INFORMATION TO USERS

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

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

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

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

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

UMI
Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600
NOTE TO USERS

Page(s) missing in number only; text follows.
Microfilmed as received.

222-229

This reproduction is the best copy available.

UMI
MOLECULAR GENETIC AND BIOCHEMICAL STUDIES OF THE 
HUMAN AND MOUSE MHC COMPLEMENT GENE CLUSTERS

DISSERTATION

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

By
Zhenyu Yang, M.S.

*****

The Ohio State University
1999

Dissertation Committee:
Chack Yung Yu, D.Phil., Adviser
George A. Marzluf, Ph.D.
Gail D. Wenger, Ph.D.
Arthur H. M. Burghes, Ph.D.

Approved by
Molecular, Cellular and Developmental Biology Program
Copyright by

Zhenyu Yang

1999
The major histocompatibility complex (MHC) is associated with more than 70
diseases. The goals of this study are to investigate the mechanisms of diseases associated
with the MHC class III region. As shown by definitive restriction fragment length
polymorphisms, the four tandemly arranged genes serine/threonine kinase RP,
complement component C4, steroid 21-hydroylase CYP21, and tenascin TNX are
organized as a genetic unit designated as the RCCX module. Variations in the number
and size of the RCCX modules can promote unequal crossover, leading to gene
duplications and deletions. A TNXB-XA recombinant resulted from an unequal crossover
between a monomodular and a bimodular RCCX was observed in a patient with
congenital adrenal hyperplasia. Elucidation of the DNA sequence for the recombination
breakpoint region enabled the designation of a diagnostic technique to detect the presence
of TNXB-XA hybrids concurrent with deletions of CYP21B and C4B.

Since the major genetic factor(s) for insulin dependent diabetes mellitus (IDDM)
reside in the MHC, IDDM patients were investigated for specific patterns of the RCCX
modular structures. The frequency of monomodular RCCX structures is found significantly higher in the diabetic population than in the normal.

A novel gene, *DOM3Z*, was discovered in the intergenic region between *SKI2W* and *RP1*. The cDNA sequence and the exon-intron structure of *DOM3Z* were determined. Immunochemical and transient expression experiments revealed that human Dom3z is a nuclear protein. *DOM3Z*-related genes are present in simple and complex eukaryotes.

Located between the complement factor *B* and *C4* genes are the four ubiquitously expressed genes, *RD-SKI2W-DOM3Z-RP1*. These four genes are configured as two head-to-head orientated gene pairs. The presence of unmethylated CpG sequences at the 5' regulatory regions of each gene pair indicates that they may be housekeeping genes. RD, Ski2w, Dom3z, and RP probably have concerted functions related to RNA transcription, translation and turnover.

The organization of the mouse MHC complement gene cluster (MCGC) is similar to that of human's. However, the duplication of mouse RCCX modules is markedly different from human as the mouse gene fragments *RP2* and *TNXA* are larger and less conserved, suggesting an independent gene duplication event during evolution.
Dedicated to My Parents
ACKNOWLEDGMENTS

First and foremost, I am grateful to my adviser, Dr. Chack Yung Yu, for all his understanding and support during my graduate studies; for his patience in correcting my scientific errors; and for his encouragement and enthusiasm, which made my research and eventually, this thesis, possible.

I want to thank Dr. George A. Marzluf, a real gentleman, for the chance he provided for me to do my first rotation (during which I ran the very first sequencing gel in my life); for sparing time and offering encouragement; and for being there whenever I need help.

I want to express my appreciation to Dr. Gail D. Wenger, for teaching me the FISH technique, the result of which inspired me to discover my gene \textit{DOM3Z}; and for bringing the toothbrush back from Baltimore for a "night person" 😊. Thank-you also goes to Dr. Arthur H. M. Burghes, for showing me how to search the databases.

I wish to thank Dr. M. Sue O'Dorisio for being my cheerleader, offering the support and encouragement I needed. It is still fresh in my memory that despite of her busy schedule, she carefully reviewed my first first-author paper.
I want to thank Dr. William B. Zipf and the nurses in the Department of Endocrinology at Columbus Children's Hospital for recruiting IDDM and CAH patients, which makes the genetic analysis possible. I also appreciate all the patients who volunteered for the study. They are the real heroes whose belief in science will eventually benefit millions of other patients.

A special thank-you goes to Professor Manfred P. Dierich at University Innsbruck, Austria. His personal acknowledgment has enabled me to be more confident in myself. His graciousness has set a perfect example for me to be a decent person.

I also wish to thank Dr. Joann Moulds from The University of Texas at Houston, not only for her help in setting up C4 allotyping experiments, but also for offering support and being a friend.

I want to thank Jan Zinaich, manager of MCDB Program, for keeping me informed of all opportunities, and for ensuring me happiness is the most important part of life (This is more than an appreciation between Gemini's).

I am grateful to members of Dr. Yu's laboratory: Brad Baker, Andrew Dangel, Anna Mendoza, Xiaodong Qu, Kristi Rupert (my English teacher), Shanxiang Zhang, and Bi Zhou. They are the people who have watched me through the ups and downs, shared my happiness and pain. They are the ones who helped me in dealing with various problems. I also thank Monica Summers and Doug Balster for sharing cell lines, and for being friends.

Thank you to all my friends, for their love, which holds me through hard times.
Finally, I thank the Presidential Fellowship from The Ohio State University for the financial support.
VITA

June 2, 1971 ......................................... Born - P.R.China

1992 ..................................................... B.S., Biology, Beijing Normal University

1992-1994 ............................................ Research Associate,

Institute of Genetics,

Chinese Academy of Sciences, Beijing, P.R.China

1994-1998 ............................................ Graduate Teaching and Research Associate,

The Ohio State University

1998 ...................................................... M.S.,

Molecular, Cellular and Developmental Biology

The Ohio State University

1998-present ........................................... Presidential Research Fellow,

The Ohio State University
PUBLICATIONS

Original Research Articles


Peer-Reviewed Abstracts


Electronic publications


**FIELDS OF STUDY**

Major Field: Molecular, Cellular and Developmental Biology
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td></td>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td></td>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td></td>
<td>Vita</td>
<td>viii</td>
</tr>
<tr>
<td></td>
<td>List of Tables</td>
<td>xiv</td>
</tr>
<tr>
<td></td>
<td>List of Figures</td>
<td>xv</td>
</tr>
<tr>
<td>1.</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2.</td>
<td>Modular variations of the human MHC class III genes RCCX: a mechanism for gene deletions and disease associations</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>2.1 Abstract</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>2.2 Introduction</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>2.3 Experimental Procedures</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>2.4 Results</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>2.5 Discussion</td>
<td>49</td>
</tr>
</tbody>
</table>
3. Detection of Msc I polymorphism and the concurrent 120 bp deletion in TNXB/XA recombinant gene
3.1 Abstract
3.2 Introduction
3.3 Experimental Procedures
3.4 Results
3.5 Discussion

4. Four ubiquitously expressed genes, RD-SKI2W-DOM3Z-RP1, are present in the class III region of the HLA
4.1 Abstract
4.2 Introduction
4.3 Experimental Procedures
4.4 Results
4.5 Discussion

5. Biochemical characterization of human Dom3z
5.1 Abstract
5.2 Introduction
5.3 Experimental Procedures
5.4 Results
5.5 Discussion

6. Organizations and gene duplications of the human and mouse MHC complement gene clusters
6.1 Abstract
6.2 Introduction
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Polymorphisms of complement component C4</td>
</tr>
<tr>
<td>1.2</td>
<td>Associations between HLA and some diseases</td>
</tr>
<tr>
<td>2.1</td>
<td>RCCX modular structures of the B and E families</td>
</tr>
<tr>
<td>3.1</td>
<td>Detection of the 6.3 kb Msc I RFLP in TNXB genes</td>
</tr>
<tr>
<td>4.1</td>
<td>Sequence identities / similarities among Dom3z related proteins</td>
</tr>
<tr>
<td>6.1</td>
<td>A comparison of human and mouse DOM3Z exon-intron structures</td>
</tr>
<tr>
<td>7.1</td>
<td>RCCX modular structures and C4 allotypes for 50 IDDM patients</td>
</tr>
<tr>
<td>7.2</td>
<td>RCCX modular variations in IDDM patients</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Map of the human MHC ................................................................. 21</td>
</tr>
<tr>
<td>1.2</td>
<td>Polymorphism of human complement component C4A and C4B proteins ...... 22</td>
</tr>
<tr>
<td>1.3</td>
<td>Polymorphism in the number of $C4$ genes expressed ......................... 23</td>
</tr>
<tr>
<td>1.4</td>
<td>Size dichotomy of human $C4$ genes ............................................... 24</td>
</tr>
<tr>
<td>1.5</td>
<td>Gene structure of 21-OHase ($CYP21$) genes ....................................... 25</td>
</tr>
<tr>
<td>1.6</td>
<td>Number of EST and non-EST entries submitted to GenBank ..................... 26</td>
</tr>
<tr>
<td>2.1</td>
<td>A molecular map of the human MHC complement gene cluster .................. 56</td>
</tr>
<tr>
<td>2.2</td>
<td>$C4$ allotypes of the B and E families ........................................... 57</td>
</tr>
<tr>
<td>2.3</td>
<td>Southern blot analysis to detect the modular variations of $RP$, $C4$, $CYP21$ and $TNX$ ................................................................. 58</td>
</tr>
<tr>
<td>2.4</td>
<td>Southern blot analysis to segregate the $RP1/RP2$ variation from the $C4$ long/short (L/S) size dichotomy ................................................................. 60</td>
</tr>
<tr>
<td>2.5</td>
<td>$Taq$ I polymorphism to detect the $TNXA$-associated 120 bp deletion in the $TNXB$ gene of CAH patient E1 ................................................................. 61</td>
</tr>
<tr>
<td>2.6</td>
<td>Recombination between RCCX modules ................................................. 62</td>
</tr>
<tr>
<td>2.7</td>
<td>Elucidation of the breakpoint region for $TNXB$-XA recombinant ............ 63</td>
</tr>
<tr>
<td>2.8</td>
<td>RCCX modular variations in the human population .................................. 68</td>
</tr>
<tr>
<td>3.1</td>
<td>The $Msc$ I map at $TNXB$ and $CYP21B$ gene region ........................... 83</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>3.2</td>
<td>Southern blot analysis to detect the nucleotide mutations at the <em>Msc</em> I site (nt. 2268-2273) in TNXB genes</td>
</tr>
<tr>
<td>3.3</td>
<td>Southern blot analysis to determine the RCCX modular structures of CAH patients</td>
</tr>
<tr>
<td>3.4</td>
<td>A schematic diagram to show variant TNX genes by <em>Pst</em> I - <em>Msc</em> I RFLP</td>
</tr>
<tr>
<td>3.5</td>
<td>Southern blot analysis to detect the concurrent 120 bp deletion and <em>Msc</em> I RFLP in TNXB genes</td>
</tr>
<tr>
<td>4.1</td>
<td>Exon-intron structure and <em>BamH</em> I RFLP of the human <em>SKI2W</em> gene</td>
</tr>
<tr>
<td>4.2</td>
<td>cDNA and amino acid sequences of human <em>DOM3Z</em></td>
</tr>
<tr>
<td>4.3</td>
<td>An alignment of the amino acid sequences for Dom3z-related proteins</td>
</tr>
<tr>
<td>4.4</td>
<td>Exon-intron structure of the human <em>DOM3Z</em> gene</td>
</tr>
<tr>
<td>4.5</td>
<td>PFGE analysis of the <em>RD-SKI2W-DOM3Z-RPI</em> complex</td>
</tr>
<tr>
<td>4.6</td>
<td>Northern blot analyses of human multiple tissue RNAs</td>
</tr>
<tr>
<td>4.7</td>
<td>A molecular map of the HLA class III region</td>
</tr>
<tr>
<td>5.1</td>
<td>Restriction analyses to determine the presence and orientations of <em>DOM3Z</em> cDNA inserts in different expression vectors</td>
</tr>
<tr>
<td>5.2</td>
<td>Production and detection of bacterial fusion protein Trx-His-Dom3z</td>
</tr>
<tr>
<td>5.3</td>
<td>Production and detection of bacterial fusion protein GST-Dom3z</td>
</tr>
<tr>
<td>5.4</td>
<td>Cellular localization of GFP-Dom3z in HeLa cells</td>
</tr>
<tr>
<td>5.5</td>
<td>Indirect immunofluorescence experiment to detect the cellular localization of Dom3z in HeLa</td>
</tr>
<tr>
<td>5.6</td>
<td>Genomic Southern blot analysis showing the incorporation of GFP-<em>DOM3Z</em> into the stable cell line genome</td>
</tr>
</tbody>
</table>
5.7 Reverse transcription (RT)-PCR of GFP-DOM3Z to examine the expression of the exogenous construct in the stable cell lines ................................................. 149
5.8 Immunoblot analysis to detect GFP-Dom3z in CHO transfectants .......... 150
5.9 Immunoblot analysis to detect the human endogenous Dom3z ............ 151
6.1 Southern blot analyses of DOM3Z in primates and mammals ............. 175
6.2 Genomic sequences linking human DOM3Z to RPI and SKI2W .......... 176
6.3 Exon-intron structure and alternative splicings of human and mouse DOM3Z .................................................................................................................. 178
6.4 An alignment of human and mouse Dom3z protein sequences .......... 179
6.5 cDNA and derived amino acid sequences of mouse RP1 .................... 180
6.6 A comparison of the exon-intron structures and gene duplications for human and mouse RPI and RP2 ........................................................................ 182
6.7 A comparison of the gene organizations of human and mouse MHC complement gene clusters .................................................................................. 183
7.1 Comparisons of C4 genes and RCCX modular structures between chromosomes from the IDDM and non-IDDM populations ....................... 202
7.2 C4 allotypes of selected IDDM patients ........................................... 203
7.3 Sequence comparison of the C4d region from Rg1-associated C4B3 and C4A ................................................................................................................. 204
7.4 RFLPs of SKI2W in IDDM patients ................................................... 207
CHAPTER 1

INTRODUCTION

As history turns the page of this century, it is time for us to review the research conducted on the major histocompatibility complex (MHC). MHC was first discovered in mouse as the H-2 region. About 10 years after, in 1958, the homologous system in human was also identified. It was named the human leukocyte antigen (HLA), because as blood group antigens, they are expressed on the white blood cells. Later, it was demonstrated that during organ transplantation, matching HLA antigens can significantly improve the fate of the transplant in the recipient. Otherwise, transplant rejection is an absolute consequence. HLA is important in both transfusion and transplantation medicine since several loci within HLA region are involved in histocompatibility. The HLA was therefore also named the major histocompatibility complex, MHC (Terasaki, 1990).
Overview of the MHC region

In humans, the MHC constitutes a 4,000 kb region on the short arm of chromosome 6 in the distal portion band 6p21.3 (Senger et al., 1993). To date, over 200 genes have been identified in the MHC region, and they are grouped into three major linked gene clusters: Class I, II and III (Fig. 1.1) (Trowsdale, 1996). The class I group lies in the telomeric half of the MHC and spans about 1,800 kb. It contains genes encoding the classical transplantation antigens, HLA-A, -B, and -C loci. The 1,000 kb class II region lies in the centromeric half of the MHC complex. The products of this region were initially identified as immune-response loci, known as HLA-D. However, it became apparent now that the products of the class II loci, HLA-DP, -DQ, and -DR, were structurally and functionally related to the class I loci as they were all members of the immunoglobulin superfamily. These two parts are separated by the class III region, which is 1,100 kb long, and consists of many densely spaced genes [reviewed in (Trowsdale, 1996)].

Much of the attention has been focused on the MHC region because the genes in this region fulfill important immunologic functions. Many class I and II genes encode highly polymorphic families of cell-surface glycoproteins that present antigenic peptides to the T cells during an immune response. On the other hand, in recent years, a detailed characterization of the MHC has revealed the presence of genes that mostly encode proteins of unknown functions or with functions unrelated to the immune response (Campbell and Trowsdale, 1993; Carroll et al., 1987; Trowsdale, 1993). Many of these genes are located in the class III region. Although the class III region also contains genes
that encode proteins with important immune related functions, such as complement component C2, C4 and factor B, the products of most other genes have a variety of different functions [reviewed in (Aguado et al., 1996; Salter-Cid and Flajnik, 1995)].

**Diversities of Complement Component C4**

The complement component C4 is an abundant blood protein (0.4 mg/ml). It is essential for the immune response triggered by antibodies. During its activation and inactivation, the C4 molecules interacts with at least eight other molecules: the antigen, the antibody, complement factors C2, C3, C5, regulatory proteins factor I, C4b binding protein, and complement receptor type I (CR1) (Porter, 1985a). The covalent binding ability through the thioester carbonyl group after activation, and the relative large size with ~200 kDa, enable C4 to link and provide surfaces for these protein interactions. At the same time, the binding abilities are different between two classes of C4 molecules: C4A (acidic) exhibits 100-fold higher affinity to form amide bonds with amino groups or peptide antigens than C4B (basic). The difference in chemical reactivity explains why C4A binds to immune complexes three times more effectively than C4B, and therefore is primarily involved in promoting the physiological disposal of immune complexes. C4B can form ester bonds with hydroxyl groups or carbohydrate antigens 10-fold more efficient than C4A. Since C4B is hemolytically more active, it participates preferentially in the clearance of microorganisms (Law et al., 1984; Yu, 1999). The differential chemical reactivities of C4A and C4B are determined by variations of four amino acids
between positions 1101-1106, by which C4A has PCPVLD, while C4B has LSPVIH (Belt et al., 1985; Yu et al., 1986).

As one of the most polymorphic serum proteins besides immunoglobulins, the complexity of C4 is manifested in many aspects other than the differential chemical reactivities (Table 1.1) (Isenman and Young, 1984; Yu et al., 1988). More than 34 allotypes of C4A and C4B have been identified (Mauff et al., 1990a, b), the locations of polymorphic residues are shown in Fig. 1.2 (Yu, 1999). They are remarkably different in electrophoretic mobilities and associations with antigenic determinants [reviewed in (Yu, 1999)]. Two blood group antigens, Rodgers (Rg1, Rg2) and Chido (Ch1 - Ch6), are located in the C4d region of C4A or C4B (O’Neill et al., 1978; Tilley et al., 1978). C4d is the factor I - mediated proteolytic degradation fragment that may be linked to various cell surfaces (Reid and Porter, 1981). Amino acid variations in C4d constitute the sequential and conformational epitopes for the Rg and Ch antigenic determinants (Yu et al., 1986, 1988). Generally, C4A is associated with Rg1 while C4B is associated with Ch1 (Giles et al., 1988; Yu et al., 1988). However, there are some exceptions, such as the association of C4A1 with Ch1 and C4B5 with Rg1 (Rittner et al., 1984; Roos et al., 1984).

In most cases, the two classes of C4 molecules, C4A and C4B, are encoded by two tandem gene loci, C4A and C4B, respectively (Carroll et al., 1984). However, there are examples for the two C4 loci coding for identical C4 isotypes or allotypes (Braun et al., 1990; Yu and Campbell, 1987). It was reported that phenotypically, 64.5-74.5% of chromosome 6 in the population contain two different C4 genes, C4A and C4B;
9.5-14.5% contain C4A gene only; 16-19% contain C4B gene only. Another 1-2% of chromosome 6 have three C4 genes, with either two C4A genes or two C4B genes (Fig. 1.3) (Carroll et al., 1986).

Besides the gene number variations, a unique feature of C4 is the size dichotomy (Carroll et al., 1984; Prentice et al., 1986; Yu et al., 1986). C4 genes could be either 21 kb (long, L) or 14.6 kb (short, S), and the difference lies in intron 9. For the long C4 gene, intron 9 is 6.8 kb in size, while for the short gene, this intron is only 416 bp in size (Dangel et al., 1994; Yu, 1991). When the sequence of the large intron 9 was determined, an endogenous retrovirus, HERV-K(C4), was discovered (Dangel et al., 1994, 1995b; Tassabehji et al., 1994; Yu et al., 1986). HERV-K(C4) contains all the hallmark structures of a retrovirus, including the gag, pol, and env genes (Fig. 1.4). However, these genes have acquired many point mutations and therefore, may not produce a live retrovirus. HERV-K(C4) is organized in the opposite transcriptional orientation with respect to the C4 gene. Therefore, transcription of C4 produces the antisense version of the retrovirus. Such configuration may have selection advantage for the host because the antisense RNA for the retrovirus could hybridize to proviral genomic RNAs and serve as a neutralizing agent to inhibit retroviral proliferation (Dangel et al., 1994; Yu, 1999).

As it is essential for the immune response triggered by antibodies, complete deficiencies of C4A and C4B are rare and highly deleterious (Hauptmann et al., 1988; Lokki and Colten, 1995; Uring-Lambert et al., 1989). Almost all patients suffer systemic lupus, immune complex disease in the kidney or overwhelming infections. On the other
hand, partial or complete deficiencies of either C4A or C4B are more common, occurring in 10-30% of the Caucasian population (Schneider, 1990). C4 deficiencies can be due to the absence or deletion of a specific C4 gene (Carroll et al., 1985a), the presence of two genes coding for identical C4 isotypes or allotypes (Braun et al., 1990; Yu and Campbell, 1987), or the presence of pseudogenes caused by deletions, insertions or point mutations [reviewed in (Yu, 1999)].

**Genes flanking complement component C4**

Downstream of C4A and C4B, a pseudogene CYP21A and a functional gene CYP21B were identified respectively (Carroll et al., 1985b; White et al., 1985). CYP21A contains three deleterious mutations and many other point mutations (Fig. 1.5) (Higashi et al., 1986; Rodrigues et al., 1987; White et al., 1986). CYP21B encodes steroid 21-hydroxylase, which is a microsomal enzyme expressed in the adrenal gland that catalyzes conversion of 17-hydroxyprogesterone and progesterone to 11-deoxycortisol and deoxycorticosterone respectively [reviewed in (White et al., 1994)]. Deficiency of the adrenal steroid 21-hydroxylase is the most common enzymatic defect of steroid synthesis. It can lead to the recessively inherited disorder of adrenal steroidogenesis: congenital adrenal hyperplasia (CAH), in which inhibition of the formation of cortisol drives the adrenal cortex to overproduce androgens. This hormonal setting affects the development of genetic females by misdirecting the differentiation of external genitalia towards the male type. In classical CAH, the most common cause of genital ambiguity in females, prenatal exposure to excess androgens results in virilization of the female fetus. It occurs
in about 1 in 14,000 live births. Postnatally, untreated patients do not effectively synthesize aldosterone and are salt-wasting, a condition that is potentially fatal [reviewed in (New, 1995; White et al., 1994)]. Complete deletions of functional gene CYP21B, large gene conversions to pseudogene CYP21A, and single point mutations in CYP21B are reported to be responsible for the disease (Strachan, 1994).

Overlapping with the 3' end of CYP21A and CYP21B genes are another pair of duplicated genes: tenascin TNXA and TNXB. The transcriptional orientations of TNX are opposite to those of CYP21 genes (Matsumoto et al., 1992b; Morel et al., 1989). The duplication of TNX is incomplete with TNXB being 68.2 kb in size and consisting of 45 exons, while TNXA starting from intron 32 (Bristow et al., 1993; Matsumoto et al., 1994; Rowen et al., 1997a). Another important difference is that there is a TNXA-specific 120 bp deletion at the boundary of exon 36 and intron 36. This deletion results in a frame shift mutation and premature termination of translation (Gitelman et al., 1992; Shen et al., 1994).

The function of TNX has not been determined, but the striking similarities of its overall structure with tenascin-related molecules cytostatin (TNC) and restrictin (TNR) suggest that it belongs to the tenascin family (Chiquet-Ehrismann, 1995). Tenascins are believed to be important extracellular matrix proteins involved in regulating numerous developmental processes (Chiquet-Ehrismann et al., 1994). The predicted TNX protein is over 400 kDa in size and consists of five distinct domains: a signal peptide, a hydrophobic domain containing three heptad repeats, a series of 18.5 EGF-like repeats, 32 fibronectin type III repeats, and a carboxyl-terminal fibrinogen-like domain. TNX is
ubiquitously expressed in the fetus with the most prominent expression in testis and muscle (Bristow et al., 1993). Data obtained from rat embryo studies indicate a role of TNX in connective tissue cell migration and late muscle morphogenesis (Burch et al., 1995). This notion is substantiated by the report that in human, TNX deficiency is associated with connective tissue disease Ehlers-Danlos syndrome (EDS) (Burch et al., 1997).

When the promoter region of C4A was characterized, a polyadenylation signal AATAAA, with the characteristic feature for the 3' end of a mammalian gene, was found 631 bp upstream of C4A transcriptional initiation sites. Novel gene RPl (also known as G11) was therefore discovered (Sargent et al., 1994; Shen et al., 1994). Northern blot analysis suggested that RPl is ubiquitously expressed. The deduced amino acid sequence of RPl does not reveal significant similarities to any known proteins, although a bipartite nuclear localization signal was identified (Shen et al., 1994; Yang et al., 1998). The function of RPl is unknown. Similar to TNX, RP is also partially duplicated. The duplicated gene segment RP2 is identical to the last 913 bp of the RPl gene and is unlikely to code for a protein product (Shen et al., 1994).

MHC and disease associations

In 1967, MHC was first reported by Frenchman Amiel to be associated with Hodgkin's disease (Amiel, 1967). Although the association is weak, this finding more than 30 years ago is certainly significant: it opened the door to a whole new area of the MHC field by triggering thousands of studies on the association of MHC with hundreds
of diseases. At least 70 diseases have now been shown to be clearly associated with MHC, including insulin-dependent diabetes mellitus (IDDM), congenital adrenal hyperplasia (CAH), systemic lupus erythematosus (SLE), as well as juvenile rheumatoid arthritis (JRA) (De Vries and Van Rood, 1992). *Table 1.2* gives a representative list of well-established MHC and disease associations (De Vries, 1994).

Autoimmune disease is obviously in the first category on this MHC associated disease list. Most of the autoimmune diseases are polygenic. As reviewed in Vyse and Kotzin (1998), this is first demonstrated by the genetic heterogeneity, i.e., the same phenotype resulted from the combined effect of different genes and / or alleles. There is no single gene being either necessary or sufficient for the disease development. Therefore, the common disease genes are referred to as susceptibility genes (Heward and Gough, 1997). The susceptibility genes may interact with one another [epistasis, (Hodge, 1981)] or act independently [additivity, (Risch, 1990)]. Secondly, the alleles that predispose to the autoimmune diseases are common in the general population. In other words, the genetic susceptibility factors for autoimmunity are not due to rare disease genes, but are instead due to common polymorphisms among apparently normal genes (Wu *et al.*, 1998). There is a pool of mutations or allelic variants affecting the expression and / or function of the genes involved in control of the immune response. Individually, most mutations in the genome have mild, if detectable at all, effects. However, in combination with "normal" alleles of other loci, they produce a measurable and occasionally lethal autoimmune phenotype (Vyse and Todd, 1996). The third puzzling fact is that the penetrance of autoimmune disease in carriers of the disease susceptibility
genes is incomplete (Vyse and Todd, 1996). Even in the presence of a full complement of susceptibility alleles at multiple loci, overt disease does not always result (Todd, 1995). The exception is classic congenital adrenal hyperplasia (CAH) caused by CYP21 deficiency, which is the only disease linked with MHC known to show complete penetrance and to be present at birth (De Vries, 1994).

Research of insulin-dependent diabetes mellitus

The most extensively studied polygenic autoimmune disease is insulin dependent diabetes mellitus (IDDM), also known as the type 1 diabetes. IDDM is characterized by the autoimmune destruction of insulin-secreting pancreatic β cells causing tissue damage, which can lead to blindness, kidney failure and reduced life expectancy (Tisch and McDevitt, 1996). The peak age of onset is about 12 years, and from then onwards, daily injections of insulin are required by affected individuals. With a frequency of about 0.4% in Caucasian population, IDDM is second to asthma as the most serious chronic childhood disease in the Western world [reviewed in (Cordell and Todd, 1995)].

A genome wide search for human genes that predispose to IDDM indicates that as many as twenty different chromosome regions have certain positive linkage to this disease. Among them, MHC is the strongest genetic factor (Davies et al., 1994). Individuals with a combination of HLA class II antigens DR3 and DR4 are also at higher risk of IDDM (Svejgaard and Ryder, 1981; Wolf et al., 1983). It has also been demonstrated that susceptibility to IDDM is most strongly determined by MHC class II DQ β alleles that encode serine, alanine, or valine at position 57 on both chromosomes.
Aspartic acid at position 57 mediates resistance to IDDM, which varies in degree with the sequence of other residues in the DQ α and β chains. One theory to explain this phenomenon is that the presence or absence of an Asp at residue 57 of the β chain may alter the peptide binding motif and the ability to bind specific immuno-pathogenic peptides (Nepom, 1990; Nepom and Kwok, 1998).

In addition to MHC class II genes, the possible candidates for IDDM susceptibility genes include C4 and several other poorly defined polymorphic genes in the class III region (Degli-Esposti et al., 1992). There have been reports describing associations of increased frequencies of the C4A null allele and C4B3 allotype with type I diabetes (Jenhani et al., 1992; Marcelli-Barge et al., 1990). Whether these associations are a consequence of linkage disequilibrium with the DR and DQ loci or result from a primary association with the disease itself is still unclear. It has also been suggested that there may be a second susceptibility gene mapping close to the C4 locus (Thomsen et al., 1988).

Important as it is, it has long been apparent that genetic susceptibility is a necessary but not sufficient predisposing factor. Even in monozygotic twins, the concordance rate is only 50% (Barnett et al., 1981), indicating the importance of a number of as yet unidentified environmental factors (Castano and Eisenbarth, 1990). Infecting viruses are most often implicated in this category (Oldstone et al., 1991). As reviewed by Conrad et al. (1994) and Moller (1998), there are several mechanisms proposed for virus-induced autoimmunity, among which superantigens can act as activators of potentially pathogenic T cell clones (Conrad and Trucco, 1994).
Superantigens are unique products of bacteria and viruses with the capacity to activate a large proportion of immunocompetent T lymphocytes (1-30% vs. ~1 out of $10^5$ of the total peripheral T cell pool) irrespective of their immunological specificity. They can cause extreme inflammation with high levels of toxic cytokines (Conrad et al., 1994; Herman et al., 1991; Jorgensen et al., 1992). It has been reported the involvement of a pancreatic islet cell membrane-bound superantigen as a diabetes aetiopathogenetic factor (Conrad et al., 1994). Superantigens can also be the products of endogenous retroviral genes. In 1997, a paper published in Cell proposed that a human endogenous retroviral superantigen constitutes a candidate autoimmune gene in type I diabetes (Conrad et al., 1997). However, this conclusion is put into serious doubt recently as several other groups failed to reproduce similar experiment results (Badenhoop et al., 1999; Jaeckel et al., 1999; Muir et al., 1999).

The research of complex polygenic diseases such as IDDM, in which multiple alleles interact to produce an autoimmune phenotype, can be facilitated by studying experimental species, such as the mouse or rat. While the genes predisposing to disease in an animal model might not be the same as those for human, the underlying genetic basis could present similarities (Cordell and Todd, 1995). Non-obese diabetic (NOD) mouse, discovered at the Shionogi Research Laboratories in Japan, spontaneously develops a disease with many of the characteristics of IDDM in human (Makino et al., 1980). Similar to human patients with IDDM, NOD mice also have cellular and humoral immune responses specific for pancreatic β cell antigens. As such, it has become the most frequently employed small animal model of IDDM, and has permitted some
important insights into the complicated pathogenesis of this disorder (Wicker et al., 1995).

NOD mice not only share pathologic features, but also the genetic predispositions with human IDDM patients. Of the 18 genetic loci identified to date in NOD mice, the strongest contribution to the disease is again from the genes of the MHC (Wicker et al., 1992; Wicker et al., 1995). The gene organizations of mouse MHC, or H2, are similar but not identical to that of the human (Chaplin et al., 1983). Generally, H2 region contains similar number of genes as the equivalent human region but is smaller in size (Gasser et al., 1994). This is due to both smaller intron length as well as smaller intergenic regions (Cho et al., 1991; Newell et al., 1996).

**Gene duplications**

An important feature of mouse genome is the tolerance of a large viral load, which was suggested to create potential for beneficial variations (Stoye and Coffin, 1988). Endogenous proviruses can affect genes in their locale either by disruption, or by influence of their regulatory features on neighboring gene expression (Copeland et al., 1983; Hayward et al., 1981). Although some of these effects are deleterious, most proviral sequences are presumably benign or would not have accumulated to their present high copy numbers. One of these examples comes from the mouse gene encoding for sex limited protein (Slp), which derived from the gene for complement component C4, but diverged in regulation and function (Karp et al., 1982; Shreffler et al., 1984). Slp requires androgen for expression and is inactive in the complement pathway, although
greater than 95% homology is maintained in coding and flanking regions (Hemenway et al., 1986; Nonaka et al., 1986a, b; Ogata and Sepich, 1985). The hormone responsive expression of Slp is conferred by an endogenous provirus which was inserted 2 kb upstream of the gene. As it imposed its regulatory features on the adjacent SLP gene, this endogenous provirus is termed imposon (IMP). That the imposon and its close relatives are stable within the mouse genome argues for some possible selective advantage provided by these elements (Stavenhagen and Robins, 1988).

Within multigene families, divergence is tolerated, so long as a functional gene is retained (Smith, 1974). Duplicated genes are allowed to acquire mutations that may eventually result in acquisition of a new function or final status as a psuedogene (Stavenhagen and Robins, 1988). In human, an increasing number of instances of gene duplications are also being revealed as enormous progress in the physical mapping and sequencing of the human genome has been achieved in recent years. The human genome research not only provides valuable approach to study individual genes; more significantly, it starts a whole new era of biomedical research by offering the feasibility of screening the entire genome to identify susceptibility genes of the polygenic diseases.

Search of disease susceptibility genes

The elucidation of the genetic basis of disease has been a goal of medicine for many decades. With the advent of molecular genetic technologies, over the past ten years, nearly 100 genes causing various genetic diseases have been identified by positional cloning. However, these diseases are for the most part monogenic in nature,
examples include cystic fibrosis and Huntington's disease. Studies of complex, polygenic diseases have proven much more difficult [reviewed in (Pratt and Dzau, 1999)].

Traditionally, linkage and association studies using candidate genes (the population based case control studies) can show the association of a gene to the disease in a disease population compared with a disease-free population (Heward and Gough, 1997). They have provided insight into the causes of polygenic diseases. However, the biggest problem here is that a positive association could result not only from a real disease causing gene, but also linkage disequilibrium or even an artifact of population admixture (Lander and Schork, 1994). Therefore, the number of genes that are consistently positive in multiple studies is small and the predictive power of these genetic variants is, unfortunately, limited (Spielman et al., 1993).

The entire genome screening approach makes use of genetic markers, such as microsatellites, from the existing genetic maps (Heward and Gough, 1997). A whole genome scan is the analysis of about 100 evenly spaced markers in mouse pedigree, or about 300 markers in human families, in which there are affected and non-affected progenies. The goal is to identify chromosome regions that are inherited by affected progenies more often than expected from Mendelian random segregation (Vyse and Todd, 1996). Once genetic dissection implicates a chromosomal region, positional cloning can be used to identify the responsible gene. The availability of densely mapped genetic markers spanning the entire genome can definitely accelerate this process, and could be applied to the analysis of multifactorial traits without prior knowledge of any gene function. This realization formed a basis for the initial promotion of the Human
Genome Project, which promises to make a tremendous contribution to the positional cloning of complex traits by eventually providing a complete catalog of all genes in a relevant region (Tomlinson and Bodmer, 1995).

The Human Genome Project

The Human Genome Project (HGP) in U.S. was initiated about 10 years ago aiming at sequencing the entire human genome, comprised of approximately 3 billion base pairs. However, it has been estimated that only 2-3% of the genome encodes proteins. Therefore, in 1991, an intermediate goal was attempted: the defining and sequencing of the regions of the genome encoding proteins by sequencing clones from cDNA libraries (Adams et al., 1991). These sequences termed as the expressed sequence tags (ESTs), account for more than half the sequence records in GenBank (Fig 1.6) (Boguski, 1995). Over the past 8 years, 1,304,811 entries of human EST sequences have been reported (dbEST release, March 19, 1999).

The reported ESTs are approximately an order of magnitude greater than the estimated number of genes in the human genome, which is between 50,000 and 150,000 (Fields et al., 1994). This redundancy allows for sequence comparisons and definition of consensus sequences. A hint to the completeness of the EST database is provided by comparing the entries from the database to independently isolated genes. For example, of the 91 genes identified by positional cloning, 83 (91%) are represented in the dbEST database. Similarly, of the 94 genes cloned as oncogenes or tumor suppressers, 94% are represented (Pratt and Dzau, 1999).
The identification of the expressed sequences in the human genome has been useful in terms of defining gene families, discovering novel genes, and determining patterns of expression in different tissues and disease states (Pratt and Dzau, 1999). Currently, only 10% of the ESTs correspond to known genes, another 20% have sequence homology to known genes, while the remaining 70% are unknown genes with no homology and with no known function (Adams et al., 1995). Nevertheless, the ability to identify genes residing within an interval linked to a particular disease, especially if expression patterns in the tissues can be determined, will dramatically increase the power of genetic analysis and more rapidly yield candidate genes for further analysis [reviewed in (Pratt and Dzau, 1999)].

In biological research, hypotheses are formulated based on the most abstract components, and then tested utilizing a large body of rules that govern the selection and use of methods and tools best suited for addressing a particular biological question. Hypothesis testing ultimately results in the accumulation of data, which in turn, reshape the world view and theory, leading to subsequent rounds of hypothesis formulation and testing, and an enhanced understanding of a biological question (Schena et al., 1998). In this specific case, the need for a well-spaced map of highly polymorphic genetic markers to study multifactorial diseases promoted the Human Genome Project, which with its progress adds new dimensions to our understanding of the causes of human diseases. The ability to rapidly genotype individuals at high density will greatly enhance the ability to determine genes playing causal roles in the development of disease on a population
basis and possibly to predict which individuals will be susceptible to disease and to better
diagnosis and treat disease (Pratt and Dzau, 1999).

**Recombination as the cause for genomic disorders**

The classical mechanism for a genetic disease is that an abnormal phenotype is
primarily a result of point mutation. While from the genome point of view, structural
characteristics of the human genome predispose to rearrangements that also result in
human disease traits. The genomic disorders are caused by an alteration of the genome
that might lead to the complete loss or gain of a gene(s) sensitive to dosage effect or
alternatively, might disrupt the structural integrity of a gene(s). It can occur through
many mechanisms, one of which is recombination between region-specific, low-copy
repeated sequences [reviewed in (Lupski, 1998)].

There is no doubt that recombination is an essential cellular process. The
resultant exchange of information is critical for the survival of species. Recombination
provides an effective means of generating genetic diversity that is important for
evolution. It also allows cells to retrieve sequences lost by replacing the damaged section
with an undamaged strand from a homologous chromosome. Although the process of
homologous recombination is used to study gene function by way of gene knockouts, it
certainly is also a source of harmful mutations and diseases [reviewed in (Purandare and
Patel, 1997)].

Under certain circumstances, recombination can lead to deletion and / or
duplication of the genetic material. It can also result in the creation of fusion genes.
Sequence homology between members of gene families or pseudogenes is the trigger for such rearrangements. Unequal crossover or unequal sister chromatid exchange between a functional gene and a related pseudogene can lead to deletion of the functional gene or formation of fusion genes containing a segment derived from the pseudogene.

Alternatively, the pseudogene can act as a donor sequence in gene conversion events and introduce deleterious mutations into the functional gene [reviewed in (Purandare and Patel, 1997)].

Recombinations do not occur randomly, but rather at site-specific hot spots (Chakravarti et al., 1986; Cullen et al., 1995; Lebo et al., 1983; Wahls et al., 1990).

Specific regions within MHC have been suspected to contain hot spots for recombination, based on the studies of linkage disequilibrium between polymorphic loci (Cullen et al., 1997; Kobori et al., 1984; Steinmetz et al., 1986). In this region, finding the class III hot spots is of importance not only for its relevance in understanding meiotic recombination within the MHC, but also because of its implications for the positional cloning of candidate disease susceptibility genes (Snoek et al., 1998).

Goals of the study

Traditionally, the study of the genetic predisposition to the large number of MHC associated diseases have been mainly focused on the polymorphic class I and class II genes. However, in many cases, these gene products cannot fully explain all the disease susceptibility. Therefore, characterization of many other genes, in particular genes in the class III region, has become increasingly important. Therefore, genetic and biochemical
studies of the class III genes were performed. The first purpose of the study presented in this dissertation is to establish the molecular genetic basis of the \textit{C4} and \textit{CYP21} gene duplications and deletions. Due to the concurrent gene duplications and gene deletions of \textit{RP, C4, CYP21} and \textit{TNX}, the genetic unit consisting of these four genes was termed the RCCX module. The RCCX modular variations may contribute to the genetic instability of the MHC class III region and lead to MHC-associated diseases. The second purpose of the study is to investigate whether there are specific patterns of the RCCX modular structures in patients with insulin dependent diabetes mellitus (IDDM). As it is important to study IDDM using mice models, the comparison of human and mouse gene structures and gene organizations become apparently important. Consequently, the third purpose is to compare the gene organization of the human and mouse MHC class III region. Finally, although many diseases are associated with \textit{C4} gene deletion or polymorphism, \textit{C4} may just serve as a genetic marker. It is possible that yet unidentified gene(s) in close proximity may actually be the disease susceptibility gene(s). Therefore, another important goal of this study is to determine the presence of novel gene(s) close to \textit{C4}. 

20
Fig. 1.1 Map of the human MHC. From centromeric to telomeric of chromosome 6p: A, Class II; B, Class III; and C, class I regions. (This figure is taken from Trowsdale, 1996.)
Fig. 1.2 Polymorphism of human complement component C4A and C4B proteins. C4 is synthesized as a single chain precursor and processed to a three chained structure (β-α-γ) linked by disulfide bonds. The C4A and C4B isotypic residues are in bold. (This figure is taken from Yu, 1999).
Fig. 1.3 Polymorphism in the number of C4 genes expressed. (This figure is taken from Carroll et al. 1986)
A. Long C4 gene, 21 kb

![Gene structure diagram](image)

B. Short C4 gene, 14.6 kb

![Gene structure diagram](image)

Fig. 1.4 **Size dichotomy of human C4 genes.** A and B illustrate the exon intron organization of a long and a short C4 gene, respectively. Filled boxes represent the 5' LTR and the 3' LTR; horizontal arrows stand for transcriptional orientation. (This figure is taken from Dangel et al., 1995b).
Fig. 1.5 Gene structure of 21-OHase (CYP21) genes. Exons are boxed and numbered 1-10. The position of the TATAA box and the polyadenylation signal AATAAAA are shown. The positions at which the 21-OHase A (CYP21A) differs from 21-OHase B (CYP21B) are marked by bars beneath the gene structure. (') over a bar indicates a codon change due to nucleotide alteration. The three deleterious mutations in the OHase A (CYP21A) gene are illustrated below the corresponding bars. The 8 bp deletion and the T insertion, which cause frameshifts, and the C — > T transition, which generates a stop codon, are underlined. (This figure is taken from Rodrigues et al, 1987).
Fig. 1.6 Number of EST and non-EST entries submitted to GenBank.

(This figure is taken from Boguski, 1995)
1. Electrophoretic mobility
   Agarose gel
   SDS-PAGE (α chain)  
   **C4A**  Fast (acidic)  **C4B**  Slow (basic)
   Mr 96 000  Mr 94 000

2. Thioester reactivity
   (a) Hemolytic activity
   (b) Rel. covalent binding affinities
       (i) Amino group  Higher  Lower
       (ii) Hydroxyl group  Lower  Higher

3. Antigenic determinants
   Blood group antigen  Rodgers, WH  Chido, WH
   Rg: 1,2  Ch: 1, 2, 3, 4, 5, 6

*Table 1.1 Polymorphisms of complement component C4* (This table is taken from Yu, 1999).
<table>
<thead>
<tr>
<th>Disease</th>
<th>HLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hodgkin's disease</td>
<td>DPB1*0301</td>
</tr>
<tr>
<td>Graves' disease</td>
<td>DQA1*0102</td>
</tr>
<tr>
<td>Pemphigus</td>
<td>DR4</td>
</tr>
<tr>
<td>Type 1 diabetes</td>
<td>DQA1*0301</td>
</tr>
<tr>
<td></td>
<td>DQB1*0302</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>DQA1*0102</td>
</tr>
<tr>
<td></td>
<td>DQB1*0602</td>
</tr>
<tr>
<td></td>
<td>DRB1*1501</td>
</tr>
<tr>
<td>Narcolepsy</td>
<td>DQA1*0102</td>
</tr>
<tr>
<td></td>
<td>DQB1*0602</td>
</tr>
<tr>
<td></td>
<td>DRB5*0101</td>
</tr>
<tr>
<td></td>
<td>DRB1*1501</td>
</tr>
<tr>
<td>Celiac disease</td>
<td>DQA1*0501</td>
</tr>
<tr>
<td></td>
<td>DQB1*0201</td>
</tr>
<tr>
<td>Sicca syndrome</td>
<td>DR3</td>
</tr>
<tr>
<td>IgA deficiency</td>
<td>DR3</td>
</tr>
<tr>
<td>Idiopathic Addison's disease</td>
<td>DR3</td>
</tr>
<tr>
<td>Idiopathic membranous nephropathy</td>
<td>DR3</td>
</tr>
<tr>
<td>Alloimmune thrombopenia</td>
<td>DR52</td>
</tr>
<tr>
<td>Leprosy</td>
<td></td>
</tr>
<tr>
<td>Tuberculoid</td>
<td>DR3</td>
</tr>
<tr>
<td>Lepromatous</td>
<td>DQ1</td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>DR2</td>
</tr>
<tr>
<td>Goodpasture's syndrome</td>
<td>DR2</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>DRB1*0404</td>
</tr>
</tbody>
</table>

(to be continued)

*Table 1.2* Associations between HLA and some diseases
(This table is taken from De Vries, 1994)
(Table 1.2 continued)

<table>
<thead>
<tr>
<th>Disease</th>
<th>HLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pernicious anaemia</td>
<td>DR5</td>
</tr>
<tr>
<td>Pauci-articular juvenile rheumatoid arthritis</td>
<td>DR5</td>
</tr>
<tr>
<td>Behcet's disease</td>
<td>B51</td>
</tr>
<tr>
<td>Congenital adrenal hyperplasia</td>
<td>(B47)</td>
</tr>
<tr>
<td>C2 deficiency</td>
<td>(DQ1-DR2-C2 NULL-B18)</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>C4A*Q0</td>
</tr>
<tr>
<td>Ankylosing spondylitis</td>
<td>B27</td>
</tr>
<tr>
<td>Reiter's syndrome</td>
<td>B27</td>
</tr>
<tr>
<td>Reactive arthritis</td>
<td>B27</td>
</tr>
<tr>
<td>Acute anterior uveitis</td>
<td>B27</td>
</tr>
<tr>
<td>Subacute thyroiditis</td>
<td>B35</td>
</tr>
<tr>
<td>Fast progression of HIV infection</td>
<td>B35</td>
</tr>
<tr>
<td>Psoriasis vulgaris</td>
<td>cw6</td>
</tr>
<tr>
<td>Idiopathic haemochromatosis</td>
<td>(A3)</td>
</tr>
</tbody>
</table>
CHAPTER 2

MODULAR VARIATIONS OF TEH HUMAN MHC CLASS III GENES RCCX:
A MECHANISM FOR GENE DELETIONS AND DISEASE ASSOCIATIONS

2.1 ABSTRACT

The frequent variations of human complement component C4 gene size and gene numbers, plus the extensive polymorphism of the proteins, render C4 an excellent marker for MHC disease associations. As shown by definitive RFLPs, the tandemly arranged genes RP, C4, CYP21 and TNX are duplicated together as a discrete genetic unit termed the RCCX module. Duplications of the RCCX modules occurred by the addition of genomic fragments containing a long (L) or a short (S) C4 gene, a CYP21A or a CYP21B gene and the gene fragments TNXA and RP2. Four major RCCX structures with bimodular L-L, bimodular L-S, monomodular L and monomodular S are present in the Caucasian population. These modules are readily detectable by Taq I RFLPs. The RCCX modular variations appear to be a root cause for the acquisition of deleterious mutations from pseudogenes or gene segments in the RCCX to their corresponding functional genes. In a patient with congenital adrenal
hyperplasia (CAH), we discovered a $TNXB-TNXA$ recombinant with the deletion of $RP2-C4B-CYP21B$. Elucidation of the DNA sequence for the recombination breakpoint region and sequence analyses yielded definitive proof for an unequal crossover between $TNXA$ from a bimodular chromosome and $TNXB$ from a monomodular chromosome.
2.2 INTRODUCTION

Besides the immunoglobulins, complement component C4 is probably the most polymorphic serum protein. There are two isotypes, C4A and C4B, which manifest remarkable differences in chemical reactivities and serological properties [reviewed in (Yu, 1999)]. More than thirty-four allotypes for C4A and C4B have been demonstrated by agarose gel electrophoresis, based on gross differences in electric charge (Mauff et al., 1990a). Similar to the protein, the complement C4 genes are unusually complex with frequent variations in gene size and gene number. In addition, the genes surrounding C4A or C4B also exhibit considerable variations. These neighboring genes include RP1 or RP2 at the 5' region, CYP21A or CYP21B, and TNXA or TNXB at the 3' region (Fig. 2.1). The complex organizations of the C4A and C4B genes, together with the extensive polymorphisms of the C4A and C4B proteins render C4 an excellent marker for MHC associated diseases (Porter, 1983; Yu, 1999). For instances, congenital adrenal hyperplasia (CAH) is mainly caused by mutations or deletions of CYP21B (Miller and Morel, 1989), systemic lupus erythematosus (SLE) is correlated with C4A deficiencies (Fielder et al., 1983). In addition, insulin dependent diabetes mellitus (Degli-Esposti et al., 1992; Jenhani et al., 1992), sudden infant death syndrome and spontaneous recurrent abortion (Laitinen et al., 1991; Schneider et al., 1989), IgA deficiency and common variable immunodeficiency (Howe et al., 1991; Schaffer et al., 1989), IgA nephropathy (Welch et al., 1989), skin
vitiligo and pemphigus vulgaris (Ahmed et al., 1991; Venneker et al., 1992), autism and
narcolepsy (Matsuki et al., 1985; Warren et al., 1992), have all been suggested to be associated with specific alleles or null alleles of C4.

The human C4 genes are either 21 kb (long, L) or 14.6 kb (short, S) in size (Yu, 1991). This dichotomous size variation is due to the presence of an endogenous retrovirus HERV-K (C4) in intron 9 of the long gene (Chu et al., 1995; Dangel et al., 1994; Tassabehji et al., 1994). There may be one, two or three C4 genes in the MHC class III region of chromosome 6 (Carroll et al., 1984). Most people have two C4 genes in the MHC with one coding for a C4A protein and the other coding for a C4B protein (Fig. 2.1). C4A has higher affinities to amino group containing targets; C4B has higher affinities to hydroxyl group containing targets. These differences are the result of four amino acid changes between positions 1101 - 1106 (Carroll et al., 1990; Dodds et al., 1996; Yu et al., 1986). A significant proportion of the population has a single C4 gene in chromosome 6 coding for C4A or for C4B. Deletion or duplication of the C4 genes are always concurrent with their downstream genes, steroid 21-hydroxylase genes, CYP21A or CYP21B (Carroll et al., 1985a; Schneider et al., 1986).

RP1 is one of the four novel genes, RD-SKI2W-DOM3Z-RP1, present in the 30 kb genomic region between complement component genes factor B (Bf) and C4 (Levi-Strauss et al., 1988; Qu et al., 1998; Sargent et al., 1994; Shen et al., 1994; Yang et al., 1998). RP2 is a partially duplicated gene segment that contains only 913 bp of the sequence corresponding to the last two and half exons of RP1. RP1 transcripts are ubiquitously expressed (Yang et al., 1998). Derived amino acid sequence suggested that RP1 codes for a
nuclear protein that is probably a serine/threonine kinase (Gomez-Escobar et al., 1998; Shen et al., 1994; Yang et al., 1998).

The cytochrome P450 steroid 21-hydroxylase genes CYP21A or CYP21B are located 3028 bp downstream of C4A or C4B, respectively (Yu, 1991). CYP21 is essential for the biosynthesis of glucocorticoid and mineralocorticoid hormones. Complete absence of CYP21 leads to salt-wasting, low activity of CYP21 causes simple virilizing, while below average CYP21 activity causes androgen excess [reviewed in (Miller and Morel, 1989)]. CYP21A is a pseudogene because it contains three deleterious mutations: an 8 bp deletion in exon 3 and a T nucleotide insertion in exon 7 that result in frame shift mutations, as well as a C to T transition in exon 8 that generates a premature stop codon. In addition, there are many other point mutations in coding and non-coding sequences (Higashi et al., 1986; Rodrigues et al., 1987; White et al., 1986). If on both copies of chromosome 6, the deleterious mutations in CYP21A are incorporated into CYP21B, or the CYP21B genes are deleted, the subject suffers CAH.

The 3' ends of CYP21A or CYP21B overlap with the 3' ends of extracellular matrix protein tenasin TNXA or TNXB by 444 bp, respectively (Morel et al., 1989). The gene configurations of TNXA and TNXB are opposite to those of RP, C4, and CYP21. TNXB gene is 68.2 kb in size, consists of 45 exons, and encodes a protein of 4,289 amino acids (Bristow et al., 1993; Rowen et al., 1997b). The derived amino acid sequence of TNXB reveals a heptad, 18.5 epidermal growth factor repeats, thirty-two fibronectin type III repeats, and a fibrinogen domain. The overall structure of TNXB shows a striking similarity to extracellular matrix proteins tenasin/cytostatin (TN-C) and restrictin (TN-R) (Gitelman et
TN-C is present in central and peripheral nervous system and in smooth muscle and tendon. It is probably involved in cell adhesion and cell morphology (Chiquet-Ehrismann, 1993). TN-R is expressed in the nervous system and implicated in neural cell attachment (Rathjen, 1993). The function of TNXB is yet to be determined. Its transcripts are ubiquitously expressed in the fetus (Bristow et al., 1993). TNXA is a partially duplicated gene segment that corresponds to intron 32 to exon 45 of TNXB. In addition there is a 120 bp deletion at exon 36-intron 36 that results in a frame shift mutation and premature termination of translation (Gitelman et al., 1992; Shen et al., 1994).

The demonstration of the endogenous retrovirus HERV-K (C4) mediating the size variation of C4 genes (Dangel et al., 1994), the elucidation of DNA sequences for RPl and RP2 (Shen et al., 1994), C4A and C4B (Yu, 1991; Yu et al., 1986), CYP21A and CYP21B (Higashi et al., 1986; Rodrigues et al., 1987; White et al., 1986), as well as TNXA and TNXB (Bristow et al., 1993; Rowen et al., 1997b; Shen et al., 1994) provide important information for resolving the fine structures and the complex organizations of the consecutive genes RP, C4, CYP21 and TNX. The concurrent deletions / duplications of C4 and CYP21 genes (Carroll et al., 1985a; Schneider et al., 1986) prompted us to investigate if RP and TNX also undergo rearrangements in normal individuals and in selected CAH patients. Diagnostic RFLPs for RPl and RP2, TNXA and TNXB have been devised. RP-C4-CYP21-TNX genes are organized in variable, modular fashions. The unusually frequent modular variation appears to be the root cause for unequal crossovers and exchange of sequences between the functional and non-functional genes of the RCCX.
2.3 EXPERIMENTAL PROCEDURES

**Oligonucleotides** — Oligonucleotides were synthesized by an Applied Biosystem Model 380B DNA Synthesis machine. The sequences added to facilitate cloning are represented in lower cases. For amplification and sequencing of **TNX** genes (Bristow et al., 1993): **RDX-5**, aga gAA TTC AGT GAA ATC AGG GAG ACC; **RDX-3**, gag gaa TTC CAG TGC AGC ACG GCG AA; **SDX-52**, GGA GCC TCA GAG TGT GCA; **SDX-32**, CAA TCG GAG CCT CCA CCA; **XB54H**, gtg gaa ttc AAG CGA GCA CCT GAC TCA; and **XA31H**, gtt gaa ttc TTT TCT TGA CTC CCT G. For amplification of **CYP21A** probe (Rodrigues et al., 1987): **21A5**, TGT GGC CAT TGA GGA GGA A; and **21A3**, TGC CAC CGA TCA GGA GGA GGT C.

**Isolation of Human Genomic DNA** — Genomic DNAs were isolated following standard protocols from cultured cell lines HepG2 (liver carcinoma) and MOLT4 (T-cell leukemia), peripheral blood of normal individuals, and a congenital adrenal hyperplasia patient (CAH-E1). Appropriate consents from blood donors were obtained according to approved protocols by the Institutional Board of the Columbus Children's Hospital.

**Complement C4 allotyping** — Complement C4A and C4B allotypes from EDTA-blood plasma were determined as described (Awedeh and Alper, 1980; Sim and Cross, 1986).
Briefly, 10 μl of plasma was digested with 0.1 U of neuraminidase (Sigma, St Louis, MO) at 4°C overnight, and with 0.1 U of carboxyl peptidase B (Sigma) at room temperature for 30 min. Two agarose gels were prepared. Four μl of digested plasma was loaded to each gel and resolved by high voltage gel electrophoresis (Schneider and Rittner, 1997). One of the gels was subjected to immunofixation using goat antisera against human C4 (Incstar, Stillwater, MN). Plasma proteins in the other gel were subjected to immunoblot analyses using anti-Ch1 or anti-Rg1 monoclonal antibodies (anti-C4B, cat no. C057-325.2, lot no. 120287; anti-Rg1, RGd1; kindly provided by Dr. Joann M. Moulds, Houston, TX) at a dilution of 1:5000 and 1:1000, respectively. Immune complexes were detected by chemiluminescence method using the ECL-plus reagents (Amersham, Piscataway, NJ).

**HLA typing** — HLA typing of the E family was kindly performed by The Ohio State University Tissue Typing Laboratory.

**PCR of Cosmid and Genomic DNA** — PCR of cosmid and genomic DNA were performed after standard procedures (Saiki, 1990). PCR products were purified from 0.8% low gelling temperature agarose and cloned into pBluescript vectors.

**DNA Probes** — RP: RP1.1, a 1.1 kb insert of RP1 cDNA (Shen et al., 1994); RP1 3' probe, a 651 bp Nhe I-EcoR I fragment of RP1.1. C4: PA, a 476 bp BamH I-Kpn I cDNA fragment isolated from pAT-A. pAT-A contains the almost full length cDNA insert for the human C4A4 allele (Belt et al., 1984; Yu et al., 1986). PB, a C4d-specific 926 bp
BamHI DNA fragment subcloned from λJM-2a that contains a C4B5 gene (Yu et al., 1986). CYP21: a 757 bp fragment of CYP21A, amplified from cos 2 using primers 21A5 and 21A3. A cosmid isolated from a human genomic library DA (Yu et al., 1992), cos 2 spans from the 5' region of C4A1 gene to the 3' region of TNXB gene. The genomic DNA of cos 2 derives from the HLA haplotype A3 B47 DR7 that contains a deletion of the CYP21B gene. TNX: a 600 bp fragment of TNXB, corresponding to exons 35-37 of TNXB, amplified from cos 2 using primers RDX-5 and RDX-3. TNX-800: an 800 bp fragment of TNXB, immediately upstream of the TNXA/TNXB breakpoint, amplified from cos 2 using primers XB-MS5 and XB-BK3.

Southern Blot Analysis — Ten micrograms of genomic DNA were digested to completion with the appropriate restriction enzymes for 16 hrs, resolved on 0.8% agarose gel, blotted onto Hybond-N membrane (Amersham, Arlington Heights, IL), and hybridized with an appropriate [α-32P] dCTP-labeled probe, as described in (Yu and Campbell, 1987).

DNA Sequencing and Sequence Analysis — Sequencing reactions were performed using a Sequenase kit (U.S. Biochemicals Corp., Cleveland, OH) and 35S-ATP following the dideoxy sequencing method (Sanger et al., 1977), or by automated sequencing method using an ABI 377 machine. DNA sequences were compiled using PC/Gene software (Intelligenetics, Mountain View, CA). Comparisons of the sequences were
performed by FASTA, BESTFIT, PILEUP and PRETTY programs in GCG package through the Pittsburgh Supercomputing Center (Genetics Computer Group, 1991).
2.4 RESULTS

**Allotyping of C4**

Allotyping of the plasma C4 proteins from the B and E family members is shown in *Fig. 2.2, panel I*. Deduced from the immunofixation experiments, B1 is homozygous for C4AQ0 C4B1 (*lane 1*); B2 and B3 are both heterozygous, with C4A3 C4B1, C4A3 C4BQ0 for B2 (*lane 2*) and C4A3 C4B1, C4AQ0 C4B1 for B3 (*lane 3*). E1 has C4A1 C4BQ0, C4A3 C4BQ0 (*lane 4*); E2 is homozygous for C4A3 C4BQ0 (*lane 5*); and E3 is heterozygous for C4A1 C4B1, C4A3 C4B1 (*lane 6*). A normal control with homozygous C4A3 C4B1 was shown in *lane 7*. Previously, it was shown by hemagglutination that the C4A3 allotype reacts with anti-Rg1 and the C4B1 reacts with anti-Ch1. In this study, immunoblot analyses confirmed that in each individual except for E1 and E2 (both of which are homozygous C4BQ0), C4B1 reacted with anti-Ch1 monoclonal antibody (*panel II*). Noticeably, C4A1 in E1 (*lane 4*) and E3 (*lane 6*) also interact with the anti-Ch1 antibody. In *panel III*, all individuals except B1 (which is homozygous C4AQ0) manifested positive reactions for the C4A3 allotype with the anti-Rg1 monoclonal antibody.
Modular variations for RP, C4, CYP21 and TNX

Whether the flanking genes RP (located 5' to C4) and TNX (located 3' to CYP21) are involved in the C4-CYP21 gene deletion events were investigated. Diagnostic RFLPs were devised to detect and distinguish the presence of the RP1 and RP2, as well as TNXA and TNXB. By using an RP1 3' probe for hybridization, the RP1 gene can be represented by a 9.6 kb BamHI I fragment, and the RP2 (and TNXA gene segments) can be represented by a 5.0 kb BamHI I fragment. By using a 600 bp TNX probe corresponding to TNXB exons 35-37, the presence of the TNXB gene can be detected by a 9.0 kb Sca I fragment, and the TNXA gene by a 4.0 kb Sca I fragment (Fig. 2.3, panel I). Previously, techniques were established for detecting the presence of the C4A and C4B genes by Nla IV RFLP, and their associations with Rg1 or Ch1 antigenic determinants by EcoO 109I RFLP (Yu and Campbell, 1987). In addition, the presence of CYP21A and CYP21B can be detected by the 3.2 kb and 3.7 kb Taq I fragments, respectively (White et al., 1986). Fig. 2.3 (panels II to IV) shows a series of Southern blot analysis of DNA samples isolated from two families, B and E. Members of the B family appear to be normal. For the E family, there is a congenital adrenal hyperplasia (CAH) patient E1.

As shown in Fig. 2.3, panel II, restriction fragments corresponding to RP1 (9.6 kb BamHI I fragment) and RP2 (5.0 kb BamHI I fragment) were detected in all individuals except B1 and E1. In these two individuals, only the RP1-specific 9.6 kb fragment is present (lanes 1 and 5).

The presence of the C4A or C4B gene was analyzed by Nla IV restriction patterns shown in panel III-A. B1 contains the 467 bp fragment corresponding to C4B genes
E1 and E2 have the 276 bp and 191 bp fragments specific for C4A genes (lanes 5 and 6). Other individuals contain fragments corresponding to both C4A and C4B genes. The associations of C4 genes with the major Chido (Ch1) and Rodgers (Rg1) antigens were revealed in panel III-B by the EcoO 109I restriction patterns. The C4B genes in B1 express the Ch1 antigen since only the 458 bp fragment can be detected (lane 1). The C4A genes in E2 express the Rg1 epitope since only the 565 bp fragment is present (lane 6). Both Rg1 and Ch1 specific fragments are present in E1 (lane 5). Therefore, one of the C4A genes in E1 expresses Ch1 (that is frequently associated with C4B), the other C4A gene expresses Rg1. All other individuals contain both C4A and C4B genes. These C4 genes express Rg1 and Ch1 because both 565 bp and 458 bp fragments were detectable. An additional 344 bp, EcoO 109I fragment exists in all individuals as the 926 bp C4d specific probe (Pb) was used. This fragment is common to C4 genes with Rg1 or Ch1 (Yu and Campbell, 1987).

The presence of CYP21A and CYP21B was determined by Taq I RFLP (panel IV). Both CYP21A (3.2 kb fragment) and CYP21B (3.7 kb fragment) were detected in all samples except B1 and E1. B1 only contains the functional CYP21B gene (lane 1), while the CAH patient E1 only contains the pseudogene CYP21A (lane 5).

The presence of TNXA and TNXB genes are determined by Sca I RFLP (panel V). Restriction fragments for both TNXA (4.0 kb Sca I fragment) and TNXB (9.0 kb Sca I fragment) were detected in all individuals except B1 and E1 (panel II-B). These two individuals have the 9.0 kb Sca I restriction fragment corresponding to TNXB but no 4.0 kb Sca I fragment corresponding to TNXA (lanes 1 and 5).
From the above results, it becomes clear that individuals who have a single locus for RP also have single gene loci for C4, CYP21 and TNX. Individuals with both RP1 and RP2 loci also have duplicated loci for C4, CYP21 and TNX. Hence, the four tandemly arranged genes RP, C4, CYP21, and TNX are duplicated or deleted together as a discrete genetic unit. This genetic unit is designated as the RCCX module.

Molecular basis of the C4 and RP Taq I RFLP and variations of the RP, C4, CYP21 and TNX (RCCX) loci in the B and E families

The Taq I RFLP is the most widely used technique to illustrate the complexities of polymorphisms in the number and size of C4 genes. Four fragments of 7.0, 6.4, 6.0 and 5.4 kb in size can be detected by Southern blot analysis of Taq I digested genomic DNA hybridized to PA, a C4 5' probe (Schneider et al., 1986; Simon et al., 1997; Yu and Campbell, 1987). At the same time, all these Taq I fragments can also be detected by an RP1.1 probe, suggesting that the Taq I RFLP is caused by both RP and C4 gene variations. There is a BamHI I site at the 5' untranslated region of the C4A and C4B genes. To segregate the RP1/RP2 variation from the C4 long/short (L/S) size dichotomy, the BamHI I-Taq I double digests of genomic DNA were performed (Fig. 2.4, panel I). As shown in panel II, the single 6.4 kb Taq I fragment (panel II-A, lane 1) was split to a 3.1 kb (panel II-B, lane 1) and a 3.2 kb (panel II-C, lane 1) Taq I-BamHI I fragments corresponding to an RP1 gene and a short C4 gene. The 7.0 kb and 6.0 kb Taq I fragments (panel II-A, lane 2) were split to 3.1 kb and 2.1 kb Taq I-BamHI I fragments corresponding to RP1 and RP2 genes (panel II-B, lane 2), and a single 4.0 kb BamHI I-
Taq I fragment corresponding to long C4 genes (panel II-C, lane 2). The 7.0 kb and 5.4 kb Taq I fragments (panel II-A, lane 3) were divided into the RPl-associated 3.1 kb and the RP2-associated 2.1 kb Taq I-BamH I fragments (panel II-B, lane 3), and the 4.0 kb and 3.2 kb BamH I-Taq I fragments associated with long and short C4 genes, respectively (panel II-C, lane 3). Therefore, the restriction patterns of the Taq I genomic Southern blot for RP or C4 genes are resulted from a combination of variations of RPl or RP2, one or two loci of C4 genes, and long or short C4 genes. The basis of the Taq I RFLP is interpreted as follows: the 7.0 kb fragment indicates the presence of an RPl gene and a long C4 gene; the 6.4 kb fragment corresponds to an RPl gene and a short C4 gene; the 6.0 kb fragment represents an RP2 and a long C4 gene; and the 5.4 kb fragment stands for an RP2 and a short C4 gene. The actual sizes of these Taq I restriction fragments derived from DNA sequences are 0.1 kb greater or smaller than the apparent sizes described above.

The specific combinations of RPl/RP2 genes with C4(L)/C4(S) genes in the B and E family members were examined by Taq I digests and RP1.1 probe as shown in Fig. 2.5, panel I. The RPl-C4(S) haplotype is present in B1, as represented by the single 6.4 kb Taq I fragment (lane 1). The RPl-C4(L) haplotype is present in E1, as shown by the single 7.0 kb Taq I fragment (lane 5). Therefore, homozygous, single RPl gene and single C4 gene are present in B1 and E1. B2 has the homozygous RPl-C4(L)-RP2-C4(L) haplotype as revealed by the presence of the 7.0 kb and the 6.0 kb Taq I fragments (lane 2). B3 and B4 (children of B1 and B2) are heterozygous for the RPl-C4(S) haplotype and the RPl-C4(L)-RP2-C4(L) haplotype since they both contain the 7.0 kb, 6.4 kb and
6.0 kb Taq I fragments (lanes 3 and 4). E2 and E3 (parents of E1) are heterozygous for the RP1-C4(L)-RP2-C4(L) haplotype and the RP1-C4(L) haplotype since they both have the 7.0 kb and 6.0 kb Taq I fragments (lanes 6 and 7).

Detection of the TNXA-associated 120 bp deletion in a TNXB gene of CAH patient E1

Compared with TNXB, TNXA has not only a 63 kb truncation of the 5' region (Bristow et al., 1993; Rowen et al., 1997b; Shen et al., 1994), but also a 120 bp deletion spanning across the junction of exon 36 and intron 36 (Shen et al., 1994). This deletion attributes to a Taq I RFLP when a TNX 3' probe is used for Southern blot analysis: TNXA associates with a 2.4 kb fragment (TNX-2.4) and TNXB with a 2.5 kb fragment (TNX-2.5). The analysis of this TNX Taq I polymorphism in the B and E families are shown in Fig. 2.5, panel III. B1 who has monomodular RCCX (S) structure exhibited TNX-2.5 (lane 1) that is consistent with the presence of TNXB genes only. B2 who has bimodular RCCX (L-L) structure revealed both TNX-2.4 and TNX-2.5 (lane 2), which is concordant with the presence of both TNXA and TNXB. B3 and B4 who are heterozygous for bimodular RCCX (L-L) and monomodular RCCX (S) also revealed both TNX-2.5 and TNX-2.4 fragments.

E3 is heterozygous for monomodular RCCX (L) and bimodular RCCX (L-L), implying that she has two TNXB genes and one TNXA gene. As expected, the relative band intensity of TNX-2.5 to TNX-2.4 is 2:1 (lane 7). E2 is also heterozygous for monomodular RCCX (L) and bimodular RCCX (L-L) with two TNXB genes and one TNXA gene in a diploid genome. However, the band intensity of TNX-2.5 is only half of
that for TNX-2.4 (lane 6), which is opposite to the expected result. The CAH patient E1 is homozygous for monomodular RCCX (L) and has only TNXB but no TNXA. Unexpectedly, he manifested both TNXB-related 2.5 kb fragment and TNXA-related 2.4 kb fragment. The relative band intensity for TNX-2.5 and TNX-2.4 is 1:1. These results suggest that in both E1 and E2, one of the TNXB genes contains the 120 bp deletion. In other words, the aberrant TNXB gene in E1 originates from the paternal chromosome.

Correlation of the HLA typing data for the E family (Table 2.1) and the Taq I RFLP data suggests that the aberrant TNXB gene is present in HLA haplotype A3 B35 DR1.

Based on the phenotypes of the C4A and C4B proteins, together with the haplotypes of the RCCX modules, the organizations of RCCX genotypes for the B and E family members are deduced and shown in Table 2.1. The haplotype c of the B family and the haplotype a of the E family both have bimodular RCCX structures coding for C4A3 protein from the two C4 loci.

An unequal crossover between monomodular and bimodular RCCX chromosomes leading to gene deletion and gene duplication

The presence of a monomodular RCCX structure in E1 with RP1-C4(L), CYP21A, and a TNXB gene with TNXA-associated 120 bp deletion leads us to hypothesize that there was an unequal crossover between TNXA and TNXB from the homologous chromosomes. This unequal crossover resulted in the formation of an XB-XA recombinant and the deletion of the RP2-C4B-CYP21B genes (Fig. 2.6, panel I). To test this hypothesis, the genomic region spanning the putative 120 bp deletion was amplified
by PCR using the CAH-E1 genomic DNA. The strategies for PCR are depicted in panel II of Fig. 2.6. When primers RDX5 and RDX3 were used, two PCR products in the size of 493 bp and 613 bp were obtained. These products were cloned and sequenced. Both sequences correspond to the exons 35 - 37 of TNX. The sequence of the 613 bp is identical to that of the regular TNXB, while the 493 bp product appears to contain a 120 bp deletion similar to that observed in TNXA. To further prove that this is an aberrant TNXB gene with a TNXA-associated 120 bp deletion, a 2.6 kb genomic DNA fragment was amplified using RYM25 and XA31H. RYM25 is a TNXB-specific primer because it is located in exon 32 and is 220 bp upstream of the gene duplication breakpoint for TNXA and TNXB. The XA31H primer spans across the 120 bp deletion and therefore is TNXA-specific. The 2.6 kb PCR product was cloned and sequenced to completion. The sequences of the 2.6 kb and the 493 bp fragments overlap and so were compiled. The resulting sequence E1XB-A was aligned with two normal TNXB sequences (TNXB-H and TNX-M) (Bristow et al., 1993; Rowen et al., 1997b) and two normal TNXA sequences (TNXA-M and TNXA-Y) (Shen et al., 1994; Tee et al., 1995) obtained from three different laboratories. In addition, we include a recombinant TNXA sequence in a pauciarticular JRA patient, L1XA-B that acquired the described 120 bp genomic DNA sequence (Rupert et al., 1999). The alignment starts at the breakpoint of gene duplications for TNXA-TNXB. The alignment reveals a picture for the original DNA recombination leading to the reciprocal recombinant sequences in E1XB-A and L1XA-B (Fig. 2.7, panel I). There are nine single nucleotide changes or informative sites (marked by asterisks) which can differentiate TNXA from TNXB. A Chi sequence related to DNA
recombination in bacteriophage λ is located at nucleotides 450-457. Five of the
TNXA/TNXB informative sites (nucleotides 14, 60, 63, 73 and 278) are 5' to the Chi site.
The other four sites (nucleotides 2234, 2268, 2273 and 2423) are 3' to the Chi site and
they are flanking the 120 bp deletion. For E1XB-A, the sequence is TNXB-specific for
the first five informative sites, but is TNXA-specific for the last four sites. It also acquires
the TNXA-specific 120 bp deletion (Fig. 2.7, panel II). This suggests that E1XB-A is the
result of a recombination between TNXB and TNXA, occurred between nucleotides 278
and 2234.

For L1XA-B, it has sequences characteristic of TNXA at the first five informative
sites, but has sequences characteristic of TNXB at the last four sites and also includes the
120 bp sequence. This aberrant TNXA sequence in L1XA-B appears to be the reciprocal
product of E1XB-A resulted by DNA recombination between TNXA and TNXB.
It is shown here that in human MHC class III region, four tandemly arranged genes serine/threonine kinase RP, complement C4, steroid 21-hydroxylase CYP21, and tenascin TNX are organized as a genetic unit designated as an RCCX module. In a monomodular RCCX haplotype, the "full length" genes RP1 and TNXB are always present, inferring relevant cellular functions for RP1 and for TNXB. In an RCCX bimodular haplotype, duplication of the RCCX module occurs by the addition of a C4 gene, a CYP21 gene together with the TNXA and RP2 gene segments. This additional, modular genomic fragment is either 32.5 kb or 26.2 kb in size, depending on whether the C4 gene contains the endogenous retrovirus HERV-K (C4). The three pseudogenes/gene segments, CYP21A, TNXA and RP2, present between the two C4 loci, probably do not encode for functional protein products. The concurrent deletions of a C4A or C4B gene with a CYP21A or CYP21B gene is a well-established phenomenon (Carroll et al., 1985a; Schneider et al., 1986), this report provides the detailed documentation for the modular deletions or duplications of RP and TNX genes together with C4 and CYP21.

Although this multiple-gene modular variation observed in RCCX is uncommon in mammalian genetics, a similar phenomenon with concurrent variations of at least three tandem genes has been observed in a genomic region at chromosome 5q12-13. The three genes are BTFp44 (a p44 subunit of transcriptional factor TFIIF), NAIP (neuronal
apoptosis inhibitory protein), and SMN (survival motor neuron). One, two or three modules of these three genes, each module spans about 500 kb in size, are present in the population (Lewin, 1995). The divergence of sequences in C4A and C4B is analogous to that in SMNT and SMNC; the presence of the pseudogene CYP21A and the functional CYP21B is analogous to the pseudogene NAIPA (with deletion of exon 5) and the intact NAIP. The absence of the SMNT gene is associated with most spinal muscular atrophy (SMA), and deletion of the NAIP gene in most severe forms of SMA (Taylor et al., 1998a).

The frequency of the RCCX modular variations has been studied in a population of 150 normal Caucasian females. It is discovered that 75.4% of the C4 genes are long, and 24.6% are short. Bimodular and monomodular RCCX organizations are present in about 71.6% and 16.2% of the chromosome 6, respectively. Trimodular RCCX haplotypes are uncommon and have a frequency about 12.2% [Hauptmann et al., 1988, C. A. Blanchong and C.Y. Yu, unpublished observation]. Excluding the trimodular haplotypes (Mauff et al., 1990a; Raum et al., 1984), there are four major RCCX modular structures in the Caucasian population: bimodular long-long (L-L), bimodular long-short (L-S), monomodular long (L), and monomodular short (S) (Fig. 2.8). These four RCCX structures can be detected conveniently by Taq I RFLPs. The widely applied Taq I RFLP analysis of C4 genes yields information on the combination of RP1 or RP2 with C4(L) or C4(S). It does not yield definitive information, however, on whether the C4 gene codes for C4A or C4B proteins. The information for the presence of C4A and C4B genes may be obtained by Nla IV RFLP analysis or by direct DNA sequencing. From
these four RCCX organizations, twelve haplotypes for $RP1/RP2$, $C4A/C4B$, $CYP21A/CYP21B$, $TNXA/TNXB$ are observable in the normal and in disease populations. Six of these haplotypes are more common in the normal population and they are highlighted (haplotypes 1, 2, 5, 9, 10 and 12). RCCX bimodular haplotypes with two $C4B$ genes (haplotypes 4 and 8) are yet to be shown definitively. Bimodular haplotypes with two $CYP21A$ genes (haplotypes 3 and 6) and a monomodular haplotype with a $CYP21A$ gene (haplotype 11) are present in CAH patients. In two Brazilian tribes, a fifth RCCX organization with two short $C4$ genes is found (Weg-Remers et al., 1997). This bimodular $C4(S)$-$C4(S)$ combination (haplotype 13) is extremely rare in other ethnic groups.

The diversities in the number and size of the RCCX modules probably contribute to the genetic variability and instability of the MHC class III region. A recombination between a bimodular RCCX chromosome and a monomodular RCCX chromosome may lead to the exchange or homogenization of polymorphic or mutant sequences between complement $C4A$ and $C4B$ loci. This has been demonstrated by the presence of the C4A-associated amino acid residues in many $C4B$ allotypes, and vice versa. The typical examples are the reverse associations of Ch1 antigenic determinant with $C4A1$ and $C4A13$, and Rg1 antigenic determinant with $C4B5$ (Giles et al., 1988; Yu et al., 1986, 1988). It was also shown that in an SLE patient, there is an acquisition of a 2 bp insertion in exon 29 from $C4AQ0$ to $C4BQ0$ (Barba et al., 1993; Lokki et al., 1999). The same type of recombination may also lead to the acquisition of mutations from pseudogene $CYP21A$ or gene segment $TNXA$ to their corresponding functional genes (Harda et al.,
1987; Tusie-Luna and White, 1995). This is manifested in many disease-associated haplotypes such as the presence of two CYP21A pseudogenes in RCCX bimodular haplotypes (Harada et al., 1987), and a CYP21A/CYP21B hybrid in CAH patients with HLA haplotype A3 B47 DR7 and monomodular RCCX (Chu et al., 1992; Donohoue et al., 1995). Another example comes from the presence of a TNXB/TNXA hybrid gene with the 120 bp deletion together with the deletion of the RP2-C4B-CYP21B gene in CAH patient(s), as demonstrated in this chapter.

In CAH patient E1, the 120 bp deletion in the TNXB-XA recombinant will cause a premature termination and therefore truncation of the carboxyl terminal sequences. The truncation includes three fibronectin type III repeats (with four N-linked glycosylation sites) and the entire fibrinogen domain. This mutation may diminish or knockout the function of TNX. The haplotype for the recombinant chromosome with monomodular RCCX characterized the presence of a CYP21A pseudogene linked to the TNXB/XA hybrid is HLA A3 B35 DR1, RP1 C4A3 CYP21A TNXB-XA.

Three independent observations on the deletions of the CYP21B genes, which could have arisen by a mechanism similar to that for haplotype b of CAH-E1, were reported. In the first study, HLA haplotype B35 DR1 was found in a salt losing CAH patient. In this case, a single C4A3 gene and a CYP21A gene were present with the deletion of C4B and CYP21B (Jospe et al., 1987), which is similar to haplotype b of CAH-E1. Whether this patient has a recombinant TNXB-XA in the monomodular RCCX was not determined.

In the second study, a de novo deletion of C4B together with CYP21B was suggested to derive from a meiotic unequal crossover between the maternal homologous
chromosomes (Sinnott et al., 1990). From our current knowledge, this *de novo* deletion was probably resulted by an unequal crossover between *TNXA* from a bimodular L-S chromosome (HLA A30 B13 DR7) and *TNXB* from a monomodular S chromosome (HLA A1 B8 DR3). The recombinant has a monomodular L chromosome (HLA A30 B13 DR3) with *CYP21A* and a 2.4 kb *Taq* I fragment at its 3' end, which is indicative of a 120 bp deletion in the *TNXB* gene.

In the third study, one of the chromosome 6 in a CAH patient appears to have a monomodular RCCX with the 120 bp deletion in *TNXB*. The other chromosome 6 has a bimodular RCCX with two *CYP21A* genes and no *CYP21B* gene, as revealed by Southern blot analyses using *Taq I*, *Bgl II* and *BssH II* digested genomic DNA. Immunoblot analysis showed that this patient did not produce any TNXB protein. Therefore, both *TNXB* genes from the homologous chromosomes were presumed to be non-functional. Since the patient suffers Ehlers Danlos syndrome in addition to CAH, it is proposed that malfunction of TNXB is associated with the connective tissue disease (Burch et al., 1997).

Another important piece of evidence for the described unequal crossover comes from studies on the molecular genetics of a pauciarticular juvenile rheumatoid arthritis (JRA) patient in our laboratory. This JRA patient has an RCCX bimodular haplotype with *two* *CYP21B* genes and a 5'-*TNXA/XB*-3' hybrid with the *TNXA*-specific truncation of exons 1-32 at the 5' region, and the presence of the *TNXB*-specific 120 bp sequence at the 3' region (Rupert et al., 1999). The *TNXA-XB* hybrid appears to be the reciprocal product in CAH-E1 5'-*TNXB/XA*-3', attributable to the genetic recombination between *TNXA* and
This notion is substantiated by the reciprocal associations of the informative sites for TNXB and for TNXA at the two ends of the hybrid sequences (Fig. 2.7).

In the bimodular (or trimodular) RCCX haplotypes, one of the duplicated genes or gene segments could afford to sequence mutations without immediate deleterious effect of knocking out the gene function. The RCCX modular variations in the population allowed sequence variations and enhanced the incorporation of diversified or mutant sequences among the paralogous genes, pseudogenes or gene segments. The selection advantage is probably the emergence of various polymorphic forms of complement C4A and C4B to tackle different microbial antigens (Kawaguchi et al., 1991). The burdens are the accompanying genetic or autoimmune diseases such as CAH, SLE and possibly EDS, caused by unequal crossovers and incorporations of deleterious mutations in the constituents of the RCCX.

In the MHC class II region between DRB1 and DRA genes, there may be 1 to 3 DRB pseudogenes. In addition, DRB3, DRB4 or DRB5 can be present (Apple and Erlich, 1996). The DR locus is about 350 kb centromeric to the RCCX modules. Between heterozygous chromosomes with different DR gene number and RCCX modules, misalignments and unequal crossovers would occur during meiosis. This may result in deletion or duplication of structural genes between these two variable regions. More than ten structural genes with important functions have been discovered between DR and RCCX (Aguado et al., 1996; Rowen et al., 1997b). Recombinant chromosomes with the essential structural genes deleted would be lethal. Therefore, the apparent productive recombination frequencies between certain MHC haplotypes would be less than
expected. This would have contributed to the linkage disequilibrium of the MHC genes, as many of the "ancestral" MHC haplotypes remain largely conserved in the population.
**Fig. 2.1** **A molecular map of the human MHC complement gene cluster.** This map represents the most common organization of the genes in the normal population from complement component C2 gene to extracellular matrix protein TNXB gene. *Horizontal arrows* represents the direction of gene transcription. Pseudogenes or gene segments are *shaded*. The negative signs for the intergenic distances between CYP21A and TNXA, CYP21B and TNXB represent overlaps at the 3' ends of these genes (This figure is taken from Yang et al., 1998).
I. Immunofixation

![Immunofixation diagram]

II. Immunoblot (anti-Chl)

![Immunoblot (anti-Chl) diagram]

III. Immunoblot (anti-Rg1)

![Immunoblot (anti-Rg1) diagram]

Fig. 2.2  **C4 allotypes of the B and E families.**  I. Human C4A and C4B allotypes detected by immunofixation; Immunoblot analysis of the C4 allotypes using (II) anti-Ch1 monoclonal antibody and (III) anti-Rg1 monoclonal antibody. The C4A1 allotype with reverse association of Ch1 epitope is marked by an *arrow*. *Lane 4* is from a control plasma with C4A3 and C4B1.
Fig. 2.3  **Southern blot analysis to detect the modular variations of** **RP**, **C4**, **CYP21** **and** **TNX**. DNA samples were isolated from B and E family members and subjected to appropriate restriction enzyme digestions. **Panel I** illustrates the basis to detect **RP1** and **RP2** by **BamH I** RFLP, and to detect **TNXA** and **TNXB** by **Sca I** RFLP. **Solid bars** represent locations of the probes used. **Panel II** shows detection of **RP1** and **RP2** genes by **BamH I** RFLP. The probe used was RP1.1. **Panel III-A** shows the variation of **C4A** and **C4B** isotypes by **Nla IV** RFLP; **panel III-B** reveals the expression of the Rg1 and Ch1 antigenic determinants in **C4** genes by **EcoO 109I** RFLP. The probe used was P_B. **Panel IV** exhibits the presence of **CYP21A** and **CYP21B** genes by **Taq I** RFLP. The probe used was a 757 bp fragment amplified from **CYP21A**. **Panel V** shows the detection of **TNXA** and **TNXB** by **Sca I** RFLP. The probe used was a 600 bp fragment corresponding to exons 35-37 of **TNXB**.
Fig. 2.3 (continued)

II.

<table>
<thead>
<tr>
<th>kb</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

III.

A. bp

<table>
<thead>
<tr>
<th>467</th>
<th>276</th>
<th>191</th>
</tr>
</thead>
</table>

B.

| 565 | 458 | 344 |

IV.

<table>
<thead>
<tr>
<th>kb</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CYP21B

CYP21A

V.

<table>
<thead>
<tr>
<th>kb</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TNXB

TNXA
Fig. 2.4  Southern blot analysis to segregate the RP1/RP2 variation from the C4 long/short (L/S) size dichotomy. Panel I is a schematic diagram for the molecular basis of the Taq I and Taq I - BamH I RFLPs. In panel II-A, genomic DNA isolated from B1 (lane 1), HepG 2 (liver carcinoma, lane 2) and Molt 4 (T-cell leukemia, lane 3) were digested with Taq I and hybridized with RP1.1 probe. In panels II-B and II-C, DNAs were double-digested with Taq I and BamH I, and subjected to RP1.1 probe (panel II-B) or C4 5' probe P_A (panel II-C).
Fig. 2.5  *Taq I* polymorphism to detect the *TNXA*-associated 120 bp deletion in the *TNXB* gene of CAH patient E1. Genomic DNA isolated from B and E families were digested with *Taq* I and hybridized with appropriate probes to demonstrate different *RP-C4* combinations (*panel I*), the presence of *CYP21A* and *CYP21B* genes (*panel II*), and the variations of *TNX* genes (*panel III*). The 2.5 kb *Taq* I fragment is usually associated with *TNXB*, while the 2.4 kb *Taq* I fragment is usually associated with *TNXA*. An *arrow* indicates the unusual association of a 2.4 kb fragment with *TNXB* in E1.
Fig. 2.6  Recombination between RCCX modules. (I) A model for an unequal crossover between a bimodular RCCX chromosome and a monomodular chromosome to generate a TNXB/XA recombinant and a TNXA/XB recombinant; (II) PCR strategy to amplify the breakpoint region of gene recombination and its corresponding exon-intron structure.
Fig. 2.7 Elucidation of the breakpoint region for TNXB-XA recombinant. (I) An alignment of the normal and recombinant TNX4 and TNXB genomic DNA sequences. Nucleotide number starts at the breakpoint region (bold and underlined) of the TNXB and TNX4 gene duplication. The normal TNXB sequences, TNXB-H (GenBank accession no. U89337) and TNXB-M (GenBank accession no. X71937) are taken from Ref. 37 and Ref. 38, respectively. The normal TNX4 sequences, TNX4-Y (GenBank accession no. L20263) and TNX4-M (GenBank accession no. S38953) are taken from Ref. 30 and Ref. 38, respectively. EIXB-XA is generated by this work. L1XA-XB (GenBank accession no. AF077974) is from Ref. 72. TNX4 and TNXB specific sequences or informative sites are marked by asterisks. (II) A schematic diagram showing the reciprocal locations of TNX4 and TNXB informative sites in the TNX recombinants of CAH-E1 and JRA-L1. Informative sites are shown as vertical strokes; an open box represents the 120 bp deletion. The continuous DNA sequence for the breakpoint region of the TNXB/TNX4 recombinant (EIXB-A) is available from the GenBank under the accession number AF086641.
**Fig. 2.7** Elucidation of the breakpoint region for TNXB-XA recombinant.

(to be continued)
(Fig. 2.7 continued)
(Fig. 2.7 continued)

II.

**E1**  
*TNXB*  

**L1**  
*RP2*  

Breakpoint of Gene Duplication

120 bp deletion

Chi

200 400 bp

to *CYP21A*

to *CYP21B*
Fig. 2.8  **RCCX modular variations in the human population.**

Monomodular and bimodular RCCX structures with 13 different haplotypes of *RP, C4, CYP21* and *TNX* gene combinations are shown. The common haplotypes are in *bold*. Disease haplotypes associated with CAH are marked with *asterisks*. Abbreviations: 21A, *CYP21A; XA, TNXA*. The presence of haplotypes 4 and 8 (indicated by question marks) has not been definitively demonstrated.
Family B

B1  
   a. \( RP1 \rightarrow C4B1 \ (S; \ Ch1) \rightarrow CYP21B \rightarrow TNXB \)  
   b. \( RP1 \rightarrow C4B1 \ (S; \ Ch1) \rightarrow CYP21B \rightarrow TNXB \)

B2  
   c. \( RP1 \rightarrow C4A3 \ (L; \ Rgl) \rightarrow CYP21A \rightarrow TNXA \rightarrow RP2 \rightarrow C4A3 \ (L; \ Rgl) \rightarrow CYP21B \rightarrow TNXB \)  
   d. \( RP1 \rightarrow C4A3 \ (L; \ Rgl) \rightarrow CYP21A \rightarrow TNXA \rightarrow RP2 \rightarrow C4B1 \ (L; \ Chl) \rightarrow CYP21B \rightarrow TNXB \)

B3  
   a. \( RP1 \rightarrow C4B1 \ (S; \ Ch1) \rightarrow CYP21B \rightarrow TNXB \)  
   d. \( RP1 \rightarrow C4A3 \ (L; \ Rgl) \rightarrow CYP21A \rightarrow TNXA \rightarrow RP2 \rightarrow C4B1 \ (L; \ Chl) \rightarrow CYP21B \rightarrow TNXB \)

B4  
   a. \( RP1 \rightarrow C4B1 \ (S; \ Ch1) \rightarrow CYP21B \rightarrow TNXB \)  
   c. \( RP1 \rightarrow C4A3 \ (L; \ Rgl) \rightarrow CYP21A \rightarrow TNXA \rightarrow RP2 \rightarrow C4A3 \ (L; \ Rgl) \rightarrow CYP21B \rightarrow TNXB \)

Family E

E2  
   a. \( RP1 \rightarrow C4A3 \ (L; \ Rgl) \rightarrow CYP21A \rightarrow TNXA \rightarrow RP2 \rightarrow C4A3 \ (L; \ Rgl) \rightarrow CYP21B \rightarrow TNXB(2.5) \)  
   b. \( RP1 \rightarrow C4A3 \ (L; \ Rgl) \rightarrow CYP21A \rightarrow TNXB* (2.4) \)

E3  
   c. \( RP1 \rightarrow C4A3 \ (L; \ Rgl) \rightarrow CYP21A \rightarrow TNXA \rightarrow RP2 \rightarrow C4B1 \ (L; \ Ch1) \rightarrow CYP21B \rightarrow TNXB(2.5) \)  
   d. \( RP1 \rightarrow C4A1 \ (L; \ Ch1) \rightarrow CYP21A \rightarrow TNXB (2.5) \)

E1  
   b. \( RP1 \rightarrow C4A3 \ (L; \ Rgl) \rightarrow CYP21A \rightarrow TNXB* (2.4) \)  
   d. \( RP1 \rightarrow C4A1 \ (L; \ Ch1) \rightarrow CYP21A \rightarrow TNXB (2.5) \)

(to be continued)

Table 2.1  RCCX Modular Structures of the B and E Families
(Table 2.1 continued)

**HLA haplotypes of Family E**

E2  
  a. HLA A2 B44 DR4  
  b. HLA A3 B35 DR1

E3  
  c. HLA A29 B44 DR14  
  d. HLA A3 B47 DR7

E1  
  b. HLA A3 B35 DR1  
  d. HLA A3 B47 DR7
CHAPTER 3

DETECTION OF Msc I POLYMORPHISM AND
THE CONCURRENT 120 bp DELETION IN TNXB/XA RECOMBINANT GENE

3.1 ABSTRACT

Four tandemly arranged genes in the human MHC complement gene cluster (MCGC), serine/threonine kinase RP, complement C4, steroid 21-hydroxylase CYP21, and tenasin TNX are organized as a genetic unit designated as the RCCX module. Malfunction of CYP21 leads to congenital adrenal hyperplasia (CAH). A rearranged monomodular RCCX with TNXB/XA and concurrent deletion of RP2-C4B-CYP21B was observed in a CAH patient. Nine informative sites were revealed to distinguish TNXA from TNXB. Two of the sites in TNXB can be recognized by the Msc I restriction enzyme. The conversion of TNXB specific sequences to TNXA specific sequences leads to the abolition of the Msc I site. When the nucleotide conversions are associated with the
120 bp deletion, i.e., the presence of $TNXB/XA$ hybrid, it also indicates the deletion of $RP2-C4B-CYP21B$ genes. Molecular genetic techniques were developed to detect the presence of the $TNXB/XA$ recombinants in the population. This method can provide an efficient approach to screen for the possible carrier of the CAH disease in the clinical study.
3.2 INTRODUCTION

In human MHC class III region, genes located in between complement C2 and extracellular matrix tenasin X (TNX) are defined as the MHC complement gene cluster (MCGC) (Yu, 1999). It is a highly variable genomic region that is characterized by polymorphisms, variations in gene size and gene number. Four tandemly arranged genes in this cluster, serine/threonine kinase RP, complement C4, steroid 21-hydroxylase CYP21, and tenasin TNX are organized as a genetic unit designated as the RCCX module. Combining the modular number variations (monomodular, bimodular, or rarely trimodular), together with the C4 gene size variations (long or short), there have been 13 different RCCX haplotypes observed. The diversities in the number and size of the RCCX modules probably contribute to the genetic variability and instability of the MHC class III region (Yang et al., 1998).

TNXB gene is 68.2 kb in size, consists 45 exons, and encodes a functional protein product (Bristow et al., 1993; Rowen et al., 1997a, b). TNXA is a partially duplicated gene segment that corresponds to intron 32 to exon 45 of TNXB. TNXA also acquires a 120 bp deletion at exon 36 – intron 36 that results in a frame shift mutation and premature termination of translation (Gitelman et al., 1992; Shen et al., 1994). In a congenital adrenal hyperplasia (CAH) patient E1, an unusual chromosome rearrangement was identified. It contains a TNXB/XA hybrid gene with the 120 bp TNXB specific
sequence deleted. Sequence comparison of this hybrid gene with normal \textit{TNXA} and \textit{TNXB} genes revealed nine single nucleotide changes or informative sites, together with the 120 bp \textit{TNXB} specific sequence, which can differentiate \textit{TNXA} from \textit{TNXB} (Yang \textit{et al.}, 1999).

The recombinant chromosome in E1 is resulted from the pairing and crossover between \textit{TNXA} from a bimodular RCCX chromosome with the \textit{TNXB} from a monomodular RCCX chromosome. It can lead to the formation of \textit{TNXB/XA} hybrid with the deletion of the intervening \textit{RP2- C4B-CYP21B} genes. Individual containing this recombinant chromosome is a CAH carrier (Yang \textit{et al.}, 1999). The reciprocal product of the unequal crossover has \textit{TNXA/XB} with two \textit{CYP21B} genes. It was detected in the JRA family L (Rupert \textit{et al.}, 1999). Reversed associations of the informative sites were observed at the \textit{TNXB/XA} and \textit{TNXA/XB} recombinants.

From the study of E1, it was found that \textit{TNXB/XA} recombinant not only implies the abolition of the \textit{TNXB} function, but also indicates the deletion of the \textit{C4B} and \textit{CYP21B} genes, the latter of which may lead to the pathogenesis of CAH. Despite the importance, there has been no simple method to identify the recombination between \textit{TNXA} and \textit{TNXB}. Therefore, molecular genetic technique was developed to detect the presence of the \textit{TNXB/XA} recombinants with the deleterious 120 bp deletion in the population. This method can provide an efficient approach to screen for the possible carrier of the CAH disease in the clinical study.
3.3 EXPERIMENTAL PROCEDURES

*Oligonucleotides* -- Oligonucleotides were synthesized by Life Technologies (Grand Island, NY). For amplification of *TNXB* specific probe: *XB-MS5*, AGA TCA CAG CCT GGC AGT GAT GGG; and *XB-BK3*, AAA CCC ACA CAA GCT GGC TTG CTA.

*Isolation of Human Genomic DNA* — Genomic DNAs were isolated following standard protocols from peripheral blood of normal individuals, congenital adrenal hyperplasia (CAH) patients, pauciarticular juvenile rheumatic arthritis (JRA) patients, and IgA nephritis (IgA-N) patients. Appropriate consents from blood donors were obtained according to approved protocols by the Institutional Board of the Columbus Children's Hospital.

*DNA Probes* — *TNX-500* is a 500 bp fragment corresponding to exons 35 – 37 of *TNXA*. It is amplified using primers RDX-5 and RDX-3 from *cos 4A3*. *TNX-800* is an 800 bp *TNXB* specific fragment located immediately upstream of the *TNXA/XB* breakpoint. It was amplified from *cos 2* using primers XB-MS5 and XB-BK3.
Southern Blot Analysis — Ten micrograms of genomic DNA were digested to completion with $Msc$ I or $Taq$ I and resolved on 0.8% agarose gel. In a separate experiment, DNAs were subjected to $Msc$ I – $Pst$ I double digestion and resolved on 1.2% agarose gel. Afterward, they were blotted after standard procedures and hybridized with an appropriate $[\alpha^{-32}P]dCTP$-labeled $TNX$ probe.
3.4 RESULTS

Two nucleotide conversions at the TNXB informative sites lead to an Msc I RFLP

Among the nine TNX gene informative sites revealed from the previous study, two of them located at nucleotides 2268 and 2273, respectively, are present at less than 30 nucleotides away from the 5' of the 120 bp TNXB specific sequence (Yang et al., 1999). It was discovered that in TNXB, these two nucleotides are in the restriction enzyme Msc I recognition site: TGGCCA. However, in TNXA sequence, they are mutated to C and G respectively, and therefore abolished the Msc I site. Moreover, if these two conversions are conjugated with the 120 bp deletion (like the case in the CAH patient E1), the abolition of the Msc I site can further serve as an indicator of the RP2-C4B-CYP21B deletion.

Detailed sequence analysis of RCCX revealed two other Msc I sites adjacent to the one located at the TNX gene informative sites (Fig. 3.1). One is 3.1 kb upstream in TNXB gene, which is 800 bp upstream of the TNXA/XB breakpoint. The other site is 3.2 kb downstream in the CYP21B gene. Therefore, it was predicted that in a genomic Southern analysis using Msc I enzyme for the digestion and a TNXB specific DNA probe, the normal TNXB gene can be represented by a 3.1 kb genomic fragment. On the other hand, if the Msc I site at the informative sites was abolished due to the two nucleotide conversions, a 6.3 kb Msc I genomic fragment will be detected.
This Msc I RFLP was investigated in 19 CAH patients, 40 IgA-N patients, 41 JRA patients and 14 normal individuals. The results are summarized in Table 3.1. Data from nineteen individuals were presented in Fig. 3.2. Among them, lane 1 represents a normal individual with standard TNXB gene. Lanes 2 and 3 are from CAH family: parent E2 and patient E1; lanes 4 – 8 are from five other CAH patients. Lanes 9 and 10 are from JRA family L: patient L1 and grandparent L3 (Rupert et al., 1999); lane 11 is from another JRA patient. Lanes 12 – 17 are from six IgA-N patients. In lanes 18 and 19, two healthy controls were included.

In lanes 1, 9 and 10, the 3.1 kb Msc I fragment representing the normal TNXB gene was observed. In lanes 2, 3, 6 – 8, 11, 13 – 17 and 19, two fragments with the sizes of 3.1 kb and 6.3 kb were detected. The 6.3 kb band indicates the mutation of the Msc I site at nt. 2268 – 2273. These twelve individuals are heterozygous for this TNXB gene mutation. There are four other individuals who are homozygous for the same mutation since they only exhibit the 6.3 kb Msc I fragment (lanes 4, 5, 12, and 18).

As shown in Fig. 3.2 and summarized in Table 3.1, there is an unexpectedly high frequency of the TNXB gene associated 6.3 kb Msc I fragment in both patient and normal population. In CAH patient E1, it has already been shown that the presence of 6.3 kb Msc I fragment is accompanied by the 120 bp deletion in TNXB/XA recombinant (Yang et al., 1999). To determine whether the Msc I RFLP is a marker for the concurrent 120 bp deletion and the presence of the rearranged monomodular haplotype, the RCCX modular structures for two CAH patients, who have homozygous TNXB 6.3 kb Msc I fragment (Fig. 3.2, lanes 4 and 5), were analyzed.
Taq I RFLP (Fig. 3.3) suggests that both patients are heterozygous for biomodular RCCX: \(RP1-C4(L)-RP2-C4(L)\) and \(RP1-C4(L)-RP2-C4(S)\) (lanes 2 and 3). It is well known that CAH patients do not have normal CYP21 function (New, 1995; White et al., 1994). In these two cases, although the 3.7 kb fragment representing \(CYP21B\) is detectable, the 3.2 kb fragment corresponding to \(CYP21A\) is three times more intense, indicating the presence of two \(CYP21A\) genes on one chromosome. From this experiment, it is concluded that the \(Msc\) I RFLP is not necessarily coupled to the presence of a rearranged monomodular haplotype with \(TNXB/XA\) recombinant.

Identification of the \(TNXB/XA\) hybrid gene

In order to distinguish the \(TNXB/XA\) recombinant containing the concurrent \(Msc\) I RFLP and 120 bp deletion from the \(TNXB\) with only \(Msc\) I RFLP but no 120 bp deletion, an informative RFLP with \(Msc\) I – \(Pst\) I double digestion was specifically designed. Sequence analysis shows that there are two \(Pst\) I sites located 263 bp upstream and 561 bp downstream of the \(Msc\) I (2268-2273) site (Fig. 3.4). In a \(Msc\) I – \(Pst\) I double digested genomic Southern blot analysis using \(TNX-500\) as the probe, the normal \(TNXB\) gene is represented by two fragments of 561 bp and 263 bp, respectively. The \(TNXB\) with the \(Msc\) I site mutated can be revealed by a single 824 bp fragment. If there is the 120 bp deletion concurrent with the \(Msc\) I mutation (similar to the case in E1), the sequence becomes a \(TNXB/XA\) hybrid and the 824 bp fragment should no longer be detected. In this case, \(TNXA\)-like characters in between the two \(Pst\) I sites should be observed. The normal \(TNXA\) gene lacks the specific \(Msc\) I site but does contain the two
*Pst* I sites at the corresponding locations. Therefore, the *TNXA* gene or the *TNXB/XA* hybrid can be shown by a 704 bp fragment.

The nineteen individuals analyzed for the *Msc I* site mutation were again examined for the concurrent 120 bp deletion (*Fig. 3.5*). The individuals in *lanes* 1, 9, and 10 were previously shown to contain the normal *TNXB* sequence represented by the 3.1 kb *Msc I* fragment. This was confirmed by the presence of the 561 bp and 263 bp *TNXB* fragment in all three lanes. In addition, in *lanes* 1 and 10, the *TNXA* specific 704 bp fragment was also exhibited, and thus indicating that these two individuals have bimodular RCCX chromosomes. All other individuals were previously demonstrated to be either heterozygous or homozygous for the mutation at the *Msc I* site. However, the 824 bp *Msc I*–*Pst I* fragment corresponding to the mutation was missing in *lanes* 2, 3, 7, 14, and 17. It suggests that in these five individuals, the 120 bp *TNXB* specific sequence was concurrently deleted with the abolition of the *Msc I* site, and therefore, these individuals contain the *TNXB/XA* hybrid genes.
3.5 DISCUSSION

In this chapter, definitive RFLP has been designed for the detection of the Msc I site mutation in the *TNXB* gene. With another informative RFLP, it is also possible to conveniently identify the concurrent deletion of the 120 bp deletion in the *TNXB/XA* recombinant. In this study, five individuals were discovered to contain the abnormal *TNXB* genes. Among them, CAH patient E1 (*lane* 2 in *Figs.* 3.2 and 3.4) and the patient’s father (*lane* 3 of *Figs.* 3.2 and 3.4) have been proven to comprise an unusual chromosome rearrangement resulting in a *TNXB/XA* hybrid gene and the deletion of the *RP2-C4B-CYP21B* genes (Yang *et al.*, 1999). For the other three individuals newly identified in this study (*lanes* 7, 14 and 17), it is likely that they also possess the rearranged monomodular RCCX module with *RP2-C4B-CYP21B* genes deleted and a *TNXB/XA* hybrid gene. Consequently, these people become carriers of the CAH disease. If only the pseudogene *CYP21A* exists on the other chromosome 6, the person suffers CAH. This can be demonstrated by the established case in E1 and a newly discovered patient presented in *lane* 7. Two other examples shown in *lanes* 14 and 17 do not manifest CAH phenotype, as they contain a functional *CYP21B* gene from the other chromosome 6. However, these two do suffer from IgA nephritis. IgA-N has been suggested to be associated with C4B deficiency (Abe *et al.*, 1993; Welch *et al.*, 1989). It is probably worthy of noting that at least one *C4B* is also possibly deleted in one of their
chromosomes, which bears the TNX hybrid gene resulting from the unequal crossover. Whether this C4B deletion has a direct or indirect effect leading to the kidney disorder is yet to be determined. It has been suggested that malfunction of TNXB may be associated with a connective tissue disorder, Elner Danlo's syndrome (EDS) (Burch et al., 1997). Therefore, these five individuals may also be carriers of the EDS.

One of the two nucleotide conversions in TNXB, T2268→C, is a silent mutation, while the other one A2273→G can lead to a Q → R substitution. It is not clear whether the latter has any effect on the function of TNXB. As the abolition of the Msc I site caused by these changes was found not only in patients, but also in normal individuals, some of whom even are homozygous for the mutation, it appears that these two nucleotide changes do not cause any obviously identifiable disease. However, as it is already known that a pool of DNA mutations or allelic variants affects the expression/function of different genes. Individually, most mutations in the genome have mild, if not undetectable, effects, but in combination with "normal" alleles of other loci, they can produce a measurable phenotype (Vyse and Todd, 1996). Whether the single nucleotide polymorphisms (SNPs) here in TNXB gene could have any potential damage in the whole genomic context needs to be identified. Obviously, the elucidation of the TNX function becomes extremely critical for the understanding of the possible role in disease association.
Fig. 3.1  The *Msc I* map at *TNXB* and *CYP21B* gene region. The transcriptional orientations are shown by *horizontal arrows*. The *vertical arrows* point out the 3' end of the two genes. The duplication breakpoint of *TNXA/TNXB* is indicated by a *dashed line*. The 120 bp *TNXB* specific sequence is represented by a *triangle*. *Msc I* restriction sites are marked with the polymorphic site *outlined*. *Solid bar* represents the location of the probe used. The sizes of the *Msc I* restriction fragments are labeled below the map.
**Fig. 3.2** Southern blot analysis to detect the nucleotide mutations at the *Msc* I site (nt. 2268 – 2273) in *TNXB* gene. Genomic DNAs were digested to completion with *Msc* I and hybridized with the *TNXB* specific probe *TNX-800*. *Lane* 1, normal control with standard *TNXA* and *TNXB* gene; *lane* 2, father of CAH patient E1; *lane* 3, CAH patient E1; *lanes* 4 – 8, CAH patients; *lane* 9, JRA patient L1; *lane* 10, grandfather of L1; *lane* 11, JRA patient; *lanes* 12 – 17, IgA-N patients; *lanes* 18 and 19, normal individuals.
Fig. 3.3  Southern blot analysis to determine the RCCX modular structures of CAH patients. Genomic DNAs were digested with Taq I and hybridized with RP1.1, CYP21 and TNX-500 probes. Lane 1, normal control with known homozygous RP1-C4(L)-RP2-C4(L); lanes 2 and 3, CAH patients with homozygous 6.3 kb Msc I TNXB fragment; lane 4, CAH patient E1 with heterozygous 6.3 / 3.1 kb Msc I TNXB fragments.
Fig. 3.4 A schematic diagram to show variant TNX genes by Pst I – Msc I RFLP. The 120 bp TNXB specific region is indicated between two vertical arrows. The polymorphic Msc I site is outlined. Four different possibilities are shown with the sizes of the restriction fragments marked. XA/XB, TNXA/XB hybrid gene, gaining both Msc I site and the 120 bp sequence; XB (M⁻), TNXB gene losing the Msc I site; XB/XA, TNXB/XA hybrid gene, losing both Msc I site and the 120 bp sequence; XA (M⁺), TNXA gene gaining the Msc I site.
Fig. 3.5  Southern blot analysis to detect the concurrent 120 bp deletion and Msc I RFLP in TNXB genes. Genomic DNAs were subjected to Msc I - Pst I double digestion and hybridized with the TNX-500 probe. DNA samples are in the same order as described in Fig. 3.3.
<table>
<thead>
<tr>
<th>Disease</th>
<th>Number of Patients</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
<td>heterozygous</td>
<td>homozygous</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.3/3.1</td>
<td>6.3/6.3</td>
</tr>
<tr>
<td>CAH</td>
<td>19</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>IgA-N</td>
<td>40</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>JRA</td>
<td>41</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>normal</td>
<td>14</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

*Table 3.1* Detection of the 6.3 kb Msc I RFLP in TNXB genes
FOUR UBIQUITOUSLY EXPRESSED GENES, RD-SKI2W-DOM3Z-RP1, ARE PRESENT IN THE CLASS III REGION OF THE MHC

4.1 ABSTRACT

The association of the MHC class III region with many diseases motivates the investigation of unidentified genes in the 30 kb segment between complement component genes Bf and C4. RD, which codes for a putative RNA binding protein, is 205 bp downstream of Bf. SKI2W, a DEVH-box gene probably involved in RNA turnover, is 171 bp downstream of RD. RP1 is located 611 bp upstream of C4. The DNA sequence between human RD and RP1 was determined and the exon-intron structure of SKI2W elucidated. SKI2W consists of 28 exons. The putative RNA helicase domain of Ski2w is encoded by nine exons. Further analysis of the 2.5 kb intergenic sequence between SKI2W and RP1 led to the discovery of DOM3Z. The full-length cDNA sequence of DOM3Z encodes 396 amino acids with a leucine zipper motif. Dom3z-related proteins are present in simple
and complex eukaryotes. In *C. elegans*, Dom3z-related protein could be involved in the development of germ cells. Human *RD-SKI2W* and *DOM3Z-RPL1* are arranged as two head-to-head orientated gene pairs with unmethylated CpG sequences at the common 5' regulatory region of each gene pair. The ubiquitous expression pattern suggested that these four genes are probably housekeeping genes.
4.2 INTRODUCTION

Many autoimmune, genetic and multifactorial disorders are associated with the
dynamic genomic region of chromosome 6 known as the MHC (Porter, 1985b; Tiwari
and Terasaki, 1985). As one of the most gene-dense regions in the human genome,
identification of genes present in the MHC class III region and elucidation of their
function are of profound interest. On the average, this region contains a gene in every
15 kb of genomic DNA (Aguado and Campbell, 1996) and the intergenic regions
between adjacent genes are minimal in size. For examples, the complement
compponent C2 gene and factor B gene are separated by only 421 bp (Wu et al., 1987),
while the extracellular matrix protein tenasin-X (TNX) gene and the cytochrome P450
21-hydroxylase gene (CYP21) overlap at their 3' ends (Morel et al., 1989). In the 30
kb genomic region between factor B (Bf) and complement C4, others have cloned RD
downstream of the Bf gene (Levi-Strauss et al., 1988); we have demonstrated the
presence of SKI2W (Dangel et al., 1995a) and RP1 (Shen et al., 1994) upstream of C4.
RP, complement C4, steroid CYP21, and tenascin TNX are duplicated together as
genetic unit termed the Rx module in 85.6% of the Caucasian population (Shen et
al., 1994; Yang, Z., Mendoza, A.R., Blanchong, C.A., Rennebohm, R. and Yu, C.Y., in
preparation).
The \textit{RD} gene is 6.7 kb in size, consists of eleven exons and is organized in a tail-to-tail configuration with \textit{Bf}. The intergenic distance between the 3' ends of \textit{RD} and \textit{Bf} is 205 nucleotides (Rowen \textit{et al.}, 1997b). The putative RD protein has 380 amino acids (Cheng \textit{et al.}, 1993; Speiser and White, 1989; Surowy \textit{et al.}, 1990). It contains the RNA recognition motif (RRM) of the ribonucleoprotein (RNP) family that is also present in poly(A) binding proteins (Hornstein \textit{et al.}, 1997) and in nucleolysin (Kawakami \textit{et al.}, 1992). However, the most striking feature of RD is the presence of twenty-four copies of alternate, positively and negatively charged residues, Arg-Asp (RD) (Levi-Strauss \textit{et al.}, 1988), which may be related to RNA binding. The cellular function of RD has not been determined.

Human \textit{SKI2W} codes for a transcript of 4 kb in size that encodes a polypeptide of 1,246 amino acids. The Ski2w protein contains an RNA helicase domain with a DEVH-box, two leucine zipper motifs that may be involved in protein-protein interaction, and an RGD motif that may be a ligand for cell adhesion molecules. ATPase activities were demonstrated from fusion proteins containing human Ski2w (Dangel \textit{et al.}, 1995a). Hydrolysis of ATP is required as an energy source for helicases to unwind helical structures of nucleic acids (Fuller-Pace, 1994). By indirect immunofluorescent experiments using Ski2w specific antibodies, it has been shown that Ski2w is present in the nucleolus and also in the cytoplasm (Qu \textit{et al.}, 1998; Lee \textit{et al.}, 1995). Immunoblot analysis further revealed that in the cytoplasm of HeLa cell extracts, Ski2w is likely associated with ribosomes and polyribosomes (Qu \textit{et al.}, 1998). The human Ski2w protein shares striking and extensive sequence similarity to
the yeast antiviral protein Ski2p. The yeast Ski2p is involved in antiviral defense probably through its general role on the RNA turnover or the regulation of RNA degradation (Jacobs et al., 1998; Masison et al., 1995; Widner and Wickner, 1993). The expression pattern and the localization to the nucleolus and polysomes would suggest that the human Ski2w shares similar functional properties with its yeast homolog. The yeast SKI2 gene is intronless but the human SKI2W gene is split into multiple exons by introns. Elucidation of the human SKI2W gene structure is necessary for molecular genetic studies to investigate the mutation and polymorphism of this gene and its disease associations.

The RPI gene is about 11 kb in size, consisted of nine exons, and located 611 bp upstream of C4 (Shen et al., 1994). There is a bipartite nuclear localization signal close to the amino terminal of the protein sequence, suggesting RPI could be a nuclear protein. The primary structure of RPI does not show significant similarities to any known protein. Two different cDNAs, RPI and GII, have been reported. RPI extends the reading frame of GII at the 5' region. The larger isoform, RPI, encodes 364 amino acids (Shen et al., 1994). The smaller isoform, GII, encodes 254 or 258 amino acids (Sargent et al., 1994).

To study the genetics and functions of the MHC class III genes, the DNA sequence between RD and RPI was completely determined. From this sequence the exon-intron structure of the SKI2W gene is elucidated. Analysis of the 2.5 kb intergenic sequence between SKI2W and RPI led to the discovery of DOM3Z, which has related genes in yeasts, a flowering plant and a nematode. RD-SKI2W and DOM3Z-RPI are arranged
as two head-to-head orientated gene pairs separated by only 59 nucleotides. The ubiquitous gene expression and structural features suggest that these four genes are probably housekeeping genes.
4.3 EXPERIMENTAL PROCEDURES

Sequence determination of the human SKI2W gene — Cos 3A3, which contains the human C4A3 gene and about 20 kb of genomic DNA upstream of the C4 gene (Carroll et al., 1984; Yu, 1991), was digested with BamHI and subcloned into pBluescript vectors. Subclones containing SKI2W genomic DNAs were screened with SKI2W cDNA probes (Dangel et al., 1995a). Four different subclones containing 3.0 kb, 3.0 kb, 2.2 kb, and 1.8 kb inserts were obtained. These clones were subjected to Exo III digestion using the Erase-a-Base kit from Promega (Madison, WI). Using cos la DNA that overlaps with cos 3A3 by 6.5 kb at the 5' region (Carroll et al., 1984) as template, a 1.4 kb fragment between RD and SKI2W was generated by PCR and cloned into pCRII vector (Invitrogen, Carlsbad, CA).

For the PCR, one of the primers, RD5UT 5'-GCA GCG ATA T TC ACG CTC TC-3', is located at the 5' untranslated region of the RD gene (Speiser and White, 1989), while the other primer from SKI2W, GW5 5'-TGT AAC CCA GTA TCT GGC CT-3', was derived from this study. Sanger's dideoxy sequence reactions were performed using ^35S-ATP and Sequenase (United State Biochemicals, Cleveland, OH). Gel readings were compiled by PC/GENE software (Intelligenetics, Mountain View, CA). Sequence analysis was performed by GCG software (Genetics Computer Group, 1991) from Pittsburgh Supercomputing Center.
Sequencing of EST clones for human DOM3Z — The EST clone ID# 271616 was purchased from Genome System Inc (St. Louis, Missouri). The cDNA insert is present between the Not I and EcoR I sites in a modified pT7T3 vector. The EST clone HBC4394 was kindly provided by Dr. J. Takeda (Gunma University, Japan). The cDNA insert is located between the EcoR I and Xho I sites of pBluescript vector. Both clones were sequenced using a thermal sequenase sequencing kit (Amersham Pharmacia Biotech, Arlington Heights, IL). The primers T7 and T3 were used to sequence across the cloning sites. Internal primers used for sequencing are $f_0, f_1, f_2, f_3, f_4, r_0, r_1, r_3,$ and $r_4$. Oligonucleotides were purchased from Life Technologies (Grand Island, NY). Sequences of the primers used in this study are listed below. Sequences added to the primers to facilitate cloning are in lower case.

- $f_0$: 5' aga gga tcc GAG ACA GAG GTA GCT 3'
- $f_1$: 5' CGA GCC CTG CGC TAG TAT AG 3'
- $f_2$: 5' aga gga tcc GGG CAC CTG ACA AAA CTG 3'
- $f_3$: 5' GAG CTT ATG TAC ATG GGA TAC 3'
- $f_4$: 5' GAA GAT GTT TGA ATA TGT GAG G 3'
- $r_0$: 5' aga gga tcc AAC CTC TTA GAA ATA TGC TTT 3'
- $r_1$: 5' GTC TCC ACT TCA CTC AGG TAT 3'
- $r_2$: 5' CGT CAG CAG TTT TGT CAG GTG 3'
- $r_3$: 5' GCA GGT GGT CCA GCC TTT CCT 3'
- $r_4$: 5' TCC CAA GAG AAG AGA TGA ACG AG 3'

96
**Determination of 5' cDNA sequences for human DOM3Z by 5' RACE** — Total RNA was isolated from HeLa cells using RNAzol™ B (Tel-Test Inc, Friendswood, TX) following the manufacturer's instruction. To obtain the 5' end of the DOM3Z cDNA, 5' RACE was performed using a kit from Life Technologies. Briefly, the first strand cDNA was synthesized from total HeLa RNA using the r1 primer. A homopolymeric tail was added to the 3' end. PCR amplification was performed using the dC-tailed cDNA, the nested r2 primer, and an anchor primer. PCR products were cloned into the pCR2.1 vector (Invitrogen). A 0.8 kb BamH I DNA fragment corresponding to the genomic region upstream of the clone 271616 sequence was used as a probe for colony hybridization. Clones containing 5' RACE products were sequenced by automated sequencing (ABI model 377, Foster City, CA) using T7 primer and M13 reverse primer.

**Sequence analyses** — The DNA sequence for the intergenic region between RP1 and SKI2W was submitted to NCBI (National Center for Biotechnology Information). BLASTN program (Altschul et al., 1990) was used to search the EST database for any related cDNA clones. Sequences of the DOM3Z EST clones and RACE products were analyzed using GCG software. Overlapping sequences were compiled to obtain the full length DOM3Z cDNA. Dom3z amino acid sequence was deduced and used to search for related or homologous proteins using the BLASTP program (Altschul et al., 1990). The Dom3z related protein sequences were aligned using PILEUP and PRETTY programs of the GCG package. The presence of structural motifs in the human Dom3z protein sequence was searched against PROSITE database (Hofmann et al., 1999). To define the
gene structure for DOM3Z, FASTA program was used to compare the EST and RACE sequences with DOM3Z genomic DNA sequences, which correspond to genomic sequences between RP1 (Shen et al., 1994) and SKI2W (this work).

**Pulsed-field gel electrophoresis (PFGE)** — Human lymphoblastoid cells were embedded in low gelling temperature agarose, digested with BssH II enzyme, resolved with Statagene rotaphor pulsed field gel machine, and blotted onto Hybond N membrane, as described previously (Yu et al., 1993). The blot was hybridized to a complement C2 probe, PC201 (D'Eustachio et al., 1986), a DOM3Z/RP1 5' probe (0.8 kb BamHI genomic fragment), and a 1.1 kb RP1 cDNA probe, consecutively.

**Southern blot analysis** — Cos M1A1 and cos M1B2 were screened from a MOLT4 cosmid library (Yu et al., 1993) using a 2.2 kb SKI2W cDNA probe (Dangel et al., 1995a). Cos KEM1 (Carroll et al., 1985a) and cos 1a were previously described (Carroll et al., 1984). Cosmid DNAs were digested with BamH I, processed after Southern's protocol (Southern, 1975) and hybridized to a 1.7 kb SKI2W cDNA probe.

**Northern blot analysis** — Multiple tissue Northern blots (MTN blots I and II, Clontech, Palo Alto, CA) were hybridized with human RD, DOM3Z and RP1 probes, separately. The RD probe was 1.4 kb full length cDNA generated by PCR. The DOM3Z probe was derived from the cDNA insert of EST clone 271616. The RP1 probe is a 1.1 kb cDNA fragment (Shen et al., 1994).
4.4 RESULTS

Gene structure and polymorphism of human SKI2W

The complete DNA sequence for the human SKI2W gene has been determined by sequencing genomic DNA from cos 3A3 and cos 1a (Carroll et al., 1984). The exon-intron structure of SKI2W was deduced by comparing genomic and cDNA sequences. The human SKI2W gene spans 11 kb and contains 28 exons (Fig. 4.1 A). The intergenic distance between SKI2W and the upstream RD gene is 171 bp. RD and SKI2W are arranged in a head-to-head configuration.

The helicase domain with the putative double leucine zipper motifs is encoded by exon 10 to exon 18. Apart from helicase boxes V and VI that are encoded by exon 18, each of the other helicase boxes is mainly encoded by a separate exon. The helicase box I is coded by exon 10. The helicase box Ia with leucine zipper motif I is coded by exon 11. The helicase box II with leucine zipper motif II is coded by exon 12. Helicase boxes III and IV are encoded by exons 13 and 14, respectively. The three putative N-linked glycosylation sites are encoded by exons 5, 14 and 22, respectively. The acidic motif is encoded by exon 6. The RGD cell adhesion motif is coded by exon 8. The two consecutive sulfation sites are encoded by exon 20. Two Alu elements are present in the SKI2W gene. Alu-C, flanked by aactatat direct repeats, is located in intron 17; Alu-J, in
reverse orientation with respect to the SKI2W gene, is present in intron 18. This Alu element is flanked by cctc direct repeats.

A comparison of the genomic and cDNA sequences of SKI2W revealed three nucleotide variations. The first is a G to A transition at codon 151 (exon 5) that leads to a Gin to Arg substitution. The second is a C to T transition at codon 1052 (exon 25) that leads to a Leu to Phe substitution. Whether these substitutions affect the Ski2w protein function remains to be determined. The third change is a T to C transition at codon 1,067 (exon 26) that is a silent mutation.

Southern blot analysis of cosmid clones of human SKI2W gene isolated from different genomic DNA libraries revealed that SKI2W gene is polymorphic. A BamH I restriction fragment length polymorphism (RFLP) was detected when a 1.7 kb SKI2W cDNA probe was used for hybridization. As shown in Fig. 4.1 B, besides a common 1.8 kb fragment that is present in all lanes, a 5.2 kb fragment was observed in cos M1A1 (lane 1) and in cos M1B2 (lane 2). However, a 2.2 kb fragment was detected in cos KEM1 (lane 3) (Carroll et al., 1985) and in cos 1a (lane 4). This polymorphic BamH I site has been mapped to the intron 18 of the SKI2W gene, where the Alu-J element is located.

Identification of a new gene, DOM3Z, located in between SKI2W and RP1

To investigate whether the intergenic region between SKI2W and RP1 contains an unidentified gene, the 2.5 kb genomic sequence was submitted to NCBI to search the EST database. Two cDNA clones, 271616 and HBC4394, were analyzed. The
published EST sequence is 359 bp for 271616 (Hillier et al., 1996), and 270 bp for HBC4394 (Takeda, 1996). Subsequently, both clones were sequenced to completion. For clone 271616, the cDNA insert is 1,227 bp in size (nucleotide number 160 - 1386, Fig. 4.2), and contains a poly(A) tail sequence of 28 nucleotides. For clone HBC4394, the cDNA insert is 1,315 bp in size (nucleotide number 60 - 1374). It extends clone 271616 cDNA sequence by 100 bp at the 5' end, but is 12 bp shorter at the 3' end and no poly(A) tail is present. Both clones contain the poly(A) signal at nucleotide number 1362 - 1367.

The cDNA inserts in the two EST clones contain a reading frame, with a stop codon TAG at the nucleotide number 1318 - 1320. To obtain the full-length cDNA sequence and to determine the open reading frame of the new gene DOM3Z, 5' RACE was performed using total RNA isolated from HeLa cells. The longest RACE product extends the 5' end of the known cDNA sequence by 59 bp. Therefore, the composite, full-length DOM3Z cDNA is 1386 bp in size.

Amino acid sequence encoded by DOM3Z cDNAs

Dom3z amino acid sequence was deduced from the full-length cDNA (Fig. 4.2). Because of the in-frame stop codon at nucleotide 19-21, the ATG at the nucleotide 130 - 132 was assigned as the initiation codon. The Dom3z protein contains 396 amino acids, of which 11.6% are proline residues. The predicted isoelectric point for Dom3z is 6.82. The amino acid sequence was used to search for structural motifs against PROSITE database. A leucine zipper motif is located from amino acids 103 to 124. Four potential
sites for N-myristoylation are located at amino acids 5, 9, 264, and 358, respectively.

There are four potential casein kinase II phosphorylation sites at positions 141, 163, 202, and 256, respectively. At amino acid number 47, there is a potential cAMP- and cGMP-dependent protein kinase phosphorylation site. In addition, six potential protein kinase C phosphorylation sites are present at amino acids 6, 130, 256, 302, 308, and 394.

Three nucleotide changes were detected between the two DOM3Z cDNA clones. The first variation is at nucleotide 211, which is an A to T transition that leads to the substitution from Thr to Ser at amino acid number 28. Another variation from His to Gln at amino acid number 261 is caused by the C to A transition at nucleotide 912. The third nucleotide variation, A→G at position 873, is a silent mutation.

*Dom3z-related proteins are present in simple and complex eukaryotes*

BLASTP algorithm was used to search for related proteins of Dom3z in the national databases. Hypothetical proteins in *Schizosaccharomyces pombe* (GenBank accession no. 2440182), *Saccharomyces cerevisiae* (accession no. 1723981), *Caenorhabditis elegans* (Dom3, accession no. 1706489), and *Arabdiopsis thaliana* (accession no. 2245120) all share significant sequence similarities and identities with human Dom3z (Table 4.1). Apart from the case in *Arabdiopsis thaliana* which has a much longer reading frame of 1,148 amino acids, the other Dom3z related proteins are similar in sizes with 393 amino acids for *C. elegans*, 387 amino acids for *S. cerevisiae*, and 352 amino acids for *S. pombe*. The overall sequence identities/similarities vary from 21.37% / 45.96% between the human and baker's yeast Dom3z related proteins, to
30.3% / 53.33% between those of baker's yeast and fission yeast. It is worthwhile to mention that the sequence similarity of the human Dom3z protein to the related protein in fission yeast (52.43%) is as close as that between the two yeast strains (53.33%).

An alignment of the five Dom3z related protein sequences is shown in Fig. 4.3. A consensus is deduced from conserved sequences from three or more species. Conserved sequences among all five sequences are highlighted in red and shaded in yellow. They probably represent relevant motifs important for protein structures and/or functions. The most notable conserved sequences are YWGYKFE at positions 231-237, EEYCSVVR at positions 266-273, GEVDC at positions 295-299, YVELKT at positions 334-339, and IIVGFRDD at positions 373-380.

The exon-intron structure of human DOM3Z and the 5' region of DOM3Z/RP1

The human DOM3Z gene is present between SKI2W and RP1. The exon-intron structure of DOM3Z was deduced by comparing DOM3Z cDNA sequence with the genomic sequences at the 3' region of SKI2W (this work) and the 5' region of RP1 (Shen et al., 1994). The DOM3Z gene contains 7 exons (Fig. 4.4). The first exon is 123 bp in size. It has an in-frame stop codon at nucleotide 19-21 and therefore is untranslated. The second exon contains six nucleotides of the remaining 5' untranslated sequence, and the coding sequence for the first 119 amino acids of the Dom3z protein. The putative leucine zipper motif is encoded by exons 2 and 3. The seventh exon encodes for the last 49 amino acids, two consecutive stop codons and an additional 56 nucleotides of the 3' untranslated region. Stretches of T-rich sequences are present downstream of the
poly(A) site, which is consistent with the 3' ends of mammalian genes (Keller, 1995; Proudfoot, 1991). Downstream of the DOM3Z gene is the last (twenty-eighth) exon of SKI2W. These two genes are arranged in tail-to-tail configuration and the intergenic region between them is only 59 bp.

Located at the 5' region of the DOM3Z gene is the RPl gene. DOM3Z and RPl are arranged in head-to-head configuration. The 5' end of GII, which is a smaller form for RPl, is 322 bp away from the 5' end of DOM3Z. The 5' end for the larger form of RPl is located at the genomic region corresponding to intron 2 of DOM3Z. Therefore, the 5' regulatory regions of DOM3Z and RPl probably overlap to some extent. There are no consensus TATA boxes but four SP1 and four AP1 sites for bindings of ubiquitous gene expression factors at the 5' regions for DOM3Z and RPl (Quandt et al., 1995).

The DNA sequence for the 5' region of DOM3Z and RPl is relatively rich in G+C content, with 61.3% of the nucleotides being G+C, in contrast to 41% normally present in a mammalian genome (Normore et al., 1976). In addition, there are numerous copies of CG dinucleotides, which are generally under-represented in the mammalian genome. Hypomethylation of the CG sequences is an indication of active transcription (Bird, 1986, 1987). Methylation of CG sequence may block the cleavage of some restriction enzymes that have CG dinucleotides in their recognition sequences (McClelland et al., 1994). A BssH II recognition site with two CG sequences, GCGCGC, is present 575 nucleotides upstream of the DOM3Z (Fig. 4.4). Restriction mapping of the genomic DNA between RD and C4 reveals that there are two consecutive BssH II sites at the
intergenic region between SKI2W and RD, which is 13.6 kb away from the BssH II site in the DOM3Z/RP1 complex. To investigate if these BssH II restriction sites are unmethylated and susceptible to enzyme cleavage, genomic DNAs isolated from two human lymphoblastoid cell lines with RCCX monomodular structures were digested with BssH II, and resolved by PFGE for high molecular weight DNA. A 65 kb fragment was detected when the complement C2 probe (pC201) was employed (Fig. 4.5, lanes 1 and 2); a ~14 kb fragment was detected when the DOM3Z/RP1 probe (B0.8) was used (lanes 3 and 4); and a 70 kb fragment was detected when an RP1 3' probe (RP1.1) was applied (lanes 5 and 6). Detection of the 14 kb BssH II fragment suggests that the BssH II sites between the RP1/DOM3Z genes and between the RD/SKI2W genes are probably unmethylated. The detection of the ~70 kb fragments when C2 or RP1 3' probes were used suggested that the BssH II sites in the C2 gene and in the CYP2I gene were methylated (and therefore resistant to BssH II cleavage) in the lymphoblastoid cells.

RD, SKI2W, DOM3Z and RP1 transcripts are ubiquitously expressed in human tissues

To study the tissue distribution patterns for transcripts of RD, SKI2W, DOM3Z and RP1, multiple tissue Northern blots were hybridized with appropriated probes (Fig. 4.6).

For RD (panel A), 1.4 kb transcripts were detected in all sixteen tissue samples. Larger transcripts about 2.0 kb in size are also detectable. The testis (lane 4), heart (lane 9), placenta (lane 11) and pancreas (lane 16) expressed large quantities of RD
transcripts. The liver (lane 13) produced RD transcripts at a relatively lower level similar to that of the lung (lane 12).

It has been shown that SKI2W transcripts of ~4.0 in size were detectable in all human cell lines and tissue samples tested (Dangel et al., 1995a, Qu et al., 1998). As shown in panel B, SKI2W is expressed in all nineteen tissue samples examined, which is consistent with the suggestion that SKI2W is ubiquitously expressed (Qu et al., 1998). In addition to the 4.0 kb transcripts, slightly larger transcripts ~4.2 kb in size were present in some of the tissue samples, including thymus, ovary, small intestine, colon, brain, placenta, liver, kidney and pancreas.

The size of the DOM3Z transcript is predominantly 1.5 kb but may vary from 1.4 kb to 2.5 kb. In the testis, liver and pancreas, smaller transcripts about 1 kb in size were also observed. In addition, transcripts about 4 kb, 5 kb and 7 kb were also detectable in the liver. The expression level for DOM3Z is high in the testis and pancreas, and low in the lung.

For RP1, the predominant transcript is 1.4 kb in size that was detectable in all tissue RNAs. In addition, larger transcripts of 3 kb and 9 kb were present in many tissues. Similar to the pattern observed for DOM3Z, the testis and pancreas exhibit higher expression levels of RP1, while low level of expression was observed in the lung. Detection of multiple transcripts in most tissues was a common feature for the DOM3Z-RP1 gene pair.
4.5 DISCUSSION

Four ubiquitously expressed genes, \(RD\), \(SKI2W\), \(DOM3Z\) and \(RP1\), are present within the 30 kb genomic region between complement components factor \(B\) and \(C4\) (Fig. 4.7). These four genes are arranged in two tandem gene pairs, each pair in a head-to-head configuration. The extremely high gene density of the MHC class III region can be reflected by the very short intergenic distances, e.g., 171 bp between \(RD\) and \(SKI2W\), 59 bp between \(SKI2W\) and \(DOM3Z\), and the overlap of the 5' regions of \(DOM3Z\) and \(RP1\), although these genes are arranged in opposite transcriptional orientations (Fig. 4.7 A). No TATA boxes are present but there are multiple SP1 and AP1 sites at the intergenic and the 5' regulatory regions between \(RD\) and \(SKI2W\), and between \(DOM3Z\) and \(RP1\), which are both relatively GC-rich with multiple copies of CG sequences. The CG sequences at the promoter regions of these gene pairs are probably unmethylated or hypomethylated, because the \(B_{ssH}\) II sites at these regions were cleavable to yield a 14 kb restriction fragment. These are hallmarks of ubiquitously expressed genes. Indeed, Northern blot analysis showed that transcripts for all of these four genes are present in all tissues tested. It would be of interest to investigate whether the expression of these four genes are coordinated or controlled by similar trans-acting transcriptional factors. Investigation on the bi-directionality of the 5' regulatory regions for \(RD-SKI2W\), and for \(DOM3Z-RP1\) would yield relevant information on the control of expression of these genes.
Pancreas is a tissue where high levels of expression are observed for all four genes. Another major site of expression is the testis, where RD, DOM3Z, and RP1 showed extraordinary high levels of expression. Transcripts from the liver for SKI2W, DOM3Z, and RP1 all contain higher molecular weight isoforms. Partially spliced and multiple isoforms of DOM3Z transcripts are detectable in Northern blot analysis (Fig. 4.6 B) and in cDNA clones (Yang and Yu, 1999). The expression pattern for RP1 and DOM3Z are very similar, with relatively higher levels in the testis, ovary and pancreas, and lower levels in the lung and peripheral blood lymphocytes (Fig. 4.6 B, C). It is also worthwhile to note that a targeted disruption of Bf in mouse resulted in the loss of expression of two flanking genes, which are the upstream C2 gene and the downstream RD gene (Taylor et al., 1998b). In the mouse MHC, C2 and Bf are separated by 641 bp, while the 3’ ends of Bf and RD overlap by 2 bp (Rowen et al., 1998; Yang and Yu, 1999). The very close gene placements in the MHC class III region would imply the presence of regulatory sequences of a gene in its neighboring genes. Therefore it is conceivable that the disruption of a gene in the MHC complement gene cluster would affect the expression of its neighbors.

Only the RP1 has no related proteins discovered in lower eukaryotes. Homologous or related proteins for RD, Ski2w, Dom3z are present from simple eukaryotes such as yeasts to highly advanced species such as humans. Dom3z related protein is also present in a flowering plant (A. thaliana). The conservation of these proteins implies that they may have very fundamental function in a biochemical pathway. Ski2w contains an RNA helicase domain and is associated with ribosomes (Qu et al., 1998). RD contains RNA binding motifs and 24 copies of Arg-Asp (RD). This gene
pair appears encoding proteins both involved in RNA metabolism.

Determination of the complete genomic sequence of the human SKI2W gene provides the foundation for future genetic studies on the polymorphism and mutation of the SKI2W gene. Unlike the yeast SKI2 gene that contains no introns, the human SKI2W gene consists of 28 exons. The helicase domain is encoded by nine exons. The coding of a functional protein domain by multiple exons is also exemplified by genes coding for serine proteinases. For example, the serine proteinase domain of factor B is encoded by eight exons (Campbell et al., 1984).

Although the function of human DOM3Z is unknown, related genes for DOM3Z are present in both simple and complex eukaryotes. In the nematode C. elegans, DOM3 is located downstream of the MES-3 gene. The MES-3 gene encodes for maternal effect component required for normal postembryonic development of the germ line. Since MES-3 and DOM3 belong to the same operon in C. elegans, their gene products may be functionally related. It would be of interest to determine if human Dom3z plays a role in growth and reproduction, particularly on the development of germ cells. The relatively high levels of transcripts detected in the testis and ovary are consistent with this notion.

The protein sequences for RD, Ski2w and Dom3z all contain leucine zipper motifs involved in protein interactions. The tight linkage of RD, SKI2W, DOM3Z and RP, the ubiquitous gene expression pattern, and the potential to bind RNAs (for RD and Ski2w) imply that these proteins may have concerted functions, probably related to RNA translation and turnover. Genetic and functional studies of these four ubiquitously
expressed genes may help elucidate the molecular basis of disease associations with the MHC complement gene cluster.
**Fig. 4.1** Exon-intron structure and *BamH I* RFLP of the human *SKI2W* gene.

A. Exon-intron structure and *BamH I* restriction map. Exons of *SKI2W* are in *filled* boxes. Exons coding for untranslated sequences are in *empty* boxes. *Vertical* arrows show the 3' end of genes. *Horizontal* arrows represent transcriptional start sites and orientations.

*BamH I* restriction sites and the sizes of restriction fragments are marked beneath exons. The polymorphic *BamH I* site is marked with an *asterisk*. B. Southern blot analysis of *BamH I* digested cosmids DNAs M1A1 (*lane* 1), M1B2 (*lane* 2), KEM-1 (*lane* 3) and 1a (*lane* 4). The probe used for hybridization is a 1.7 kb *SKI2W* cDNA fragment.
**Fig. 4.1** Exon-intron structure and *BamH* I RFLP of the human *SKI2W* gene.
(Fig. 4.1 continued)

B.

<table>
<thead>
<tr>
<th>kb</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Fig. 4.2  cDNA and amino acid sequences of human DOM3Z.** The first 59 bp of the cDNA was obtained by 5' RACE. The remaining sequence was derived from EST clones HBC 4394 and 271616. The initiation codon ATG and the poly(A) signal are underlined. Asterisks indicate the positions of the stop codons. The leucine zipper motif is indicated with the Leu residues highlighted. The polymorphic sites (nt 211, 873, and 912) are shown with the polymorphic nucleotides above the nucleotide sequence, and the resulted amino acid changes below the amino acid sequence. (GenBank accession number: AF059252).
Fig. 4.2 eDNA and amino acid sequences of human DOM3Z.
Fig. 4.3 An alignment of the amino acid sequences for Dom3z-related proteins from human (Hosa, accession no. AF059252), nematode (Cael, accession no. 1706489), fission yeast (Scpo, accession no. 2440182), baker's yeast (Sace, accession no. 1723981), and Arabdiopsis (Arth, accession no. 2245120). Gaps inserted into sequences are shown as dots; residues conserved among three or more species are in upper case and shown in the consensus. Conserved residues among all species are in red fonts and shaded in yellow. Unconserved residues are in lower case and shown as hyphens in the consensus. Asterisks indicate the positions of stop codons.
Fig. 4.3 An alignment of the amino acid sequences for Dom3z-related proteins.
Fig. 4.4  **Exon-intron structure of the human DOM3Z gene.** Exons coding for protein sequences are in *shaded* boxes. Exons coding for 5' or 3' untranslated regions are in *empty* boxes. Transcriptional orientations are marked by *horizontal* arrows. An *inverted, vertical* arrow indicates the position of the 3' end of the *SKI2W* gene. The transcriptional orientations of *RP1/G11* and *SKI2W* are opposite to that of *DOM3Z.*
Fig. 4.5 **PFGE analysis of the **RD-SKI2W-DOM3Z-RP1 complex**. Genomic DNAs from two individuals with homozygous, monomodular RCCX structure were digested with BssH II, resolved by PFGE and processed after Southern's procedure. Probes used are: *lanes* 1 and 2, complement C2 (pC201); *lanes* 3 and 4, 0.8 kb BamH I fragment corresponding to the 5' regions of the RP1 and DOM3Z gene (B0.8); *lanes* 5 and 6, RP1 cDNA corresponding to the 3' region of the RP1 gene (RP1.1).
Fig. 4.6 Northern blot analyses of human multiple tissue RNAs hybridized to (A) RD, (B) SKI2W, (C) DOM3Z and (D) RPI probes.
Fig. 4.7 A molecular map of the HLA class III region with (A) RP, complement C4, steroid CYP21, and tenascin TNX (RCCX) bimodular structure and (B) RCCX monomodular structure. Transcriptional orientations are indicated by horizontal arrows; pseudogenes or partially duplicated gene segments are shaded.
<table>
<thead>
<tr>
<th></th>
<th>S. pombe</th>
<th>C. elegans</th>
<th>A. thaliana</th>
<th>S. cerevisiae</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. sapiens</td>
<td>30.10/52.43</td>
<td>28.85/53.50</td>
<td>28.19/50.00</td>
<td>21.37/45.96</td>
</tr>
<tr>
<td>S. pombe</td>
<td>-</td>
<td>25.63/47.15</td>
<td>29.37/58.33</td>
<td>30.30/53.33</td>
</tr>
<tr>
<td>C. elegans</td>
<td>-</td>
<td>-</td>
<td>24.26/47.87</td>
<td>23.88/48.06</td>
</tr>
<tr>
<td>A. thaliana</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25.89/46.45</td>
</tr>
</tbody>
</table>

*Table 4.1*  % Sequence Identities / Similarities among Dom3z Related Proteins
CHAPTER 5

BIOCHEMICAL CHARACTERIZATION OF HUMAN Dom3z

5.1 ABSTRACT

Human DOM3Z is the most recently identified member of the RD-SKI2W-DOM3Z-RPL gene cluster located between complement Bf and C4 genes in the MHC class III region. It has related proteins in lower eukaryotes including baker's yeast, fission yeast, a nematode, and even a flowering plant. The function of human Dom3z has not been determined. Preliminary studies have been conducted on expressing and purifying bacterial fusion proteins, generating polyclonal antibodies and establishing long-term mammalian transfectant cell lines. It has been shown by indirect immunofluorescence using Dom3z antisera, and by green fluorescent protein technique, that human Dom3z is mainly a nuclear protein. In immunoblot analysis using the polyclonal antibodies against the Dom3z fusion proteins, a polypeptide with the apparent size about 67 kDa was detected in several tissue samples and cell lysates of culture human cell lines.
5.2 INTRODUCTION

As described in the previous chapter, a novel human gene DOM3Z was discovered by combining the conventional bench work with the state-of-art bioinformatics database searching. Part of the credit in this gene identification process should be given to the Human Genome Project, especially the EST databases. The DOM3Z gene is located in one of the most gene dense region in the human genome, the MHC class III region. It belongs to a cluster of four ubiquitously expressed genes RD-SKI2W-DOM3Z-RPI. These four genes are arranged as two head-to-head orientated gene pairs (Yang et al., 1998).

RD encodes for a putative RNA binding protein with 380 amino acids (Cheng et al., 1993). The cellular function of RD has not been determined. However, RD protein contains the RNA recognition motif (RRM) of the ribonucleoprotein (RNP) family (Surowy et al., 1990). Moreover, it shares amino acid sequence similarity with the U1 snRNP 70 kDa protein (Speiser and White, 1989). The presence of the RNP-like domain and the sequence similarity with the U1 snRNP suggest a possible function related to the RNA splicing (Taylor et al., 1998b). The biochemical and cellular characterization of Ski2w has been extensively discussed in the last chapter. RPI does not have any homolog in lower eukaryotes (Shen et al., 1994). It shares limited sequence similarity with a number of different protein families, with the tyrosine kinase transforming protein
from Fujinami virus being the most significant one (18.7% identity, 32.5% similarity over the C-terminal 157 amino acids) (Sargent et al., 1994). In many cases, the homology is indicative of the protein function, and indeed RP1 was recently suggested to be a nuclear serine / threonine protein kinase (Gomez-Escobar et al., 1998).

As a brand new member in the four-gene cluster, the function of human Dom3z has not been determined yet. However, it is already known through sequence alignment that human Dom3z has related proteins in both simple and complex eukaryotes including the baker’s yeast, fission yeast, a nematode, and even a flowering plant. The conservation of these proteins implies that they have very fundamental function in an organism (Yang et al., 1998). In C. elegans, Dom-3 is possibly involved in the postembryonic development of the germ line cells (Paulsen et al., 1995). It is possible that human Dom3z also plays a role in growth and reproduction. Studies in rat have revealed that deletions of this region are associated with increased perinatal mortality, decreased body size, male infertility, and reduced female fertility (Kunz et al., 1980). Taken all results together, they indicate that the MHC class III region may indeed play an essential role in growth, development, and reproduction, and DOM3Z may be the primary candidate for this function.

To examine the biochemical and cellular functions of human Dom3z, bacterial fusion proteins for Dom3z were expressed and purified. Antisera against these Dom3z fusion proteins were generated. Short-term and long-term transfected mammalian cell lines of human Dom3z were established. It was determined using exogenous GFP-Dom3z protein and confirmed by the indirect immunoflourescence of endogenous protein
that human Dom3z is mainly a nuclear protein. Immunochemical studies have shown that endogenous Dom3z protein is expressed in different mammalian cell lines.
5.3 EXPERIMENTAL PROCEDURES

*Oligonucleotides* -- Custom primers were synthesized by Life Technologies (Grand Island, NY). Human *DOM3Z* gene specific primers *f0, r0, r2,* and *r3* were described in (Yang *et al.*, 1998). GFP 5′ primer was located at 218 bp upstream of the *Bam HI* site in the mammalian expression vector pEGFP-C1 (Clontech, Palo Alto, CA). It was designed for amplifying the GFP-*DOM3Z* fusion gene. The sequence of GFP 5′ primer is: 5′ CCG ACC ACT ACC AGC AGA ACA C 3′.

**PCR of human DOM3Z cDNA** — Primers *f0* and *r0* were used to amplify the near full-length human *DOM3Z* cDNA from the EST clone ID #271616. The Expand™ high fidelity PCR system (Roche Molecular Biochemicals, Indianapolis, IN) was utilized to ensure the accuracy of the sequence. PCR products were digested with *Bam HI* and purified using PCR purification kit (Qiagen Inc., Valencia, CA).

**Cloning of DOM3Z cDNA into bacterial and mammalian expression systems** — The purified *DOM3Z* cDNA PCR product was ligated in frame with bacterial expression pETa *Bam HI* vector (Novagen, Madison, WI), pGEX(2T) *Bam HI* vector (Amersham Pharmacia Biotech, Piscataway, NJ), and mammalian expression pEGFP-C1 *Bam HI* vector, respectively. *DOM3Z* cDNA probe (as described in the previous chapter) was
used for colony hybridization to screen for positive clones with the insert, which were confirmed by *Bam* HI digestion. Afterwards, the correct configured cloning products were identified using appropriate restriction enzyme digestions. The precision of the ligations at the 5' cloning sites was determined by automated sequencing using the *r3* primer.

Expression and purification of Dom3z bacterial fusion proteins — Bacterial fusion proteins Trx-His-Dom3z (from pETa-*DOM3Z*) and GST-Dom3z (from pGEX(2T)-*DOM3Z*) were induced following the standard protocols at 30° or 37° using 0.05, 0.5, or 1 mM IPTG for 30, 60, 90 or 120 min, respectively. Harvested cell pellets were resuspended in PBS and sonicated. The supernatant and the inclusion bodies were separated after centrifugation and analyzed in 10% SDS-PAGE gel. Electroelution was applied to purify Trx-His-Dom3z directly from the gel. The appropriate protein band was excised and placed in a dialysis bag with buffer (0.2 M Tris-acetate, pH 7.4, 1% SDS, 10 mM DTT) and then electroeluted in a gel machine containing 50 mM Tris-acetate, pH 7.4, 0.1% SDS at 4°, 100 V for 3 hours. GST-Dom3z was purified using Sepharose 4B beads (Amersham Pharmacia Biotech). Briefly, French Press was applied to lyse the cells. Soluble proteins were recovered from the supernatant and incubated with the Sepharose 4B beads. The bound fusion protein was eluted from the beads using 10 mM glutathione.
Generation and purification of antisera against human Dom3z bacterial fusion proteins — Purified bacterial fusion proteins Trx-His-Dom3z and GST-Dom3z were used as antigen to immunize rabbits, respectively. For each rabbit, about 50 μg of fusion protein was primarily injected into multiple subcutaneous sites, followed by four more boost injections, each time using about 20 to 10 μg antigen, with the dose gradually reduced. The rabbits were bled after the second boost, and the antisera were examined using ELISA. The specificities and avidities of the antisera were again checked after the terminal bleeding. The antisera were subjected to Immobilized E. coli lysate Kit (Pierce, Rockford, IL) to remove the part cross-reacting with bacterial proteins for any further immuno analysis.

Transient expression of GFP-Dom3z in HeLa cells — HeLa cells were cultured on coverslips until ~50% confluent, transfected with GFP vector only and GFP-DOM3Z, respectively, using lipofectin reagent (Life Technologies). Seventy-two hours after the transfection, the green fluorescence in cells was observed under a UV microscope with appropriate fluorochrome-specific filter.

Development of long-term GFP-DOM3Z transfected mammalian cell lines — GFP-DOM3Z was digested with EcoR I and purified. HeLa and CHO cells were transfected with the linearized DNA using electroporation method (Ausubel et al., 1999). The G418 selection began 48 hrs after the transfection. The concentration of G418 starts
at 500 μg/ml, and was gradually increased to 800 μg/ml within two weeks time. The stable transfected cell lines were established by limiting dilutions and clonal expansions.

**Isolation of DNA, RNA from mammalian cells** — Genomic DNAs were isolated from normal HeLa and CHO cells as well as from the stable GFP-DOM3Z transfectants using DNA isolation kit (Gentra System, Inc., Minneapolis, MN). Total RNA were isolated using RNAzol® kit (Tel-Test, Inc., Friendswood, TX) after the manufacturer’s instructions.

**Extraction of protein lysates from mammalian cell lines** — Cells were lysed using 20 mM Tris, pH 7.4, 150 mM NaCl and 0.1% SDS, followed by sonication. The total cellular proteins were obtained from the supernatant after centrifugation. Cell lines used include: 293, kidney; Molt4, lymphoblastoid; HT 29, colon carcinoma; IMR32, neuroblastoma; SKNSH, neuroblastoma; Raji, Burkitt lymphoma; lung carcinoma.

**Southern blot analysis** — Ten micrograms of genomic DNA were digested to completion with restriction enzyme Acc I, resolved on a 0.8% agarose gel, blotted onto Hybond-N membrane and hybridized with the human DOM3Z cDNA probe.

**Reverse transcription (RT) PCR** — The RT-PCR experiments were executed using GeneAmp RNA PCR kit (Perkin Elmer Cetus, Norwalk, CT) after the manufacturer’s protocol. Specifically, random hexamer was used as the downstream primer for
synthesizing the first strand cDNA in reverse transcription reaction. The subsequent PCR reactions were carried out using the GFP-5' primer and the DOM3Z r2 primer.

**Indirect immunofluorescence experiment** — HeLa cells were cultured on coverslips until ~70% confluent followed by an additional 3 hours incubation in fresh medium. Cells were then rinsed with PBS, fixed with 4% paraformaldehyde and permeated with methanol at −20°C. After blocked using 10% goat serum, cells were incubated with primary antibodies against Dom3z (1:50, 1:100, and 1:200 dilutions) at 4°C overnight. The secondary antibody used was 1:80 diluted goat anti-rabbit IgG conjugated with FTC (Sigma Chemical Company, St. Louis, MO). Fluorescence was observed using a UV microscope.

**Immunoblot analysis** — Total cell lysates from bacteria and mammalian cells were resolved in 10% SDS-PAGE gel, and transferred onto nitrocellulose membranes by electroblotting. After blocked using 5% non-fat milk in PBS, the membranes were incubated with the primary antibodies with the appropriate dilutions for 1 hour at room temperature. GFP monoclonal antibody was purchased from Clontech (#8362-1). The signals were detected by chemiluminescence method using horseradish peroxidase (HRP) conjugated secondary antibodies and ECL-plus reagents (Amersham Pharmacia Biotech).
5.4 RESULTS

Cloning of human DOM3Z into bacterial and mammalian expression vectors

The 1.2 kb human DOM3Z cDNA fragment was amplified and digested with Bam HI as the two primers used, f0 and r0, both have Bam HI sites incorporated to facilitate cloning. The PCR fragment was cloned into bacterial pETa Bam HI vector, pGEX(2T) Bam HI vector, and mammalian pEGFP-C1 Bam HI vector. The scheme to determine the presence and orientation of the insert in each vector were analyzed in Fig. 5.1, panels I-III. The results for restriction digestion of clones containing DOM3Z cDNA insert were shown in panel IV.

In panel IV-A, the presence of DOM3Z in pETa vector in two clones was demonstrated by the cloning enzyme Bam HI digestion (lanes 1 and 2), which could release the insert. Meanwhile, since both clones exhibit the 7 kb Nco I fragment (lanes 3 and 4) as analyzed in panel I-B, they were therefore concluded to both contain the correctly configured pETa-DOM3Z. Similarly, Bam HI was also used to examine the presence of DOM3Z in pGEX (lanes 5 and 6) and pEGFP (lanes 9 and 10) respectively. For these two constructs, Sma I digest was chosen to determine the orientation of the insert. The insert in pGEX-DOM3Z clones are correctly configured as both exhibit 5.2 kb and 960 bp Sma I fragments (lanes 7 and 8), which are consistent with the analysis in panel II-B. One clone contains the correctly constructed GFP-DOM3Z as it shows the
5.7 kb fragment (lane 11) predicted in panel III-B; while the insert in the other clone was in the opposite direction with the 5.1 kb and ~800 bp Sma I fragments (lane 12) as analyzed in panel III-C.

Expression and purification of Trx-His-Dom3z fusion protein

Bacteria transformed with correctly configured pETa-DOM3Z were induced with IPTG to express the fusion protein Trx-His-Dom3z. The induced protein can be easily observed in the 10% SDS-PAGE gel at the apparent size of around 65 kDa (Fig. 5.2, panel I). Since the calculated fusion protein size is about 54 kDa, it becomes necessary to prove that the induced protein with the apparent size of 65 kDa is indeed the Trx-His-Dom3z. In immunoblot analysis, the induced 65 kDa protein was detected using polyclonal antibody against GST-Dom3z (Fig. 5.2, panel II). The antibody used should only react with the Dom3z polypeptide because it was raised against a different fusion protein and also the minimal cross-reactions with bacterial proteins were removed through purification. Therefore, the 65 kDa protein was confirmed to be the Dom3z fusion protein. It was speculated that besides the possible inaccuracy of the protein marker, the discrepancy between the observed and calculated protein sizes may be due to the structural reason, since the Dom3z protein is relatively proline-rich (Yang et al., 1998). The ring structures of the proline residues may retard the mobility of the Dom3z fusion protein and cause the appearance of a larger protein.

It appears that the induced Trx-His-Dom3z protein mainly exists in the pellet of the bacterial cell lysate. To obtain soluble fusion protein of Dom3z, different
combinations of IPTG concentration (1, 0.5, or 0.05 mM), inducing duration time (120, 90, 60, or 30 minutes) and temperature (37°, 30°, or room temperature) were tried to optimize the expression. Even under 0.05 mM IPTG, at room temperature for 30 minutes, the induced protein already became insoluble and could only be detected in the inclusion bodies. The insolubility of the fusion protein makes its purification by affinity chromatography very difficult. Therefore, Trx-His-Dom3z was purified directly from SDS-PAGE gel using electroelution method. The eluted protein was shown to be the main component in the Coomassie blue stained SDS-PAGE gel (Fig. 5.2, panel I, lane 5).

Expression and purification of GST-Dom3z fusion protein

Bacteria with the correctly orientated pGEX(2T)-DOM3Z were induced to express the GST-Dom3z fusion protein using IPTG. The apparent size of the induced protein in SDS-PAGE gel is about 67 kDa (Fig. 5.3, panel I). By means of immunoblot analysis using antibody against Trx-His-Dom3z, the 67 kDa induced protein was proved to be indeed the Dom3z fusion protein (Fig. 5.3, panel II).

GST-Dom3z was also mainly present in the inclusion bodies under conventional lysis and sonication procedure. One possibility leading to the protein precipitation is that during the isolation process, the protein may be denatured because of harsh temperature and mechanical effect. To avoid this problem, harvested cells with the exogenously expressed GST-Dom3z protein were subjected to French Press. The solubility of the fusion protein was indeed significantly improved. The soluble fusion protein was purified using Sepharose-4B beads and eluted with glutathione (Fig. 5.3, panel I, lane 5).
Nuclear localization of Dom3z

To determine the cellular localization of Dom3z, HeLa cells transiently transfected with GFP-DOM3Z were observed under a UV microscope. It was shown clearly that the fusion protein was mainly localized in cell nuclei (Fig. 5.4). Control experiment using cells transfected with GFP vector only shows that GFP is expressed in cytoplasm as well as in nuclei. It has been reasoned that the fluorescence of GFP is observed throughout the cell because the protein can freely pass the nuclear pore due to its small size (27 kDa).

To further substantiate the nuclear localization of Dom3z, indirect immunofluorescence microscopy was performed. Using the purified antisera against Trx-His-Dom3z as the primary antibody, the endogenous Dom3z protein was detected mainly in HeLa cell nuclei (Fig. 5.5), which is consistent with the previous observation through the GFP experiment.

Establishment of stable GFP-DOM3Z transfectant cell lines

Stable GFP-Dom3z transfected HeLa and CHO cell lines were established under G418 selection by limiting dilutions and clonal expansions. By genomic Southern blot analysis (Fig. 5.6), it was demonstrated that GFP-DOM3Z construct has been incorporated into all nine CHO (lanes 1-9) and nine HeLa (lanes 10-18) cell lines with possibly different copy numbers. Lane 19 is the genomic DNA from untransfected HeLa cells, it clearly does not have the GFP-DOM3Z construct.
To examine the expression of the exogenous GFP-DOM3Z, RT-PCR was specifically designed using the upstream GFP 5' primer in the GFP vector while the downstream r2 primer in the DOM3Z cDNA. As shown in Fig. 5.7, all except one HeLa cell line (panel I, lane 9) exhibit a product at 560 bp as expected. It was therefore concluded the exogenous DOM3Z gene was expressed in these cell lines.

The GFP-Dom3z fusion protein expression in stable CHO transfectants was detected by immunoblot analysis using commercially available GFP monoclonal antibody. A polypeptide with an apparent size about 93 kDa was shown in Fig. 5.8.

**Immunoblot analysis to detect the endogenous Dom3z protein in tissue samples and mammalian cell lines**

The first step leading to solve the cellular and physiological functions is the detection of the endogenous protein. Endogenous Dom3z protein was revealed using immunoblot analysis as shown in Fig. 5.9. When antisera against Trx-His-Dom3z was used, a polypeptide around 67 kDa can be detected in all the samples tested, including adrenal and liver tissue samples and 6 cell line extracts. Similar result was obtained from immunoblot analysis using antisera against GST-Dom3z. However, the calculated molecular weight from the deduced amino acid sequence is only about 44 kDa. Whether this 70 kDa protein is a product after post-translational modifications is yet to be determined.
5.5 DISCUSSION

Human DOMZ is the most recently discovered gene in the RD-SKI2W-DOMZ-RP1 gene cluster located between complement Bf and C4 genes in the MHC class III region (Yang et al., 1998). The exact function of Dom3z has not been determined yet. As demonstrated in this chapter, preliminary studies have been performed to help further investigations. Dom3z has been shown to be mainly a nuclear protein. The endogenous protein is detectable in cell line lysate with an apparent size of 65 - 70 kDa. The development of Trx-His-Dom3z and GST-Dom3z fusion proteins, the generation of polyclonal antibodies, the establishment of stable HeLa and CHO cell lines transfected with GFP-DOMZ are of significance for future research. They may provide the basic materials for in vitro biochemical, immunochemical and cellular studies.

The endogenous Dom3z protein detected by immunoblot analysis has much higher apparent molecular weight (65 - 70 kDa) than the predicted size of the deduced protein (44 kDa). The discrepancy is hard to explain at this moment. From the deduced amino acid sequence, it was shown by PROSITE analysis that there are four potential sites for N-myristoylation located at amino acids 5, 9, 264, and 358, respectively (Yang et al., 1998). Therefore, it is possible that post-translational modifications of the endogenous Dom3z make the size of the mature protein larger than the predicted molecular weight. Meanwhile, it is worthy of noting that Dom3z has a leucine-zipper
motif that may be important for protein-protein interactions. It is possible that Dom3z may be dimerized as it could interact with itself, or it may be tightly associated with some other cellular protein and therefore exist as a complex. Although most non-covalently associated protein complex can be dissociated in the presence of mercaptoethanol and SDS in the protein sample buffer, the 67 kDa band would not disappear even after prolonged boiling in up to 200 mM β-mercaptoethanol or 4% SDS (data not shown). Therefore, it becomes important to determine whether Dom3z interacts with itself or any other protein(s), and if it does, which protein(s) is involved since it would be one relevant aspect to look into the function of Dom3z.

It has been implied that Dom3z, together with RD, Ski2w and RP1 may have concerted functions, probably related to RNA translation and turnover (Yang et al., 1998). The nuclear localization of human Dom3z is consistent with the suggestion of its involvement in the RNA metabolism. Consequently, whether human Dom3z can bind RNA becomes an obvious target to be addressed. The improved solubility and purification of GST-Dom3z may provide the essential material for biochemical studies.

Recently, in a personal communication with Dr. Arlen Johnson from the University of Texas at Austin, it was found that the yeast version of Dom3z interacts with Rat1p, a nuclear 5'-3' exoribonuclease required for RNA turnover (Johnson, 1997). Human Ski2w and its yeast homolog Ski2p have been suggested to be involved in the 3'-5' RNA degradation (Jacobs et al., 1998; Qu et al., 1998). However, since Ski2w has been shown to be present in the nucleolus and cytoplasm of the cells (Qu et al., 1998), while Dom3z is mainly a nuclear protein, it is unlikely that Dom3z would interact with Ski2w.
directly. It is plausible that Ski2w is involved in a cytoplasmic exonuclease pathway (Qu et al., 1998), while Dom3z is involved in a nuclear exonuclease pathway.

The yeast two-hybrid system would be one of the ways to fish out the protein(s) that interacts with Dom3z. Besides screening the human cDNA library, it is more feasible to test the interaction of Dom3z with a handful of known candidate proteins. Preliminary results from this study are promising with the indication that Dom3z may actually be interacting with itself. It would be exciting to investigate whether human Dom3z would interact with any exoribonuclease, especially the human homolog of yeast Ratlp. With the commercially available GFP antibodies, which are reportedly designed for this purpose, immunoprecipitation experiment using as the long-term transfectants of GFP-DOM3Z may yield relevant information.

It is logical to speculate that as yeast Dom3z interacts with 5'-3' exonuclease Ratlp, the Dom3z mutants would be defective in the 5'-3' RNA degradation pathway. One way to examine the human Dom3z function is to determine whether human Dom3z can rescue the yeast Dom3z mutants in the exoribonuclease pathway. If the human Dom3z can complement the yeast protein function, which suggests that human Dom3z also interacts with exonuclease and is involved in RNA turnover, it would be of interest to use site-directed mutagenesis to find out the protein domain that is responsible for the protein-protein interaction. Deletion of the leucine-zipper motif is another obvious experiment to test. Since Ski2w is possibly involved in the 3'-5' degradation pathway, it appears that Dom3z and Ski2w would both play critical role in RNA degradation. It
cannot be overemphasized that RNA turnover is one of the most important approaches to achieve the regulation of gene expression.
Fig. 5.1  **Restriction analyses to determine the presence and orientations of DOM3Z cDNA inserts in different expression vectors.** Schemes for digestions were illustrated with pETa-DOM3Z in *panel* I, pGEX-DOM3Z in *panel* II, and pEGFP-DOM3Z in *panel* III. *Bam* HI digestion was used to examine the presence of the insert (*panels* I-III, A). *Nco* I (*panel* I, B and C) and *Sma* I (*panels* II and III, B and C) digestions were used to determine the configurations of the inserts. The enzymes chosen were marked in each analysis, and the approximate sizes of the fragments were revealed. The results of the digestions are presented in *panel* IV. The sizes of the insert and each vector were labeled.
Fig. 5.1 Restriction analyses to determine the presence and orientations of DOM3Z cDNA inserts in different expression vectors.
(Fig 5.1 continued)

<table>
<thead>
<tr>
<th>IV.</th>
<th>pETa</th>
<th>pGEX</th>
<th>pEGFP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bam HI</td>
<td>Nco I</td>
<td></td>
</tr>
<tr>
<td>kb</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A  B  C
I. kDa

| 93 | 67 | 56 | 42 | 28 |

II.

| 93 | 67 | 56 | 42 |

**Fig. 5.2** Production and detection of bacterial fusion protein Trx-His-Dom3z. I. Expression and purification of Trx-His-Dom3z. *Lanes* 1 and 2, soluble proteins from uninduced and induced cell lysates; *lanes* 3 and 4, insoluble proteins from uninduced and induced cell lysates; *lane* 5, Trx-His-Dom3z purified using electoolution method. II. Immunoblot analysis of uninduced and induced cell lysates from Trx-His (*lanes* 1 and 2), as well as from Trx-His-Dom3z (*lanes* 3 and 4). The antiserum used was against GST-Dom3z.
Fig. 5.3 Production and detection of bacterial fusion protein GST-Dom3z. I. Expression and purification of GST-Dom3z fusion protein. Lanes 1 and 2, soluble proteins from uninduced and induced cell lysates; lanes 3 and 4, insoluble proteins from uninduced and induced cell lysates; lane 5, GST-Dom3z purified using Sepharose-4B beads and eluted with glutathione. II. Immunoblot analysis of GST-Dom3z. Soluble proteins from uninduced and induced total cell lysates were shown in lanes 1 and 2; results from insoluble proteins from uninduced and induced total cell lysates were presented in lanes 3 and 4. The antiserum used was against GST-Dom3z.
Fig. 5.4 Cellular localization of GFP-Dom3z in HeLa cells. HeLa cells were transiently transfected with GFP-DOM3Z for 72 hrs. The green fluorescence was observed under UV microscope.
Fig. 5.5  Indirect immunofluorescence experiment to detect the cellular localization of Dom3z in HeLa. The primary antiserum is against Trx-His-Dom3z.
Fig. 5.6 Genomic Southern blot analysis showing the incorporation of GFP-DOM3Z into the stable cell line genome. The genomic DNAs were subjected to Acc I digestion and probed with the human DOM3Z cDNA. Lanes 1 to 9 are from CHO transfectants, lanes 10 to 18 are from HeLa transfectants. Lane 19 is genomic DNA from untransfected HeLa DNA as negative control. The position of the incorporated DNA is indicated by an arrow.
I. RT-PCR of HeLa transfectants

Fig. 5.7  Reverse transcription (RT) – PCR of GFP-\(DOM3Z\) to examine the expression of the exogenous construct in the stable cell lines. Total RNA from HeLa (panel I) and CHO (panel II) cell lines were amplified using GFP 5’ primer and \(DOM3Z\ r2\) primer. Lane 1 of each panel is 100 bp ladder. The size of the product is labeled. Lane 11 of panel I is a negative control using untransfected HeLa total RNA.
Fig. 5.8  **Immunoblot analysis to detect GFP-Dom3z in CHO transfectants.** Monoclonal antibody against GFP (Clontech, #8362-1) was used.
Fig 5.9  **Immunoblot analysis to detect the human endogenous Dom3z** in:
*lane* 1, adrenal; *lane* 2, liver; *lanes* 3-8, cell lines of 293, Molt4b, HT29, IMR32, SKNSH, Raji, and A549, respectively. Antiserum against Trx-His-Dom3z was used.

<table>
<thead>
<tr>
<th>kDa</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>93</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 6

ORGANIZATIONS AND GENE DUPLICATIONS OF THE
HUMAN AND MOUSE MHC COMPLEMENT GENE CLUSTERS

6.1 ABSTRACT

The MHC complement gene cluster (MCGC) in most people contains thirteen structural genes, pseudogenes and gene segments. Novel genes $RD$, $SKI2W$, $DOM3Z$ and $RPl$ are organized as two head-to-head gene pairs between complement gene $Bf$ and the first locus of $C4$. Southern blot analysis shows that single copy genes for $DOM3Z$ are detectable in primates and other mammals. Sequence analyses revealed that the exon-intron structures of human and mouse $DOM3Z$ genes are identical. Both human and mouse $DOM3Z$ transcripts exhibit splice variants at the 5' regions, although the open reading frames remain identical. Cloning and characterization of the mouse $RPl$ cDNA revealed a reading frame for 254 amino acids with a bipartite nuclear localization signal close to the amino-terminus. The mouse $RPl$ gene consists of 7 exons and spans 12.9 kb.
Located in intron 4 of mouse *RP1* is an endogenous retrovirus that probably confers the androgen responsive expression of the Slp protein in certain male congenic mice. The availability of the complete human and mouse MCGC genomic and cDNA sequences allows further deliberate analyses of gene duplications and evolution. The intergenic region between mouse *SLP* and *C4* genes is more than six times larger than the corresponding region in human. It contains the functional gene steroid *CYP21A*, long stretches of repetitive DNA elements, and three partially duplicated gene segments *TNXA, SKI2W2* and *RP2*. The modular duplications of human and mouse *RP-C4-CYP21-TNX* (RCCX) are sharply different as *SKI2W2* is absent in the human MCGC, and *TNXA* and *RP2* are much smaller in size but higher in sequence conservation in humans.
6.2 INTRODUCTION

The MHC complement gene cluster (MCGC) is a genomic region in the class III region of the major histocompatibility complex (MHC). This gene cluster has many intriguing features, namely, the very close gene placements with minimal intergenic regions or with overlapping 3' ends, the complete and partial duplication of genes in a modular fashion, the presence of endogenous retroviruses that modulate gene size variation, the high degree of protein polymorphism, and the disease associations (Yu, 1998). In humans the MCGC includes four genes, C2, Bf, C4A and C4B, that code for subunit proteins of the complement C3 and C5 convertases (Carroll et al., 1984). Also located in this cluster are nine other functional genes, pseudogenes and gene segments, which are present between the complement genes and/or undergoing concurrent gene duplications or gene deletions with complement C4A or C4B in a modular fashion (Rupert et al., 1999; Yang et al., 1999; Yang et al., 1998). The products of the functional genes include an enzyme involved in the hydroxylation of steroids that is essential for hormonal homeostasis of body salts and for sex hormones (cytochrome P450 21-hydroxylase CYP21) (Miller and Morel, 1989), an extracellular matrix protein (tenascin TNX) (Bristow et al., 1993), and four ubiquitously expressed proteins which are probably engaged in the various stages of the regulation of gene expression (RD, Ski2w, Dom3z
and RPI) (Levi-Strauss et al., 1988; Sargent et al., 1994; Shen et al., 1994; Yang et al., 1998).

Deficiencies and polymorphisms of complement C4 have implicated associations with many genetic and immune disorders (Porter, 1983) such as systemic lupus erythematosus (SLE) (Fielder et al., 1983), insulin dependent diabetes mellitus (Lhotta et al., 1996; Mijovic et al., 1985), recurrent spontaneous abortion (Laitinen et al., 1991), and IgA deficiency (Volanakis et al., 1992). Except SLE, the deficiency or malfunction of C4 itself does not necessarily lead to these disorders, suggesting the presence of disease susceptibility genes close to C4. It is possible that some of the recently identified genes in the MHC class III region could be involved in these disorders.

The novel genes RD, SK12W, DOM3Z and RPI are located in the 30 kb genomic region between human complement genes Bf and C4. RD (or D6S45) is located 205 bp downstream of Bf. It encodes a 46 kDa polypeptide that contains an RNA recognition domain and 24 copies of repeating Arg-Asp or Arg-Glu (Levi-Strauss et al., 1988). The RD protein is a subunit of NELF that is a negative factor for transcriptional elongation (Yamaguchi et al., 1999).

Sharing the 5' regulatory region with RD, SK12W encodes a putative RNA helicase (Dangel et al., 1995). We have shown that human Ski2w proteins are not only present in the nucleoli, they are also associated with polysomes and probably the 40S subunit of ribosomes, suggesting that Ski2w proteins are located at the sites of ribosome biosynthesis and protein synthesis (Qu et al., 1998). The yeast homologue, Ski2p, is
involved in mRNA turnover and antiviral defense (Jacobs et al., 1998; Masison et al., 1995).

The human SKI2W gene is located 2.4 kb upstream of the RPI gene (Yang et al., 1998). RPI (or G11, D6S60E) probably encodes a nuclear, Mn$^{2+}$-dependent, Ser/Thr protein kinase (Sargent et al., 1994; Shen et al., 1994). DOM3Z, the latest identified gene, is located between SKI2W and RPI. Dom3z-related proteins are present in simple eukaryotes such as baker's yeast, as well as complex eukaryotes like humans. The homologue in C. elegans, Dom-3, may have a related function to Mes-3, that is a maternal effect component required for normal postembryonic development of germ cells (Paulsen et al., 1995). Northern blot analysis revealed that human DOM3Z transcripts are heterogeneous in size and are highly expressed in reproductive tissues and pancreas (Yang et al., 1998).

The human genes for nuclear protein kinase RP, complement C4, steroid CYP21 and tenasin TNX (RCCX) are duplicated in a modular fashion. RCCX modules have a spectrum of variations in sequence polymorphisms, number of modules, and C4 gene size (Rupert et al., 1999; Yang et al., 1999). Although four genes were involved in the RCCX modular duplication process, only complement C4 evolved to yield two functional protein products, C4A and C4B. The human CYP21A is a pseudogene because of an 8-bp deletion in exon 3 and two other deleterious mutations in the coding sequence (Higashi et al., 1986; White et al., 1986). The duplications for human RP and TNX are incomplete as the 5' ends of the duplicated genes, RP2 and TNXA, are truncated.
Between human C4 locus I and locus II are CYP21A and the partially duplicated gene segments TNXA and RP2 (Bristow et al., 1993; Yang et al., 1999).

The mouse MHC, or H2, has a class III region with a gene organization similar but not identical to that of the human (Chaplin et al., 1983). The duplicated gene for mouse C4, SLP, is marked by its androgen responsive expression and by coding a C4-like protein in certain H2 haplotypes such as H2^d (Shreffler et al., 1984). The intergenic distance between the SLP and C4 loci is about 65 kb (Chaplin et al., 1983), in contrast to a size of 10 kb in the orthologous region of humans (Carroll et al., 1984).

The duplication of mouse C4 is also accompanied by CYP21. The CYP21A gene is located downstream of the SLP gene; the CYP21B gene is located downstream of the C4 gene. In contrast to human, the mouse CYP21A is a functional gene, while CYP21B is a pseudogene because of the deletion of exon 2 and multiple mutations that cause frame shifts and premature termination codons (Chaplin et al., 1986). It is not known whether the mechanism leading to the duplication of SLP and C4, CYP21A and CYP21B in mouse H2 class III region is identical to that in human.

In this study we attempt to provide a detailed genetic analysis of the genes in the human and mouse MCGC, particularly DOM3Z and RP. The human genomic DNA sequence linking three novel genes RPl, DOM3Z and SKI2W is presented. The mouse DOM3Z and RPl cDNAs and genes are investigated and compared with those of human's. While the molecular genetic characterization of the human and mouse MCGC was in progress, genomic sequences of a human MCGC with a monomodular RCCX structure, and a mouse MCGC from the H2^b haplotype became available (Rowen et al.,
1997; Rowen et al., 1997; Rowen et al., 1998; Rowen et al., 1997; Rowen et al., 1997). These sequences allowed a deliberate comparison of the gene structures and gene duplications of the MCGCs between the two mammalian species.
6.3 EXPERIMENTAL PROCEDURES

**Southern Blot Analysis** — Genomic DNAs from African green monkey (COS7), pigtail monkey, Rhesus macaque, Owl monkey (OMK), cotton top tamarin (NPC-LC) and Squirrel monkey (DPSO) were isolated as described in (Dangel et al., 1995b). Genomic DNAs from sheep, dog and pig were purchased from Clontech (Palo Alto, CA). Genomic Southern blot analyses were performed following standard procedures. Briefly, 10 µg of each genomic DNA was digested to completion with \( Sst \) I, resolved in 0.8% agarose gel, blotted onto Hybond N\(^+\) membrane (Amersham Pharmacia Biotech, Arlington Heights, IL), and hybridized with \([\alpha-32P]dCTP\)-labeled human \( DOM3Z \) 0.5 kb 3' genomic DNA probe. The membrane was subsequently washed and exposed to X-ray film.

**Determination of 5' cDNA Sequences for Human \( DOM3Z \) by 5' rapid amplification of cDNA ends (RACE)** — Total RNA was isolated from HeLa cells using RNAzol™ B (Tel-Test Inc, Friendswood, TX) following the manufacturer's instruction. 5' RACE of human \( DOM3Z \) was performed using a kit from Life Technologies (Grand Island, NY). Briefly, the first strand of cDNA was synthesized from total HeLa RNA using the \( rI \) primer (5' GTC TCC ACT TCA CTC AGG TAT 3'). A homopolymeric tail was added to the 3' end. PCR amplification was performed using the dC-tailed cDNA,
the nested r2 primer (5' CGT CAG CAG TTT TGT CAG GTG 3'), and an anchor primer. PCR products were cloned into the pCR2.1 vector (Invitrogen, Carlsand, CA). A 0.8 kb BamHI DNA fragment corresponding to the genomic region upstream of DOM3Z EST clone 271616 sequence was used as a probe for colony hybridization (Yang et al., 1998). Clones containing 5' RACE products were sequenced by automated sequencing (ABI model 377, Foster City, CA) using T7 primer and M13 reverse primer.

Sequence Analyses — The human DOM3Z gene sequence is obtained from our previously published genomic DNA sequences for RPI (GenBank accession no. L26261) (Shen et al., 1994) and for SKI2W (GenBank accession no. AF059675) (Yang et al., 1998). DOM3Z EST clones were identified by searching the National Center for Biotechnology Information (NCBI) EST database with the human Dom3z amino acid sequence, using the "tblastn" program (Altschul et al., 1997). The mouse Dom3z amino acid sequence was deduced from the mouse DOM3Z EST sequence (Marra et al., 1996) and DOM3Z gene sequence (Rowen et al., 1998), with reference to the human Dom3z amino acid sequence (Yang et al., 1998). The human and mouse DOM3Z cDNA sequences for coding regions were compared using the GCG program (Genetics Computer Group, 1991) FASTA, accessed through the Pittsburgh Supercomputing Center. The human and mouse Dom3z protein sequences were aligned using GCG programs PILEUP and PRETTY. Sequence identities and similarities of orthologous proteins from human and mouse were determined by GCG program BESTFIT (Rowen et al., 1998).
Isolation and Sequence Determination of Mouse RP1 cDNA — Using the human RP1.1 cDNA probe, the mouse cDNA clone for RP1 was obtained by screening a brain, λgt10 cDNA library prepared from female C57 Black B6 RNA (Clontech, Palo Alto, CA). The cDNA insert was subcloned into pBluescript (KS) vector and sequenced manually by primer walking method using T7 Sequenase (US Biochemical, Cleveland, OH). Mouse RP1 and RP2 genomic DNA sequences were analyzed through published genomic DNA sequences upstream of mouse C4 (Nonaka et al., 1986b), SLP (Ogata and Zepf, 1991), and mouse genomic DNA sequence from C2 to TNXB (Rowen et al., 1998).
6.4 RESULTS

**DOM3Z is conserved in mammals**

To study whether the **DOM3Z** gene is conserved among mammals, Southern blot analyses of **Sst I**-digested genomic DNAs were performed (Fig. 6.1). Panel I shows results on the analysis of Old World primates and New World primates. Using a 0.5 kb genomic DNA fragment corresponding to the 3' end of human **DOM3Z** as a probe, a single, 1.6 kb **Sst I** fragment was detected in African green monkey (COS7), pigtail monkey, Rhesus macaque and human (HT29) (*lanes* 1-3 and 7, respectively). These identical restriction patterns suggest the presence of a single copy **DOM3Z** gene in the genomes of Old World primates. Similarly, restriction patterns are identical for **DOM3Z** genes in New World primates, as shown by the presence of a 3.5 kb **Sst I** fragment in cotton top tamarin (NPC-LC), Owl monkey (OMK), and Squirrel monkey (DPSO) (*lanes* 4-6, respectively). Panel II shows results on the analysis of non-primate mammals. Single **Sst I** restriction fragments are detected in rabbit (*lane* 1), porcine (*lane* 4), rat (*lane* 5) and mouse (*lane* 6). Double restriction fragments for **DOM3Z** are detectable in ovine (*lane* 2) and canine (*lane* 3), which are most likely due to the presence of additional **Sst I** sites at the 3' region of **DOM3Z** or, less likely, the presence of two genes.
Genomic DNA sequence of human DOM3Z

Human DOM3Z was discovered between SKI2W and RPI (Shen et al., 1994; Yang et al., 1998). A 3.0 kb genomic DNA sequence covering the 5' end of RPI, the entire human DOM3Z gene, and the 3' end of SKI2W is shown in Fig. 6.2. The transcribed region of human DOM3Z spans from nt 601 to 2841 and is split into seven exons (Fig. 6.3, panel I), as compared with its full length cDNA sequence. The entire exon 1 and the first six nucleotides of exon 2 delineate the 5' untranscribed region. The sizes of the exons vary between 95 bp (exon 6) and 362 bp (exon 2). The sizes of the introns vary between 84 bp (intron 6) and 247 bp (intron 1). The poly(A) site of human DOM3Z (nt 2841) is located 59 bp away from the poly(A) site of SKI2W (nt 2901), which is transcribed from the complementary DNA strand. The most 5' nucleotide of DOM3Z mRNA, as identified by 5' RACE, is 267 bp away from one of the 5' transcriptional initiation sites of RPI/G11 (Sargent et al., 1994; Yang et al., 1998).

The intergenic 5' region of DOM3Z and RPI is relatively G+C rich. Of the first 1200 nucleotides shown in Fig. 6.2, 61.3% consist of G+C nucleotides, in contrast to 41% normally present in a mammalian genome (Normore et al., 1976). Remarkably, 75 copies of CpG dinucleotides, which are generally under-represented in the mammalian genome, have been found in this shared intergenic region. No TATA boxes are present but two SP1 sites configured in the same orientation of DOM3Z are located at nt 64-76 and nt 413-450, and two SP1 sites arranged in the same orientation of RPI/G11 are found at nt 577-581, and at nt 1536-1549. Three API sites are present, two of which are in the DOM3Z configuration and one is in RPI/G11 configuration.
Heterogeneous 5' ends of human and mouse DOM3Z mRNAs

The 5' ends of human DOM3Z mRNAs: analysis of the 5' RACE products of DOM3Z performed using total RNA isolated from HeLa cells yielded complex data for the 5' sequences of DOM3Z mRNA. It appears that there are three groups of 5' sequences (Fig. 6.3, panel II). Clones in the first group are similar to that of EST clone HBC4394 and have been reported (Yang et al., 1998). This cDNA was designated as cDNA1 and its full-length sequence is 1,386 bp. Sequence of the second group clones starts at nucleotide 46 of cDNA1, and there is a 44 bp insertion between nt 123 and 124 of cDNA1. A comparison of this cDNA sequence, cDNA2 (accession no. AF059253) with the DOM3Z genomic sequence revealed that the 44 bp insertion is also present in the DOM3Z gene (Fig. 6.2, nt 927-970). Therefore, the addition of this 44 bp sequence is likely the result of an alternative splicing using a different acceptor site in intron 1 (Fig. 6.2). However, this transcript does not change the open reading frame of DOM3Z because an in-frame stop codon (nt 967-969) is present six nucleotides upstream of the initiation codon.

The third group of the RACE products (cDNA3, accession no. AF059254) retains the "intron 1" sequence, with respect to cDNA1 and the DOM3Z gene sequence (panel II, Fig. 6.3). Again the open reading frame is not changed because of the same in-frame stop codon as described in cDNA2. This group of sequences may be the results of partially spliced or unspliced RNAs, or the presence of multiple initiation sites for DOM3Z transcripts.
The 5' variants of mouse DOM3Z cDNA: analyses of the mouse DOM3Z cDNAs identified from the EST database also revealed three groups of variants with different 5' untranslated regions (Fig. 6.3, panel III). In comparison with the regular transcript mcDNA1, one of the variants (mcDNA2; accession no. AA510930) is the result of a different donor site at intron 1 that is 25 bp downstream for that of clone mcDNA1, the other variant (mcDNA3; accession no. AA471915) is caused by an unspliced intron 1. All of these variants share the same initiation codon and therefore encode for an identical protein. However, differences in the 5' untranslated regions of the mRNAs might have variable translational efficiency for the Dom3z protein.

A comparison of human and mouse DOM3Z genes and proteins

The intron-exon boundaries of the mouse DOM3Z gene were defined by comparing sequences from mouse EST clones with the genomic sequence of DOM3Z (Marra et al., 1996). Sequence analyses revealed that the gene structures for human and mouse DOM3Z are identical (Table 6.1). The mouse gene also contains 7 exons with an in-frame stop codon in the first exon. Similar to the human gene, the first exon and six nucleotides of the second exon are untranslated. Except for the untranslated regions from exon 1 and exon 7, the sizes of exons 2 to 6 are identical between the two species, as are the intron phases (Patthy, 1987). The sizes of the six introns in the human and mouse DOM3Z genes are similar.

The coding regions of the human and mouse DOM3Z cDNA sequences are 86.6% identical to each other. The deduced human and mouse Dom3z proteins contain 396 and
397 amino acids, respectively. The two protein sequences share 89.7% identities and 93.7% similarities. There are relatively more dissimilar amino acid residues at the regions close to the amino and the carboxyl termini of the proteins (Fig. 6.4). Both proteins are proline-rich, with 11.6% and 10.8% of the amino acids being proline in the human and the mouse protein, respectively. Leucine is the second most frequent amino acid residue in both proteins, with 9.5% in the human and 10.3% in the mouse protein. In human Dom3z, there is a leucine zipper motif at amino acid no. 103-124 (Fig. 6.4); at the corresponding position in mouse protein, one of the leucine residues is substituted by a valine. The leucine zipper motif may be involved in protein-protein interactions.

An analysis of the cDNA and genomic sequences derived from our laboratory and from the GenBank database for human DOM3Z revealed a polymorphism changing codon 28 from Thr to Ser (Figs. 2 and 4), and codon 261 from His to Gln (Yang et al., 1998). When the derived amino acid sequences from seven mouse DOM3Z EST clones (accession nos. AA218230, AA716917, AA178524, AA509363, AA267188, AA510930, AA471915) (Marra et al., 1996) and from the mouse DOM3Z gene (Rowen et al., 1998) were compared, two amino acids substitutions were found in multiple clones and therefore likely to be polymorphic residues. These substitutions are: a His to Asn change at codon 20 that is the result of a C to A transversion at nt 199, and a Leu to Ser alteration at codon 28 that is the result of a T to C transition at nt 224.
Elucidation of the mouse RPl cDNA and amino acid sequences

A cDNA clone, MRP1.0, for mouse RPl was isolated from a BALB/c brain cDNA library using a human RPl cDNA probe. Sequence determination revealed a 1,028 bp insert that contains a polyadenylation signal and a poly(A) tail. Screening of the related cDNA sequence from the mouse EST database yielded an EST clone, AA036570 (Marra et al., 1996), with a sequence of 571 bp in length that overlaps with MRP1.0 at the 3' end and extends its 5' end by 102 bp. Fig. 6.5 shows a compiled cDNA sequence for mouse RPl that has a 5' untranslated region of 63 nt, an open reading frame for 254 amino acids, and a 3' untranslated region of 284 nt. A bipartite nuclear localization signal, which consists of two basic tetrapeptides separated by nine amino acid residues, is present at the amino terminal region. The derived amino acid sequence shares 85.8% sequence identities or 88.2% sequence similarities with human RPl/G11. Searching of the related protein family or sequence motifs by BLASTP (Altschul et al., 1997) and PROSITES (Bairoch et al., 1997) did not yield information on the possible protein structures or functions of mouse RPl.

Duplication of mouse RPl and RP2 genes

The availability of the genomic DNA sequence extending from the 3' region of mouse complement C2 gene to the tenascin TNXB gene (Rowen et al., 1998) allows a deliberate analysis to characterize the duplication of the mouse RP sequences. In comparison with the cDNA sequence described above, the mouse RPl located upstream of SLP is split into 7 exons spanning 12,946 bp (Fig. 6.6, panel II). An endogenous
retrovirus of 6050 bp in size is present in intron 4 of the \textit{RPI} gene. This endogenous retrovirus was originally discovered while investigating the mechanism leading to androgen-dependent expression of the sex-limited protein (Slp) and was termed the "imposon" (IMP) (Stavenhagen and Robins, 1988). The IMP is located 1891 bp upstream of the \textit{SLP} gene. It is flanked by a 5 bp target site repeat sequence (5'-ACATT-3'), and is configured in the reverse transcriptional orientation with respect to that of \textit{RPl} (and \textit{SLP} gene).

\textit{RP2} is located upstream of mouse \textit{C4}. Similar to the case in human, the duplication of \textit{RP} is incomplete. The breakpoint for the mouse \textit{RPl/RP2} gene duplication occurred in intron 2 3,194 nt downstream of the donor site of exon 2, or 715 bp upstream of the acceptor site of exon 3. The size of the duplicated region for \textit{RP2} is 3,404 bp. The corresponding region in \textit{RPl} from intron 2 to exon 7 spans 9264 bp. In contrast to \textit{RPl}, there is no trace of IMP integration in intron 4 of \textit{RP2}, suggesting that the gene duplication of \textit{RPl} and \textit{RP2} in mouse occurred prior to the integration of IMP to \textit{RPl}. The exonic sequences from \textit{RPl} and \textit{RP2} were compared. Besides the truncation of exons 1 and 2, there are 20 point mutations in the remaining five exons of \textit{RP2}. In addition, there are a single nucleotide deletion in exon 3, a 10-nucleotide deletion in exon 4, and an insertion of a C nucleotide in the 3' untranslated sequence of exon 7 (shown in \textit{Fig. 6.5}). Therefore, it is unlikely that \textit{RP2} would code for a functional protein product.
**Duplication of Mouse TNXA and TNXB Genes**

Similar to the case in human, mouse TNX and CYP21 overlap at the 3' end. The mouse TNXB gene is 59.3 kb in size. It consists of 43 exons that encode a protein of 4,114 amino acids (Matsumoto *et al.*, 1994; Rowen *et al.*, 1997). Analysis of the mouse MCGC genomic sequence from a H2b haplotype (Rowen *et al.*, 1997; Rowen *et al.*, 1998) revealed that 10 kb of the 3' region TNX sequence are present downstream of the (functional) CYP21A gene. This sequence, TNXA, corresponds to exon 26 to exon 43 of mouse TNXB with the sequence for exons 34 and 35 being deleted. A nonsense and a frame-shift mutation are present in exons 28 and 30, respectively. In humans, TNXB consists of 45 exons and spans 68.2 kb. The partially duplicated sequence for human TNXA spans 4.5 kb, which corresponds to exon 33 to exon 45. Human TNXA fused directly to RP2. In mouse, there is a large stretch of highly repetitive DNA sequence spanning 28.7 kb between TNXA and RP2. Unexpectedly, a 1.8 kb DNA sequence is present in this region that is similar to mouse SKI2W from intron 12 to intron 17 (including exons 13 - 17). This partially duplicated gene fragment, SKI2W2, is 94.3% identical to the corresponding region in the intact SKI2W gene. The major mutations in SKI2W2 include an 11 bp deletion in exon 16, and a 6 bp deletion in intron 16.
6.5 DISCUSSION

Analysis of the mouse genomic DNA sequence at the MCGC reveals the partially duplicated sequences for TNX and RPl between the CYP21A and C4 (Fig. 6.7). The intergenic sequence between SLP and C4 in the mouse H2\textsuperscript{b} haplotype is 65.5 kb (Rowen \textit{et al.}, 1997), which is much larger than the orthologous region in human that is 10 kb in size. In humans the duplicated sequence for RP2 is 913 bp that spans the last two and half exons. The breakpoint of gene duplication is located in exon 5 (Fig. 6.6). The sequence of RP2 is exactly identical to that of RPl. In mouse, 3.4 kb of genomic DNA sequence corresponding to the last five exons of RPl are present upstream of mouse C4 gene. The breakpoint of gene duplication occurred in intron 2. This duplicated DNA sequence for mouse RP2 contains multiple point mutations and minideletions and therefore, mouse RP2 is unlikely to be functional even though it is relatively larger than human RP2. The duplicated region for mouse TNXA is 10 kb with 16 exons, compared with 4.5 kb and 13 exons for human TNXA. Again, there are extensive sequence variations/mutations in mouse TNXA gene fragment. Unexpectedly, an 1.8 kb partially duplicated gene fragment for SKI2W is present between RP2 and TNXA, which is not observed in the human MHC. Therefore, the modular duplications of RP, C4, CYP21 and TNX genes in human and in mouse are substantially different and likely occurred independently. In some mouse haplotypes there are multiple copies of SLP/C4 and
CYP21 genes (Levi-Strauss et al., 1985; Pattanakitsakul et al., 1990; Rosa et al., 1985). The mechanism(s) leading to these gene duplications have not been determined.

In both human and mouse MCGC, there are two sets of gene duplications. The first set is a single gene duplication giving rise to C2 and factor B (Bf). In both human and mouse, C2 and factor B proteins share 41.3% amino acid sequence identities. The second set is the modular gene duplication for RP1/RP2, C4A/C4B (SLP/C4 in mouse), CYP21A/CYP21B, and TNX1/TNX2. In human, the duplicated RCCX modules are highly similar (99% identical) in sequence. There is overwhelming evidence for homogenization of polymorphic or mutant sequences in the RCCX, modules due to misalignments between monomodular and bimodular/trimodular structures. This homogenization process is not only a driving force for functional diversity of C4, but may also be one of the root causes for MHC disease associations (Kawaguchi et al., 1991; Yang et al., 1999; Yu, 1998). In mouse, the RCCX modules have much more sequence deviations that include selective integrations of numerous repetitive DNAs, retroelements, multiple point mutations and minideletions. These gross sequence changes would discourage misalignments. Therefore, homogenizations of the diversified sequences in the duplicated gene modules would be less efficient in the mouse MCGC than that of human’s.

The genomic sequence of SLP from the H2^b haplotype has a G-nucleotide deletion in exon 24 (Rowen et al., 1997). Exon 24 encodes the thioester residues of the protein. This single nucleotide deletion would lead to frame shift mutation and therefore truncation of the Sip protein. This is of particular interest because it has in fact been
noted that no Sip proteins are produced in both male and female mice with \( H2^b \) haplotype (Atkinson et al., 1982). Hence, the SLP in \( H2^b \) is a mutant or a pseudogene. Frameshift mutations have also been found in a human \( C4AQ0 \) gene caused by a 1-bp deletion in exon 20 (Fredrikson et al., 1998), and in human \( C4AQ0 \) and \( C4BQ0 \) genes by a 2-bp insertion in exon 29 (Barba et al., 1993; Rupert et al., 1999).

An endogenous retrovirus IMP is present in reverse orientation in the intron 4 of mouse \( RPI \). The 5' LTRs of the IMP are shown to contain glucocorticoid responsive elements that confer the sex limited expression of SLP in male mice (Adler et al., 1992; Ramakrisman and Robins, 1998). Whether the LTRs would bestow a differential expression of \( RPI \) and/or \( DOM3Z \) in male and female mice has yet to be determined. In humans, an endogenous retrovirus HERV-K(C4) is present in the intron 9 of the long \( C4 \) gene. The orientation of the endogenous retrovirus is opposite to that of the resident gene \( C4 \). The IMP present in intron 4 of mouse \( RPI \) is also organized in the reverse orientation with respect to the resident gene. It has been suggested that the integration of endogenous retroviruses into the intronic sequence of structural genes in the reverse configuration might confer protective function on the host cell. This is because antisense RNA molecules for the retrovirus are synthesized (as part of the intervening sequences) whenever the resident genes are transcribed. These antisense molecules could anneal to the genomic RNA and transcripts of the retrovirus and lead to their degradation (Dangel et al., 1994).

The size of human \( RPI \) transcript was estimated to be about 1.6 kb in size. The cDNA clone R1.1 yielded a sequence of 1.1 kb (Shen et al., 1994). An additional cDNA
sequence of 520 bp in size was obtained by multiple rounds of RT-PCR using DNA primers based on the genomic sequence at the 5' region of \textit{RP1}. It was suggested that human \textit{RP1} gene contained 9 exons and encoded 364 amino acids. However, it was not anticipated that there would be another gene, \textit{DOM3Z}, located less than 300 bp upstream of \textit{RP1} but in the opposite transcriptional orientation. \textit{DOM3Z} also codes for a ubiquitously expressed transcript of similar size to that of \textit{RP1} (Yang et al., 1998). Screening of the 5' cDNA sequence previously assigned for \textit{RP1} with the EST database yielded entries for \textit{DOM3Z} rather than \textit{RP1}. In addition, the derived protein sequence from this region is not present in mouse \textit{RP1}. Therefore, it is possible that (the major versions of) the human or mouse \textit{RP1} transcript is encoded by 7 exons with an open reading frame for 254 amino acids. If this is the case, the distance between the transcriptional start sites of human \textit{RP1} and \textit{DOM3Z} will be 267 bp, or 616 bp between translational initiation codons of \textit{DOM3Z} and \textit{RP1} (Fig. 6.2).

\textit{RD, SK12W, DOM3Z} and \textit{RP1}, are all present between the factor \textit{B} and the \textit{C4A/SLP} genes (Dangel et al., 1995; Levi-Strauss et al., 1988; Sargent et al., 1994; Shen et al., 1994; Yang et al., 1998). These four genes are arranged in two head-to-head configuration gene pairs, i.e., \textit{RD-SK12W} and \textit{DOM3Z-RP1} (Fig. 6.7). The transcriptional start sites are separated by less than 171 bp for of human \textit{RD-SK12W}, 267 bp between human \textit{DOM3Z} and \textit{RP1}, 302 bp between mouse \textit{RD} and \textit{SK12W}, and 144 bp between mouse \textit{DOM3Z} and \textit{RP1}. The absence of TATA-box sequences at the 5' regulatory regions of these gene pairs in both species, the presence of CpG islands and
the ubiquitous expression are consistent with the hypothesis that these genes have housekeeping functions (Yang et al., 1998).
Fig. 6.1 **Southern blot analyses of DOM3Z genes in (I) primates and (II) mammals.** Genomic DNAs were digested with *Sst* I and a 0.5 kb 3' human DOM3Z genomic DNA was used as a probe. In (I), DNA in lane 1 is from human, lanes 2 to 4 are DNAs from Old World primates, lanes 5 to 7 are DNAs from New World primates. HT-29, human colon carcinoma cell line; COS7, African green monkey; NPC-LC, cotton top tamarin; OMK, owl monkey; DPSO, squirrel monkey.
Fig. 6.2  Genomic DNA sequences linking human DOM3Z to RPI and SKI2W.

Sequences are obtained from previously published RPI genomic sequence (accession no. L26261) and SKI2W genomic sequence (accession no. AF059675), with reference to HLA class III sequence (accession AF019413). Exon-intron structure of DOM3Z is derived by comparing genomic sequence with DOM3Z cDNA sequences. The derived amino acid sequences are shown for Dom3z only. Intergenic and intron sequences are in lower cases. The locations of 5' splice variants for DOM3Z transcripts, cDNA2 and cDNA3 are indicated. Polymorphic residues are placed above sequence for nucleotides, and below for amino acid sequences. Arrows indicate directions of transcription, translation, or orientations of cis-acting DNA regulatory motifs. An asterisk represents a stop codon. Nucleotide sequences for regulatory motifs are italicized.
Fig. 6.2 Genomic DNA sequences linking human *DOM3Z* to *RPI1* and *SKI2W*.
Fig. 6.3 Exon-intron structure and alternative splicings of human and mouse DOM3Z. I. Exon-intron structure of the human DOM3Z (Yang et al., 1998); II and III, alternative splicings to generate different 5' untranslated regions of human and mouse DOM3Z transcripts, respectively.
Fig. 6.4  *An alignment of human and mouse Dom3z protein sequences.* The human Dom3z sequence is shown. The mouse amino acids identical to those of human are represented by *dash lines*. Dissimilar sequences are shown at the corresponding positions, with conserved residues in *upper case* and unconserved residues in *lower case*. Polymorphic residues are shown above for the human sequence, and below for the mouse sequence. *Hosa, Homo sapiens; Mumu, Mus musculus.*
Fig. 6.5  **cDNA and derived amino acid sequences of mouse RP1.** The first 113 bp is obtained from EST clone AA036570 and the rest from cDNA clone MRP1.0. The bipartite nuclear localization signal for RP1 protein is *underlined*. The location of exon-intron boundaries in the RP1 gene is marked by *vertical strokes*. Mutant sequences at the "coding region" of mouse RP2 are shown *above* the DNA sequence; deletions in RP2 are marked by a *hyphen*; insertion is marked by ^ and the inserted nucleotide(s).
Fig. 6.5  cDNA and derived amino acid sequences of mouse RP1.
I. HUMAN

II. MOUSE

Fig. 6.6 A comparison of the exon-intron structures and gene duplications for human (I) and mouse (II) RPI and RP2. The coding exons are in solid, filled boxes; non-coding exons are in empty boxes. Retroelements are shown as gray, shaded boxes. Horizontal arrows represent configurations of genes or retroelements. Vertical arrows indicate the positions of RPI/RP2 gene duplications.
Fig. 6.7 A comparison of the gene organizations of I. human and II. mouse MHC complement gene cluster. Arrows represent transcriptional orientations of functional genes. Pseudogenes and gene segments are shaded.
### Table 6.1 A comparison of human and mouse DOM3Z exon-intron structures

<table>
<thead>
<tr>
<th></th>
<th>exon</th>
<th>intron</th>
<th>intron phase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(human/mouse, bp)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>123/135</td>
<td>247/176</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>362/362</td>
<td>172/146</td>
<td>2/2</td>
</tr>
<tr>
<td>3</td>
<td>236/236</td>
<td>86/89</td>
<td>1/1</td>
</tr>
<tr>
<td>4</td>
<td>220/220</td>
<td>127/102</td>
<td>2/2</td>
</tr>
<tr>
<td>5</td>
<td>136/136</td>
<td>139/129</td>
<td>0/0</td>
</tr>
<tr>
<td>6</td>
<td>95/95</td>
<td>84/79</td>
<td>2/2</td>
</tr>
<tr>
<td>7</td>
<td>214/199</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
7.1 ABSTRACT

Insulin-dependent (type 1) diabetes mellitus (IDDM) is an immune-mediated multifactorial autoimmune disease. A genome wide search for IDDM susceptibility genes indicated that the major histocompatibility complex (MHC) is the most important locus of the disease association. Besides class II genes \( DR \) and \( DQ \), possible candidates also include \( C4 \) and some other poorly defined polymorphic genes in the class III region. Variations of the genetic unit RCCX module, consisting of \( C4 \) and neighboring genes \( RP1, CYP21 \) and \( TNX \), can result in the genetic instability of the class III region, and consequently, lead to disease associations. It has been discovered that in the diabetic patient population (\( n=50 \)), the frequency of monomodular RCCX structures (42%) is significantly higher than that in the normal population (16.2%, \( n=150 \)). In the disease group, 68% of the patients have at least one chromosome with monomodular RCCX,
compared with 30.9% in the control group. In addition, 11 C4B3 alleles were identified in the IDDM population, which is significantly higher than in the normals (1 in 150 patients). In one of the patients, an abnormal C4B3, which is associated with Rg1 was identified. Determination of the genomic sequence of the C4d region from this patient confirmed that it encodes Rg1 antigenic determinant, with the C4B specific amino acids at the isotypic region. Two ubiquitously expressed genes, SKI2W and DOM3Z, located upstream of C4, were screened for possible polymorphism in IDDM patients. A Taq I RFLP and an Nla IV RFLP in SKI2W were detected. The significance of these polymorphisms has yet to be determined. This study further consolidates the association of IDDM with the MHC class III region.
7.2 INTRODUCTION

Insulin dependent (type I) diabetes mellitus (IDDM) is an immune-mediated, multi-genetic autoimmune disease with worldwide incidence. Environmental factors are also necessary to precipitate the disease. IDDM is characterized by destruction of the insulin-secreting β cells of the pancreas. Although IDDM has become one of the most extensively studied polygenic autoimmune diseases, the fundamental genetic defects leading to IDDM still remain undetermined (Davidson, 1998).

A genome wide search for human type 1 diabetes susceptibility genes indicated that the major histocompatibility complex (MHC) is the major locus for the disease association (Davies et al., 1994). The human MHC is also known as the HLA. Individuals with a combination of HLA class II antigens DR3 and DR4 are found to have higher risk of IDDM (Harrison et al., 1990; Svejgaard and Ryder, 1981; Wolf et al., 1983). It has also been suggested that mutation at position 57 of HLA DQ-β is strongly associated with IDDM (Nepom et al., 1987; Todd et al., 1987). However, the onset of IDDM cannot be fully explained by such a mutation. Therefore, the presence of susceptibility genes in the MHC other than the class II molecules cannot be ruled out. It is likely that the specific alleles of the HLA DR or DQ are markers for the presence of nearby susceptibility genes or elements for IDDM.

In addition to HLA class II, possible candidates include C4 and some other poorly defined polymorphic genes in the class III region (Degli-Esposti et al., 1992). Several
groups have reported abnormalities of the complement system in type I diabetic patients (Jenhani et al., 1992; Marcelli-Barge et al., 1990), including low levels of plasma C4 that may be due to hypercatabolism or to a reduction in protein synthesis (Jacob et al., 1986; Senaldi et al., 1988). Alternatively, the increased frequency of C4 gene deletion may be responsible (Caplen et al., 1990). Another C4 allele that has been reported to be associated with IDDM is C4B3 (Bertrams et al., 1984; Thomsen et al., 1988). Whether these associations of C4 with IDDM are actually a consequence of linkage disequilibrium with the DR and DQ loci or indeed result from a primary association with the disease itself is unclear. It is also possible that certain C4 alleles serve as markers for the real susceptibility gene(s) nearby, as it has already been suggested that HLA DR3/4 heterozygous individuals may have a second susceptibility gene that maps close to the C4 locus (Caplen et al., 1990; Thomsen et al., 1988).

**SKI2W** and **DOM3Z** are located between Bf and C4 (Yang et al., 1998). Both genes are ubiquitously expressed but the levels of transcripts are significantly higher in pancreas, the organ producing insulin. The exact functions of these two genes have not been determined yet. **SKI2W** belongs to the helicase gene family (Dangel et al., 1995a), which have been suggested to be associated with many inherited human disorders, including diabetes (Ellis, 1997). It is of interest to investigate whether there are any abnormalities of these genes in the IDDM patients.

In order to elucidate the susceptibility genes or genetic elements in the MHC class III region associated with IDDM, C4 gene deletions and RCCX modular variations were examined. The two housekeeping genes **SKI2W** and **DOM3Z** were also explored to hunt
for any possible mutations in the IDDM patient samples. The studies conducted here will help pave the way to better understanding on the genetic basis of IDDM.
7.3 EXPERIMENTAL PROCEDURES

*Patient recruitment* — Fifty juvenile onset IDDM patients were recruited from the Columbus Children’s Hospital for this study. Five to Fifteen milliliters of peripheral blood was taken from each person based on the patient’s age. Appropriate consents were obtained according to approved protocols by the Institutional Board of the Columbus Children’s Hospital.

*Isolation of genomic DNA, total RNA and plasma* — Genomic DNAs were isolated from 3 ml of peripheral blood using DNA isolation kit (Gentra System, Inc., Minneapolis, MN). Four ml of blood was subjected to Ficoll gradient centrifugation to separate white blood cells, plasma and red blood cells. Total RNAs were isolated from white blood cells using RNAzolB kit (Tel-Test, Inc., Friendswood, TX) after the manufacturer’s instruction. One-fourth of the white blood cells were used to create an immortalized cell line for each patient sample.

*Establishment of immortalized cell lines* — B lymphoblastoid cell lines were developed for each patient by transformation of B lymphocytes in the white blood cells using Epstein Barr viruses (EBV)(Walls and Crawford, 1987). Basically, white blood cells were incubated in 24-well cell culture plates with EBV, which are in the dilutions of
10^1, 10^2, 10^3, and 10^4 in RPMI 1640 medium. The cells were fed weekly until the clumps of the transformed B lymphoblastoid cells were observed. Afterwards, cells were transferred to T25 flasks and the cultures were expanded.

**Southern blot analysis** — Ten micrograms of genomic DNA were digested to completion with an appropriate restriction enzyme for 16 hrs, resolved on an agarose gel, blotted and hybridized with an appropriate [α-^{32}P] dCTP-labeled probe. Probes used in this study include RP1.1; C4 Pa and P. CYP21, TX (Yang et al., 1999), DOM3Z and SKI2W (Yang et al., 1998).

**Complement C4 allotyping** — Complement C4A and C4B allotypes from EDTA-blood plasma were determined as described (Awedeh and Alper, 1980; Sim and Cross, 1986). Briefly, 10 µl of plasma was digested with 0.1 U of neuraminidase (Sigma, St Louis, MO) at 4°C overnight, and with 0.1 U of carboxyl peptidase B (Sigma) at room temperature for 30 min. Plasma proteins were resolved by high voltage gel electrophoresis (Schneider and Rittner, 1997). C4 proteins were detected by immunofixation using goat antisera against human C4 (Instar, Stillwater, MN). The association with Ch1 or Rg1 antigenic determinants were examined by immunoblot analyses using anti-Ch1 or anti-Rg1 monoclonal antibodies (anti-C4B, cat no. C057-325.2, lot no. 120287; anti-Rg1, RGd1; kindly provided by Dr. Joann M. Moulds, Houston, TX) at a dilution of 1 in 5000 and 1 in 1000, respectively. Immune complexes were detected by chemiluminescence method using the ECL-plus reagents (Amersham, Piscataway, NJ).
Oligonucleotides — Oligonucleotides for amplifying and sequencing the C4d region were synthesized by an Applied Biosystem Model 380B DNA Synthesis machine. The sequences of the primers are: C4E22.5 GAA GGG GCC ATC CAT AGA GA; C4E25.3 CAG GTG CTG CTG TCC CGT GA; C4E26.5 GCT CAC AGC CTT TGT GTT GAA; C4E27.3 CAC TCT CTG CTT CAA TGG CT; C4E28.5 GAA GCC TCC ATC TCA AAG GC; C4E29.3 TTG GGT ACT GGG GAA TGG GG; C4E30.5 GAC ACG GTG ATT GCCC CTG GA; and C4E31.3 CTT CAG GGT TCC TTT GCT GT.

Genomic PCR of the C4d region — To ensure the accuracy of the sequence, PCR of genomic DNAs were performed utilizing the Expand™ high fidelity PCR system (Roche Molecular Biochemicals, Indianapolis, IN) following manufacturer’s protocol. The PCR products were purified and cloned into the pCR2.1 vector (Invitrogen, Carlsbad, CA). C4E22.5 and C4E31.3 are the primers used for the amplification.

Allotype specific PCR — Clones containing the C4B-specific C4d region were identified by allotype specific PCR, which were performed using allotype specific primers described in (Barba et al., 1994). Clones that give the product of expected size were chosen for further analysis.

DNA Sequencing and Sequence Analysis — Sequences of the C4d region were determined by automated sequencing method using an ABI 377 machine. Sequence
comparisons were performed by FASTA program in GCG package (Genetics Computer Group, 1991) through the Pittsburgh Supercomputing Center.
7.4 RESULTS

**C4 gene deletions and RCCX modular variations in IDDM patients**

Low serum concentrations of C4 are found in about one-fourth of all IDDM patients, and may be important in the etiology of the disease (Jacob et al., 1986). One direct cause for low protein levels is the gene deletion. Therefore, to determine the role of \( C4A \) and \( C4B \) gene deletions, as well as the RCCX modular variations in IDDM patients, genomic \( Taq \ I, Nla \ IV, \) and \( EcoO \ 109I \) Southern blot analyses were performed according to the established techniques described in Chapter 2 (Yang et al., 1999).

The RCCX modular structures of the 50 IDDM patients studied were listed in Table 7.1. Among the chromosome 6 analyzed, only 50% have bimodular RCCX structure (33% are with two \( C4 \) long genes L-L, 17% are with one long and one short \( C4 \) gene L-S), while 42% have monomodular RCCX, and 8% have trimodular RCCX. In comparison, 71.6% of the chromosome 6 from non-IDDM population (n=150) have bimodular RCCX (47.7% L-L, 23.8% L-S), only 16.2% have monomodular RCCX, and the other 12.2% have trimodular RCCX (Fig. 7.1, panels I and II). The extensive deletion of \( C4 \) genes can also be demonstrated by the even more significant fact that in the diseased population, only 32% have bimodular and/or trimodular RCCX structures on both copies of chromosome 6, or bimodular RCCX on one chromosome and trimodular RCCX on the other. In comparison, the frequency is 69.1% in the non-IDDM
population. In other words, 68% of the IDDM patients have a deletion of at least one $C_4$ and its neighboring genes in the RCCX.

An intriguing aspect on the RCCX modular variations among the 50 IDDM patients is that, there is no single patient who has homozygous, bimodular RCCX structures with two $C_4$ long genes (L-L / L-L). In the non-IDDM population, the frequency of homozygous L-L / L-L RCCX is 19.5%. Since $C_4$ long gene is due to the presence of the endogenous retrovirus HERV-K(C4), the significantly lower percentage of $C_4$ long gene in the diseased population suggests that this endogenous retrovirus in $C_4$ gene could have a protective function against IDDM. The results of these comparisons are summarized in Table 7.2.

$C_4$ protein variations in IDDM patients

Extensive $C_4$ protein polymorphisms have been demonstrated in the IDDM population (Jenhani et al., 1992), and various allotypes of $C_4A$ and $C_4B$ are determined by immunofixation and immunoblot analyses. The allotyping results for all 50 IDDM patients studied are listed in Table 7.1. A few examples are presented in Fig. 7.2. Deduced from the immunofixation experiments (panel I), patient 46 in lane 1 appears to have normal $C_4A3 C_4B1$; patient 20 (lane 2) is homozygous for $C_4A3 C_4BQ0$; and patient 24 (lane 3) is homozygous for $C_4AQ0 C_4B1$. Patient 11 (lane 4) has $C_4A3 C_4B2, C_4A3 C_4BQ0$. Patients 13 and 14 (lanes 5 and 6) both have total $C_4A$ deficiency, with $C_4AQ0 C_4B1, C_4AQ0 C_4B2$ and $C_4AQ0 C_4B1, C_4AQ0 C_4B3$, respectively. Both
patients 4 and 15 (lanes 7 and 8) exhibit similar allotyping pattern, and are assigned to be C4A3 C4BQ0, C4A3 C4B3. Patient 28 (lane 9) shows C4AQ0 C4B1, C4A3 C4B6.

In immunoblot analyses, C4B1 in lanes 1, 3, 5, 6, and 9 react with anti-Ch1 monoclonal antibody (panel II). C4B2 in lanes 4 and 5, C4B3 in lanes 6 and 7, as well as C4B6 in lane 9 also show positive reactions with this anti-Ch1 antibody. Nothing was detected in lane 2 since this patient is homozygous for C4BQ0. Unexpectedly, nothing was detected in lane 8 either, although the patient has C4B3 according to the immunofixation experiment. It turned out that this C4B3 positively reacts with anti-Rg1 monoclonal antibody (panel III, lane 8). Panel III also confirms the presence of C4A3 associated with Rg1 in lanes 1, 2, 4, 7, 8 and 9. No signal was detected in lanes 3, 5 and 6 because all three patients are homozygous for C4AQ0.

Among the 50 patients examined, 11 have C4B3 identified. In other words, 11 C4B3 alleles were present on the 100 chromosomes in the patient population. This frequency is significantly higher than that in the normals (1 in 300 chromosomes).

Illustration of Rg1-associated C4B3 sequence

As described above, judging from the mobility, one of the C4 proteins in patient 15 was assigned as C4B3. Intriguingly, it positively reacts with anti-Rg1 instead of anti-Ch1 antibody. The association of C4B3 with Rg1 has never been reported before. To definitively prove the reverse association, genomic PCR was performed to amplify the C4d region of this patient. The PCR products were cloned and the sequence of the C4B3-specific C4d region was determined. The sequence comparison of Rg1-associated C4B3
with C4A was presented in Fig. 7.3. At nucleotide 1119, 1122, 1130, 1133, and 1135, the cloned genomic PCR product obviously contains C4B specific sequence, and encodes for LSPVIH, the C4B isotypic amino acids. However, unlike most other C4B, which are associated with Ch1, this C4B3 in an IDDM patient encodes for Rg1 antigenic determinant characterized by VDLL at amino acid position 1188-1191. There are other nucleotide changes observed in the sequence, some of which cause amino acid substitutions. For example, the C → T change at nucleotide 802 leads to amino acid Ala 1049 → Val substitution; the A → G change at nucleotide 817 results in a Asp 1054 → Gly substitution.

Screening for possible polymorphism of DOM3Z and SKI2W in IDDM patients

As C4 is possibly a marker for the presence of real disease susceptibility gene(s) in close proximity, DOM3Z and SKI2W, two novel genes located between C4 and Bf, were studied for any possible RFLP associated with IDDM.

Genomic Taq I, Nla IV, and EcoO 109I Southern blots used for determining RCCX modular structures were hybridized with DOM3Z and SKI2W probes separately. No RFLP was observed in EcoO 109I blots when hybridized with either probe. When hybridized with DOM3Z probe, in addition to the 2.7 kb Taq I fragment, patient no. 3 shows an extra fragment with the size about 1.8 kb (Fig. 7.4, panel I-A, lane 4). Sequence analysis indicates that the Taq I polymorphism may actually lie in SKI2W, the gene in tail-to-tail configuration with DOM3Z. This was indeed the case as confirmed by
the fact that when hybridized with SKI2W probe, the 1.8 kb extra band was also detected in the same patient (Fig. 7.4, panel I-B, lane 4).

In Nla IV blots, no RFLP was detected in DOM3Z gene. However, when SKI2W was used as the probe, four IDDM patients (no. 28, 38, 41 and 49) exhibited the same abnormal restriction digestion pattern (Fig. 7.4, panel II, lanes 1-4).
MHC is the most significant genetic factor that is associated with the pathogenesis of IDDM [reviewed in (Tisch and McDevitt, 1996)]. At the same time, environmental factors, one of which is viral infection, are also important for the disease development (Conrad et al., 1994; Yoon, 1995). There have been reports describing the associations of the MHC class III region with IDDM, including increased frequency of the C4A null allele (Caplen et al., 1990; Mijovic et al., 1985). In this chapter, preliminary studies have been presented in searching for any possible effect of the C4 gene size and gene number on IDDM. It appears that in the diseased population, C4 gene deletions occur much more frequently than in the non-IDDM population. Gene deletion is the most direct cause for the low C4 protein levels in the serum, which are found in about 25% of all IDDM patients (Senaldi et al., 1988). The association between C4 null alleles and IDDM could be due to the linkage disequilibrium with susceptibility loci HLA DR or DQ. Alternatively, a deficiency of C4 could directly predispose to the disease by impairing the efficiency of the complement system. As a major player in the immune response, C4 has a key role in virus neutralization and its deficiency might enable a virus to initiate the process leading to IDDM (Senaldi et al., 1988).

There has been some controversial demonstration suggesting that endogenous retroviruses can code for superantigens that would lead to the activation of β cell
destructions in the pancreas (Conrad et al., 1994; Conrad et al., 1997; Murphy et al., 1998). However, in this study, the statistics is significant that the frequency of C4 long genes present is much lower in the IDDM population than in the normal, which indicates the possible protective role of HERV-K(C4) against IDDM. This endogenous retrovirus exists in intron 9 of C4 long gene with the opposite transcription direction (Dangel et al., 1995a). Therefore, antisense viral RNA is synthesized during the C4 transcription. Since the antisense product is complementary to the viral genomic RNA, it can consequently block the translation of viral proteins and limit the production of superantigens. Eventually this may reduce the chance for the induction of IDDM by viral superantigens (Yu, 1999).

DOM3Z and SKI2W, which are located upstream of C4, were also studied for possible mutation(s) in the IDDM population. There is no DOM3Z RFLP observed in the patient population. This result is consistent with the notion that Dom3z plays an essential role in a biochemical pathway (Yang et al., 1998). Dom3z may have such fundamental function that any mutation could possibly have very severe consequences or even be lethal. There are some SKI2W RFLPs detected in some patients, whether they have any impact on the disease, or are just some examples of the single nucleotide polymorphism (SNP) with no obvious phenotype by themselves, still need to be determined.

Attentions need to be paid to two important aspects: first, many MHC loci are in linkage disequilibrium with one another (Tomlinson and Bodmer, 1995). It is therefore extremely difficult to distinguish the primary "disease allele" from the associated allele. From this point of view, HLA typing becomes so necessary that any conclusion drawn
about the IDDM "disease gene" before combining the typing data would seem premature. The HLA typing of the 50 IDDM patients are being conducted as this chapter is written. Secondly, for any polygenic disease like IDDM, the sample sizes of the diseased population as well as the well-matched control studied need to be large enough for any valid conclusion. Too often, false positive can be observed in a small subject group, and it needs to be aware that this project also faces the same potential limitation.

Recently, there is another interesting development on the IDDM and human MHC genetics: the mapping of allograft inflammatory factor 1 (AIF1) in the class III region. AIF1 appears to inhibit insulin secretion at low glucose concentrations and stimulate insulin secretion at high glucose concentration (Chen et al., 1997). It would be informative to investigate whether there is any mutation in AIF1 gene or any variation in the gene expression in the IDDM population. Meanwhile, it has been shown extensively that MHC molecules are related to each other functionally and evolutionarily (Salter-Cid and Flajnik, 1995). Thus, it is possible that other genetic factors or elements located in the MHC class III region may also be involved in the secretion of insulin or destruction of the β cells in the pancreas. Take in consideration of the high RD-SKI2W-DOM3Z-RP1 transcription levels in the pancreas, these four genes remain attractive of our attention for further exploration.
RCCX modular structures

IDDM     normals
bi-     50     71.6
mono-   42     16.2

Bimodal RCCX

IDDM     normals
L-L     33     47.7
L-S     17     23.8

Fig. 7.1  Comparisons of C4 genes and RCCX modular structures between chromosomes from the IDDM (n=100) and non-IDDM (n=300) populations. 

n is the number of chromosomes studied.
I. Immunofixation

Patient # 46 20 24 11 13 14 4 15 28

II. Immunoblot (anti-Ch1)

III. Immunoblot (anti-Rg1)

Fig. 7.2  **C4 allotypes of selected IDDM patients.**  *I.* Human C4A and C4B allotypes detected by immunofixation; Immunoblot analysis of the C4 allotypes using: *II.* anti-Ch1 monoclonal antibody; and *III.* anti-Rg1 monoclonal antibody.
Fig. 7.3  **Sequence comparison of the C4d region from Rg1-associated C4B3 and C4A.** The C4A sequence was taken from Yu, 1991 and used as the standard. The coding regions were shown in *uppercase* and the introns in *lowercase*. The isotypic region and the antigenic determinant region were *underlined*, with the amino acid sequences *bolded* or *bolded* and *italicized*, respectively. The nucleotide changes were labeled *above* the sequence and the amino acid changes *below*. Insertions and deletions of nucleotide were marked.
Fig. 7.3  Sequence comparison of the C4d region from Rg1-associated C4B3 and C4A.
Fig. 7.4  RFLPs of *SKI2W* in IDDM patients. *I.* Taq I RFLP detected in one IDDM patient (*lane* 4) using A. *DOM3Z* or B. *SKI2W* probe. *II.* *Nla* IV RFLP detected in four IDDM patients (*lanes* 1-4) using a *SKI2W* probe.
<table>
<thead>
<tr>
<th>Patient #</th>
<th>C4A</th>
<th>C4B</th>
<th>allotypes</th>
<th>Rg1</th>
<th>Ch1</th>
<th>RP - C4</th>
<th>21A/21B</th>
<th>XA/XB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>A3 B1 A3 BQ0</td>
<td>2</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L-RP1-C4L</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>A3 BQ0 A3 B1</td>
<td>1</td>
<td>1</td>
<td>RP1-C4L-RP1-C4S</td>
<td>21B</td>
<td>XB</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>A3 BQ0 A3 B1</td>
<td>1</td>
<td>1</td>
<td>RP1-C4L-RP1-C4S</td>
<td>21B</td>
<td>XB</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>A3 B3 A3 BQ0</td>
<td>2</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L-RP1-C4L</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
<td>A3 B1 A3 BQ0</td>
<td>2</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L-RP1-C4L</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>A3 B3 A3 BQ0</td>
<td>1</td>
<td>2</td>
<td>RP1-C4L-RP2-C4L-RP1-C4S</td>
<td>1:2*</td>
<td>1:1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1</td>
<td>A3 B1 A3 BQ0</td>
<td>2</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L-RP1-C4L</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>1</td>
<td>A3 B1 A3 BQ0</td>
<td>2</td>
<td>1</td>
<td>RP1-C4L-RP2-C4S-RP1-C4L</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>2</td>
<td>3A3 2B1</td>
<td>3</td>
<td>2</td>
<td>RP1-C4L-RP2-C4L-RP2-C4L-RP1-C4L-RP2-C4L-RP1-C4L-RP2-C4S</td>
<td>3:2</td>
<td>3:2</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>2</td>
<td>A3 B1 A3 BQ0</td>
<td>1</td>
<td>2</td>
<td>RP1-C4L-RP2-C4L-RP1-C4S</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>1</td>
<td>A3 B2 A3 BQ0</td>
<td>2</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L-RP1-C4S</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>3</td>
<td>A3 B1 A3 B1B1</td>
<td>2</td>
<td>3</td>
<td>RP1-C4L-RP2-C4L-RP2-C4S-RP2-C4S-RP1-C4L-RP2-C4S-RP1-C4L-RP2-C4S</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>2</td>
<td>AQ0 B2 AQ0 B1</td>
<td>0</td>
<td>2</td>
<td>RP1-C4L-RP1-C4S</td>
<td>21B</td>
<td>XB</td>
</tr>
</tbody>
</table>

(to be continued)

Table 7.1 RCCX modular structures and C4 allotypes for 50 IDDM patients
(Table 7.1 continued)

<table>
<thead>
<tr>
<th>Patient #</th>
<th>C4A</th>
<th>C4B</th>
<th>allotypes</th>
<th>Rg1</th>
<th>Ch1</th>
<th>RP - C4</th>
<th>21A/21B</th>
<th>XA/XB</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>0</td>
<td>2</td>
<td>AQ0 B3 AQ0 B1</td>
<td>0</td>
<td>2</td>
<td>RP1-C4L RP1-C4S</td>
<td>21B</td>
<td>XB</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>1</td>
<td>A3 B3 (Rg1+) A3 BQ0</td>
<td>3</td>
<td>0</td>
<td>RP1-C4L-RP2-C4L RP1-C4L</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>3</td>
<td>A3 A2 2B1 1B3</td>
<td>2</td>
<td>3</td>
<td>RP1-C4L-RP2-C4L RP1-C4L-RP2-C4S RP1-C4L</td>
<td>3:2</td>
<td>2:3</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>1</td>
<td>A3 A2 AQ0 B1</td>
<td>2</td>
<td>1</td>
<td>RP1-C4L-RP2-C4S RP1-C4L</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>2</td>
<td>A3 B1 AQ0 B1</td>
<td>1</td>
<td>2</td>
<td>RP1-C4L-RP2-C4S RP1-C4S</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>1</td>
<td>A3 A3 A3 B2</td>
<td>3</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L RP1-C4L-RP2-C4S RP1-C4L</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>0</td>
<td>A3 A3 A3 A3</td>
<td>5</td>
<td>0</td>
<td>RP1-C4L-RP2-C4L RP1-C4L-RP2-C4S-RP2-C4L</td>
<td>3:2</td>
<td>3:2</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>1</td>
<td>A3 BQ0 AQ0 B1</td>
<td>1</td>
<td>1</td>
<td>RP1-C4L RP1-C4S</td>
<td>21B</td>
<td>XB</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>1</td>
<td>A3 B1 A3 BQ0</td>
<td>2</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L RP1-C4L</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>2</td>
<td>A3 B1 AQ0 B1</td>
<td>1</td>
<td>2</td>
<td>RP1-C4L-RP2-C4L RP1-C4S</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>2</td>
<td>AQ0 B1 AQ0 B1</td>
<td>0</td>
<td>2</td>
<td>RP1-C4S RP1-C4S</td>
<td>21B</td>
<td>XB</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>1</td>
<td>A3 B1 A3 B1</td>
<td>1</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L RP1-C4L-RP2-C4S RP1-C4L</td>
<td>1:1</td>
<td>1:2</td>
</tr>
</tbody>
</table>

(to be continued)
(Table 7.1 continued)

<table>
<thead>
<tr>
<th>Patient #</th>
<th>C4A</th>
<th>C4B</th>
<th>allotypes</th>
<th>Rg1</th>
<th>Ch1</th>
<th>RP - C4</th>
<th>21A/21B</th>
<th>XA/XB</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>1</td>
<td>2</td>
<td>A3 B1</td>
<td>1</td>
<td>2</td>
<td>RP1-C4L-RP2-C4L</td>
<td>1:2</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>1</td>
<td>2</td>
<td>A3 B6</td>
<td>1</td>
<td>2</td>
<td>RP1-C4L-RP2-C4S</td>
<td>1:2</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>0</td>
<td>2</td>
<td>A3 B6</td>
<td>0</td>
<td>2</td>
<td>RP1-C4S</td>
<td></td>
<td>21B</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RP1-C4S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>2</td>
<td>A3 B3</td>
<td>1</td>
<td>2</td>
<td>RP1-C4L-RP2-C4L</td>
<td>1:2</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>2</td>
<td>1</td>
<td>A3 A2</td>
<td>2</td>
<td>1</td>
<td>RP1-C4L-RP2-C4S</td>
<td>1:2</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>1</td>
<td>2</td>
<td>A3 B3</td>
<td>1</td>
<td>2</td>
<td>RP1-C4L-RP2-C4L</td>
<td>1:2</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>2</td>
<td>1</td>
<td>A3 A3</td>
<td>2</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L</td>
<td>1:2</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>4</td>
<td>1</td>
<td>A3 A3</td>
<td>4</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L</td>
<td>3:2</td>
<td>3:2</td>
</tr>
<tr>
<td>35 (?)</td>
<td>1</td>
<td>1</td>
<td>A3 B3</td>
<td>1</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L</td>
<td>2:1</td>
<td>1:2</td>
</tr>
<tr>
<td>36</td>
<td>3</td>
<td>0</td>
<td>A3 A2</td>
<td>3</td>
<td>0</td>
<td>RP1-C4L-RP2-C4L</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>37</td>
<td>2</td>
<td>1</td>
<td>A3 B1</td>
<td>2</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>38</td>
<td>2</td>
<td>3</td>
<td>2A3 B1</td>
<td>2</td>
<td>3</td>
<td>RP1-C4L-RP2-C4S-RP2-C4S</td>
<td>3:2</td>
<td>3:2</td>
</tr>
<tr>
<td>39</td>
<td>2</td>
<td>1</td>
<td>A3 B3</td>
<td>2</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L</td>
<td>1:2</td>
<td>1:2</td>
</tr>
</tbody>
</table>

(to be continued)
(Table 7.1 continued)

<table>
<thead>
<tr>
<th>Patient #</th>
<th>C4A</th>
<th>C4B</th>
<th>allotypes</th>
<th>Rg1</th>
<th>Ch1</th>
<th>RP - C4</th>
<th>21A/21B</th>
<th>XA/XB</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>2</td>
<td>1</td>
<td>A2 A3 AQ0 B1</td>
<td>2</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L-RP1-C4S</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>41</td>
<td>0</td>
<td>2</td>
<td>AQ0 B1 AQ0 B1</td>
<td>0</td>
<td>2</td>
<td>RP1-C4S-RP1-C4S</td>
<td>21B</td>
<td>XB</td>
</tr>
<tr>
<td>42</td>
<td>1</td>
<td>1</td>
<td>A3 B3 A3 B1</td>
<td>1</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L-RP1-C4S</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>43</td>
<td>2</td>
<td>1</td>
<td>A3 B1 A2 BQ0</td>
<td>2</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L-RP1-C4L</td>
<td>1:2</td>
<td>1:2</td>
</tr>
<tr>
<td>44</td>
<td>3</td>
<td>1</td>
<td>A3 A3 A3 B1</td>
<td>3</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L-RP1-C4L-RP2-C4S</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>45</td>
<td>1</td>
<td>1</td>
<td>A3 B1 A2 B1</td>
<td>1</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L-RP1-C4L-RP2-C4S</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>46</td>
<td>1</td>
<td>1</td>
<td>A3 B1 A3 B1</td>
<td>1</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L-RP1-C4L-RP2-C4S</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>48</td>
<td>1</td>
<td>1</td>
<td>A3 AQ0 B1 B2 A- mutation?</td>
<td>1</td>
<td>2</td>
<td>RP1-C4L-RP2-C4L-RP1-C4L-RP2-C4S</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>49</td>
<td>1</td>
<td>1</td>
<td>A3 B1 A3 B1</td>
<td>1</td>
<td>1</td>
<td>RP1-C4L-RP2-C4L-RP1-C4L-RP2-C4S</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>2</td>
<td>A3 B3 AQ0 B1</td>
<td>1</td>
<td>2</td>
<td>RP1-C4L-RP2-C4L-RP1-C4S</td>
<td>1:2</td>
<td>1:2</td>
</tr>
</tbody>
</table>
I. A comparison of the RCCX structures and $C4$ genes between chromosomes of IDDM and non-IDDM population.

<table>
<thead>
<tr>
<th></th>
<th>IDDM (n*=100)</th>
<th>non-IDDM (n=300)</th>
<th>$p$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bimodular RCCX</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$C4$ L-L</td>
<td>50</td>
<td>71.6</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>$C4$ L-S</td>
<td>33</td>
<td>47.7</td>
<td>0.012</td>
</tr>
<tr>
<td>Monomodular RCCX</td>
<td>17</td>
<td>23.8</td>
<td>0.163</td>
</tr>
<tr>
<td>Trimodular RCCX</td>
<td>42</td>
<td>16.2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>$C4B3$</td>
<td>8</td>
<td>12.2</td>
<td>0.305</td>
</tr>
</tbody>
</table>

*n: number of chromosomes studied.

II. A comparison of individuals with bi-/bi- (tri-) modular RCCX and homozygous $C4$ L-L/L-L between IDDM and non-IDDM population.

<table>
<thead>
<tr>
<th></th>
<th>IDDM (n*=50)</th>
<th>non-IDDM (n=150)</th>
<th>$p$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>bi-/bi- and bi-/tri-</td>
<td>16</td>
<td>103</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>$C4$ L-L/L-L</td>
<td>0</td>
<td>58</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

*n: number of individuals studied.

*Table 7.2* RCCX modular variations in IDDM patients
THE MHC, as one of the most important genomic regions associated with many human diseases, has received a lot of attention. Over the last two decades, enormous amounts of research conducted has indeed made MHC one of the most extensively studied regions of the human genome. Individuals at high risk to develop autoimmune and other diseases would be most informative for investigation of the basis of MHC disease associations.

In particular, class III region has become a rising star in this area, partly due to many new genes identified, which have turned it to be one of the most gene-dense regions. Of the human genome, whose size is estimated at $3 \times 10^9$, only 2-3% encodes proteins. In contrast, in the 1,100 kb class III region, there is a gene every 10-15 kb (Aguado et al., 1996; Yang et al., 1998). One of the mechanisms to achieve such gene density is by gene duplication.

Duplications of large regions of DNA, including duplications of whole genes, provide substrates for genetic evolution. As analyzed by Ohno more than 30 years ago
(Ohno, 1970), these duplication events liberate copies of the gene to diverge and take up new functional roles in the organism, while the master gene is constrained to preserve its original role. On the other hand, duplications of smaller regions involving parts of genes are now recognized as contributors to the mutation spectrum that result in genetic diseases (Hu and Worton, 1992). Complement component gene C4 has been duplicated in the human genome. While there are a small proportion of the population (16.2%) who have single C4 gene in monomodular RCCX, most of the population (83.8%) have duplicated C4 with two, three or more copies in bi-, tri- or more modular RCCX. The RCCX modular variation is one of the major mechanisms leading to genetic instability of the MHC class III region and disease associations.

A variety of specialized selective pressures, such as the need to increase the expression level, could promote the development of gene duplications. In MHC, gene duplication is one of the mechanisms to generate and maintain polymorphism to confer resistance to a variety of infectious diseases. The C4 molecule, C4A and C4B, are specified by duplicated C4 loci. C4A is primarily involved in promoting the physiological disposal of immune complexes. C4B participates preferentially in the clearance of microorganisms (Law et al., 1984; Yu, 1999). In order to function properly, C4 interacts with at least eight other molecules: the antigen, the antibody, complement factors C2, C3, C5, regulatory proteins factor I, C4b binding protein, and complement receptor type I (CR1) (Porter, 1985a). C4A and C4B differ in two of the eight interactions: with the antigen and with the antibody. The reason for C4 gene duplication could thus be the need to maintain identity of six interaction sites, while allowing
functional differentiation at two other sites. In other words, C4 is under pressure to retain the C4A-C4B difference in the C4d region and identity in the rest of the molecule (Horiuchi et al., 1993).

The other side of the coin is that the same MHC polymorphism created by gene duplications is also the price that some people have to pay through increased susceptibility to genetic and autoimmune diseases (De Vries, 1994). For example, between the duplicated copies of CYP21, CYP21B is functional, while CYP21A is a pseudogene. Individuals with only CYP21A in the genome would suffer CAH, which may have fatal consequences. Therefore, the concerted duplication of C4-CYP21 in RCCX must be occurring because of a strong selection advantage for C4 polymorphisms, with which CYP21 genes accidentally became locked into a permanent partnership by the primigenial duplication (Horiuchi et al., 1993). The parallel C4 gene duplications in mice, and the different inactivation of the CYP21 genes are consistent with this notion.

Another potential problem resulted from the concerted duplication / deletion of the RCCX module is the unequal crossover. Normally, there is a strong selection pressure in natural populations for polymorphism at certain MHC loci. Recombination (i.e., homologous equal crossover) within the MHC shuffles alleles among haplotypes and thus could help to maintain useful polymorphic alleles. However, as RP2 and TNXA in the RCCX module are both partially duplicated gene segments, unequal crossover could occur between RP1 and RP2, or TNXA and TNXB, from homolgous chromosomes or sister chromatids. In chapter 2, it has been demonstrated that in a CAH patient, the improper pairing of TNXB from a monomodular RCCX and TNXA from a bimodular
RCCX has led to the unequal crossover, and consequently the deletion of \textit{RP2-C4B-CYP21B} genes (Yang \textit{et al}., 1999).

The Williams Syndrome is one of some other human diseases caused by the unequal crossover. It is an autosomal dominant syndrome based on 7q11.23 (OMIM no. 194050). Clinical features of WS include characteristic "elfin" faces, mental retardation, growth deficiency, and hypertension (Morris \textit{et al}. 1988). About 90\% of WS patients are hemizygous for the elastin gene, having deleted the copy from one chromosome (Nickerson \textit{et al}., 1995). The mechanism of the deletion is suggested to involve the improper pairing of homologous sequences and an unequal crossover that loses a 2 Mb intervening DNA (Mazzarella and Schlessinger, 1998). Recent evidence indicates that the homologous sequences may involve the \textit{GTF2I} gene and its pseudogene. The \textit{GTF2I} gene encodes the transcription initiator binding protein TFII-I, and maps to the telomeric breakpoint of the 2 Mb deletion; its pseudogene \textit{GTF2I} maps to the centromeric breakpoint (Perez \textit{et al}., 1998).

Homologous recombination between repeated sequences has also been implicated as the mechanism by which a common deletion is produced in Prader-Willi syndrome (PWS; OMIM no. 176270) and Angelman syndrome (AS; OMIM no. 105830) patients in chromosome 15q11-q13 (Christian \textit{et al}., 1995; Huang \textit{et al}., 1997). Seventy percent of the individuals affected with PWS and AS acquire a 4 Mb deletion in the paternal and maternal genomes, respectively. In addition, this region is also subjected to duplications. The recent construction of a detailed YAC map of the region should help to clarify which
element(s) may be responsible for this chromosomal rearrangement (Christian et al., 1998).

As discussed in Chapter 1, homologous recombination is the major mechanism from which genomic disorders are resulted. This is in sharp contrast to the classical mechanisms of diseases, in which point mutations in a gene specifically affect the structure and expression of the proteins encoded. The introduction of genomic diseases should be credited to the recent advances of the Human Genome Project (HGP) and the completion of total genome sequences for C. elegans, yeast and many bacterial species. It is these progresses that have enabled scientists to view genetic information in the context of the entire genome and recognize that the mechanisms for some diseases are best understood at a genomic level (Lupski, 1998).

With HGP's promise to complete an accurate, high-quality sequence of the human genome by the end of 2003, the door has opened wide to the era of whole genome science (Collins et al., 1998). With the free access to various databases, the process of gene discovery has been revolutionized. The birth of the novel gene DOM3Z is one of the typical examples. After a new gene is identified, it is possible to determine the sequence of the full length transcript and alternative splicing patterns, from which primary sequence and structure of the protein produced can be deduced. It is also possible to detect sequence variations in normal individuals as well as those associated with disease.

However, there will be a compelling need to better understand the function of human genes and their roles in health and disease. Therefore, the development of new research strategies for the post-genome era, even before the entire sequence of human
genome is finished, has become increasingly urgent. At such a turning point of biological sciences, functional genomics, the study of gene function using the tools of the genome program, has emerged. Functional genomics requires the development of analytical strategies that are comprehensive and hierarchical. Comprehensive, because we aim to uncover the action and interaction of all of the genes in a given species. Hierarchical, because this daunting task is only possible if we find ways of grouping genes of related function in order to limit the total number of experiments to be performed (Oliver et al., 1998).

Functional genomics can be best defined as the continuum from a gene's physical structure to the regulation and to its role in the whole organism (Woychik et al., 1998). Understanding of gene regulation include the study of the expression patterns in development and in all adult tissues as well as in tissue undergoing pathogenic changes. The insights into these aspects would likely derive from the development of innovative new technologies for examining gene expression. Before the genome program, it was possible only to study the temporal and spatial expression of individual genes or small groups of cloned genes under a given experimental setting. However, this has changed with the progress of the HGP and the development of new technologies: microarray and DNA chips. For the first time in history, it becomes possible to evaluate the temporal/spatial patterns of expression of all genes of an mammalian organism, both under normal conditions or in response to environmental or genetic changes that ultimately lead to a disease phenotype. Consequently, it comes the ability to decipher the genetic pathways
for normal development and to understand how these pathways are altered in specific disease states.

For many diseases, especially polygenic diseases such as IDDM, it is important to determine all the contributing genetic factors. Therefore, a genome wide search becomes obvious. However, it is extremely difficult to carry out these projects using the traditional analysis as it requires the family studies and together with genetic mapping, both are dreadfully time consuming. The advancement of the microarray technology, in which the expression patterns of all genes from the whole genome can be compared between normal and diseased tissues, can address this problem in a much more accurate and efficient way.

Understanding of the role of genes in the whole organism can be clarified by studying the model systems. The completion of the yeast and C. elegans genomes, together with the ongoing effort to elucidate the Drosophila and mouse genomes will shed lights on the eventual identification of all human genes. Generating mutations in the animal models and investigation on the resulting phenotypes can provide invaluable data for our understanding of the gene function in the whole organism. Coupled with the transgenic and knock-out techniques, animal models can also be used to regenerate phenotypes and confirm discoveries made from genetic studies. Therefore, model systems have tremendous impact on our understanding of disease etiology, and will also help in designing new treatments in the future.

The Human Genome Project, the ultimate foundation on which new technologies are building, new concepts are formulating, new discoveries are making, will have
unprecedented impact and long-lasting value for basic biology, biomedical research, biotechnology, and health care. They will improve our understanding in gene-environment interactions and facilitate the development of highly accurate DNA-based medical diagnostics and therapeutics.
REFERENCES


NOTE TO USERS

Page(s) missing in number only; text follows.
Microfilmed as received.

222-229

This reproduction is the best copy available.

UMI


Chen, Z.W., Ahren, B., Ostenson, C.G., Cintra, A., Bergman, T., Moller, C., Fuxe, K.,
Mutt, V., Jornvall, H. and Efendic, S. (1997). Identification, isolation, and
characterization of daintain (allograft inflammatory factor 1), a macrophage polypeptide
with effects on insulin secretion and abundantly present in the pancreas of prediabetic

of the protein encoded by RD, a gene located in the class III region of the human major


Chiquet-Ehrismann, R. (1993). Tenascin. In "Guidebook to the extracellular matrix and
dhesion proteins" (Kries, T. and Vale, R., Eds.), Oxford University Press, Oxford. pp
93-95.


proximal deletion breakpoint regions in both Prader-Willi and Angelman syndrome

Christian, S.L., Bhatt, N.K., Martin, S.A., Sutcliffe, J.S., Kubota, T., Huang, B.,
YAC contig map of the Prader-Willi/Angelman region on chromosome 15q11-q13 with

recombination site within the steroid 21- hydroxylase gene (CYP21) of the HLA-B47,

Immunogenet.* **12**: 74-81.


Johnson, A.W. (1997). Ratlp and Xrn1p are functionally interchangeable exoribonucleases that are restricted to and required in the nucleus and cytoplasm, respectively. Mol.Cell Biol. 17: 6122-6130.


Muir, A., Ruan, Q.G., Marron, M.P. and She, J.X. (1999). The IDDMK(1,2)22 retrovirus is not detectable in either mRNA or genomic DNA from patients with type 1 diabetes. *Diabetes* **48**: 219-222.


Rowen, L., Qin, S., Lasky, S., Loretz, C., Dors, M., Mahairas, G., and Hood, L. (1998) Mus musculus major histocompatibility locus class III region: complement C4 (C4) and cytochrome P450 hydroxylase A (CYP21OH-A) genes, complete cds; lip pseudogene, complete sequence; NG6, SKI, and complement factor B (Bf) genes, complete cds; and complement factor C2 (C2) gene, partial cds. GenBank accession no. AF049850.


polymorphisms of major histocompatibility complex class III genes. 


251


