1 and 4)
Unpurified antisera at 1:100 dilution (lanes 2 and 5)
Pre-immune antisera at 1:100 dilution (lanes 3 and 6).
Molecular weight marker were purchased from Bio-Rad. Note the size discrepancies of these markers as described in Figure 27.
Figure 29
Analysis of six antisera with immunoprecipitation and the use of Ab6 on TEJES and MOLT4 cell lysates

(A) Human fibroblast cell line TEJES (lane 1) and T-lymphocyte MOLT4 (lane 2) were immunoprecipitated with antisera 6. The preimmune sera control was tested with MOLT4 cell lysates (lane 3).

(B) T-lymphocyte MOLT4 cells were labeled with $^{35}$S-[Met] for 4 hours. The labeled cell lysates were then precipitated with six individual antisera (lane 1 to 6). One unrelated antisera and the polyclonal antibody against CD1 were used as controls (lanes 7 and 8, respectively). Molecular weight markers were from Gibco-BRL.
Figure 30
Detection of a 55 kDa polypeptide for RP by immunoprecipitation
Three independent experiments of immunoprecipitation from [\textsuperscript{35}S]-labeled MOLT4 cells by using Ab6 (lanes 1, 2 and 3), a 55 kDa polypeptide was clearly shown. These antisera failed to obtain the similar result as described above: Ab4 (lane 4), Ab2 (lane 5), did not detect the same 55 kDa protein. Molecular weight markers were from Gibco-BRL.
CHAPTER IV
CHARACTERIZATION OF THE 5' REGULATORY REGIONS OF THE HUMAN RP 1 GENE

INTRODUCTION

The complete RP1 gene has been characterized and more than 4.7 kb genomic sequence corresponding to the 5' regulatory region has been determined in this study. They are 1) a 1.6 kb BamHI fragment including the first exon and intron, 2) one 1.8 kb BamHI fragment located upstream of 1.6 kb fragment, 3) a 1.3 kb BamHI-SalI fragment (derived from pSH13, Figure 9), which is further upstream of 1.8 kb fragment. Due to the ubiquitous expression manner of the RP gene based on the observation in the Northern blot analysis (Chapter 2); and the detection of endogenous RP proteins from various cell lines and tissues (Chapter 4), it has been suspected that the RP gene encodes a housekeeping protein. Hence, it is of interest to analyze the sequence data and to characterize the upstream region of RP gene in order to understand the transcription regulation of the RP gene and its transcription initiation site(s).

A further observation draws our attention is that for both mouse and human RP1 genes about 1.6 kb upstream of transcriptional start site are highly conserved (see genomic map of Chapter 3). Such high homology (more than 90% identities) suggests the presence of regulatory elements at this 5' region of the RP gene, which may play a significant role in the RP function for both species. Such highly conserved region may also imply the presence of extra RP coding sequence or another novel gene at this region.
In this study, questions such as the location of the transcription initiation site for the RP gene, regulation of RP gene expression, and the functional significance for the sequence conservation within the upstream region of the RP genes in different species, were investigated. Since many obtained cDNA clones from the human or the mouse cDNA libraries did not provide us the full-length RP transcript, we employed "anchor-PCR" and RT-PCR techniques to isolate the 5' region of the RP cDNA. These 5' region cDNA clones will allow us to understand more about the RNA splicing for the upstream transcripts. Furthermore, in order to define more precisely the transcriptional initiation start site(s), the RNase protection analysis and primer extension technique were employed. A preliminary study for the putative promoter activity of the RP gene was also performed by using a transient expressing assay, the chloramphenical acetyltransferase (CAT) reporter system.

Presumably, the RP gene encodes a housekeeping protein based on its ubiquitous expression from various cell lines and tissues (see Chapters 2 and 3). The absence of TATA-box sequence in the upstream region of the RP gene could further support this observation. Those genes without TATA-box at their promoter region is a relatively common feature for the housekeeping genes (Smale et al., 1988; Dudov et al., 1984). Thus, the RP gene expression through the RNA polymerase II machinery requires an additional binding element for accurate anchorage onto the transcription initiation site. Through sequence analysis on the upstream region of the RP gene and determination of the transcriptional initiation site, we could understand more about the regulatory cis-elements located upstream of the RP gene.

In the search for the biological function of RP, extensive literature review regarding the functional-structural features of RP were conducted (see the discussion in Chapter 2). A putative tumor suppressor gene(s) has been suggested
to be present in the MHC region. This gene locus is also involved in the cell
growth and development, inbred rats with this gene-defect are more susceptible to
many carcinogens and predisposed to a variety of tumors (Melhem et al., 1993).
The RP gene product appears to be a nuclear protein with a bipartite nuclear
localization signal (NLS). A RP fusion protein (Chapter 3) has been shown to bind
genomic DNA specifically by cross-linking experiments and electrophoretic
mobility shift assays (unpublished data from Dr. Wu). As a result, it is tempting
to speculate that RP may be a candidate for the described tumor suppressor gene
in the MHC region.

MATERIALS AND METHODS

Ribonuclease Protection Assay

The DNA templates for RNA probe were prepared by inserting a 0.8 kb or
a 1.6 kb BamHI genomic DNA fragment into Bluescript (KS+) vector which
carries the bacteriophages T7 and T3 promoters. DNA was digested with a suitable
restriction enzyme that cleaves immediately downstream of the RP sequence to
generate unique sized runoff transcripts by a specific RNA polymerase. The
digested DNA were treated with proteinase K to a final concentration 50 μg/ml at
37°C for 30 minutes, followed by two times of phenol/chloroform extraction. After
ethanol precipitation, DNA was resuspended at about 10 mg/ml into TE buffer
made with DEPC-treated water.

A RNA transcription kit (Stratagene) was used for the transcription reaction.
1 μg proteinase-treated DNA was added to the following reagents in order: 5 μl
5X transcription buffer (200 mM Tris pH 8.0, 10 mM Spermidine 40 mM MgCl₂,
250 NaCl), 1 μl 10 mM ATP, 1 μl 10 mM CTP, 1 μl 10 mM GTP, 1 μl 0.75 M
DTT, 1 U RNase-block II, 5 μl 400-800 Ci/mmol α-³²P-UTP, 10 U of T3 or T7
RNA polymerases, and DEPC treated water to a final volume of 25 µl. The mixtures were incubated at 37°C for 30 minutes. 10 U of RNase-free DNase was then added and incubated at 37°C for 15 minutes. 2 µl of tRNA (10 mg/ml) was added as a carrier, and the volume was adjusted to 50 µl. After phenol/chloroform extraction and ethanol precipitation, the RNA probe was redissolved in 10 µl of loading buffer and was heated at 85°C for 5 minutes to denature the RNA. This denatured sample was resolved in a 4% native polyacrylamide gel. The gel covered with plastic wrap was exposed to X-Ray film for 30 seconds, the radioactive RNA probe was eluted from the excised polyacrylamide gel strip with 400 µl elution buffer at 37°C for 4 hours. After ethanol precipitation, the RNA probe was redissolved in hybridization buffer, and its radioactivity was counted with a scintillation counter. The probe was ready for hybridization with sample RNA isolated from MOLT4 and HT29 RNAs.

About 5 X 10⁵ cpm of the RNA probe were mixed with 10 µg total RNA in 30 µl of hybridization buffer, heated to 85°C to denature the RNA, and subjected for hybridization at 30°C to 60°C for 1, 4, 8 and 16 hrs. 350 µl of ribonuclease buffer containing 40 µg/ml ribonuclease A and 2 µg/ml ribonuclease T1 were added to the reaction mixture, incubated at 30°C for 30 to 60 minutes. 10 µl of 20% SDS and 2.5 µl of 20 mg/ml proteinase K were added, and the mixture was incubated at 37°C for 15 minutes to inactivate the ribonuclease. The sample was extracted with phenol/ chloroform, supplemented with 10 µg of tRNA and alcohol precipitated. The dried RNA sample was redissolved in 5 µl of RNA loading buffer, denatured at 85°C for 3 minutes, the protected fragments were resolved in a 4% polyacrylamide/urea gel.
Primer Extension

Labeling of oligonucleotide

Oligonucleotide RP35 (5'-caggaattcGCGGGAACAGCTGCATGAG-3') and 0.8F-R1 (5'-CGCAGACGTCCTCTCATAACCTCAG-3') were labeled with \([\gamma-\text{32P}]\)ATP by T4 polynucleotide kinase. The reaction was performed at 37°C for 30 minutes with the following reagents: 20 µl \([\gamma-\text{32P}]\)ATP (10 mCi/ml, 200 µCi total), 1 µl 100 µg/ml primer (100 ng, preferably 30-40 mer), 0.5 µl 10X polynucleotide kinase buffer, 4U T4 DNA kinase. The mixture was then heated at 65°C for 5 minutes to inactivate the kinase. The labeled primers were precipitated with ethanol for three times and resuspended in 100 µl 0.3 M sodium acetate. 1 µl of the solution was subjected to scintillation counting.

Hybridization

5 x 10⁴ cpm of labeled oligonucleotide was mixed with up to 50 µg of test RNA. The salt concentration of the solution was adjusted to 0.3 M sodium acetate, followed by ethanol precipitation. The pellet was dried in an inverted vial on a Kimwipe and dissolved in 30 µl 1X aqueous hybridization solution (3 X buffer: 3 M NaCl, 0.5 M HEPES, pH 7.3, 1 mM EDTA pH 8) thoroughly by vortexing and drawing liquid through pipet tip. Then the mixture was allowed to hybridize at 30°C overnight. The next day, 170 µl 0.3 M sodium acetate and 500 µl ethanol were added to the hybridization mixture to precipitate RNA and oligonucleotide. The pellet was rinsed with 75% ethanol / 25% 0.1 M sodium acetate pH 5.2, and air-dried.

Primer extension reaction

Reverse Transcriptase Mix was prepared fresh on ice: 3.5 µl 4 mM dNTPs, 2.5 µl 10X RT buffer, 1.25 µl RNasin, 18 µl H₂O, and added to RNA-oligo mixture. Finally, 40 U of reverse transcriptase was added to each reaction mixture. The reaction was performed at 42°C for 90 minutes. 1 µl of 0.5 M EDTA and 1
μl of 1 mg/ml pancreatic ribonuclease A were added to stop the reaction and degrade the RNA template. After ethanol precipitation, the sample was resuspended in 3 μl TE buffer and 4 μl of formamide loading buffer, and boiled for three minutes. 3-4 μl of chilled sample was loaded into 4-6% sequencing gel for analysis.

**RT-PCR for the cloning of RP 5'-end**

Total RNAs were obtained from cell lines BJAB (B-lymphocyte) and HT29 (colon carcinoma). The first strand cDNA synthesis was performed with RNA PCR kit (Perkin Elmer Cetus), 1 μg of total RNA was used for each reverse transcription reaction. The reagents and oligo-nucleotide used in this experiment are as following:

<table>
<thead>
<tr>
<th>Volume (μl)</th>
<th>Final concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>MgCl₂ solution</td>
<td>4</td>
</tr>
<tr>
<td>10X PCR Buffer II</td>
<td>2</td>
</tr>
<tr>
<td>Sterile distilled water</td>
<td>1</td>
</tr>
<tr>
<td>dNTP mixture</td>
<td>8</td>
</tr>
<tr>
<td>RNase inhibitor</td>
<td>1</td>
</tr>
<tr>
<td>Reverse transcriptase</td>
<td>1</td>
</tr>
<tr>
<td>Random Hexamers</td>
<td></td>
</tr>
<tr>
<td>or oligo d(T)₁₆</td>
<td>1</td>
</tr>
<tr>
<td>or downstream primer</td>
<td></td>
</tr>
<tr>
<td>HRP3 or RP35</td>
<td></td>
</tr>
<tr>
<td>RNA sample</td>
<td>2</td>
</tr>
<tr>
<td>Total volume : 20 μl</td>
<td></td>
</tr>
</tbody>
</table>

The RT reaction was incubated at 42°C for one hour. Afterwards, the reverse transcriptase was inactivated at 99°C for 5 minutes, soaked at 4°C for 5 minutes.
For the PCR reaction, 4 µl of MgCl₂ solution, 8 µl of 10X PCR buffer and 0.5 µl AmpliTaq DNA polymerase were added into the RT-reaction mixture. Two RP-specific upstream and downstream primers were used to amplify the RT-reaction products

1.6K-F2: 5'-CGTCAGCAGTTTTGTAGGTGCCC-3';
HRP3: 5'-TCCGAATTCTAGGTCTGAGCGGCG-3'.

The conditions for PCR were: 1 cycle at 94°C for 5 minutes; 30 cycles at 94°C for one minute, 54°C for one minute, 72°C for one minute: one cycle at 72°C for 10 minutes. Primers sets: 1.6K-F2, RP-R5; and 0.8K-F1, RP-R5 were chosen for nested-PCR in order to obtain more specific products.

RP-R5: 5'-AAGAGGATCCGACTCCACAGGCCC-3'
0.8K-F1: 5'-TCCTCCAAATGCAGTGAGTT-3'

RESULTS

4.1 Cloning of RT-PCR products and analysis of aberrant spliced fragments

Total RNAs from B-lymphocyte, BJAB, and colon carcinoma cell line, HT29, were used for the amplification and cloning of the RP 5'-end cDNA by RT-PCR method. The RT-PCR products were shown in Figure 31A and 32A. Since multiple bands or a smear of PCR products were observed on agarose gels, size fractionation were performed and subcloned into various cloning vectors: Smal restricted Bluescript, Smal cut M13mp19, BamHI-EcoRI digested M13mp19 and pBS, EcoRI digested pBS and PCR cloning vector TA. The choice of these vectors for subcloning depended on the primers used for nested-PCR reactions: whether they were kinase-treated (cloned into Smal vector); containing EcoRI, and/or BamHI restriction sites (cloned into EcoRI-BamHI site); or direct cloning from the
PCR products (use of a TA vector).

All insert-containing clones (Figure 31B, 32B) were partially sequenced in order to identify the upstream region of the RP cDNA. These Bluescript subclones contained 0.3 kb, 0.5 kb and 1.5 kb inserts, (Figure 31B) generated by RT-PCR with primers 1.6K-F2 and RP-R5. Clone #12 in figure 31B with 0.5 kb insert revealed partial exon 1 (due to the location of primer 1.6K-F2) and exon 2. Hence, intron 1 sequence is 594 bp in size and agrees with the GT-AG intron splicing consensus. Most of the sequenced clones contain part of the RP cDNA sequences on both ends, while they were spliced in an unusual manner. The putative introns were generated without the GT-AG flanking junction when the sequences were compared with the genomic sequence at the corresponding region. Interestingly, the breakpoints of all unusually spliced products were the positions with four or five short nucleotide repeats (Figure 33). Presumably, these aberrant RT-PCR products originated from the "slippage" reaction from the reverse transcriptase reaction due to unmelted RNA structure.

Another set of nested-PCR products using primers 0.8K-F1 and RP-R5 was subcloned into PCR cloning vector, TA (Figure 32A and 32B). The clones contained inserts with 300 bp, 500 bp and 600 bp in size. TA10.1, the only clone without an abnormal splicing pattern, was about 500 bp in size and contained exon 2 and portion of exon 3. Hence, based on the sequence of RT-PCR products and also the genomic DNA sequence of RP1, it can be concluded that the intron 2 is 84 bp in size.

As for the construction of full-length RP cDNA, we made an appropriate fusion of RP1.1 clone (cDNA library clone) in pBluescript with RP-PCR fragment derived from TA10.1 clone (RT-PCR clone). Although TA vector has one BamHI site and RT-PCR product should provide another BamHI site, the trial of BamHI digestion failed to release the expected 540 bp insert. It was likely that one of the
BamHI site was destroyed during the cloning experiments. Thus, we adopted PCR to amplify the 540 bp fragment from TA-10.1 using primers SP60 (a TA vector primer) and RR-R5. The PCR product was purified and digested with BamHI, and then ligated with BamHI digested and phosphatase-treated vector pBS-RP1.1 (Figure 34). The resulting construct was designated BS-hRP (full), with 1.5 kb insert containing exon 2 to the poly-A tail of RP cDNA.

4.2 Characterization of RP gene transcription initiation sites by primer extension

In order to determine the transcription initiation site, a primer extension analysis was performed. Total RNAs were isolated from B-lymphocytes, Raji, MOLT4 and colon carcinoma HT29 cell-lines, and used as templates for reverse transcription. Two primers which are complementary to the RP mRNA sequence were used in these studies. The first is a reverse primer, RP35, corresponding to the 5'-end of the cDNA clone RP1.1. The second primer, 0.8K-R1, corresponds to the nucleotide sequence starting at nucleotide 36 of the exon 2 of the RP1 gene. They were labeled by using T4 polynucleotide kinase and [γ^{32}P]-ATP, annealed with 50 μg of test RNAs from human cell lines. The extended products of both primers by using reverse transcriptase were analyzed by electrophoresis in 4% polyacrylamide / 8M urea gels (Figure 35).

There were four major transcription initiation sites of RP for HT29 and Raji cell RNA: 120 bp, 185 bp, 190 bp, and 260 bp, when the primer 0.8K-R1 was used for extension (Figure 35B). The reactions for both RNA templates isolated from HT29 and Raji cell lines resulted in the identical primer extension fragments. The extension reaction from the primer RP53 gave rise to larger products. They were 285 bp, 370 bp, 450 bp, 470 bp and also another major fragment larger than
500 bp. Beside the possibility for the early stop for RT reaction and due to the resolution of the polyacrylamide sequencing gel and the sizes of these fragments, it was difficult to resolve the molecular weight accurately and localize the positions at corresponding genomic DNA sequence (see Figure 35A). However, the extension products from the experiment of 0.8K-R1 primer allowed us to map the putative transcription start sites of RP (Figure 36). In conclusion, both experiments indicated the presence of multiple initiation sites, which was consistent with the RNase protection analysis (summarized in Figure 38).

4.3 Characterization of the RP transcripts by the RNase protection assay

To further confirm and localize the putative initiation site(s) of the RP gene, the RNase protection assay was applied. Antisense RNA probes were synthesized using the T7 RNA polymerase and [\(^{32}\)P]-UTP from two different constructs. Firstly, the pBS-0.8K clone spanning part of intron 1 to part of exon 3 was linearized with EcoRI enzyme, which gave rise to 800 bp antisense RNA probe after transcription synthesis. The second clone, pBS-1.6K containing the putative exon 1 and part of intron 1 was linearized with EcoRI, AccI and SstI (Figure 37A and 37B), and the sizes of these labeled RNA probes were 1659 bp, 1005 bp and 226 bp, respectively. Both pBS-0.8K and pBS-1.6K constructs were BamHI fragments derived from the RP1 genomic clone pSH13 (see Chapter 2).

Total RNAs from MOLT4 were hybridized with various RNA probes described as above. The mixtures were subjected to RNase digestion, the protected fragments were resolved with a 4% polyacrylamide sequencing gel. The result of RNase protection by using the 800 bp probe derived from pBS-0.8K did not reveal any distinct protected fragments except a smear signals at about 90 bp in size. Presumably, they were nonspecific or degraded products from the RNase digestion. However, multiple protected fragments were consistently detected in two other
independent experiments by using two RNA probes, AccI and EcoRI-transcripts, derived from the pBS-1.6K. Three prominent bands were shown: 134 bp, 198 bp, and 232 bp from the EcoRI-transcript probe (lanes 1 and 2 of Figure 37A). Five distinct fragments, 175 bp, 198 bp, 201 bp, 225 bp and 300 bp, were the protected products derived from the AccI probe (lanes 3 and 4). Furthermore, there were several less prominent fragments on the autoradiography film, and their origins were not fully characterized.

4.4 The use of GAL 4 hybrid system to determine the transcription factor activity of the RP protein

This hybrid system consists of two DNA constructs: First, a reporter plasmid (pMCV-110) contains a CAT gene under the transcription regulation of SV40 TATA box and GAL4 recognition element UAS (Sadowski et al., 1988). Second, an effector construct encodes GAL4 DNA binding domain and RP coding region which produce GAL4$_{1-147}$-RP fusion proteins under the regulation of a SV40 promoter (Figure 39). Chinese hamster ovary (CHO) cells or COS-1 cells were chosen as transfecting hosts. CAT activities were determined along with negative controls of transfecting a plasmid with GAL4 DNA binding domain alone and a reporter plasmid alone (Gorman et al., 1985). As a positive control, we transfected effector plasmid with GAL4 DNA binding domain fused to part of the herpes simplex virus VP16 which is a strong activator of transcription (Sadowski et al. 1988).

Two GAL4-RP constructs were tested for their transcription activity. The first clone contained the full-length RP cDNA linked with the vector pSG424 encoding GAL4 DNA binding domain and 364 amino acid residues of RP protein, pSG-hRP(full). This construct was made by a ligation between a PCR-amplified 1.1 kb fragment and a pSG424 EcoRI-digested vector (Figure 40). The 1.1 kb
fragment was obtained from PCR procedure by using templates BS-RP(full) and primers, RP-Met and HRP3. The second construct, pSG-hRP was a truncated RP fragment encoding amino acid residues 111 to 364 including the putative nuclear localization signal (Figure 40). A 0.7 kb fragment was derived from the PCR product by using HRP5 and HRP3 primers. These constructs were designed in order to generate GAL4<sub>1-147</sub>-RP fusion protein in frame. All transfected cell lysates were analyzed by incubating acetyl-CoA and <sup>14</sup>C-chloramphenicol for CAT activity assay. By transfecting both COS-1 and CHO cells with various plasmids of hybrid constructs, the GAL4-VP16 plasmid was the only construct which could give rise to positive transcription activity. The resulted activity for both entire or partial RP hybrid constructs revealed a similar level of CAT expression as that of the plasmid with GAL4 DNA binding domain.

**DISCUSSION**

Because of the difficulty of obtaining 5' end of the RP cDNA from the screening of several cDNA libraries, we employed the RT-PCR method to isolate the upstream region of the RP cDNA based on the complete RP genomic sequence and putative promoter region in the RP gene. We synthesized oligonucleotides flanking the possible transcribed region according to the RP genomic sequence. The total RNA samples isolated from colon carcinoma HT29 cell line and T-leukemia MOLT4 cell line were used as templates for amplification of the RP 5'-end. Almost all the clones of the RT-PCR products which have been sequenced revealed a slippage-splicing pattern which are similar to chromosomal translocation breakpoints in genomic DNA. This phenomenon might be due to an inappropriate condition for primer-RNA hybridization in the reverse transcription reactions. However, a possibility could not be ruled out in which a special folding structures
of the RP RNAs reflecting on their transcriptional regulation. In spite of these unusual cDNA structures, it appeared that certain RP transcripts might start somewhere upstream of the primer 1.6K-F1 sequence from which the farthest 5'-end region primer was available for the RT-PCR amplification. Due to the multiple initiation feature of the RP transcripts, the 5' boundary region of the exon 1 was undefined, while 191 bp of the exon 1 and 70 bp for portion of the exon 2 were confirmed in this study.

The DNA sequence upstream of putative translation initiation codon of the RP gene is located in a CpG rich region. The distribution of CG dinucleotide are non-random in mammalian genomes. Under-methylated tracts of CG rich DNA (CpG islands) are frequently associated with the 5' region of many housekeeping genes and also tissue-specific genes (Bird, 1987). Indeed, three restriction enzyme sites specific for rare CpG sequences were located in the 5' region of the RP gene: BssHII, EglI and KspI based on pulsed-field gel electrophoresis analysis. This led to the mapping of the novel gene, G11 (Sargent et al., 1989), based on the presence of CpG islands and their cosmid mapping.

Several GC-boxes, potential binding sites for a transcription factor Sp1, were found at the 5' flanking sequence of the RP gene. They are at nucleotide positions 81, 220, 675, 776 of the RP1 gene (Figure 38). In addition, potential CAP-box, (nucleotide 44, 128, 353, 388, 451, 934) and octamer site ATTTGTAT (at nucleotide 858), specific for Oct-1 were also located within this region. It was noteworthy that a consensus sequence, CTCAACCT, a potential initiator site for TFIID (transcription factor IID) binding of non-TATA box containing genes, is present at the upstream region (at nucleotide 789) of the RT-PCR primer, 1.6K-F1. This reflected the fact again that the RP transcription could start near to the 5' end region of 1.6K-F1 primer. Nonetheless, multiple initiations of RP gene transcription were shown both from RNase protection assays and primer extension analysis.
Multiple initiations are common for those genes without TATA or CAAT boxes, and especially, for "housekeeping gene" (Ohgi et al., 1992).

The promoters of many genes associated with cell growth have been found to lack the canonical TATA box. Some genes can still transcribe their mRNA from a single site (Smale et al., 1988 and Dudov et al., 1984), while others direct the transcription from multiple start sites (Dynan, 1986). Most of the TATA-less genes discovered to date contain the housekeeping functions. And these genes usually are expressed in all cell types at relatively low levels in contrast to some tissue-specific genes which are expressed at high level at selected cell types (Dynan, 1986 and 1989). Despite the lack of the TATA box in many housekeeping genes, high GC content is one of the common features for the regulatory region (Dynan, 1986 and Bird, 1986; 1987). Most of the studies on the role of GC boxes in the transcriptional initiation were performed on the mammalian dihydrofolate reductase (DHFR), one of the well-characterized housekeeping genes. A study of the hamster DHFR gene has indicated that GC boxes not only are required for efficient transcription but also regulate the relative utilization of transcriptional start sites (Blake et al., 1990).

RNA polymerase II-transcribed genes without a TATA box are speculated to consist of alternative initiation elements (Initiator, \( Inr \)) to anchor the transcription complex to the Cap site (Roeder et al., 1991; Carcamo et al., 1991). It was first demonstrated that 17 bp element of \( Inr \) containing the transcriptional start site of the terminal deoxynucleotidyl-transferase (TdT) gene is sufficient for accurate basal level of transcription (Smale, et al., 1989). There are no obviously defined consensus sequence near the Cap site, however, it was believed that the initiation took place at the proximity region of these \( Inr \) elements, either upstream or downstream of the cap site (Zawel et al., 1992). One of the downstream promoter element found in Drosophila, termed downstream initiation element, has
the conserved 4-bp sequences ACA(G/C) or CGCT, within 30 bp downstream of the CAP site (Arkhipova et al., 1991). Interestingly, three such "ACAG motifs" are present in the RP gene at positions 995, 1089, and 1115 (Figure 38). All three are located at the 20-45 bp downstream region of the putative Cap sites predicted by primer extension experiments.

To determine the regulatory activity for putative RP promoter, four different DNA fragments were constructed to link with vector SV0-CAT, a vector with the expression of reporter gene, CAT. However, none of the constructs showed significant promoter activity. There are several possibilities for this negative result. First, the lacking of promoter activity from these constructs might originate from the improper junction of the putative promoter region with the CAT gene. For instance, the existence of the intron sequence could interfere with the proper splicing and expression of CAT gene. Second, the inclusion of a negative regulatory element in the upstream region might down-regulate the expression of the CAT gene. Hence, promoter DNA constructs with more diverse region and serial deletions from the plasmid DNA in this study are essential to locate the RP promoter and to understand the RP gene regulation. A more sensitive expression system such as the luciferase assay might be utilized (Ausubel et al., 1987). However, the transcription initiation site of the RP gene could not be accurately defined, the DNA constructs for the promoter region with reporter gene could be more tricky. Another alternative approach could be employed such as the "in vitro competition assay" using synthesized oligonucleotide to pinpoint critical regulatory elements around the transcription initiation site. Such studies have been reported recently for characterizing the transcription activation of mouse complement C4 gene which is also a TATA-less gene (Miyagoe et al., 1994). This technique will be useful for confirming the promoter elements such as the putative ACAG motifs found in the downstream region of the RP Cap-site.
Evolutionary conservation has been considered a key indicator of functional importance. The alignment of two sequences from the mouse genomic DNA (2.2 kb BamHI fragment in clone mRP5.3 of chapter 5) and 5' region of human RP1 gene (nucleotide 1-2414 of Figure 10 in Chapter 2) indicated that they are highly conserved between human and mouse RP gene in the promoter regions. Hence this conservation might imply the presence of regulatory elements with a significant function for RP expression of both species. The characterization of the upstream region of the RP gene is in progress in our laboratory.

To explore the possibility of RP is a transcription factor, a GAL4 hybrid system was employed. High transcription activity of the VP16 was demonstrated by acetyl products which were increased dramatically in the GAL4-VP16 control experiment. However none of the RP fusion constructs revealed any transcription factor activity. Beside the possibility of being unrelated to the transcription mechanism for the RP protein, this system did not resolve the question such as whether the GAL4-RP fusion proteins maintain structural-functional properties as in the physiological conditions. In other word, the GAL4 domain of the fusion protein might interrupt the proper conformation of RP protein, or the essential domain for biological function of the RP protein was blocked by steric hindrance of the extra 147 residues derived from the GAL4. However, these GAL4-RP fusion constructs can be further tested by two-hybrid system to identify the putative protein-protein interaction domains in RP and RP-associated proteins (Chien et al. 1991 and Fields et al. 1989).
**Figure 31**

**RT-PCR analysis and subcloning for 5'-end cDNA of RP**

(A) The RT-PCR was performed with primers, 1.6K-F2 and RP35. RNA templates used were: Raji poly-A⁺ (lane 1); Nalm6 polyA⁺ (lane 2); BJAB polyA⁺ (lane 3); BJAB total RNA (lane 4); and HT29 total RNA (lane 5). RT-PCR products were resolved in a 0.8% agarose gel. The 1 kb ladder (Gibco-BRL) was shown at the left.

(B) The RT-PCR product from BJAB cell line (lane 4 of Figure 31A) was fractionated and subcloned into Bluescript vector. All clones were analyzed with restriction enzyme digestion: *Eco*RI and *Bam*HI, and partially sequenced on their ends. Lanes 1-18 represented individual clones with inserts of various sizes: 300 bp insert (lanes 1, 4, 8); 500 bp insert (lane 12); 1 kb insert (lanes 6, 18); 1.5 kb insert (lanes 2, 9, 13, 14, and 16), whereas lanes 3, 7, 10, 11, and 17 were empty clones without insert DNA.
Figure 32
RT-PCR analysis and subcloning for 5'-end cDNA of RP

(A) The second (nested) RT-PCR was performed with different sets of primers, 0.8K-F1 and RP-R5. RNA templates were HT29 (lane 1), BJAB (lanes 2 and 3). The products were resolved in 0.7 % LGT agarose gel. Three fractions were isolated at about 200 bp, 200-400 bp and 400-550 bp, and subcloned into TA PCR cloning vector individually.

(B) 12 transformed clones were isolated and subjected to EcoRI restriction enzyme digestion to diagnose the subcloned fragments. Various sized-inserts were revealed: 500 bp (lane 1); 550 bp (lanes 2, 3, 4, 5, 7, 8, and 9); 450 bp (lanes 6, 10, and 11); 200 bp (lane 12).
Figure 33
Sequences comparison among RT-PCR products and genomic DNA on their spliced junctions
Cloned RT-PCR products (from figure 31) were partially sequenced and compared with the 5' region of RP1 genomic DNA sequence. Eight representative clones (underlined in the middle) were analyzed and revealed that the artificial junctions contained short repeats (double-underlined, 3-5 bp). The nucleotide numbers above or below the repeats are after Figure 10 of Chapter 2.
Figure 33
Figure 34
Construction of RP full-length cDNA by combining R1.1 and the RT-PCR product

The human RP cDNA clone, BS-hRP1.1 (or R1.1 in Chapter 2) was a subclone of Bluescript vector. A 540 bp RT-PCR product (TA-RP10.1 in lane 1 of Figure 32) was subcloned into TA vector. A 540 bp fragment was amplified by using SP6 and RP-R5 primers in order to release the RT-PCR product. This 5'-segment of the RP cDNA was subsequently subjected to BamHI digestion and ligated into BS-hRP1.1 (BamHI-digested and phosphatase-treated) to produce a 1.5 kb cDNA insert. This DNA construct spans part of the exon 2 to the end of RP cDNA (1.5 kb in size) and is designated BS-hRP (full).
Figure 34
(A) Primer extension analysis using primer RP53
About 50 μg of MOLT4 total RNA was hybridized with $^{32}$P-labeled oligonucleotide of $5 \times 10^4$ cpm (lane 1); $10^3$ cpm (lane 2); and the blank control (lane 3). The major reverse transcription products were 285 bp, 370 bp, 450 bp, 470 bp, and ~500 bp. Dideoxy DNA sequencing ladders were loaded as a molecular weight markers.

(B) Primer extension analysis using primer 0.8K-R1
Total RNAs from HT29 cell line (lane 1) and from Raji cell line (lane 2) were hybridized with radioactive primer, 0.8K-R1. The extension products was resolved on a 4% acrylamide sequencing gel along with four-lane dideoxy sequencing products. The size corresponding to the primer was shown at the right. Despite the multiple extension products, four prominent fragments 120 bp, 185 bp, 190 bp and 260 bp, were observed.
Figure 36
Mapping of transcription start sites from primer extension experiments
A schematic diagram summarizing the putative transcriptional initiation sites upstream of the primer 0.8K-R1 (arrowhead). Four major primer extension products are indicated as lines under the RP gene structure. The putative upstream region extended from exon 1 to exon 2 of the RP1 gene are derived from Figure 10 of Chapter 2.
Figure 37
RNase protection assay to determine the RP transcript
(A) Three RNA antisense probes were derived from the BS-1.6k construct using the T7 RNA polymerase to synthesize three RNA probes: 226 bp, 1005 bp, and 1659 bp transcripts (the orientations are indicated as arrowheads). The putative position of the exon 1 is according to the Figure 10 of Chapter 2, while its 5' end is unknown (hatched box). Three restriction enzymes used in the probe production were Styl, AccI, and EcoRI. The lines under each transcripts were the RNase protected fragments.

(B) All samples followed by RNase digestion and DNA sequencing of Bluescript vector (a size marker shown at the right) are resolved on 4% polyacrylamide-8 M urea gel, and visualized by autoradiography. About 10 μg of Raji (lanes 1, 3, and 5), and MOLT4 (lanes 2, 4, and 6) total RNAs were incubated with 32p-UTP-labeled antisense RNA probes. The protected fragments derived from antisense Styl transcript exhibited a doublet (lanes 1 and 2). Five major RNase protected fragments (lanes 3 and 4) and three protected fragments (lanes 5 and 6) were produced from the AccI transcript and the EcoRI transcript, respectively.
Figure 37
Figure 38
The nucleotide sequence of the 5' regulatory region of the human RP gene
The possible positions of transcriptional initiation sites based on primer extension experiments are shown, where the locations of the prominent and less prominent extension fragments are represented as * and :, respectively. One putative initiation site, CTCAACCT for TFIID binding of non-TATA box containing genes is in bold type. Potential CAP-box are underlined. Three downstream initiation elements ACAG/C or CGCT are italicized. Double underline denotes the Octamer site specific for numerous distinct transcription factors. e.g., Oct1-6, NF3. Potential GC-box and CACCC-box are boxed.
Reporter

![Diagram of reporter construct]

Effectors

![Diagram of effector constructs]

**Figure 39**

Reporters and effectors constructs for the GAL4 hybrid system

The reporter gene has the MMTV-LTR fused to the chloramphenicol acetyl transferase (CAT) gene. A GAL4 recognition element, UAS\(_g\), and TATA sequence are located upstream of the CAT gene. The effector constructs include a negative control plasmid, GAL4\(_{1-147}\) (encoding DNA binding domain, amino acids 1 to 147), positive controls plasmids, GAL4 and GAL4-VP16. The two RP hybrids constructs, GAL4\(_{1-147}\)-RP\(_{364}\) and GAL4\(_{1-147}\)-RP\(_{111-364}\), encode GAL4 DNA binding domain and the entire or partial RP proteins (with their amino acid residues indicated).
Figure 40
Cloning strategy for the construction of two GAL4-RP hybrid proteins
The pSG-hRP (full) clone was a full-length RP with GAL4 hybrid, while a 1.1 kb PCR product was derived from clone BS-hRP (full) (see Figure 34) using primers, RP-Met and HRP3. The pSG-hRP clone was the ligation product of 0.7 kb PCR product and the GAL4 DNA binding domain which was originated from a pSG424 clone.
CHAPTER V
GENETICS OF THE MOUSE RP1-SLP AND RP2-C4 GENE
STRUCTURES IN THE H-2 REGION

INTRODUCTION

Mouse complement C4 and sex-limited protein (Slp) gene are closely linked and both are located within the S region of H-2 complex. The C4-related genes of the mouse have provided an excellent model system to study structure-function relationship of a multigene family within the MHC. In addition to the allelic variation, mouse C4 and Slp show a striking genetic difference in serum levels among inbred and wild mice (Shreffler et al., 1984) and definitive functional variation. Mouse C4 shows complement activity and is involved in the activation of the classical complement pathway as a subunit of the C3 convertase (Reid et al., 1988), while the Slp is hemolytically inactive.

In contrast to the human C4A and C4B genes which are constitutively expressed in the liver, in mice of standard inbred strains, only C4 is expressed constitutively, while the expression of Slp is under the strict control of testosterone. Hence Slp is present only in male adult mice (the reason termed sex-limited protein). This variation of gene expression has become an interesting system for the study of eukaryotic gene regulation. In spite of more than 95% sequence identities of coding and flanking regions between C4 and Slp genes up to 1.9 kb upstream of translation initiation codon, there are two major deletions in the 5'-flanking regions between these two genes. The Slp sequence lacked a 31-nucleotide
segment containing ACACCC repeats and a 60-nucleotide segment with ACAC repeats (Nonaka 1986). These C+A-rich regions which are present only in the C4 gene were previously suggested to be responsible for the difference in the gene regulation between mouse C4 and Sip genes. However, Robin and colleagues further characterized the Sip regulatory element 2 kb upstream of the gene by DNase I hypersensitivity in vivo (Hemenway et al., 1987) and by transfection experiments showing the ability to confer androgen response (Loreni et al., 1988). It was marked that they discovered an endogenous retrovirus (or provirus, IMP) insertion which comprised the major structural difference of the 5' regulatory segments between the C4 and Sip genes (Stavenhagen et al., 1988). Long terminal repeat (LTR) of retrovirus has long been shown to associated with hormone-responsive enhancer. For example, mouse mammary tumor virus (MMTV) is responsive to several steroid hormones, including androgen (Roeder et al., 1983); moloney murine sarcoma virus also responds to glucocorticoid (Miksicek et al., 1986). Present data led to the hypothesis that the 5' LTR serves as a hormone-responsive enhancer and confers androgen response on the adjacent Sip gene (Stavenhagen et al., 1988).

The striking similarity of the gene organization for C4-CYP21 between human and mouse prompted us to investigate whether there is a similar RCCX modular structure in the mouse MHC. Since the mouse C4 and Sip manifest striking differences in gene expression, it is of considerable interest to determine the pattern of RP gene expressing in the mouse. The characterization of the mouse RP genomic structure would shed light on the molecular evolution of the gene duplication of C4-CYP21. The information about mouse RP gene structure is essential for the construction of an animal model: gene-targeting experiment to elucidate the RP biological function.
MATERIALS AND METHODS

Southern Blot Analysis of Cell Lines and Mouse Strains

Mouse myeloma cell line NSO, T-lymphocyte EL4 and fibroblast NIH3T3-K1 were obtained from Dr. Cesar Milstein and Dr. Ing-Ming Chiu. Three different mouse strains were used in this study: female and male of 129SVJ, male of BALB/C, and female of FVB. Genomic DNAs were isolated as following standard protocols (Ausubel et al., 1987), subjected to complete digestion with appropriate restriction enzymes, resolved in 0.8% agarose gels, hybridized with appropriate DNA probes. A RP 5'-end probe was derived from a cDNA clone, MRP1.0, with 302 bp of EcoRI-StuI fragment containing sequences from exon 3 to exon 5, while a cDNA probe was derived from a 728 bp fragment from the same digestion. A 2.2 kb probe was derived from 5' region of the RP1 gene from the Lambda genomic clone MRP5.3. The probe with sequences of exon 3 to exon 4 regions was a 320 bp fragment derived from EcoRI-BamHI digestion of human RP1 gene (Chapter 2).

Northern Blot Analysis

Total RNAs were isolated from mouse cell lines or various tissues: brain, gut, kidney, liver, muscle, and spleen after standard protocols. About 25-50 μg total RNAs were resolved by a 0.8% formaldehyde agarose gel, blotted onto Hybond N membrane (Amersham, Arlington Heights, Il). The membrane was hybridized with mouse cDNA mRP1.0k probe, washed twice at room temperature with 2 x SSC, 0.1% SDS and twice at 65 °C with 1 x SSC, 0.5% SDS (Ausubel et al., 1987).
Determination of RP Transcripts by RT-PCR

Reverse transcription reaction was performed by using 1-3 µg of total RNAs and oligo-(dT) for the first strand DNA synthesis. The PCR reaction was taking place with a set of primers: RP-S1 (exon 4, 5'-CTGCCGCCCATCGTGCTG-3'), mRP-R2 (exon 8, 5'-TGAGGTGCTGCACATGGTAGG-3') with an expected product of 532 base pairs. PCR conditions were: 1 cycle at 94°C for 5 min; 35 cycles at 94°C for 1 min, 54°C for 1 min, and 72°C for 1 min; and 1 cycle at 72°C for 10 min.

Isolation of Mouse RP cDNA Clones and Sequence Determination

A mouse brain cDNA library (Clontech, Palo Alto, CA) from female C57 black B6 mouse strain was screened with [α-32P]dCTP-labeled human RP cDNA probe R1.1 (Shen et al., 1994). The cDNA library screening was performed by Drs Yung Yu and Lai-chu Wu. Positive clones were subcloned into PCR cloning vector, TA (Invitrogen, San Diego, CA) or Bluescript (KS+) vector. The cDNA sequences were determined by primer walking approach and double-stranded dideoxy sequencing.

Mouse RP Genomic DNA Screening

A Lambda FIX II genomic library derived from 129SVJ mouse was screened with mouse RP cDNA probe after manufacturer's suggested protocols (Stratagene, La Jolla, CA). Mouse Gene X clones were obtained by screening with probes: a 150 bp PCR products derived from human Gene XB; and a 541 bp fragment of 3'-end Gene XB genomic DNA by Scal-XbaI digestion. Genomic walking was performed by screening the library with subcloned probes derived from preceding Lambda clones. For example, the 2.2 kb fragment obtained from clone mRP5.4 was used for the screening of 5'-region of the RP1 gene. A 0.7 kb
BamHI fragment from clone GX13 was employed to probe the Gene XA or Gene XB for the duplication breakpoint studies.

The primers used for the DNA sequencing to the exon-intron structure are:

- mRP-S5, 5'-AGAAGGAGCTTCAAGAGC-S* (exon 5, sense);
- mRP-R7, 5'-GGTCCTGTAGTCTTCCGT-3' (exon 5, antisense);
- RP-S1, 5'-CTGCCGCCCCATCGTGCTG-3' (exon 4, sense);
- mRP-R6, 5'-GTACACCTTGGCTCTCAG-3' (exon 4, antisense).

**RESULTS**

5.1 Cloning and characterization of the mouse RP cDNA

To obtain the mouse RP cDNA clones, a mouse λ gt10 brain cDNA library was screened by using a human RP cDNA probe. The clone MRP1.0, with 1.0 kb DNA insert, was isolated and was completely sequenced by primer walking method. This cDNA clone contained an 18 bp polyadenylation tail which is 15 bp downstream of the polyadenylation signal. The MRP1.0 clone was not full-length with its 5'-end missing, when it was compared with human cDNA clone. A 3'-untranslated sequence of 284 bp in size is present in the mouse RP gene in contrast to 396 bp in the human RP (Figure 41). The deduced protein sequence consists of 241 amino acid residues that corresponds to exon 3 to exon 9 of the human RP gene.

The partial mouse protein sequence shares 87.6% sequence identity and more than 93% similarity with the human RP protein (Figure 41). The 5' end of this clone consists of one half of the putative bipartic Nuclear Localization Signal, (Robbins et al., 1991), GVKRRR, which is identical to that of human RP. The high sequence identity between human and mouse RP proteins infers an important
function that is conserved during the species evolution. Most of the amino acid diversities between the mouse and human RP proteins are conserved changes (for instance, residue #71 Asp to Glu). There are two notable amino acid differences: at residues 108 and 174, both Tyr residues from the human RP were changed to Cys residues in mouse RP. Whether these alternations affect the structural or functional properties of the RP protein are unknown. There are no N-linked glycosylation sites in the predicted RP protein sequence. Two potential protein kinase C phosphorylation sites at residues 145 and 277, two potential casein kinase II phosphorylation sites at residues 183 and 277, are conserved between the human and mouse RP protein. However, the potential tyrosine kinase phosphorylation site at residue 231 of the human RP has been changed to a cysteine residue. The putative post-translational modifications of the mouse RP protein also remain to be determined.

5.2 The RP genes are present in the mouse and other mammalian genomes

To determine whether the RP gene is also present among various mammalian genomes. A genomic Southern "zoo-blot" (including primates, rodents, dog, pig, cattle and sheep) was hybridized with a human RP cDNA probe, R1.1 (unpublished data). Multiple hybridized fragments were revealed, hence the RP gene is probably conserved in all mammalian genomes tested. This observation also implied the RP genes may be duplicated in human, mouse, rat, cattle, and sheep but not in the marmoset and pig genomes.

In order to further analyze if two RP genes are present in the mouse genomes, three mouse genomic DNAs were subjected to restriction enzyme digestions and hybridized with the RP cDNA probes. These studies revealed two fragments consistently. The RP cDNA probe hybridized to 9 kb and 20 kb fragments from EcoRI digestion (Figure 42); 5.5 kb and 2.5 kb for BglII digestion,
suggesting that two RP genes were clearly present in various mouse genomic DNAs (Jessica Philport, unpublished).

Two observations suggested that the RP2 duplication pattern was different from that of primates. Firstly, the genomic distance between the Slp and C4 genes in mouse is about 80 kb in size. This is the region where putative mouse RP2 and Gene XA reside (Chaplin et al., 1983). Secondly, by using primers YMR1 and HRP3 flanking the putative breakpoint we failed to obtain any product from the PCR experiment as shown in primates (Shen et al., 1994). Thus, it is possible that a full-length RP2 gene (complete duplication of the RP gene) may be located at the Slp-C4 intergenic region, in contrast to a partially duplicated RP2 gene that is only 913 bp in size in humans.

5.3 Isolation and identification of mouse RP genes

A mouse genomic library was screened to isolate λ clones to determine the genomic structures of RP1 and RP2. Two groups of overlapping clones, λMRP8.1, 5.3, 5.4, and 18; and λMRP6.2, 8.2 were obtained from 129SVJ genomic library by probing with the mouse RP cDNA fragments. The restriction enzyme map for these clones largely agreed with that of cosmid clones generated from the BALB/C strain (Chaplin et al., 1985). Based on the partial sequence of a Bluescript clone with a 3 kb insert derived from λMRP8.1 (Figure 43), and comparison of the mouse RP cDNA sequence with the published genomic sequence at the putative region of RP1 gene located upstream of Slp gene (FM mouse strain, Nonaka et al., 1986), we were able to identify three exons (exons 7, 8, and 9) downstream of retrovirus IMP (Figure 43). The poly-(A) signal of RP1 gene is located 603 bp 5' to the Slp translational initiation codon. DNA sequences similar to human RP exon 3-5 were detectable in clones λMRP5.3, 5.4, and 18.
Two overlapping clones, \( \lambda \text{MRP}6.2 \) and 8.2 were assigned to the genomic region upstream of C4, which is the location for the RP2 gene (Figure 44). Despite the divergent sequences from RP1 at their 3'-end region, five exons (exon 5 to exon 9) can be mapped from the genomic sequences available (B10.WR mouse strain in Ogata et al., 1991 and FM strain in Nonaka et al., 1986).

**Subcloning of RP genomic DNA restriction fragments**

**Mouse RP1 clones**

MRP8.1 clones with both orientations ligated to *NotI* digested BS vector were digested with *BamH1* restriction enzyme and released the following fragments: 0.5, 2.2, 2.5, 3.0, 3.5, and 5 kb of clone #1, while the clone #2 released 0.5, 2.2, 2.5, 3.0, 2.0, 6.5 kb (lanes 1 and 2, Figure 45A). Southern blot analysis revealed that only the 3 kb fragment hybridized to RP cDNA (MRP1.0) indicating the presence of exon sequences (lanes 1 and 2, Figure 45B). Based on the restriction patterns, The 1.0 and 3.5 kb *BamH1* fragments belong to the ends of MRP8.1. The 2.0 kb, 0.5 kb, 2.2 kb, 3.5 kb *BamH1* fragments were subsequently subcloned into pUC and BS vectors.

**MRP5.3 clones**

MRP5.3 clones ligated into BS vector with the inserts of both orientations were subjected to *BamH1* restriction analysis. Clone 5.3-#3 released 2.7, 3.2, 3.8, 5.2, and 2.9 kb fragments, while clone #4 released 2.7, 3.2, 3.8, 5.9, and 2.2 kb fragments. The Southern blot analysis with RP cDNA probe (MRP1.0) showed that 2.7 and 3.8 kb fragments were commonly shared by both clones and contained exon sequences (Figure 45A and 45B). In addition, the 2.9 kb fragment of clone #3 and the 5.9 kb fragment of clone #4 contained RP exon sequences. As a result, the 2.9 and 2.2 kb were assigned to 5'- and 3'-end of the MRP5.3 clones, respectively, and RP exons were located to the 2.7, 3.8, and 2.9 kb *BamH1* fragments. Four *BamH1* fragments: 2.2 kb, 2.7 kb, 3.8 kb, and 3.2 kb, were
subcloned into pUC and BS vectors. The sequencing of the both ends from pUC-3.8k, pUC-3.2k and the 3.0 kb subclone from MRP8.1 confirmed the presence of retrotransposon, IMP. The 2.2 kb subclone in BS-vector was further analyzed and further subcloned into six different fragments with restriction enzymes: Smal, Apal, BglII, and SstI digestion. Combined with several clones derived from ExoIII-nuclease digestion technique, the 2.2 kb fragment was complete sequenced.

**MRP5.4** The DNA insert from this clone was transferred to the BS vector in both orientations. Clone 5.4-#3 was digested with BamHI and released 2.2, 2.7, 3.8, 5.8 and 0.4 kb fragments, while the 5.4-#4 gave 2.2, 2.7, 3.8, 2.8, and 3.4 kb fragments (lanes 3 and 4, Figure 45C). Both 2.7 kb and 3.8 kb fragments hybridized with RP cDNA probe in Southern blot analysis (lanes 3 and 4, Figure 45D). Hence, MRP5.4 clone overlaps with clone MRP5.3 at 2.2, 2.7, and 3.8 kb BamHI fragments. The Southern blot was further probed with human genomic DNA probe: 1.3 kb from upstream region of RP1 gene. The 2.8 kb fragment from #2 clone and the 5.8 kb fragment from #1 clone gave rise to the positive signals (lanes 3 and 4, Figure 45E). Therefore, 2.8 kb and 0.4 kb fragments appear to be the 5'-end and the 3'-end of MRP5.4, respectively.

**MRP18** clone contained BamHI digested fragment of 7, 4.3, 2.2, and 2.7 kb (lane 1 of Figure 45C). From Southern blot analysis, the 2.7 kb fragment hybridized with RP cDNA probe, while the 7 kb hybridized with the RP upstream probe i.e. 1.3 kb human DNA probe as described in clone MRP5.4 (lane 1 of both Figure 45D and 45E). Thus, the 2.2 kb and 2.7 kb fragments overlapped with the clone of MRP5.4, and the 1.3 kb fragment was assigned to the far 5'-end of this clone. A BamHI-digested 7 kb fragment and a 2.3 kb BamHI-EcoRI was subcloned into pUC vectors.
**Mouse RP2 clones**

MRP8.2 clone was subcloned into BS NotI vector. These clones of one orientation when digested with BamHI enzyme, gave rise to 2.6, 2.0, 0.7 and 9.7 kb fragments. These clones with opposite orientation released 2.6, 4.7, 7 and 0.7 kb fragments. Southern blot analysis revealed that the #4, #6 and #8 clones hybridized with 2.6 kb and 2.0 kb fragments by RP cDNA probe, while the #1, #2, #3, #5 and #7 clones hybridized with 2.6 kb and 4.7 kb fragments (figures 46C and 46D). Thus fragments of 7 kb and 2.0 kb were assigned to both ends of these clones, and the RP2 exon-containing region was present in the 2.2 kb and 2.6 kb fragments.

MRP6.2 clones in NotI-treated BS vector were digested with BamHI. The resulting fragments from the clone of one orientation were 8.5, 6.2, and 2.5 kb in size, while the clone with another orientation were 8.5, 3.5, and 5.3 kb in size (lanes 1 and 2, Figure 46A). The Southern blot analysis showed that only the 2.6 kb in one clone and the 5.5 kb in the other hybridized with RP cDNA probe (Figure 46B). Therefore, the MRP6.2 clone overlapped with MRP8.2 clone at the 3' end and shared the 2.5 kb exon-containing fragment.

**5.4 The duplication structure of mouse RP1-Slp and RP2-C4 gene loci**

In order to determine whether the mouse genomic structure in the C4-RP locus resembles the human RCCX modular duplication, serial Southern blot analyses were performed. In this study, genomic DNAs were digested with EcoRI, a 9 kb fragment and a 20 kb fragment hybridized to a 700 bp, 3'-end cDNA probe suggesting two copies of RP-related genes (Figure 42C). It was noteworthy that the hybridization intensity among these two fragments exhibits no difference in contrast to that of figure 42B (with Exon 3-5 probe). It implied that both RP genes
hybridize equally with this probe DNA. However, with specific probes containing either exon 3 (2.2 kb fragment derived from MRP5.3) or exon 4, only one fragment was detected on various restriction enzyme digested genomic DNAs (Figure 42A, 42B). This results suggested that the sequence located at the upstream region of the RP1 gene is absent from the RP2 gene. It is worthwhile to point out a previous Southern blot analysis of mouse genomic DNA using an Exon 3-5 probe revealed two restriction fragments with different intensity on the autoradiography (Figure 42A). This lighter fragment is probably the RP2 because DNA sequence corresponding to Exon 1-4 is probably truncated in this gene, therefore, RP2 has a lower efficiency of hybridization to an exon 3-5 probe, when compared with RP1. Hence, the RP gene duplication breakpoint is likely located at intron 4 with respect to mouse RP1 gene, unlike the human RP2 gene by which the breakpoint of gene duplication is located at exon 7 (Shen et al., 1994).

To identify the presence of mouse Gene X sequence in the putative RCCX module as in the human genome, a 540 bp probe corresponding to the 3'-end of human gene XB was used to screen the 129SVJ genomic library. Partial sequence determination for one of the clones obtained, λMRP13 (Figure 43) revealed fibronectin type III repeat (A. Mendoza unpublished data). This clone also hybridized to CYP21-specific probe in a Southern blot analysis data. Therefore, the Gene X is probably present at the proximity of the CYP21 gene. This would imply that RP, C4, CYP21 and Gene X are linked together in the mouse MHC. Whether these four genes are duplicated in a pattern similar to that of human and apes remains to be determined.
5.5 The exon-intron structures and sequence variations of mouse RP1 and RP2 genes

The exon-intron structures of the RP1 and RP2 genes were derived from two approaches: (a) comparison of the published genomic sequences corresponding to the 5' regulatory regions of the mouse Slp and C4 genes (for exons 6-9); (b) DNA sequencing using a series of subclones generated by this study (upstream of exon 6). The exon 4 of the RP1 gene is located in the 2.7 kb \textit{BamHI} fragment of the clone \textit{\lambda}MRP5.4, while the exons 5 and 6 are present in the 3.8 kb \textit{BamHI} fragment of the clone \textit{\lambda}MRP5.3 (Figure 43). The exon 3 is located around the \textit{BamHI} junction of clones 2.2 kb and 2.7 kb of the clone \textit{\lambda}MRP5.4. The exon-intron boundaries were determined by specific primers: RP-S1, mRP-S5, mRP-R6, and mRP-R7. Determination of the precise exon-intron boundaries for RP1 exons 1-2 is progress in our laboratory. The hypothetical boundaries for RP2 exon 5-9 were obtained using similar approaches.

A comparison was made between the RP1 and RP2 gene structures from the exon 5 to exon 9. The sequences upstream of C4 and Slp genes appear to be highly divergent with major insertions and deletions. In the case of intron sequences, the RP1 gene has a characteristic proviral integration in intron 6, while the RP2 sequence has insertions of 141 bp, 31 ACACCC repeats and a 60 bp AC repeats. The first two insertions occurred in intron 6 and the last insertion is present in the intergenic region between RP2 and C4 (Figures 47 and 48). The provirus IMP was suggested to mediate the differential gene expression of Slp in many strains of male mice. It is located 1972 bp upstream of the Slp initiation codon. Its integration site is present immediately upstream of the mouse repetitive element B2 in the RP1 intron 6 (Nasara et al., 1986 and Figure 48).

No deletions in the coding sequence for the RP1 gene have been detected, on the other hand, the RP2 gene consists of various non-synonymous amino acid
changes (replacement) as well as synonymous mutations (silence). Therefore, the RP2 resembles a pseudogene and is likely without any biological function. If we assume that expression of the RP2 is taking place, some of the dramatic amino acid changes: Lys to Glu residue in exon 7, Arg to Trp changes at two positions of in exon 8, would change the properties of hypothetical products. Remarkably, if the transcriptional reading frame for RP2 gene has remained as in RP1, there is a 10-nucleotide deletion in exon 6 and a single nucleotide deletion in exon 5 of RP2 gene (Figure 49). These minideletions would result in frame-shifts and premature termination of translation. As a result, these mutations are another strong indication of the RP2 being a pseudogene.

5.6 The expression of mouse RP gene

Because of the complex RP genetic structure among mouse strains and the sequences diversity, at least, at 3' region of both RP1 and RP2 genes, it is unlikely that both genes encode proteins of the same function. Hence it is of interest to investigate whether the RP1 or RP2 gene contribute and express the mouse RP transcript. To determine the functional RP transcript and identifying the RP expression among different mouse strains and tissues distribution of individual mouse, Northern blot and RT-PCR analyses were applied in this studies.

A mouse Northern blot analysis (Dr. Lei-chu Wu, unpublished data) has indicated that the mouse RP transcripts are about 1.6-1.8 kb in size. It also suggested that the expression distribution among various tissues is ubiquitous, the same result as in the human RP expression studies (Chapter 2). Furthermore, using RT-PCR analysis with oligo-(dT) primer for reverse transcriptase reaction and two primers corresponding to exon 4 and exon 8 sequences, we were able to amplify 532 bp products from one cell line EL4, and three different mouse strains: 129SVJ, BALB/C and FVB (Figure 50) but not the negative controls (lane 1, 2, and 3).
These tissues which express RP varied from the liver, kidney, spleen to brain, and the RP transcripts were detected from both female and male mice, at least in liver of 129SVJ strain (lanes 4 and 5 of Figure 50).

**DISCUSSION**

The human RP genes are located in the upstream region of the complement component C4A and C4B genes in the MHC class III region. Various autoimmune or genetic diseases have been associated with the null alleles of the C4 (Schaffer et al., 1989, Fielder et al., 1983). Hence the characterization of the neighboring genes: Gene X, steroid 21-hydroxylase CYP21, and novel gene RP, is of interest to increase our understanding of MHC-linked diseases.

Here we reported the characterization of mouse RP genes. Using human RP cDNA probes, we were able to isolate mouse RP cDNA clones from a female C57/B6 mouse brain library. Although the deduced amino acid sequence does not show significant identity to any known protein, there is low degree of sequence similarities to two DNA repair proteins: bacteria *Micrococcus uvrA* (Shiota et al., 1989) and yeast RAD7 (Perozzi et al., 1986). Furthermore, the presence of bipartite nuclear localization signal and with more than 88% of sequence identity between human and mouse, all strongly suggest of a ubiquitously expressed protein which may play an important role in the biological function inside the nucleus. As discussed previously, this protein may be involved in the mechanism of DNA repair or associated with the regulation of transcription activity.

Mouse brain cDNA clone MRP1.0 was screened and cloned from female BALB/C cDNA library using a human RP cDNA probe. Based on available genomic sequence from the GenBank database and the comparison of the cDNA
sequence, this DNA transcript from a mouse brain cDNA library was aligned perfectly with upstream region of Slp gene. Exon 7 to exon 9 were mapped to the genomic sequence derived from FM mouse strain (Nonaka et al., 1986), and all of these exons were flanked by AG-GT splicing consensus sequence. RP1 gene which is located upstream of Slp gene is then expressed at least in mouse brain as a cDNA clone derived from RP1 locus, as well as all other tissue tested from Northern blots. Current data could map the mouse RP2 gene only up to the exon 5 with respect to RP1. Thus the breakpoint of gene duplication is likely located in intron 4, whereas the partial duplication and the breakpoint was present at the exon 7 for the human RP2. A comparison of the mouse RP1 and RP2 sequences revealed numerous point mutations and two minideletions at the partially duplicated RP2. There is no evidence to suggest that RP2 is functional. The sequence changes in RP2 could serve as a pool for generation of sequence variations for RP1 through unequal cross-over or gene conversion events.

The C4 gene is constituitively expressed in mice, while the Slp gene is expressed only under the regulation of androgens in the male mice of a subset of standard inbred strains. In contrast, in mice bearing wild-derived H-2w7, H-2w16, and H-2w19 haplotypes, the Slp is constituitively expressed and present at high levels in both sexes (Klein et al., 1975). This finding was consistent with previous discussion about the sequence analysis: these multiple Slp genes of the H-2w7 haplotype were actually C4/Slp hybrids being linked with C4 regulatory sequences (Nakayama et al., 1987; Stavenhagen et al., 1987). In addition, the expression of mouse C4 and Slp genes were also proposed to be regulated by non-H2-linked transregulatory factors (Bruisten et al., 1989). For instance, expression of Slp gene in mouse strains FM was testosterone inducible, but the expression in female mice was androgens independent. Presumably, it was controlled by non-H2 trans-acting gene in female mice (Yu et al., 1988; Brown et al., 1980).
Partial mouse RP1 and RP2 genomic structures were derived from genomic Lambda clones through chromosomal walking procedures. Partial characterization of these clones and mouse cDNA sequence, seven exons were assigned to RP1 gene. A proviral element, IMP, was present in the intron 6 of RP1 gene. The LTR of this retrotransposon was speculated to confer the androgen-responsiveness of the Sip gene expression (Stavenhagen et al., 1989). The actual presence of IMP in the RP1 gene prompted us to investigate whether the expression of RP1 gene was under the influence of the endogenous retrovirus enhancer in a strain-specific or sex-oriented manner. However, the RT-PCR and Northern blot analysis showed that this was not the case: both sexes of the mouse strain 129SVJ, a strain with H-2b haplotype without Sip gene expression (Chaplin et al., 1983), both expressed RP. Further supportive evidence was the isolation of RP cDNA clones from a female C57B6 cDNA library, C57B6 is another mouse strain with the H-2b haplotype. Therefore, the expression of RP gene does not appear to be under the same regulatory control as for the Sip gene.

In contrast to the human RP2 gene, the size of mouse RP2 gene appeared larger and the RCCX modular duplication was slightly different from that of human locus. Southern blot analysis and sequence data indicated that partial duplication of RP2 gene occurred at the intron 4 region corresponding to RP1 gene organization. Partial duplication pattern for the comparison between Gene XA and XB was not been determined.

For those wild-derived mice (i.e. H-2w7 haplotype mouse) containing multiple copies of Sip/C4 genes between the Sip and C4 genes, multiplication of C4-CYP21 could also involve RP and Gene X. However, the molecular structures of these multiplied genes have not yet been determined. Characterization of the genomic structures of the Sip/C4 hybrids genes in one H-2w7 mouse revealed RP2-like sequences upstream of the C4/Sip (unpublished data), and a mosaic
organization on these Slp/C4 hybrids (Pattanakitsukul et al., 1990). Hence it is tempting to speculate a RCCX modular multiplications in these strains.

Three different mammalian orders, primates, ungulates (cattle), and rodents contain duplicated C4-CYP21 structures in their genomes (reviewed by Miller et al., 1992). This could imply duplications of the RCCX modules. Two other observation suggested that the duplication of the RCCX modules of these mammalian species evolved independently. Firstly, the intergenic region between the CYP21A and C4B in the mouse genome is about 60 kb, instead of 7 kb at the corresponding region in the human genome. Secondly, the sequences of the mouse C4 or CYP21 genes are more similar to the duplicated counterparts than they are to any of the human C4 or CYP21 genes (Tosi et al., 1985). The duplication then occurred after the ancestors of each species separated. This post-speciation event of unrelated duplication for different species but at the same locus suggests that this dynamic locus is maintaining the continuous plasticity of genomic structure.
Figure 41

Partial sequence of the mouse RP cDNA and its deduced amino acid sequences

The cDNA and its derived amino acid sequences of MRP1.0 obtained from a mouse brain cDNA library are presented. The polyadenylation signal AATAAA is underlined. All amino acid residues are in one-letter codes. The residues placed below the mouse RP protein sequence are those of the human RP sequence which are diverged from the mouse sequence. The potential nuclear localization signal is doubly underlined.
Figure 42
Genomic Southern blot analysis of the mouse RP and RP2 genes

(A) Mouse genomic DNA from gut (lane 1), NSO (lane 2), EL4 (lane 3) were digested with EcoRI, resolved by 0.8% agarose gel electrophoresis, blotted to Hybond N membrane, and hybridized with a probe derived from the upstream region of the RP1 (exons 1-3, a 2.2 kb BamHI fragment).

(B) The same blot as in panel A but the membrane was hybridized with RP 5'-cDNA probe (a 300 bp StuI-EcoRI fragment of cDNA clone TA-MRP1.0).

(C) The same blot as in panel A but the membrane was hybridized with mouse RP 3'-cDNA probe (a 700 bp StuI-EcoRI fragment)
Figure 43

(A) Molecular organization of the mouse RP1 and RP2 gene locus
The relative location of the genes in mouse H-2 S region is shown (data taken from Chaplin et al., 1985, and from this work). Arrows represent the possible transcriptional orientations of these genes. Genomic clones corresponding to the RP1 and RP2 loci are shown under the gene map.

(B) The gene structure and the BamHI restriction map of the mouse RP1 gene
Four λ-genomic DNA clones (λMRP8.1, 5.3, 5.4, and 18) were obtained from the mouse 129SVJ library. Subcloned fragments (ligated to BS or pUC vector) are shown at the corresponding regions and the sizes of BamHI restriction fragments are indicated. The RP1 exon 3 to exon 9 are shown by dark-boxes. The retrotransposon, IMP (hatched-box) is present at intron 6.
Figure 43
The genomic organization of the mouse RP2 gene

Two λ-genomic DNA clones (λMRP8.2 and 6.2) were obtained from the screening of mouse 129SVJ genomic library. In addition to these full-length ~15 kb clones, three BamHI fragments: 2.6 kb, 8.5 kb, and 3.5 kb were also subcloned into Bluescript and pUC vectors. The dark boxes are the hypothetical exons assigned according to the similar sequence of the RP1 gene (although no RP2 transcript has been detected).
Figure 44
Figure 45
Analysis of RP1 genomic DNA clones from 129SVJ Lambda library
Four mouse RP1 genomic clones in Bluescript vector with inserts of MRP8.1 (lanes 1 and 2 of panels A and B); MRP5.3 (lanes 3 and 4 of panels A and B); MRP1.8 (lanes 1 and 2 of panels C, D, and E); and MRP5.4 (lanes 3 and 4 of panels C, D, and E). All clones were digested with *Bam*HI restriction enzyme, resolved in a 0.8% agarose gel (panels A and C). The Southern blots were probed with RP cDNA fragment, MRP1.0 (panels B and D), probed with 1.3 kb human genomic DNA fragment RP1.3 (panel E). All four overlapping clones have inserts of both orientations in the pUC vectors.
Figure 46
Analysis of RP2 genomic DNA clones from 129SVJ Lambda library
Two overlapping clones of RP2 genes: MRP6.2 (panels A and B) and MRP8.2 clones (panels C and D) with both orientations were subjected to *BamHI* restriction digestion, resolved in 8% agarose gels (panels A and C). Lanes 1, 2, 3, 5 of panels C and D are identical clones with one orientation of the MRP8.2, while the lanes 4, 6, 7, and 8 are from clones of the other orientation. The Southern blots were hybridized with RP cDNA, MRP1.0, and their autoradiographs are shown (panels B and D).
Figure 47
A comparison of the nucleotide sequences of mouse RP1 and RP2 from exon 7 to exon 9 region

The nucleotide sequence of the RP1 gene spanned from the 5'-LTR of IMP to Slp gene is presented, (modified from Nonaka et al., 1986). The residues placed below the RP1 sequence are those of the RP2 sequence which are diverged from the RP1 sequence. Dashed lines represent deletions. Underlines indicated the ACACCC repeats and ACAC repeats. The RP1 poly-(A) signal AATAAA is boxed.
Figure 47
Figure 48
A schematic diagram to compare the genomic organization of mouse RP1 and RP2 genes at their 3' regions

The structural comparison between RP1 and RP2 genes as the result from the data of figure 1 and figure 11. Five major insertions or deletions are indicated above and below the gene structures, respectively. Exons of RP1 and hypothetical exons of RP2 are represented with solid boxes. Positions for the transcriptional start sites of Slp and C4 are shown by arrows.
<table>
<thead>
<tr>
<th>Nucleotide:</th>
<th>Mouse RP1</th>
<th>Mouse RP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>exon 5</td>
<td>GAG</td>
<td>GAA</td>
</tr>
<tr>
<td></td>
<td>Glu</td>
<td>Glu</td>
</tr>
<tr>
<td></td>
<td>GCG</td>
<td>GGA</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Gly</td>
</tr>
<tr>
<td></td>
<td>ACG</td>
<td>AAG</td>
</tr>
<tr>
<td></td>
<td>Thr</td>
<td>Lys *</td>
</tr>
<tr>
<td></td>
<td>GNA</td>
<td>AA</td>
</tr>
<tr>
<td></td>
<td>Glu</td>
<td></td>
</tr>
<tr>
<td></td>
<td>92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CGG</td>
<td>CAG</td>
</tr>
<tr>
<td></td>
<td>Arg</td>
<td>Gln *</td>
</tr>
<tr>
<td>exon 6</td>
<td>TGTGATGGCCGA</td>
<td>T ........ A</td>
</tr>
<tr>
<td></td>
<td>103-106</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCG</td>
<td>GTG</td>
</tr>
<tr>
<td></td>
<td>Ala</td>
<td>Val *</td>
</tr>
<tr>
<td></td>
<td>GTG</td>
<td>GTA</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Val</td>
</tr>
<tr>
<td></td>
<td>GCC</td>
<td>GAC</td>
</tr>
<tr>
<td></td>
<td>Ala</td>
<td>Asp *</td>
</tr>
<tr>
<td></td>
<td>117</td>
<td>ACT</td>
</tr>
<tr>
<td></td>
<td>Thr</td>
<td>ACG</td>
</tr>
<tr>
<td></td>
<td>116</td>
<td>TAT</td>
</tr>
<tr>
<td></td>
<td>Tyr</td>
<td>TAC</td>
</tr>
<tr>
<td></td>
<td>112</td>
<td>CCG</td>
</tr>
<tr>
<td></td>
<td>Pro</td>
<td>CCA</td>
</tr>
<tr>
<td></td>
<td>142</td>
<td>exon 7</td>
</tr>
<tr>
<td></td>
<td>CAG</td>
<td>CAC</td>
</tr>
<tr>
<td></td>
<td>Gln</td>
<td>His *</td>
</tr>
<tr>
<td></td>
<td>CGA</td>
<td>ARA</td>
</tr>
<tr>
<td></td>
<td>Arg</td>
<td>Arg</td>
</tr>
<tr>
<td></td>
<td>156</td>
<td>AAG</td>
</tr>
<tr>
<td></td>
<td>Lys</td>
<td>GAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>exon 8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CGC</td>
<td>CAC</td>
</tr>
<tr>
<td></td>
<td>Arg</td>
<td>His *</td>
</tr>
<tr>
<td></td>
<td>CGG</td>
<td>TGG</td>
</tr>
<tr>
<td></td>
<td>Arg</td>
<td>Trp *</td>
</tr>
<tr>
<td></td>
<td>192</td>
<td>Ttg</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Ctg</td>
</tr>
<tr>
<td></td>
<td>196</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CGG</td>
<td>TGG</td>
</tr>
<tr>
<td></td>
<td>Arg</td>
<td>Trp *</td>
</tr>
</tbody>
</table>

**Figure 49**

Variations of the coding DNA sequences and its associated alterations in amino acid residues between RP1 and RP2

The putative exons of RP1 and RP2 genes and their derived amino acid sequences extended from exon 5 to exon 9 were compared. The positions of amino acid residues are after numberings in Figure 41. Two regions of RP2 gene with nucleotide deletions are indicated by dots. Nine residues which may lead to non-synonymous amino acid changes are asterisked.
Figure 50

RT-PCR analysis of various mouse strains and tissues

(A) The amplification products from RT-PCR using primers RP-S1 and mRP-R2. The RNA templates are from female 129SVJ liver (lanes 2 and 5), male 129SVJ liver (lane 4), EL4 cell line (lanes 3 and 6), BALB/C liver (lane 7) and FVB brain (lane 8). Lane 1 is a negative control without adding RNA, lanes 2 and 3 represent the PCR products without the reverse transcriptase.

(B) Autoradiography of the above Southern blot, hybridized with a RP cDNA probe (mRP-5D). Note that the RP gene transcripts are absent in the negative control experiment (lanes 1-3).
Figure 51
The gene-targeting strategy for mouse RP gene
The restriction map of the targeting construct and wild-type RP locus are shown. Exons are indicated by closed boxes, the retrotransposon, IMP is shown as a gray box. A Bluescript vector is shown as wavy-line, while the neo cassettes is indicated as an opened box. The positions of probe DNAs (a-e) are placed below restriction maps. The expected homologous recombination locus is shown at the bottom, its detection can be achieved by the alternation of the EcoRI-restricted fragments.
Probes:
a: 2.3 Kb Bm/R1
b: 400 bp RU/Bm
c: 2.2 Kb Bm
d: 300 bp cDNA containing exon 3, 4, 5
e: 2.8 Kb
f: 500 bp Bm/BglII

Enzymes:
B: BamHI
R: EcoRI

Figure 51
<table>
<thead>
<tr>
<th>Splice Acceptors</th>
<th>exon No.</th>
<th>exon bp</th>
<th>Splice Donors</th>
</tr>
</thead>
<tbody>
<tr>
<td>ND</td>
<td>3</td>
<td>210</td>
<td>AGCAG gtaaccatggcaa</td>
</tr>
<tr>
<td>tttatctgccccag</td>
<td>4</td>
<td>137</td>
<td>AACTG gtagacaccttggt</td>
</tr>
<tr>
<td>tctgtgagagcag</td>
<td>5</td>
<td>96</td>
<td>CCCGG gtgatctagtgt</td>
</tr>
<tr>
<td>acctgtgtcttag</td>
<td>6</td>
<td>140</td>
<td>ATCAC gtagatgtggtg</td>
</tr>
<tr>
<td>ctaccacctgcag</td>
<td>7</td>
<td>98</td>
<td>TAAAG gtagatgctctga</td>
</tr>
<tr>
<td>ccttcctctccag</td>
<td>8</td>
<td>148</td>
<td>GACTG gtagatctagctc</td>
</tr>
<tr>
<td>ttctgtttacacag</td>
<td>9</td>
<td>46</td>
<td>ATACA TGA (3'−UT 284 bp)</td>
</tr>
</tbody>
</table>
CHAPTER VI
POLYMORPHISM AND GENOMIC REARRANGEMENT OF THE RP GENE

INTRODUCTION

Searches for the alteration of a gene and for the loss of function through gene mutation are the important approaches for the molecular genetic studies of RP. These may provide insight of the RP function and lead to future research of the RP gene. As discussed in the previous chapter, gene duplication does not always have beneficial effects on organisms. Since one of the duplicated genes maintains the biological function, the other gene copy tends to accumulate mutations due to the release of selection pressure. In a positive manner, a new gene function might be evolved. In a negative manner, the deleterious mutations could be incorporated into the functional gene through gene conversion or recombination events and lead to genetic diseases. The duplication of the human C4-CYP21 gene cluster is a typical example. Apparently, there were unusually frequent gene conversion between CYP21A and CYP21B. CYP21A is a pseudogene with many point mutations and minideletions. Genetic exchanges between CYP21A and CYP21B would render the CYP21B gene defective and lead to pathogenesis of congenital adrenal hyperplasia (Harada et al., 1987; Chu et al., 1992). It is estimated that CYP21 gene conversion events account for about 85% of 21-hydroxylase deficiency (Miller et al., 1989; Morel et al., 1991).

The gene conversion mechanism was also suggested to be responsible for the DNA polymorphism of human complement component C4A and C4B (Yu et
al., 1986; Braun et al., 1990; Fredrikson et al., 1991). A mosaic structure, particularly within 1 kb of polymorphic C4d region, is due to the genomic shuffling of four discrete regions (Yu et al., 1986). Instead of causing diseases, the genomic exchanges create a combination of serological phenotypes between C4A and C4B, which show distinct class-specific, chemical reactivity and antigenicity.

Because of the high incidence of gene conversion between CYP21A and CYP21B and between C4A and C4B, it is tempting to believe that two other pairs of duplicated genes, RP1 and RP2, Gene XA and XB might be subjected to genomic recombination as well. Unpublished data from our lab by using Southern blot analysis demonstrated that one CAH patient with one allele of gene XB acquired the characteristic 121 bp deletion from gene XA. Gene XB encodes a protein with high sequence similarity to the extracellular matrix protein, tenascin (Matsumoto et al., 1992; Bristow et al., 1993) and is therefore also named Tn-XB. Gene XA is probably transcribed in an adrenal-specific manner with unknown function. The DNA sequence of gene XA has a >90% identity to that of gene XB. However, it is truncated at its 5'-end and bears a deletion causing a frameshift and the presence of premature translation stop signal (Gitelman et al., 1992). Due to these structural variations, the XA gene appears to be a pseudogene. Therefore the putative gene conversion between gene XA and XB occurring in this individual might have enormous effects on the genetics and pathology of the locus.

The duplicated RP1 and RP2 genes reported in this study might also suggest that the gene conversion between these two genes may well occur, given the high sequence similarity between these two genes. Moreover, mouse genomic data at the C4-CYP21 locus might also support the hypothesis that genomic recombination frequently occurred among various strains. One dramatic example was the identification of C4/Slp gene clusters in H-2w7 mice. The C3H.W7 mouse bearing H-2w7 haplotype has five C4-related genes (i.e. it has three extra copies besides the
sequences. Presumably, these hybrid genes and their mosaic structures were generated through multiple recombination or gene conversion events (Pattanakistsakul et al., 1990). Thus mouse RP1 and RP2 genes which are located immediately upstream of the S1p gene and the C4 gene, respectively, are likely to acquire similar hybrid structures resulting from genomic exchanges.

The generation of multiple copies of genes such as RP, Gene X or CYP21 would not be beneficial to the organisms. The presence of multiple hybrid genes (e.g., those individuals with three or more RCCX modular structures) would facilitate the occurrence of further recombinations, which would lead to pathological manifestations. In this preliminary study, we aim to investigate the polymorphism and rearrangement of the RP genes. Correlation of these genetic abnormalities with phenotypic changes would shed light on the long-searching and yet elusive RP gene function.

The sequence similarity of RP to RAD7 and uvr-A infers that RP could be related to DNA repair proteins. The presence of nuclear localization signal, characteristic CpG islands at the 5' region and the ubiquitous expression pattern of RP1 gene suggested that RP may encode a nuclear, housekeeping protein. Thus we hypothesized that RP may be involved in carcinogenesis. In human cancers, genetic data suggested the presence of susceptibility loci linked to HLA for the nasopharyngeal carcinoma (NPC) (Lu et al., 1990), Hodgkin's lymphoma (Klitz et al., 1994), cervical carcinoma, testicular teratocarcinoma and renal cell carcinoma (Apple et al., 1994). In rodents, it also has been shown that some congenic rats (Melhem et al., 1993) and strains of inbred mice (reviewed by Demant et al., 1989) with different MHC haplotypes differ greatly in their susceptibility to the development of spontaneous, chemical, and viral induced tumors. MHC genes of non-immunological function were implied in the involvement of carcinogenesis, especially from the class III region of the human and mouse MHC.
The nasopharyngeal carcinoma (NPC) was first chosen for the studies on the genomic alteration of RP in this study. NPC is one of the most common malignancies in the Southeast Asian population, occurring at a rate of 18 per 100,000 in Kwangtung. The rates of incidences per 100,000 in Hawaiians and in Alaskan natives are 7.8 and 13.5, respectively (Saemundsen et al., 1980). The gene(s) responsible for the tumor susceptibility in the NPC and other MHC-linked malignancies has not been identified. Therefore, it is intriguing to investigate whether there are mutations or rearrangement of the RP1 gene in these MHC-associated tumors.

MATERIALS AND METHODS

Sequence analyses

Human RP1 and RP2 sequences were described in Chapter 2 (GenBank accession numbers: L26260-L26263). One segment of Gene XA-RP2 was retrieved from GenBank and the database accession number is S38953. Six entries containing portion of mouse RP1 and RP2 sequences were fetched from database where accession numbers are M17440, M17442, M17443, M64933, M14225, and D90052. The comparison of the DNA sequences was performed by the Wisconsin GCG FASTA program from Pittsburgh Supercomputer Center. BESTFIT, PILEUP, and PRETTY sequence analysis programs were applied.

Cell line and genomic Southern analysis

Nasopharyngeal carcinoma cell line CNE was obtained from Dr. R. Glaser. Cultured cells were maintained in DMEM culture medium as described in Chapter 2. Genomic DNA was isolated according to standard protocols. Genomic Southern analysis was performed (Dangel et al. unpublished data). Probe DNA fragments
were derived from RP1 gene (Figure 10 of Chapter 2): P1, 1.6 kb BamHI fragment (nucleotide 1-1597); P2, 0.8 kb BamHI fragment (nucleotide 1594-2418); P3, 320 bp BamHI-EcoRI fragment (nucleotide 2415-2731); P4, 300 bp fragments (a doublet) from NcoI-StuI (nucleotide 9028-9037) and StuI digestion (nucleotide 9335-9646).

Cloning of rearrangement fragment from CNE cell line

About 100 µg of genomic DNA from CNE cell line was digested with restriction enzyme Ssfl and resolved in a 0.8% LMP agarose gel. DNA fragments of about 1.5-1.9 kb were isolated from the agarose gel. Approximately one-fifth of the fractionated DNA was ligated with ~1 µg of Ssfl-digested and phophatase-treated Bluescript (KS+) vector at 50 µl of reaction volume. 5 µl of the ligated products were then subjected to PCR amplification.

The first cycle reaction was carried out with only one RP-specific primer: RP-R7 (5'-GTGACAGGGAATTCCAGGAGT-3') at the following conditions: one cycle at 94°C for 5 min, 54°C for 2 min, and 72°C for 30 sec. Subsequently, a second primer BS-T3 (5'-ATTAACCTCCTCACTAAAG-3') was added and PCR was carried out under the following conditions 35 cycles at 94°C for 1 min, 54°C for 1 min, and 72°C for 1 min. About 1/5 of the PCR product was then used for nested-PCR with primers: RP-R6 (5'-CTTCACCAGCTGCGTGCGGC-3') and BS-KS (5'-CGAGGTCGACGGTATCG-3') at the same conditions as the first PCR for 35 cycles. The resulting products were resolved in a 0.8% agarose gel and the presence of RP sequence was confirmed with Southern blot analysis with RP cDNA R1.1 probe. The 450 bp PCR products were cloned into TA cloning vector (Invitrogen San Diego, CA) and sequenced by dideoxy sequencing method.
RESULTS

6.1 Sequence comparison between 3'-end of mouse RP1 and RP2 genes

Six available sequences derived from the GenBank database corresponding to the 3' regions of the RP1 or RP2 loci were aligned together for sequence comparison. These sequence variations shown in Figure 52, span about 1500 bp and start from the exon 7 of the RP gene to the transcriptional start sites of the C4 and Slp genes (Nonaka et al., 1986; Rosa et al., 1987). Sequence I (M17443) represents a transcribed functional RP gene due to perfect match with mouse RP cDNA sequence MRP1.0 (see Chapter 5), while sequence VI (M17440) is the RP2 gene located upstream of C4 gene. Thirteen nucleotides with RP1-specific sequences are observed in the exon 7 to exon 9 region (indicated by the relative positions in Figure 52). Seven RP1-like nucleotide substitutions were observed in these RP1/RP2 hybrids (underlined). These included three positions in sequence II (M14225) and four positions in sequence III (D90052). They represented a mosaic structure with the interchange of RP1-like or RP2-like sequences. It is interesting to note that about 700 bp sequences downstream of polyadenylation signal of hybrid RP gene, are identical to those of the C4 gene. These results suggested that the genetic events which generated the hybrid RP1/RP2 structures were similar to that of the C4/Slp hybrid (sequences comparison of C4/Slp genes was described in Pattanakitsakul, et al., 1990). Thus multiple genetic recombination or gene conversion events might play a role in the diversification of the extra copies of the RP1/RP2 gene in the H-2\textsuperscript{w7} mouse.
6.2 Sequence alignment between 3'-end of RP1 and three RP2 genes: Evidence for a gene conversion event at the human RP locus

In an effort to characterize the breakpoint between RP2 and Gene X and to distinguish the duplicated sequences among RP1 and RP2 genes, two subcloned fragments from two different individuals were analyzed. A 2.1 kb *BamHI*-*TaqI* restriction fragment was derived from clone λ-JM2a corresponding to the 5'-end region of C4B5 gene of the C4A4 C4B5 haplotype. This fragment along with another RP2-specific *BamHI* 5 kb fragment derived from cos 4A3 with unusual haplotype C4A2 C4B1 C4B2 were completely sequenced (Chapter 2). The genomic organization for their incomplete duplication pattern at RP2 and Gene X loci was described previously (Figure 16 of chapter 2). However, the sequence alignment between these two RP2-GeneX fragments indicated that they were identical within a 2.1 kb region except three nucleotides changes. Interestingly, the sequence alignment together with one database sequence entry #S38953 (Gene XA) from GenBank and the 3'-end region of RP1 gene (also see Figure 10 of Chapter 2) revealed four RP2-specific sequence substitutions (Figure 53). It was notable that at position # 10646 (RP1 gene numbering) of clone λ-JM2a (RP2 gene), we identified an RP1-specific nucleotide, T instead of the C residue shared by other two RP2 clones. It appears to be an example of a genetic mimic-conversion between the RP1 and RP2 genes.

Further evidence showing the occurrence of the gene conversion event was the presence of putative chi-like sequence (CCCAGC, located 54 to 60 nucleotide downstream), and a pair of inverted heptanucleotide repeats at the proximity of the mutation (at 13 bp upstream, and 42 bp downstream of putative converted residue, Figure 53). A similar observation was reported for a *de novo* mutation in the 21-hydroxylase locus (Collier et al., 1993).
The single example at this RP locus, unlike the great body of information regarding the genetic exchanges of CYP21 and C4 genes, does not conclusively demonstrate the gene conversion event. We then should interpret this data cautiously since the possibility for a random mutation at this RP2 gene could not be ruled out for this one nucleotide change.

6.3 Genomic rearrangement of one NPC cell line, CNE

The discovery of the human RP gene rearrangement was originated from the extensive studies of whether RP gene is associated to various MHC-linked autoimmune and malignant genetic diseases. One of the nasopharyngeal carcinoma cell lines, CNE, was obtained for the screening on possible RP gene mutation. A genomic rearrangement was detected from Southern analysis by using RP cDNA as a probe.

The genomic DNA was subjected to digestion with six different restriction enzymes and the blot was probed with RP cDNA. The data was summarized in Figure 54 (unpublished data). Extra DNA fragments of 1.7 kb, 6.5 kb, 8.5 kb, 8.8 kb, and 23 kb were detected from SstI, Accl, NcoI, ScaI, BglII, and BamHI digested DNAs, respectively. To determine the putative rearrangement breakpoint, a variety of genomic DNA fragments and RP cDNA segments were applied as probes to close in the site of rearrangement (Figure 54B). One of the probes extended from exon 3 to intron 4 (P3) could detect all of the rearrangement fragments, whereas the probe derived from RP 3' region (P4 at intron 5) could only detect large rearranged fragments such as the 23 kb of BamHI fragment and the 7 kb of ScaI fragment. When DNA fragments further upstream of RP gene: a BamHI 0.8 kb fragment relative to RP exon 2 and 3 (P2) and a 1.6 kb fragment derived from RP exon 1 to intron 1 region (P1), were chosen as probes, no rearranged fragments were detected. The same negative results were obtained when DNA fragments
derived from C4 gene was used to hybridize the Southern blots. Hence, all these data suggested that the putative rearrangement breakpoint was located in a region between exon 3 to intron 4 of RP1. More specifically, it was within a 320 bp BamHI-EcoR1 fragment of RP gene. Therefore, the cloning of the breakpoint and the analysis of the sequence around this junction became essential to understand the molecular mechanism of the rearrangement and the possible disease association of RP with the NPC tumor.

6.4 Cloning and determination of rearrangement breakpoint

A genomic library construction of CNE was undertaking by using the DNA isolated from CNE cells. The genomic clones with the rearranged RP will allow us to construct a detailed map of truncated RP gene and identify the rearranged DNA segment which fused with the RP gene. Here we adopted a second approach: Rapid Amplification of Genomic End (PAGE), a technique enabled us to obtain an unknown genomic DNA segment without the screening and subcloning from genomic library.

100 µg of CNE genomic DNAs were digested with SstI enzyme, subjected to LMP agarose. A fraction of 1.5-1.9 kb which corresponded to the rearrangement fragment detected by a RP probe was isolated and purified. Both un-fractionated and isolated DNA fragments were ligated with SstI-digested, phosphatase treated Bluescript vector. Followed by two sets of PCR, a 450 bp product was obtained (Figure 55A) and subcloned into a TA PCR cloning vector. Having shown that this PCR product hybridized with RP cDNA probe in Southern blot analysis (Figure 55B), we concluded that the vector-anchored PCR technique had successfully cloned the rearrangement fragment at the proximity of the breakpoint. Determination of DNA sequences from this 450 bp clone revealed portion of RP exon 4 sequence as expected (Figures 56 and 57). The rearranged breakpoint was
located at nucleotide position 2590 (Shen et al., 1994) of RP1. About 300 bp of novel recombination partner (REP) linking to exon 4 was obtained. The sequence flanking the breakpoint and a schematic representation of the putative mechanism for this recombination event (see Discussion) were shown in Figure 58.

**DISCUSSION**

**Gene conversion event in RCCX locus**

The RCCX (RP, C4, CYP21 and Gene X) modular organizations for human and mouse genomes are described in Chapter 2 and Chapter 5, respectively. In humans, the multiple haplotypes of RCCX modules vary from one to four copies have been reported (Kawaguchi et al., 1992). The breakpoints of the four-gene duplication for RCCX bimodular and trimodular structures are present at exon 7 of the RP1 genes (Shen et al., 1994). In mouse, the RCCX modular structures are more complex, especially for those with the wild-derived haplotypes (i.e. the H-2w7 mice). The breakpoint of RCCX duplication appears different from that of the humans, although the detailed structure has not been fully characterized (Chapter 5).

The modular duplication/deletion pattern involving these four genes may have an impact on the evolution of the MHC. Moreover, partial gene duplication has been suggested to be one of the major mechanisms leading to genetic diseases (Hu and Worton, 1992). Because of the high degree of sequence similarity between the tandemly repeated RCCX modules, a nonreciprocal exchange of homologous region is a common event at this four-gene cluster. Most of 21-hydroxylase deficiencies appear to be due to gene-pseudogene exchanges (reviewed by Miller, 1992). In addition to the gene conversion-like event accounting for a majority of the 21-hydroxylase deficiencies, many other clustered gene families are also prone
to such gene-pseudogene exchange. For example, gene clusters of globins (Slightom et al., 1980), immunoglobulin (Bentley et al., 1983), haptoglobin (Maeda et al., 1986), α1-acid glycoprotein (Merrit et al., 1990), complement C4 (Braun et al., 1990), class I and II MHC antigens (Kuhner et al., 1991), glutathione S-transferase (Taylor et al., 1991), glycoporphins (Huang et al., 1992), red-green visual pigments (Deeb et al., 1992) and CYP2D genes (Hanioka et al., 1990) were speculated to associate with gene conversions. Moreover, it was reported that the Gaucher's disease was the result of a gene-pseudogene exchange at the acid β-glucosidase (He et al., 1992). Therefore, gene conversion may be a general phenomenon in the mammalian genome.

The molecular mechanism for gene conversion has not been determined completely. Although there were extensive studies in lower eukaryotes, the term gene conversion has been loosely applied to the genetic exchanges observed in mammalian genome. Basically, the genomic exchanges such as the alteration of an internal portion at a specific locus has been replaced by a homologous segment copy from another allele or locus in a nonreciprocal manner (reviewed by Kourilsky, 1986). The presence of chi-like sequence (GCTGGGGG) located at the proximity of the conversion mutation of CYP21 gene has been reported (Amor et al., 1988). Palindromic sequences are also associated with sites of DNA breakage during gene conversion (Krawinkel et al., 1986). Alternatively, the acquisition of CYP21A-associated mutations could be explained by multiple recombination, double crossover events, or multiple independent point mutations. In fact, the product of gene conversion (nonreciprocal recombination) was impossible to be differentiated from that of two independent, sequential, reciprocal crossing-over (Kawaguchi et al., 1992).

Almost all examples such as 21-hydroxylase, the β-globin gene, the MHC genes, and others, gene conversions occur between multiallelic genes or between
genes and pseudogenes located on the same chromosome. In contrast, a genetic
defect between chromosomes 12 and 22 of von Willebrand factor (vWF) responsible for von Willebrand disease (vWD), the gene-pseudogene conversion event was inter-chromosomal (Eikenboom, et al., 1994). Remarkably, two chi-like sequences are also presented in proximity to the putative site of gene conversion. Thus the presence of chi sequences may facilitate the relatively frequent conversion events of the vWF gene.

Due to the frequent events for CYP21 functional gene-pseudogene exchange and the C4A-C4B genes polymorphism at the RCCX locus, we hypothesized that a similar gene conversion between RP1 and RP2 or between Gene XA and XB might as well occur. A segment of 900 bp and about 5 kb duplicated sequences are present at RP gene and Gene X, respectively. Three nucleotide changes leading to non-synonymous amino acid changes at exon 8 of RP1 and corresponding region of RP2 might be the candidates for the RP gene conversion.

There was only one incident for the human RP2 acquiring a RP1-specific nucleotide. Due to limited sequence data on the RP1 gene, we could not identify any RP1 locus with RP2-like sequence. If any genomic exchange is present at RP1 gene, two substitutions of the amino acid residues in exon 8 might occur. They are at the nucleotide # 10585 (from Ser to Gly); nucleotide # 10646 (from Ala to Val) (Figure 53). Hence, studies of resulted genomic mutations from these gene conversion events (if they do happen) may facilitate the understanding of the RP gene function.

Moreover, a gene conversion or genomic recombination event occurring at the Gene X (described previously, manuscript in preparation) was identified in a CAH patient. Despite the diversity of mouse RP-C4-CYP21-GeneX locus, the duplicated RCCX locus also represents one of the most dynamic regions of the human genome. Taken together, the ongoing duplication and contraction
(recombinational deletions), in addition to the frequent gene conversion events for every gene member in this locus are an interesting topics for the study of the human genetics.

**Genetic rearrangement in a NPC cell**

It has been consistently shown that the tumor suppressor gene p53 has frequent alteration in a wide variety of human tumors (Levine et al., 1991; Hollstein et al., 1991). Half or more of all human tumors examined to date have detectable p53 genetic lesions. Most of the somatic mutations in p53 gene in human tumors consists of a point mutation in one allele accompanied by the loss of the second allele (review by Harvey, 1993). The loss or rearrangement of both p53 alleles or mutations in only one allele have also been observed but at a less frequency. Sometimes the duplication of a mutant allele occurs spontaneously in human tumors. That was suggested to be the result of mitotic recombination, nondisjunction, reduplication, or gene conversion (Hansen et al., 1987).

The discovery of the RP1 gene deletion/duplication in NPC cell is a reflection of genomic instability within tumor cells. Approximately 10-40% mutations of a tumor suppressor gene for retinoblastoma (Rb), are gene deletions. Certain regions of Rb gene were speculated to be hotspots for deletion breakpoints (Susan et al., 1989). Such breakpoint cluster region were commonly involved in chromosomal translocations in malignancies such as chronic myelogenous leukemia (Groffen et al., 1984). In the human α-globin gene cluster (Nicholls et al., 1987) and in the hamster adenine phosphribosyl transferase (APRT) locus (Nalbantoglu et al., 1986), frequent breakpoint clusters were also observed. The knowledge of why certain regions of a genome are prone to deletions may have relevance to the human oncogenesis. However the molecular mechanism of deletion is not well understood. It was suggested that those sequences containing such as dyad
symmetries, short direct repeats, and interspersed repetitive elements were particularly susceptible to genomic recombination events. The slipped mispairing (a model described frameshift mutation proposed by Streisinger et al., 1966) during DNA replication might lead to the formation and excision of a single-strand loop between the repeats. However, two segments of DNA might also be physically close to each other in the cell nucleus, and joint together by random breakage and reunion.

The RP1 deletion/duplication in the CNE cell was rather unusual. First, RP gene is located in either single or double modular RCCX structure. The location provide a large repetitive sequences (two C4 genes, two CYP21 genes, and partial duplicated RP and Gene X sequences) allowing homologous recombination to occur. Second, the RP gene share sequence homology to yeast DNA repair protein RAD7, any dysfunctional mutations might predispose to oncogenesis similar to mutant tumor suppressor genes. Hence the gross rearrangement of one allele of RP gene might imply the loss of heterozygosity for the tumor cells. Hence, it is important to identify the sequence integrity of another allele: as if the cells consist any pre-existing point mutation at this allele. Otherwise, the RP translocation may be merely a random rearrangement as the result of genomic instability of tumor cells. Further characterization of the recombination partner in the RP locus is required to understand the molecular basis for the rearrangement in this NPC cell line.
Figure 52
A summary of the variations for six mouse RP sequences at their exon 7 to exon 9 region

The putative RP exon-intron structure is shown at the top, and the transcriptional direction of C4 or Slp gene is marked by an *arrow*. The nucleotide A in the C4 initiator codon ATG is denoted +1, and residues preceding it are given negative numerals. The positions and DNA sequence variations of six mouse RP genes are indicated. The GenBank accession number are: M17443 (I), D90052 (II); M14225 (III); and M64933 (IV); M17441 (V). I and IV are the genine sequences for RP1 and RP2, respectively (RP1 is *bold-typed*). The others are RP1/RP2 hybrids with mosaic sequences. RP1 substitutions in the RP1/RP2 sequences are *underlined*. The single nucleotide deletion at nucleotide -525 upstream of Slp-specific gene is denoted by *dash line.*
Figure 52
Figure 53
A sequence alignment of the human RP1 and RP2 genes at the exon 8 flanking region

One of the RP2 gene sequences (λ-JM2a) is shown on the top. The dotted lines indicate identical nucleotides for the other two RP2 genes, designated cos 4A3 and Xa. The RP1 gene sequence is shown on the bottom. The numbering is adopted from RP1 sequence (Figure 10 of Chapter 2). Four nucleotide variations are presented as boxes. Asterisk indicates the position where RP1 gene starts to deviate from RP2 sequences. Overlined sequences denote chi-like sequences: CCCAGC and GCTTGG. A pair of inverted heptanucleotide repeats are shown as underlines.
Figure 54
Cloning of the recombination partner fragment using RAGE (Rapid Amplification of Genomic End) technique
(A) The organization of human RP1 gene (modified from Figure 11 of Chapter 2).
(B) A restriction map of human RP1 gene. Four different restriction enzymes are shown in bold-type which have been described in the text. P1, P2, P3, and P4 represent four DNA probe fragments for genomic Southern analyses. Their corresponding locations are indicated as bars above the map.
(C) The putative restriction map of the rearranged fragment in NPC cell line. Thick line denotes the recombination partner which fuses with RP gene at the breakpoint. The 1.7 kb SstI restriction fragment from CNE genomic DNA (flanked by dashed lines) was ligated into BS (KS+) vector (wavy line). The relative positions of two primers: RP-R6 and BS-KS are shown as arrowheads.
Figure 54
The PCR products for the cloning of CNE rearranged fragment

(A) The PCR products of the vector-anchored Ss/I fragment from CNE cells were resolved in a 0.8% agarose gel. The first PCR amplification was performed by using primers, RP-R7 and BS-T3, and Ss/I digested genomic DNAs which were fractionated (lanes 1 and 2) or unfractionated (lane 3). The nested-PCR was performed with primers RP-R6 and BS-KS. The DNA templates used in the nested-PCR were from the lane 2 products (the resulted reaction was shown in the lane 4), and from the lane 3 products (shown in the lane 5).

(B) An autoradiograph of the Southern blot analysis from (A), which was hybridized with RP cDNA probe, R1.1.
Figure 56
Sequence analysis of the vector-anchored PCR product for the NPC rearrangement breakpoint

The PCR product of about 450 bp was amplified with primers RP-R6 and BS-KS and subsequently subcloned into a TA vector (see RAGE procedure at Figure 55). This clone was sequenced by dideoxy method with oligonucleotide BS-T7 and resolved in a 6% polyacrylamide gel. The sequencing revealed the breakpoint for the recombination partner, indicated as an arrowhead.
Figure 57
The sequence of the PCR product for the NPC rearrangement study
Partial sequences of the recombination partner and partial exon 4 sequences of the RP gene are shown. The rearrangement breakpoint is represented as an arrowhead. Encoded amino acid residues and the exon-intron junction (underline) of the RP gene are placed under the nucleotide sequences (the numberings are after Figure 10 of Chapter 2). The SstI site used in the vector-PCR is bolded.
Figure 58
A schematic representation of the putative mechanism of the slippage synthesis for the rearrangement of the CNE cells
Transcriptional orientation of the REP and RP1 gene are indicated as arrows. The complementary sequences are paired with bars. The recombination breakpoint is presented between A to C nucleotide.
Chapter VII
Discussion

In this thesis, I described the identification, molecular cloning and characterization of the novel gene RP located upstream of the complement C4A and the C4B genes in the MHC class III region of chromosome 6. Many immunological disorders such as systemic lupus erythematosus, immunoglobulin IgA deficiency and common variable immuno-deficiencies, and malfunctions in reproduction have been linked the human MHC, particularly, the close proximity of the C4 gene (Chapter 1). The association of the MHC to these autoimmune and genetic diseases prompts us to further investigate the structure, genetics and function of RP. This study will provide a model in general for the characterization a novel gene, since enormous amount of genes of unknown function are identified through the human genome project (s).

In the past few years, improvement in the physical and genetic mapping and new techniques have a significant impact on the efficient identification of disease genes. Traditionally, there was structural or functional information, such as protein sequence or available antibodies, for identifying a disease gene with known biochemical defect. However, "position cloning" has been proved successful to clone numerous disease genes on the basis of their chromosome locations, rather than prior knowledge about the defective protein (reviewed by Collins, 1992). For instances, the genes responsible for Duchenne muscular dystrophy (DMD) or fragile X syndrome were discovered through genomic rearrangements detectable at the Southern blot level (Monaco et al., 1986; Fu et al., 1991). In the case of
the RP research, a "candidate gene" approach can be referred as the method for characterize a disease gene without exhaustive cloning by genetic or physical mapping. A well-known example of the "candidate gene" approach is the association of p53 gene to Li-Fraumeni syndrome (Malkin et al., 1990).

There is a common scheme to characterize a novel gene such as RP. First, the complete cDNA and genomic structures were determined. Second, by the comparison of deduced RP protein sequence with the current protein databases to search for genes with related/similar structures or possibly functional domains. Third, the expression pattern of RP was studied to understand the expression regulation. Fourth, genetic aberrations or mutations of RP was investigated and its probable disease-association could be analyzed. Fifth, anti-RP sera were generated and used to investigate the biochemical properties of the endogenous protein. Last, but not the least, the RP "gene-knockout" mice are being produced, which will provide a important tool in the studies for biological function in vivo.

Intelligenetics and GCG software packages were used in the RP sequence analysis. It revealed little significant identities with other known proteins. However, low degree of similarities were found in genes corresponding to nucleotide excision repair proteins in yeast RAD7 and Micrococcus uvrA gene (Shiota et al., 1989; Perozzi et al., 1986). The DNA-repair is a complex biochemical process requiring coordination interaction among many components and the damaged DNA (Grossman et al., 1988). The DNA excision repair system has been most extensively studied in Escherichia coli. At least five gene products are involved in the repair process: uvrA, uvrB, uvrC, uvrD and polA (reviewed by Sancar et al., 1988). These proteins interact with bulky DNA-adducts, such as cyclobutane pyrimidine dimers or 6-4 photo products, in a strict sequential manner. These involve DNA recognition, excision, and post-excision events such as translocation, turnover or release of protein complex and repair synthesis (Friedberg, 1988). The
Eukaryotic DNA repair systems are more complex and less well characterized. Recent advances have been made in the study of nucleotide excision repair process in eukaryotes and in the understanding its relationship to a group of human inherited diseases. These disorders are caused by mutations in DNA-repair genes, including xeroderma pigmentosum (XP), Cockayne's syndrome (CS), and trichothiodystrophy (TTD) (reviewed by Aboussekhra et al., 1994). Most affected individuals are hypersensitive to UV light, and have various types of neurological and developmental abnormalities. XP patients, in particular, are highly susceptible to the development of skin cancer.

The RAD7 gene is involved in nucleotide excision repair system in the yeast Saccharomyces cerevisiae. While human homologs for many components (RAD1, 2, 3, 4, 14 and 25) involved in the DNA-repair mechanism of the yeast have been isolated and linked to human inherited diseases, the human gene analog to the RAD7 has not been cloned. The yeast with mutation in RAD7 were shown to be partially defective in the excision repair of UV-damaged DNA (Perozzi et al., 1988; Miller et al., 1982).

Recent report hypothesized that the RAD7 may have a putative functional domain relating to specific protein-protein interaction (Schneider et al., 1991). The putative protein-protein interaction domain, namely, the leucine-rich motif (LRM). Unlike other protein-protein interacting domains such as EGF-, kringle-, or complement repeats, the LRM has a less conserved structure and its recognition could be easily overlooked through search algorithms for functional domains. LRM has amino acid consensus residues, $\alpha\alpha\alpha\alpha\alpha\alpha xxLxxLxxLxx\alpha\alpha\alpha xxN/CxLxx\alpha\alpha\alpha xx$, where $\alpha$ represents hydrophobic amino acid. The yeast adenylate cyclase (Suzuki, et al., 1990), Lutropin-choriogonadotropin receptors (McFarland et al., 1989), 5T4 onco-trophoblast (Myers et al., 1994), and two DNA repair proteins, RAD1 and RAD7, were recently found to consist of tandemly repeated LRM (Schneider et al., 1991).
Presumably, the LRM s in these proteins are involved in specific protein-protein interactions.

The derived protein sequence of the RP gene has high Leucine content similar to RAD7. Leucine consists of 13% of the C-terminal 220 amino acids, where at least five tandemly LRM-like structure similar to these of RAD7 are present (Schneider et al., 1991). Another indication for possible role of RP for the protein-protein interaction is that the region of uvrA which is also homologous to the RP has been suggested to associate with protein oligomerization (Shiota et al., 1989).

In addition to the LRM s, RP and RAD7 both contain the bipartite nuclear localization signal (NLS) at the analogous position region. These sequences are perfectly matched the current best definition for nuclear localization signal (Dingwall et al., 1991). About 56% of the known nuclear proteins contain this bipartite signal. Only 4.2% of non-nuclear proteins in the SWISSPROT database contain the similar signal. However, most of them were categorized with proteins targeted to specific locations: secretory, membrane-bound, mitochondrial and chloroplast proteins. In other words, the non-nuclear proteins are directed to organelles within the cell with dominant signal peptides (reviewed by Pugsley, 1990).

The possible role of RP as a DNA-repair protein lead us to explore the involvement of RP in diseases with DNA-repair defects. It was found that a patient with type C xeroderma pigmentosum exhibited neurological disorders and systemic lupus erythematosus (SLE) (Hananian et al., 1980). SLE is strongly associated with the null alleles at the C4 or C2 loci (Fielder et al., 1983). The RP gene is closely linked to these disease-susceptibility loci and it has sequence homology to DNA repair proteins. Therefore, it is of interest to study whether RP is involved in the pathogenesis of these diseases.
In addition to the MHC-linked autoimmune disorders and malignant diseases described in Chapter 1, many studies have inferred the presence of genes involved in recurrent spontaneous abortion (RSA). The early observation showed an increased sharing of HLA antigens in couples having recurrent spontaneous abortion (Gill et al., 1983, Coulam et al., 1987, Schacter et al. 1984 and McIntyre et al., 1984). Couples who have recurrent spontaneous abortions may have other types of reproductive failures (Coulam et al., 1989; 1991). There is also an increased prevalence of RSA, cancer, and congenital anomalies in the families of couples with RSA or gestational trophoblastic tumors (Ho et al., 1991). They serve as genetic markers for the sharing of MHC-linked genetic defect, and the defect genes are responsible for the normal growth and reproduction (Gill et al., 1984 and 1987). In addition, these genes may also be involved in the pathogenesis of congenital anomalies and in the increased susceptibility to cancer (Gill et al., 1984). This strong association between reproduction defect and the tumor susceptibility implied that two distinct genes (one is developmentally important, while the other is a tumor suppressor gene) are closely linked to the MHC. The RCCX region of human MHC is a candidate locus responsible for these defects. Gene X encodes a tenascin-like protein, an extracellular matrix protein. Tenascin has been suggested to be involved in embryogenesis (Matsumoto et al., 1992), and therefore it is logical to investigate if tenascin-X is involved in the growth and reproduction defects, such as RSA. The similarity to a DNA-repair protein, on the other hand, leads us to hypothesize RP is a putative tumor suppressor protein as described above.

Similar observation is also present in congenic strains of rat with a deletion in the MHC-linked region having growth and reproductive defects (grc-). These rats are associated with increase prenatal mortality, decreased body size, male
infertility, reduced female fertility and more importantly, increased susceptibility to chemical carcinogens. These MHC-linked recessive genes may suggest the presence of the tumor suppressor gene(s) playing a broad role in the embryogenesis and carcinogenesis (Melhem et al., 1993).

The MHC class III region is unusually packed with genes. For examples, complement C2 and factor B is only 421 bp apart, RP1 and C4 are only separated by 611 bp. There is about 11 kb region un-explored between RP and another newly characterized gene, RD. To our surprise, a novel gene is located upstream of the RP gene. We are currently characterizing this new gene, W. The preliminary data show that this new gene encodes a nuclear protein and is probably involved in DNA or RNA metabolism. The RD gene has been suggested to encode a nuclear protein related to pre-mRNA splicing (Speiser and White, 1989). Collectively, all three nuclear proteins, RP, W and RD are located within a 30 kb region and may function as mediators for the DNA or RNA interaction (DNA-repair, RNA splicing etc.). There are examples for members of a gene cluster coordinately performing related functions in MHC class II region: the LMP2 and LMP7 genes encode subunits of the proteasome, involved in the protein degradation (Monaco, 1992). TAP1 and TAP2 genes encode peptide transporters, which together with the class II gene products, are all associated with antigen processing and presentation (Trowsdale, 1993). Whether the new gene cluster RP-W-RD may perform related function will be an interesting research topic. Combined with various genetic analysis and biochemical characterization, the function of the individual genes (e.g., the RP gene) could be inferred from the study of neighboring genes.
REFERENCES


Clarkson, R., Bete, A. S., Grennan, D. M., Chattopadhyay, C., and Sanders, P.


Cell. Biol. 12, 2124-2131.


Ho, H. N., Gill, T. J. 3d, Klionsky, B., Ouyang, P. C., Hsieh, C. Y., Seski, J., and


Immunogenetics 2, 297-299.


Nalbantoglu, J., Hritley, D., Phear, G., Tear, G., and Meuth, M (1986). Spontaneous deletion formation at the APRT locus of hamster cells: the presence
of short sequence homologies and dyad symmetries at deletion termini. EMBO J. 5, 1199-1204.


Ono, M., Yasunaga, T., Miyata, T., and Ushikubo, H. (1986). Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor
virus genome. J. Virol. 60, 589-598.


