THE EVOLUTION OF THE ADAPTIVE IMMUNE SYSTEM

by

HAO CHEN

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF IMMUNOBIOLOGY

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

WITH A MAJOR IN MICROBIOLOGY AND IMMUNOLOGY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2010
As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Hao Chen entitled THE EVOLUTION OF THE ADAPTIVE IMMUNE SYSTEM and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Date: 01/15/2010

Harris Bernstein

Date: 01/15/2010

David Harris

Date: 01/15/2010

Richard Ablin

Date: 01/15/2010

Samuel F. Schluter

Date: 01/15/2010

Janko Nikolich-Zugich

Final approval and acceptance of this dissertation is contingent upon the candidate’s submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Date: 01/15/2010

Dissertation Director: Harris Bernstein
STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the author.

SIGNED: Hao Chen
ACKNOWLEDGEMENTS

I would like to thank my mentors, Dr. Samuel F. Schluter and Dr. John J. Marchalonis, for their guidance, support, and patience. I am honored to have benefited from their expertise. I would also like to thank the members of my committee: Dr. Harris Bernstein, Dr. David Harris, Dr. Richard Ablin, and Dr. Janko Nikolich-Zugich.

I would also like to thank all the members of the Schluter Lab, past and present, whose support, encouragement and advice was invaluable.
DEDICATION

I dedicate this work to my family, my wife Jing Li, my lovely son Jackson Chen, my father Wenliang Chen, and my mother Chenguang Cai, from whom I get my strength and inspiration.
TABLE OF CONTENTS

LIST OF ABBREVIATIONS ...................................................................................... 8
LIST OF FIGURES ................................................................................................. 9
LIST OF TABLES .................................................................................................. 11
ABSTRACT ............................................................................................................ 12
CHAPTER I: INTRODUCTION .............................................................................. 15

Part I: The origin of adaptive immune system .................................................... 15
Part II: Characteristics of the γδ T cells ............................................................... 18
Part III: Somatic hypermutation ......................................................................... 22

CHAPTER II: CHARACTERIZATION OF ARRANGEMENT AND EXPRESSION OF T CELL RECEPTOR GAMMA LOCUS IN THE SANDBAR SHARK ........................................... 33

2.1 Summary ....................................................................................................... 33
2.2 Introduction .................................................................................................. 34
2.3 Materials and methods ............................................................................... 36
  2.3.1 Preparation of DNA and mRNA ............................................................. 36
  2.3.2 Degenerate primer design ..................................................................... 36
  2.3.3 Rapid amplification of cDNA ends (RACE) .......................................... 36
  2.3.4 Computer analysis ................................................................................. 37
  2.3.5 Genomic PCR ......................................................................................... 37
  2.3.6 Chromosome walking ............................................................................ 38
  2.3.7 Shotgun sequencing of large DNA fragments ........................................ 38
  2.3.8 Southern blots ........................................................................................ 38
  2.3.9 Data deposition footnote ........................................................................ 39
2.4 Results .......................................................................................................... 40
  2.4.1 Cloning of shark TCR γ chain ................................................................. 40
  2.4.2 Southern blot analysis ............................................................................ 42
  2.4.3 Sequencing of the sandbar TCR γ Locus ................................................ 42
  2.4.4 Generation of diversity .......................................................................... 45
  2.4.5 Phylogeny of TCR γ ............................................................................. 50
2.5 Discussion ..................................................................................................... 55

CHAPTER III: ANALYSIS OF SOMATIC HYPERMUTATION IN SANDBAR SHARK TCR GAMMA LOCUS ........................................................................ 65

3.1 Summary ....................................................................................................... 65
3.2 Introduction .................................................................................................. 66
3.3 Materials and methods ............................................................................... 68
  3.3.1 Animal .................................................................................................... 68
  3.3.2 Genomic and cDNA sequence of sandbar shark TCR γ gene .............. 68
  3.3.3 Genomic DNA and cDNA sequence alignment ................................... 69
  3.3.4 Calculation of mutability indices ............................................................ 69
  3.3.5 Statistical analysis ................................................................................. 69
3.4 Results .......................................................................................................... 70
PART IV CHARACTERIZATION OF ARRANGEMENT OF EXPRESSION OF THE BETA-2 MICROGLOBULIN LOCUS IN THE SANDBAR AND NURSE SHARK

4.1 Summary.................................................................110
4.2 Introduction............................................................111
4.3 Materials and method..................................................113
4.3.1 Sequence .........................................................113
4.3.2 Preparation of DNA and mRNA .................................113
4.3.3 Cloning of sandbar shark β2m cDNA ..........................113
4.3.4 Cloning of sandbar shark β2m genomic DNA ..................114
4.3.5 Shotgun sequencing of large DNA fragments ..................114
4.3.6 Screening the nurse shark BAC library ........................115
4.3.7 Computer analysis..................................................115
4.3.8 DNA sequencing...................................................115
4.3.9 Southern blots ......................................................115
4.4 Results.......................................................................117
4.4.1 Cloning and sequence analysis of sandbar shark β2m cDNA .........................................................................................................................117
4.4.2 Phylogenetic analysis..................................................119
4.4.3 Protein structure prediction.........................................119
4.4.4 Southern blot analysis..............................................121
4.4.5 Sequencing of the β2m locus of sandbar and nurse shark ..........................................................................................................................124
4.5 Discussion....................................................................128
4.5.1 The evolution of MHC and β2m ..................................128
4.5.2 Genomic organization.................................................129
4.5.3 CpG sites ..............................................................130
4.5.4 SINE and LINES ....................................................131

APPENDIX A: PUBLICATIONS..............................................133

REFERENCES.................................................................134
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIS</td>
<td>Adapative immune system</td>
<td></td>
</tr>
<tr>
<td>Ig</td>
<td>Immunoglobulin</td>
<td>T cell receptor</td>
</tr>
<tr>
<td>TCR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHM</td>
<td>Somatic hypermutation</td>
<td></td>
</tr>
<tr>
<td>AID</td>
<td>activation-induced cytosine deaminase</td>
<td>recombination activator genes</td>
</tr>
<tr>
<td>RAG</td>
<td>gene-conversion</td>
<td></td>
</tr>
<tr>
<td>LLR</td>
<td>Leucine-Rich Repeat</td>
<td></td>
</tr>
<tr>
<td>GC</td>
<td>gene-conversion</td>
<td>class switch recombination</td>
</tr>
<tr>
<td>CSR</td>
<td>base excision repair</td>
<td></td>
</tr>
<tr>
<td>MMR</td>
<td>mismatch repair</td>
<td></td>
</tr>
<tr>
<td>RPA</td>
<td>replication protein A</td>
<td></td>
</tr>
<tr>
<td>UNG</td>
<td>uracil DNA glycosylase</td>
<td></td>
</tr>
<tr>
<td>MRN</td>
<td>Mre11/rad50/Nbs1</td>
<td>framework region</td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td>complementarity determining region</td>
</tr>
<tr>
<td>CDR</td>
<td></td>
<td>mutation index</td>
</tr>
<tr>
<td>MI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>β2m</td>
<td>beta 2 microglobulin</td>
<td></td>
</tr>
<tr>
<td>Figure</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>The emergence of adaptive immune system</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>Current model of activation-induced deaminase (AID) induced somatic hypermutation (SHM)</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>Sequence comparisons of the TCR γ C, J and V regions identified in cDNA clones</td>
<td>41</td>
</tr>
<tr>
<td>4</td>
<td>Southern blots of the shark TCR γ locus</td>
<td>43</td>
</tr>
<tr>
<td>5</td>
<td>A) Genomic map of the shark TCR γ locus. B) Diagram comparing the layout of the shark and human TCR gamma chain loci</td>
<td>46</td>
</tr>
<tr>
<td>6</td>
<td>Sandbar shark γ locus sequencing scheme</td>
<td>46</td>
</tr>
<tr>
<td>7</td>
<td>Junctional diversity at the CDR3 segments of GV1 containing V region</td>
<td>48</td>
</tr>
<tr>
<td>8</td>
<td>Phylogenetic tree, derived from a constant region comparisons, showing the relationship between TCR γ, Ig l light chains and TCR β chains</td>
<td>52</td>
</tr>
<tr>
<td>9</td>
<td>Phylogenetic tree generated by different programs</td>
<td>53</td>
</tr>
<tr>
<td>10</td>
<td>Phylogenetic tree showing relationship between the five shark and six skate GV families</td>
<td>54</td>
</tr>
<tr>
<td>11</td>
<td>Scheme of the evolution of TCR γ gene</td>
<td>64</td>
</tr>
<tr>
<td>12</td>
<td>The cDNA sequences of the GV families</td>
<td>71</td>
</tr>
<tr>
<td>13</td>
<td>Somatic hypermutation in sandbar</td>
<td>75</td>
</tr>
<tr>
<td>14</td>
<td>Comparison of the parental genomic sequence with the GV2, 3 and 4 clones</td>
<td>76</td>
</tr>
<tr>
<td>15</td>
<td>The mutation rates in individual domains of V genes</td>
<td>78</td>
</tr>
<tr>
<td>16</td>
<td>Pol μ based tandem mutation model</td>
<td>98</td>
</tr>
<tr>
<td>17</td>
<td>Comparison of the TCR γ V gene cDNAs of different species with corresponding parental genomic sequence</td>
<td>104</td>
</tr>
<tr>
<td>18</td>
<td>Amino acid sequence alignment of sandbar shark β2m with those of other vertebrate species</td>
<td>118</td>
</tr>
<tr>
<td>19</td>
<td>Phylogenetic tree, derived from full length β2m protein comparisons, showing the relation between sandbar shark β2m and that of other species</td>
<td>120</td>
</tr>
<tr>
<td>20</td>
<td>Predicted three-dimensional structure of β2m protein of sandbar shark and human</td>
<td>122</td>
</tr>
<tr>
<td>21</td>
<td>Southern blotting of the sandbar shark β2m locus</td>
<td>123</td>
</tr>
</tbody>
</table>
Figure 22 Schematics of shark β2m loci
LIST OF TABLES

Table 1: Known Genome of TCR γ gene of different species ............................................. 63
Table 2. Summary of nucleotides distribution in V genes .................................................. 79
Table 3. The single nucleotide mutation profiles of TCR γ V genes .................................. 79
Table 4. The overall mutation profile of TCR γ V genes (including both single nucleotide and tandem mutations) ....................................................................................... 79
Table 5. Nature of the base substitution in the V genes of TCR γ mutants ......................... 82
Table 6: Statistical analysis of mutability index of single nucleotide mutations (A) and tandem mutations (B) .................................................................................................................. 83
Table 7: Nature of base substitutions in TCRGV1 .............................................................. 83
Table 8: Nature of base substitutions in TCRGV2 .............................................................. 83
Table 9: Nature of base substitutions in TCRGV3. ............................................................ 84
Table 10: Nature of base substitutions in TCRGV4 ........................................................... 84
Table 11: Statistical analysis (χ2 test) of the transition to transversion ratio in sandbar shark TCR γV genes .................................................................................................................. 86
Table 12: Comparison of purine and pyrimidine mutations in transition and transversion in single nucleotides and tandem mutations a normalized after correction for base composition ...... 87
Table 13: Tandem mutations in the V genes of TCRγ of sandbar shark ............................. 89
Table 14: Tandem mutations in the V genes of Ig H(A) and Ig L(B) of nurse shark .......... 89
Table 15: Ratio of replacement to synonymous mutations in V genes of TCR γ of sandbar shark (including both tandem and single nucleotide mutations) .......................................................... 91
Table 16: Ratio of replacement to synonymous single nucleotide mutations in V genes.... 91
Table 17: Statistical analysis of the ratio of replacement to synonymous mutation in V genes .................................................................................................................................................. 91
Table 18: Clones used in analysis of SHM in TCRγ of different species ......................... 100
Table 19. Database search for somatic diversification (SHM and allelic difference) of TCR γ gene in other species .................................................................................................................. 105
ABSTRACT

Models proposed for the molecular evolution of the immune system are based on comparative studies of living species. Sharks are critical in this regard since they belong to an ancient clade (chondrichthyes) that can be traced in the fossil record to the time of the earliest vertebrates. Approximately 450 million years ago, the gnathostomes diverged into two groups, the chondrichthyes (sharks and rays) and the osteichthyes (line leading to modern teleosts and tetrapods). It can be concluded that the molecular components of the immune system are ancient and arose prior to this divergence. This follows from studies showing that all the defining elements of the immune system, antibodies, T cell receptors (TCRs), MHC products and recombination activator genes (RAG), are present in chondrichthyes (4, 5). Thus, continued studies of sharks and rays, the most distant living relatives of mammals with a vertebrate type (VDJ-C recombination) immune system, should provide insights into the molecular origins and evolution of the immune system.

In this research, 1) I report the sequence of the sandbar shark TCR γ chain genomic locus and confirm that it has a prototypical translocon arrangement. 2) I also show that in the sandbar shark TCR γV regions undergo somatic hypermutation, in addition to DNA recombination and junction addition and deletion, to generate TCR diversity.3) I report the sequence of the sandbar shark beta-2 microglobulin (β2m)genomic locus.
These findings certainly have functional implications for γ/δ T cells, β2m and MHCs in sharks, and may have phylogenetic significance for understanding the evolutionary origins of diversity in the immune system.
Figure 1. The emergence of adaptive immune system.
CHAPTER I: INTRODUCTION

As the most distant living relative of mammals with an adaptive (V-D-J recombination) immune system, sharks occupy a unique position in the research of the evolution of the adaptive immune system. Comparison of the immune system of sharks with that of humans and mice provides insight into the molecular origins and evolution of the immune system.

Part I: The origin of adaptive immune system: “big bang” and “gradual evolution” model.

The adaptive immune system (AIS), also called the antibody based immune system (6), is characterized by the presence of molecules such as RAG, MHC, TCR and antibody, and by the use of gene rearrangement (VDJ-C recombination) to generate almost unlimited receptor diversity with only a limited number of genes. All the defining elements of the immune system, antibodies, T cell receptors (TCRs), MHC products and recombination activator genes (RAG), are present in sharks and rays, while no such a system can be found in Agnathan. Instead, agnathans, such as lamprey and hagfish, developed a completely different Leucine-Rich Repeat (LRR) immune receptor system, which utilizes a “gene-conversion (GC) like” mechanism to generate receptor diversity. Apparently this LRR system is as efficient as our AIS system, in terms of generating receptor diversity and memory responses. This rather abrupt appearance of the adaptive immune system and the existence of the LRR system in Agnathan poses an interesting question about the
origin of this adaptive system: where does it come from and what is the selective or evolutionary pressure?

Two different models have been proposed to explain the origin of this rather “energy efficient” system. The “big bang” theory (5, 7-9) is mainly based on the abrupt appearance of the AIS in Gnathostomata. The “gradual evolution” model (6, 10) views the origin and evolution of AIS as a gradual accumulation of small changes over a long period. This model is supported by the finding of T and B lymphocyte-like cells (11, 12), homologs of genes involved in activities of T lymphocytes (13), proto-TCR and CD4-like genes (14), pre-BCR like genes (15), and AID gene (16) in lamprey and hagfish (both belonging to the Agnathan family). In addition, many common features of the adaptive immune system, such as multi-gene complexity, somatic hypermutation and gene conversion, have been identified in jawless vertebrates, protochordates and invertebrates (10, 17, 18). However, these two models are not exclusive. The gradual evolution model may produce a platform for a later “big bang” by providing proto-type genes for further change. What is responsible for the occurrence of the “big bang” is still controversial. The recruitment of the RAG gene into this system is believed to be one of the key events, after which the immune system gained the capacity for gene rearrangement.

However, the appearance of MHC genes can’t be explained by this model since no MHC precursor has been identified in jawless vertebrates. A different model was proposed to explain the abrupt appearance of MHC genes (8, 9). Gene duplication is believed to be
another essential event for the emergence of the AIS. It is believed that large-scale gene duplication and subsequent reshuffling of exons is essential for the emergence of MHC genes. The appearance of RAG and gene duplication is believed to have occurred after the divergence of Agnathans and before the divergence of Gnathostomata(8, 19).
PART II: Characteristics of γδT cells.

In human and mice, αβ T cell is the dominant T cell population, for which the antigen recognition depends on the TCR-MHC complex interaction. Unlike αβ T cells, γδT cells only comprise about 1~5% of total T cells in peripheral blood, while in the gastrointestinal tract and reproductive epithelia they are the predominant T cell population in mice, as well as in humans (20). Although this group of cells have not been characterized as well as αβ T cells and B cells, recent research has highlighted their unique features, especially their key role in bridging the innate and adaptive immune systems to allow their classification as an independent T cell population. They are very important in limiting the infection by bacteria (21, 22), virus (23), fungi (24), as well as tumor surveillance, regulation of inflammation and other immune responses (25). Intensive cross talk between γδ T cells, dendritic cells (DCs) and macrophages also helps the maturation of these cell populations (26). γδ T cells also regulate skin homeostasis and wound healing by local secretion of cytokines, chemokines and insulin-like growth factor-1 (27).

Like αβ T cells, γδ T cells are also involved in the typical adaptive immune response. It has been showed that γδ T cells undergo a major expansion during BCG infection and exhibit features of immune memory after BCG re-challenge (28, 29). However, what makes the γδ T cells distinct from αβ T cells is their role in innate immunity. Therefore, γδ T cells should be grouped as a component of the innate immune system that bridges the adaptive system.
In 3D structure, γδ TCR is more like an immunoglobulin molecule rather than an αβ TCR. Furthermore, γδ TCR and immunoglobulin both have a long and a short CDR3 region. Unlike αβ T cells, γδ T cells are very restricted in tissue localization and TCR usage. In fact, the mechanism of antigen recognition is completely different from that of αβ T cells: they recognize antigen in a B cell like way, in which no antigen processing and presentation by major histocompatibility complex is required (30-32).

The actual ligand spectrum of γδ TCR hasn’t been fully understood yet. The ligands that have been already identified can be classified into two categories:

1) So-called phosphoantigens that are released during microbial infection, including mycolyllarabionogalactan peptidoglycan (33), alkylphosphate, alkylamine and aminobiophosphonant (34). Cell-cell contact seems to be required for the recognition of such ligands. However, it was also found that intermediates in bioprenoid biosynthesis pathways such as Isopententenyl diphosphate (IPP) (35), (E)-4-hydroxy-3-methyl-but-2-enyl pyrophosphate (HMB-PP)(36-38) can also activate γδ T cells. Certain studies have also suggested the possibility of small non-peptide antigens activating the γδ T cell in an indirect way. Instead of using antigens themselves, some self-antigens, which were involved in the process of those non-peptide antigens, are the real activator. Small molecule pyrophosphates self-antigens such as IPP and HMB-PP were suggested in the activation of γδ T cells.

2) Stress proteins, including MHC Ib antigens- T10& 22 (39), MICA & MICB(40), CD1 (41), F1 ATPase (42), Hsp -60 (43), phospholipids such as cardiolipin (CL) and
apolipoprotein H complex (44). All these molecules are generally up-regulated or exposed during cellular stress, indicating the potential role of γδ T cells in surveillance of cell stress and cell death.

Very interesting is the recent study about how γδ TCR recognize these antigens. This study showed that murine γδ TCR binds to its ligand T22 using germ line encoded residues of its δ chain CDR 3 loop, while junction encoded residues play an ancillary role, also pointing to the unique position of γδ T cells distinct from adaptive αβ T cells recognition, which junctional residues form the key contact with peptide-MHC complex (45, 46). Another study about TCR Vγ2 Vδ2 also reveals a potential binding site to nonpeptide prenyl pyrophosphates in germ line region of Vγ2 Jγ1.2 CDR3 and Vδ2 CDR 3 loop (47). Interestingly, it seems that γ and δ chain have different function, whereby the antigen specificity of TCR is more likely decided by δ chain, while the function of TCR is decided by γ chain (48). All the studies suggest that γδ T cells function predominantly in the innate immune system, utilizing the TCR as a pattern recognizing receptor, similar to other innate receptors such as TLR, FcR and CR.

Though MHC presentation of non-peptide small antigens is not required, cell-cell contact is required, implying that either other non MHC molecules present small antigens to γδ TCR or that co-stimulation is required (34). Other surface receptors that are usually expressed on the surface of the innate immune cells such as DCs and macrophage are also expressed on the surface of γδ T cells including the toll-like receptor (TLR) (49), scavenger receptor WC1 (50) as well as NK cell receptor NKJ2D (51). It was also found
that TCR/CD3 interaction and signaling pathways of \( \gamma\delta \) T cells is different from \( \alpha\beta \) T cells(52). It is possible the final activation of \( \gamma\delta \) T cells is a comprehensive result of the information from all these receptors.

Recently, another function of \( \gamma\delta \) T cells was characterized, as that of an antigen presenting cell. After being activated, these T cells can actually take up and process antigen, and also display phenotypic characteristics of professional APCs. They can also induce proliferation and differentiation of naïve CD4 T cells (53). It has also been shown that after activation, \( \gamma\delta \) T cells can home to lymphoid organs, using a migration program similar to that of dendritic cells to stimulate immunoglobulin response in lymph nodes(54).

\( \gamma\delta \) TCRs have provided an insight into the evolution of the adaptive immune system. Put together, the biological, structural and gene analysis evidence suggest strongly that \( \gamma\delta \) TCRs were the common ancestors of both \( \alpha\beta \) TCR and immunoglobulin(55-57). Therefore, research on the \( \gamma\delta \) TCRs gene organization and expression in Elasmobranchii species is important for the understanding of the evolutionary origin and the evolution of adaptive immune system.
PART III: Somatic hypermutation.

“Somatic hypermutation (SHM) introduces mutations in the variable region of immunoglobulin genes at a rate of \( \sim 10^{-3} \) mutations per base pair per cell division, which is \( 10^{6} \)-fold higher than the spontaneous mutation rate in somatic cells.” (58)

The machinery of SHM remained as a mystery until the discovery of enzyme activation-induced cytosine, or AID (59). AID is a member of APOBEC family of polynucleotide cytidine deaminase, performing hydrolytic deamination of cytidine (C) to uridine (U).

AID has been identified as the key player in SHM, class switch recombination (CSR), as well as Gene conversion (60)(61). Exogenous expression of AID can lead to SHM in non-B cells (62)(63). In vitro studies showed that ssDNA is the preferred substrate for AID, but not ssRNA or dsDNA (64, 65). In human and mice, the expression of AID is localized specifically in germinal centre B cells. AID has been found in other species (66), including in species that have no germinal centres such as bony fish (67). At least in chicken and bony fish AID has been shown to have functional activity (61). Furthermore, two AID-like molecules were identified in lamprey. Therefore, the expression of AID seems universal and evolutionarily not localized only in germinal centres. A genetic analysis study on bony fish AID yielded some interesting information about the evolution of AID (68). In this study, the amino acid sequence of AID from bony fish was compared with that of human and mice. An amino acid deletion and extensive substitutions in the C terminal end of AID from bony fish were identified, which may suggest that the molecule may not yet have developed a capacity to recruit the specific cofactor(s) needed to initiate CSR.
**Current model of AID induced SHM in immunoglobulin loci.**

In the current model, SHM is a two steps process (Figure1). Namely: 1) AID mediated conversion of C to U in the initiating stage. 2) Initiating U:G lesion lead to mutation in V region via two major mechanisms: base excision repair (BER) for mutation at C:G basepairs, and mismatch repair (MMR) for mutations at A:T base pairs, which are detailed below.

**Initial stage**

In the initiating stage, transcription dependent ssDNA is recognized by replication protein A (RPA), a single stranded DNA binding protein. RPA then target AID to its appropriate substrate(69). In this step, the function of AID has been shown to be regulated by protein kinase A dependent phosphorylation of serine-38 in AID molecules(70-72). Furthermore, it has been shown that serine-38 phosphorylation is essential for the association between AID and RPA(69). Other factors may also be involved in the targeting of AID to the mutation sites. For example, the interaction between AID and spliceosome –associated factor CTNNBL1 has been shown to be important for its function (73).

**BER mediated mutation at C:Gbase pairs**

In hypermutating B cells, uracil DNA glycosylase (UNG) is required for the removal of the U based induced by AID. UNG mediated cleavage of the U base from the DNA backbone results in a non-instructive abasic site(74)(75). Then,error prone polymerases
are recruited the abasic site to generate C/G mutations. At least two error prone – polymerases (Rev1 and DNA polymerase theta) have be suggested to be involved in this process(76)(77). In addition, an alternative UNG mediated pathway has also been proposed. Mre11/rad50/Nbs1 (MRN) complex which normally functions in DNA break sensing and repair has been implicated in SHM. The overexpression of the Nbs1 subunit in a hypermutating B cell line results in increased level of mutations, suggesting that MRN promotes DNA cleavage and/or mutagenic repair of lesions initiated by activation-induced deaminase, acting in the shared pathway of immunoglobulin gene diversification(78).

**Mismatch repair is responsible for the A:T mutations**

MMR heterodimer MSH2/MSH6 has been shown to bind to the U: G mismatches but not to other DNA intermediates produced during base excision repair of dUs. Polymerase eta is associated with MSH2(79). Furthermore, defects in MSH2 or MSH6 lead to reduction of A: T mutations(80)(81, 82). In this model exonuclease 1, a known partner of MSH2(83), is responsible for deleting of nucleotide and extending the mutation to the A:T base pair(84). There is still some uncertainty about which error prone polymerase is responsible for generation of the mutations. Currently low fidelity DNA polymerase η is thought to be the major player(85, 86)(87), although a role for other low fidelity polymerases’ (such as pol ζ(88) and pol θ (89)) are also indicated.

It has been proposed that Rad6/Rad18 mediated recognition of AID induced DNA lesions trigger the ubiquitination of proliferating cell nuclear antigen (PCNA), a major switch controlling the fidelity of DNA lesion bypass in eukaryotes, which in turn signal
the recruitment of error prone polymerases (90). In this research, the inactivation of Rad18 in the DT40 B cell line lead to the change the profile of the somatic hypermutation in these cells, indicating that the Rad6/Rad18 pathway is involved in somatic hypermutation in these cells. Other research in which same cell line shows that the PCNA(K164R) mutation not only renders cells sensitive to DNA-damaging agents, but also strongly reduces AID-dependent single-nucleotide substitutions in the immunoglobulin light-chain locus (91). This regulatory model for recruitment of error prone polymerase can fit into the MMR mechanism for A:T mutations very nicely. However, apparently this model cannot account for the recruitment of error-prone polymerase in the BER mechanism for C:G mutations.
Figure 2. Current model of activation induced deaminase (AID) induced SHM.
Targeting of the SHM

SHM is associated with transcription. Or put in another ways, transcription serves as the triggering event of SHM. High transcription is necessary but not sufficient, indicating there must be a targeting system(92). Somatic hypermutation (SHM) is restricted to VDJ regions and their adjacent flanking regions in immunoglobulin (Ig) genes, whereas constant regions are spared. Mutations occur after about 100 nucleotides downstream of the promoter and extend to 1-2 kb(93)(94, 95). SHM has been found to be linked with transcription initiation and Ig promoters(96). If the V sequence with is replaced with non-Ig sequences, these non-Ig genes start to gain the capacity of SHM(97), indicating that V gene itself is not necessary for recruiting hypermutation, but the promoter matters. Similarly, if the distance between the promoter and V gene is changed, the mutation rate can also be changed accordingly(98). Furthermore, it was found that increased transcription levels induce higher mutation rates in a hypermutating cell line (99).

Research comparing the distribution of mutations between uracil DNA glycosylase (UNG)-deficient and wild-type mice in endogenous Ig genes and in a trans-gene suggested that the very 5' end and the constant region of Ig genes are spared from somatic mutation because AID does not access these regions(100). This research suggests that the specific targeting of AID to the Ig might be the key for localization of SHM specifically to V regions. Again there must be a factor or complex involved. As in the cis-acting element regulation model, the specific factors haven’t been identified. Moreover, it is reasonable to speculate that in sharks, which represent the other end of the evolution of adaptive immune system, this targeting system is different from that of mammals.
Regulation of hypermutation

**Cis-acting element model:** cis-acting elements can function as a platform on which transcription factor assemble. The assembled protein factors in the case of SHM of immunoglobulin may have regulatory functions, such as enhancer and promoter interactions and recruiting the SHM machinery to the Ig loci. Although the exact mechanisms haven’t been clarified, research using an Igκ trans-gene have shown that at least a minimum set of acting element are required for targeting SHM to the Igκ V region(94, 101). In addition to this cis-acting element model, the intron enhancer (3’ κ enhancer), as well as the flanking matrix attachment region have been shown in an Ig κ trans-gene model to be crucial for the targeting of SHM machinery to immunoglobulin loci(102-104). Furthermore, these intronic regulatory elements can exert a complex influence on SHM that is separable from their role in regulating transcription(105).

Research on endogenous immunoglobulin loci, however, lead to complete different conclusions that the deletion of the intron enhancer or 3’ enhancer did not reduce the frequency of SHM in immunoglobulin loci(106, 107). These apparent discrepancies between study of transgenic and endogenous immunoglobulin loci raise the possibility that enhancer elements have redundant functions in targeting SHM in vivo. So far, transcription factor E2A (including E47 and E12) has been suggested as one of the trans-acting factors that regulate the SHM of immunoglobulin loci via interacting with enhancers(108).
Negative regulation model: it has been suggested that a high transcription rate alone may predispose any gene to mutation by AID, but there may be specific B cell factors that account for the SHM of immunoglobulin loci in vivo and its targeting to V region (109). There is indirect evidence also supporting this B cell specific regulator model. For example, ectopic expression of AID alone can induce hypermutation in an artificial GFP substrate in NIH 3T3 murine fibroblast cells. However, a different profile of mutations in fibroblast suggests the existence of specific B cell regulators(63, 110). Interestingly, it has been reported that the constitutive over-expression of AID lead to T cell lymphoma rather B cell lymphoma(111). This surprising finding indicated that there must be specific B cell factors involved in the regulation of SHM in immunoglobulin loci. Research by the same group also showed that abundant AID protein accumulated by constitutive expression is inactivated in B cells(112). On the other hand, loss of this checkpoint May lead to overexpression of AID and overproduction of autoantibody(113). Besides all of the above indirect evidence, it is not until last year that microRNA-155 was identified for the first time as a negative regulator of AID(114).

Selection on immunoglobulin
Earlier genetic analysis on the V region sequence of human Ig H chain showed that CDR and FR sequences can differ significantly in their inherent susceptibility to amino acid replacement given any single nucleotide change(115). The authors of this research contribute this difference in inherent susceptibility to replacement mutation between CDR and FR to the selection pressure imposed on this region. The selection pressure is represented as the high replacement mutation rate in CDR and the scarcity of replacement
in FR. This theory provides a fertile structural substrate of hypervariability for antigen selection, while still maintaining the structural integrity of the FRs. Furthermore, this selection pressure model has been proven to be as the rule in both human and mice (115-120). Furthermore, genetic analysis shows that R/S ratio bias may come from intrinsic mutability among different regions of an antibody gene even in absence of any antigenic selection (93, 121). This selection pressure on the immunoglobulin V region must come from evolution rather than ongoing antigen stimulations. Interestingly, the findings that affinity maturation can occur in the absence of germinal center formation also support this model (122-124).

Research on nurse shark shows that there is no significant difference in the R/S ratio between CDR and FR in both Ig L and IgH (125, 126). Similar results also occur forteleost Ig H (127) and T cell receptor γ of sandbar shark (our own data). All of above findings suggests that there is none, or very little selection pressure on immunoglobulin and TCR in bony fish and shark, and SHM in immunoglobulin and TCR in bony fish and shark are mainly for increasing repertoire diversity rather than increasing affinity. These findings fit nicely into the concept that somatic mutation may have originally evolved as a mechanism to principally increase repertoire diversity (125, 126, 128-130).

**Tandem hypermutation**

A very unique feature of SHM in sharks is the presence of tandem mutations. So far, this unique pattern of mutation has been found in Ig H of horn shark, Ig L and H chain of nurse shark, NAR of nurse shark, and TCR gamma of sandbar shark. A close look at
published Ig gene sequences of skate (131), a close relative of shark, reveals that tandem mutations also occur in skate Ig. Other than shark and skate, tandem mutation is rarely found in other species. The mechanism of tandem mutations has been explored. A computerized randomization (monte carlo) trial has been done to test the possibility that tandem mutations came from sequential, independent point mutation. Results of the trial shows that it is almost impossible that the theoretic number of tandem mutations generated by random point mutations can reach the level of what has been observed in the shark immunoglobulin (125). Furthermore, based on the sequence analysis the same group also suggested that gene conversion is very unlikely to be the mechanism, or at least not the classic pathway of gene conversion that has been found in chicken and rabbit (132). Furthermore, since all the found V regions of TCR gamma have been shown to be functional, it is very unlikely that gene conversion, which usually occurs between pseudo-genes and functional template, is the mechanism utilized for generation of tandem mutations in TCR gamma.

Besides the obvious appearance difference, tandem mutation is different from single point mutation in many aspects. For example, there is less transition in tandem, compared to that of single nucleotide mutations. Transition bias is almost a universal phenomenon in SHM of immunoglobulin, except in the tandem mutation of TCR of shark. Although it is lower than found in single nucleotide mutation, the transition sequences in tandem mutation of Ig light chain are still significantly higher than the theoretical transition frequency (33%) if the mutations occur randomly. However, the transition frequency
(34%) in tandem mutation of TCR is close to 33%, suggesting that there is much less selection pressure on TCR of shark.

Interestingly, Ig tandem mutation can also be found mice deficient for the DNA mismatch repair protein PMS2 and in some elderly people who were believed to have defective DNA repair system. The authors of these papers speculated that in the germinal centers, there might be high frequency of tandem mutations. However, since the tandem mutations are easier to recognize than single nucleotide mutations, they are prone to be corrected immediately unless the DNA repair machinery is defective. Therefore, the tandem mutations are rather due to defective DNA repair than due to the involvement of a different mutator. Although this might not be the case in sharks, it will be still of interest to see whether there is a difference in DNA repair machinery between sharks and mammals.
CHAPTER II: CHARACTERIZATION OF ARRANGEMENT AND
EXPRESSION OF T CELL RECEPTOR GAMMA LOCUS IN THE SANDBAR
SHARK

2.1 Summary

Immunoglobulin (Ig) and T cell receptor (TCR) genes consist of separate genomic elements, which must undergo rearrangement and joining before a functional protein can be expressed. Considerable plasticity in the genomic arrangement of these elements has occurred during the evolution of the immune system. In tetrapods, all Ig and TCR chain elements are arranged as translocons. In teleosts, the Ig heavy and TCR chains are translocons, but light chain genes may occur as clusters. In chondrichthyes, however, all the Ig light and heavy chain genes are arranged as clusters. These clusters vary in number from less than ten to several hundred, depending on isotype and species. Here I report that the germline gene for the TCR \( \gamma \) chain in a chondrichthyan, the sandbar shark, is present as a single locus arranged in a classic translocon pattern. Thus the shark utilizes two types of genomic arrangements, the unique cluster organization for Ig genes and the "conventional" translocon organization for TCR genes. The TCR \( \gamma \) translocon contains at least five V region genes, three J segment genes, and one C segment. As expected, the third hypervariable segment (CDR3), formed by the rearrangement of the \( V_\gamma \) and \( J_\gamma \) segments, contributed the major variability in the intact V region structure. However, significant diversity is generated from hypermutation in the V regions.
2.2 Introduction

Models proposed for the molecular evolution of the immune system are based on comparative studies of living species. Sharks are critical in this regard since they belong to an ancient clade (chondrichthyes) that can be traced in the fossil record to the time of the earliest vertebrates. Approximately 450 million years ago, the gnathostomes diverged into two groups, the chondrichthyes and the line leading to modern teleosts and tetrapods. It can be concluded that the molecular components of the immune system are ancient and arose prior to this divergence. This follows from studies showing that all the defining elements of the immune system, antibodies, T cell receptors (TCRs), MHC products and recombination activator genes (RAG), are present in chondrichthyes (4, 5).

A surprising feature of the organization of shark immunoglobulin genes was that the V-J-C of light chains (135-137) and V-D-D-J-C of heavy chains (138-140) are arranged in individual clusters. This is in contrast to the translocons of higher vertebrates that consist of large arrays of multiple V segments distantly linked to several J and D segments, as well as C domains (141). Although the genes for the TCR chains α/β and γ/δ have been shown to be present in the skate (142) and horned shark (143, 144), the germline loci for these molecules have only been partially characterized. Analyses of cDNA sequences in the skate revealed multiple V region families and Southern blot analysis of genomic DNA strongly indicated that the skate TCR genes were arranged as translocons (142). This was surprising in light of the cluster organization of the Ig genes. Here I report the sequence of the sandbar shark TCR γ chain genomic locus and confirm that it has a prototypical translocon arrangement. I also show that the TCR γ V regions undergo
somatic hypermutation. This is significant since it is generally accepted that
hypermutation does not occur, or is very rare (145-147), in TCR V regions. This finding
certainly has functional implications for γ/δ T cells in sharks.
2.3 Materials and methods

2.3.1 Preparation of DNA and mRNA

Sandbar shark spleen was provided by Dr. Carl Luer (Mote Marine Laboratories, Sarasota, FL). Sandbar shark spleen genomic DNA was prepared using the QIAGEN Blood and Cell Culture DNA Kit. Sandbar shark spleen mRNA was prepared using the Invitrogen Micro-FastTrack 2.0 mRNA Isolation Kit. All genomic DNA and mRNA were collected from same animal.

2.3.2 Degenerate primer design

Based on sequence analysis of the C region and transmembrane region from multiple species, including skate, chicken, and mammals, a conserved stretch of amino acid sequence (FFPDVI) was identified in the TCR chains. Based on this conserved amino acid sequence, gene specific degenerate primers (GPS) were designed.

2.3.3 Rapid amplification of cDNA ends (RACE)

5’ RACE was performed using the SMART™ RACE cDNA Amplification Kit (Clontech) according to the manufacturers instructions. Degenerate primers as well as adaptor specific primers were used for PCR reactions using the following conditions: 50 μl reaction total, 34.5 μl sterile H2O, 5 μl 10X Advantage 2 PCR buffer, 1 μl dNTP mix, 1 μl 50X Advantage 2 polymerase mix, 2.5 μl 5’ RACE ready cDNA (1), 5 μl UPM (universal primer), 1 μl GSP. PCR settings: 5 cycles at 94°C 30 sec, 72°C 3min, 5 cycles: 94°C 30 sec, 70°C 30 sec, 72°C 3min.
Agarose gel electrophoresis of PCR reactions were performed on a 1% agarose gel, and the band of proper size carefully excised. The PCR products were purified using the QIA quick gel purification kit (QIAGEN) and cloned using the TOPO4 TA vector cloning system (Invitrogen). Mach1-T1 chemically competent cells (Invitrogen) and QIA prep miniprep kit (QIAGEN) were used for transformation & plasmid purification.

2.3.4 Computer Analysis

The following software packages were used: MAFFT (148, 149) for sequence alignment, MacVector 10.5 (MacVector, Inc) for assembly of sequencing projects, sequence alignment and phylogenetic analyses, ClustalW2 (150, 151) for sequence and phylogenetic analyses and MEGA4 for phylogenetic analysis (152). Recombination signal sequences (RSS) were identified using methods of Cowell et al (153).

2.3.5 Genomic PCR

Multiple DNA polymerase kits, including Expand High Fidelity Plus PCR system (Roche), Expand 20 kb Plus PCR system (Roche), and iProof High Fidelity DNA polymerase (Bio-RAD), were used to eliminate possible PCR errors, and to increase the chances of getting the right DNA fragment. Multiple pairs of primers were designed based on the cDNA sequence of every V region and C region. To rule out the possibility of PCR bias, at least two primers were designed for each V region. Normal as well as “touchdown” PCR procedures were performed to get the DNA fragment of interest. Every PCR was performed following the manufacturer instructions for each DNA polymerase.
2.3.6 Chromosome walking

The Universal Genome Walker kit (Clontech) was used in this study. For each library construction, purified genomic DNA was digested with blunt-end restriction enzymes (DraI, EcoRV, PvuII, ScaI, and StuI) independently. An adaptor was added to the end of the digested DNA fragments. Two gene specific primers (GSP1 & GSP2) were designed based on the cDNA sequence. The nested PCR primer (GSP2) annealed to sequences beyond the 3’ end of the primary primer (GSP1) in 5’ walk (5’end of GSP in 3’ walk). iProof high fidelity DNA polymerase “touchdown” PCR procedure was performed. In the primary reaction GSP1 and adaptor primer 1(AP1) were used as primers. In the secondary reaction, 1 µl of primary reaction was used as template, and GSP2 & AP2 as primers to perform nested PCR.

2.3.7 Shotgun sequencing of large DNA fragments

For genome sequencing, shotgun-sequencing strategy was applied. The GPS-1 genome primer system (New England Biolab, Inc.) was used for this purpose. Individual sequences were assembled using MacVector.

2.3.8 Southern blots

Sandbar shark spleen DNA was digested by Bam HI, Eco RI, and Hind III independently. Digested DNA was transferred to Immobilon-Ny+ transfer membrane (Millipore) by capillary flow under alkaline conditions (1.5 M NaCl/0.5 M NaOH) and fixed to the membrane with UV cross-linking (5000 microjoules). A $^{32}$P labeled full length C
segment was used as the probe. After an overnight hybridization, high stringency washes were performed in 0.1× standard saline citrate (SSC)/0.1% SDS solution at 65°C.

2.3.9 Data deposition footnote
The sequences reported in this paper have been deposited in the GenBank database (Accession numbers TCRGV1: FJ854417-FJ854438, TCRGV2: FJ854439-FJ854457, TCRGV3: FJ854458-FJ854470, TCRGV4: FJ854471-FJ854487, TCRGV5: FJ854488-FJ854491, TCR gamma genomic locus: FJ854492.
2.4 Results

2.4.1 Cloning of shark TCR γ chain: I performed 5’RACE using degenerate primers based on a conserved C domain amino acid sequence to isolate putative sandbar shark TCR γ chain cDNA clones. A total of 25 cDNA clones were selected from the cDNA library with all of the selected clones sharing identical C region sequences. Searches of the NCBI databases using BLAST showed that the clones had clear identity to known TCR γ chains, the best match being to skate TCR γ chain. The complete sequence of the C region was assembled using sequence from 5’ and 3’ RACE. Comparison of the derived amino acid sequence of this C region with skate and human sequences, as well as the location of the IgC1 domain, trans-membrane (Tm) domain and cytoplasmic domain are shown in Figure 3A. The identity with human gamma is only 23%, a value much lower than comparable comparisons of immunoglobulin and other TCR sequences. This value is too low to indicate specific relatedness to higher vertebrate TCR gamma chains and shows that the TCR gamma chains have diverged significantly in evolution. The identification of the shark sequence as TCR γ chain is based on homologies with gamma chain in BLAST searches, the high identity with the skate sequence (50%) and the conservation of a gamma chain specific motif in the transmembrane domain (Figure 3A).
Figure 3: Sequence comparisons of the TCR \(\gamma\) C, J and V regions identified in cDNA clones. A) Derived amino acid sequence of the C region and comparison with skate and human sequences. Only one constant region was found in the shark TCR \(\gamma\) cDNA. Identities with the shark sequence are shaded. Domains are indicated by arrows above the diagram and were identified by comparison to the human sequence. Conserved residues in the TM domain are underlined. Percent identities are listed on the right. B) Three J families were identified by cDNA sequence analysis. C) Five V families, designated as GV1, GV2, GV3, GV4 and GV5, were identified by cDNA sequence analysis. Derived amino acid sequences are shown. Putative leader segments are shown in lower case and were identified using the SignalP 3.0 program (3). Identical residues are indicated using both shaded and open boxes. The percent identities are shown in the table on the right.

2.4.2 Southern blot analysis
The finding that the cDNA clones isolated above shared identical C region sequences (>99% identity) indicated the presence of only a single copy of the C region gene. To further confirm this, southern blots were performed with spleen DNA using the full-length C segment as the probe. As shown in Figure 4, the BamHI and HindIII DNA preparations each gave rise to one strong hybridizing band, indicating the presence only one TCR γ locus in the sandbar shark genome. This is buttressed by the finding that there are two equally dense hybridizing bands in the EcoRI digested DNA preparation, which is explained by the fact that there is one EcoRI restriction site in the C region. Interestingly, several weak hybridizing bands are also seen in each restriction endonuclease DNA preparation (Figure 4). Since the genomic PCR and sequencing results indicated that lymphocytes in the spleen have undergone gene rearrangement, the multiple weak hybridizing bands are most likely the result of gene rearrangement.

2.4.3 Sequencing of the sandbar TCR γ Locus: Sandbar shark spleen DNA was purified to perform genomic sequencing of the TCR γ loci. The use of spleen DNA for this purpose has advantages and disadvantages. It is expected that the T-cells and B-cells in the spleen would be mature cells and have undergone gene rearrangement. Therefore during the sequencing procedure, some of the clones contained rearranged V-J segments. However, other cell types in the spleen (e.g., epithelial cells) yielded clones containing un-rearranged DNA. Multiple rearranged and un-rearranged segments were identified, and careful analysis of the sequences was performed in order to obtain the complete germline DNA sequence. On the other hand, spleen DNA is a good target to
Figure 4: Southern blots of the shark TCR γ locus. Sandbar shark spleen DNA was digested by BamHI (lane a), EcoRI (lane b), and HindIII (lane c) independently. The probe was full length C region isolated by PCR of cDNA.
study the rearrangement events that occur at the genome level of the sandbar shark. Both genomic PCR and Genome Walker DNA walking strategies have been used here. For each genomic PCR reaction, at least two pairs of gene specific primers and multiple high fidelity polymerases were used to minimize any possible PCR bias and/or PCR error.

An overview of the TCR $\gamma$ locus is shown in Figure 5. The sequence was assembled from 6 PCR products and 10 chromosome walking fragments (Figure 6). The final sequence was derived from at least 2 independent products in most cases. A few sequence mismatches between different PCR products are present that probably represent allelic differences or hypermutation events, although a small fraction may be due to PCR error.

As shown in Figure 5A, at least 5 V regions, 3 J regions and 1 C region are arranged in a typical translocon pattern. This is probably the total number of V genes present since only 5 V region families were discerned in the 73 cDNA sequences analyzed and these can be unequivocally matched to the genomic V gene sequences. However, there are possibly more V genes upstream of GV5 that, although rarely expressed in spleen, may be expressed in other tissues. Use of the word “families” to label the different V regions is strictly speaking incorrect here, since each family has only one gene member. However, we adopted this terminology since it is consistent with its use in immunology to describe groups of V regions with less than 75% amino acid sequence identity.

Consistent with all the known Ig and TCR V regions, sandbar shark TCR $\gamma$ V regions also have leader sequences that contain an intron. As mentioned above, several genomic PCR products contained V genes joined to J segments. Since the J to C intron sequences of
these clones were identical, I concluded that they were the result of rearrangement events from a single locus, and were not indicative of multiple loci.

RSS 23 sequences were identified at the 3’ end of each V region and RSS 12 sequences at the 5’ end of each J region. Thus it appears that rearrangement in the shark γ loci follows the “23/12” rule (154) defined in primates and rodents. The RSS segments were analyzed by calculating “RSS information content” (RIC) Scores using models developed by Cowell et al (153). The statistical models used for calculating “informational content” were learned using mouse Ig and TCR sequences (153), yet the RIC scores for the shark TCR γ locus clearly discriminate the physiologic RSSs. Thus, the “informational content” of RSS segments is yet another feature of the immune system that has been conserved during vertebrate evolution.

2.4.4 Generation of Diversity: Based on TCR γ chain C region sequence, three different specific primers were designed and used in 5’RACE to examine the V region repertoire. A total of 73 different clones were obtained and sequenced. All sequences appeared functional as there were no frame shift or stop codon mutations. Analysis of the sequences showed that all 3 J region sequences and all 5 V region families identified in the genomic sequence were expressed. The cDNA sequences of the J regions (Figure 3B) and amino acid sequence of each the V regions are shown in Figure 3C. Although the J segment sequences are similar, they have different lengths and are clearly distinguishable.
Figure 5: A) Genomic map of the shark TCR γ locus. V segments, J segments and C segments are arranged in a translocon. RSS sequences at the 3’ end of V segments and the 5’ end of the J regions are highlighted by white (RSS 23) and black triangles (RSS 12). The RSS sequence at the 5’ end of J1 region is shown in gray since it has not been possible to derive an accurate sequence for an approximately 15 bp segment in this region as multiple PCR products show variation in length and sequence. I conclude that this represents PCR error. The genomic structure showing the intron present in all the V segments is illustrated using the GV2 segment. B) Diagram comparing the layout of the shark and human TCR gamma chain loci. Dark boxes are functional V regions, grey boxes are pseudo-genes, vertical lines are J segments and open boxes are constant regions. The depiction of the human gene is adapted from the Gene View page at the ImMunoGeneTics Web site (http://imgt.cines.fr).

Figure 6: Sandbar shark γ locus sequencing scheme. Solid lines show PCR products, while open lines show chromosome walk fragments. PRC products shown with dashed lines represent rearranged gene products, with the dashed lines indicating deleted regions.
For each V region family, examples of all J region sequences were found and conversely, each J region was found associated with all V regions. There are approximately equal numbers of clones containing GV1 (22), GV2 (18), GV3 (12) and GV4 (17) indicating that there is no bias in the rearrangement of these V segments. Similarly, no bias is apparent in the expression of the 3 J segments. However, I found only 4 clones expressing GV5, suggesting that there may be significantly fewer rearrangements of this most 5’ distal V region segment.

Some CDR3 region sequences of cDNA clones containing GV1 are shown in Figure 7A. The CDR3 region is formed by the joining of V and J regions, and extensive sequence variability is generated at the junction during this process. This diversity is similar to that of higher vertebrates indicating that the sharks use similar mechanisms of deletion (by exonuclease) and addition (terminal deoxynucleotidyl transferase (TdT)) to generate TCR γ diversity. The length of CDR3 loops varied from 6 to 14 amino acids, a range similar to that in mice and humans (155). CDR3 loops are usually longer when a V region is joined to the J2 segment, since this is larger than the J1 and J3 segments. The same CDR3 sequence variability was seen in genomic PCR fragments generated using a GV1 specific primer and a C region specific primer. Four different sized PCR products were obtained (Figure 7B) corresponding to unrearranged TCRγ, and 3 rearrangement events in which the GV1 segment was joined to any of the 3 J region segments. Each of the 3 PCR products containing rearranged DNA was separately cloned and a total of 18 were sequenced for analysis of CDR3 diversity (Figure 7B). Deletion and addition patterns in
Figure 7: Junctional diversity at the CDR3 segments of GV1 containing V regions. A) Sequences of shark TCR γ cDNA at the V/J junction of GV1 joined to all 3 J segments are shown. The parent genomic sequences are underlined above each group. B) Junctional diversity at the CDR3 segments of rearranged genomic DNA. Deletions (by endonuclease) are shown by asterisks and nucleotides in lower case indicate additions (by terminal deoxytransferase).
this region were identical to that of cDNA sequences, except that about 50% of the
genomic sequences show non-productive rearrangements while 100% of cDNA products
are productive. The number of nucleotide additions ranged from 0 to 13, which could
theoretically result in a diversity of nearly $10^8$ sequences. The actual number is probably
significantly lower since the addition of G residues is approximately 2 fold higher (37%)
compared to A, T and C residues (approximately 20 % each).

Surprisingly, comparisons of the cDNA sequences with the parent genomic sequences
reveal a high degree of base changes in the cDNA (see chapter III). Thirty-three of the 73
cDNA sequences had mutations, and 22 of these had 2 or more mutations. Because the
mRNA and DNA samples were prepared by using spleen tissue from a single shark, the
observed sequence diversity is not due to allelic variation. And it is not due to PCR error,
because I used a high-fidelity polymerase. The rare occurrence of changes in C region
sequences, and the fact that 33 of the V region sequences had no changes, confirms the
fidelity of the polymerase. The average mutation rate of the V region (0.017 per base pair)
is much higher than that of the C region (0.7 x $10^{-4}$ per bp) and the expected PCR error rate
($4.4 \times 10^{-7}$ per bp per cycle; in this research all of the PCRs were done with 25–30 cycles),
but is comparable with mouse and shark Ig light chain V regions(125). Thus, I conclude
that these data demonstrate that somatic mutation occurs in the TCR $\gamma$ V region. A striking
feature is the presence of tandem mutations containing 2–5 nucleotides. Tandem
mutations appear to be a unique feature of somatic hypermutation in sharks, because
they do not occur in higher vertebrates, and have been documented and characterized only
for light and heavy chain V regions in the nurse shark (125, 156-158). Also, there are other similarities between the somatic mutations observed here and those described for the nurse shark.

2.4.5 Phylogeny of TCR $\gamma$

Although the identity of sandbar shark TCR $\gamma$ chain with those of higher vertebrates is low, I was surprised that the best matches in GenBank searches are mostly with light chains of other species. To further ascertain the relationship of sandbar shark $\gamma$ with other $\gamma$ chains, $\beta$ chains and $\lambda$ light chains, I constructed an unrooted phylogenetic tree (Figure 8) based on C region sequences. Using the Neighbor Joining method, the skate and shark TCR $\gamma$ chains do not group with the $\gamma$ chains of higher vertebrates, but rather are in the $\beta/\lambda$ clade. I also used UPGMA and Maximum Parsimony methods with very similar results (Figure 9). My interpretation is that ancestral $\gamma$ chains were closely related to $\lambda$ and $\beta$ chains, and diverged in higher vertebrates because of a change in function of $\gamma/\delta$ T-cells.

Similar analysis of the V regions shows that they do not group with any V region family of higher vertebrates. However, skate and shark sequences do cluster together, with the five shark and six skate families forming a tree with four groups (Figure 10). Thus, there has been conservation of V$\gamma$ families in elasmobranches since the divergence of skates and sharks 250 million years ago. The tree indicates that the TCR $\gamma$ chain locus was well established prior to this divergence and contained four V$\gamma$ genes. The repertoire was subsequently expanded by duplication of these genes. For example, the number of
sandbar shark V regions was expanded to five by an early duplication, giving rise to GV1 and GV5.
Figure 8: Phylogenetic tree, derived from a constant region comparisons, showing the relationship between TCR γ, Ig λ light chains and TCR β chains. Sequences were aligned using MAFFT (148, 149) and the tree was constructed with MEGA4 (152) using the neighbor joining method (159). Percent bootstrap values (1000 replicates) are shown at the major interior branch points. The horizontal length is proportional to the distance score generated by the computer program. The IgC1 domain amino acid sequences, corresponding to residues 1 to approximately 125 using the IMGT numbering scheme (160), of λ light chains and TCR γ and β chains from different species were used to build this tree. The shark TCR γ sequence is from this dissertation; all other sequences were obtained from Genbank.
Figure 9: Phylogenetic tree generated by different programs, derived from constant region comparisons, showing the relationship between TCRγ, Ig light chains and TCR β chains. A) UPGMA (Unweighted Pair Group Method with Arithmetic Mean); B) Maximum parsimony.
Figure 10: Phylogenetic tree showing relationships between the five shark and six skate GV families. The shark TCRγ sequences are from this dissertation, all the skate TCRγ sequences were obtained from Genbank.
2.5 Discussion

A surprising feature of the organization of shark immunoglobulin genes (161) and the teleost immunoglobulin light chain genes (162) is that they are arranged in individual clusters, rather than the translocon arrangement found in higher vertebrates. Here I show that the sandbar shark TCR γ gene is arranged in a “classic” translocon pattern consisting of at least 5 V regions, 3 J segments and a single constant region. In higher vertebrate species, the translocon arrangements are not as “pure” as that of shark TCR γ. They are usually more complicated and contain numerous V, J and C regions as illustrated in Figure 5B. Furthermore, most contain certain cluster characteristics, such as that of the C λ region in mammals where there are at least four separate copies of the J-C λ gene cluster in mice (163) and at least nine such clusters in human (164). A similar arrangement is also found in the mammalian TCR γ and β gene loci (165-167) and teleost fish light chain genes (162).

The RSSs in the γ locus are typical. As is true for all vertebrates, recombination follows the 12/23 rule since the V regions have RSS23s and the J segments have RSS12s. Analysis of the RIC of shark RSSs and comparison with human and mouse show that these have been highly conserved. This is not surprising since RAG1 and RAG2, which mediate recombination, are also very highly conserved between shark and human. It is apparent that the core mechanisms of rearrangement, the recognition of RSSs and precise cleavage of the DNA by the RAG complex, are highly conserved features of the adaptive immune system.
The presence of both the cluster and translocon type arrangements in sharks and rays raises questions as to the ancestral arrangement pattern. Given the indications that TCR $\gamma$ chains are more ancient than Ig chains (see below), the most parsimonious explanation is that a simple translocon, as described here, was the ancestral pattern for both TCRs and Igs. Alternately, both patterns may be ancestral and evolved independently from a common precursor after the split of lymphoid cells into separate B and T cell lineages.

In this study, I provide direct evidence for the presence of gene mechanisms used by sharks to generate TCR$\gamma$ diversity that are similar to human immunoglobulin and TCR genes. The sandbar shark TCR $\gamma$ gene has the capacity to generate enormous diversity using three basic mechanisms. These are combinatorial diversity, the introduction of N region diversity during rearrangements and hypermutation in the V regions. The first two are common for TCR and immunoglobulin in all species, but the third is unique to immunoglobulins in higher vertebrates. The basic starting point is the 15 different arrangements possible through the five V and three J segments. With the exception of the decreased representation of TCRGV5, there was no bias in the utilization of these elements. As the sequence identity of the V regions ranged from 31% to 46%, the rearrangement products represent quite different and distinct backbones on which to build additional diversity. A major component of shark TCR$\gamma$ variability is contributed by the CDR3 segment formed by the junction of the V/J region. Extensive deletion and addition of nucleotides occur at the V/J junction at the genome level. Therefore, I conclude that the rearrangement mechanisms for generating CDR3 combinational diversity in sharks are essentially no different from those of higher vertebrates. It also seems clear that the
incorporation of combinatorial and junctional mechanisms for the generation of diversity in the binding repertoire occurred very early in the evolution of the vertebrate immune system.

Analysis of cDNA sequences clearly shows that somatic hypermutation is a third mechanism utilized to generate diversity in the expressed TCR γ repertoire. This is remarkable since hypermutation does not occur in TCR V regions, or is rare (145-147). It is generally accepted that this is because mutation in the CDR 1 and 2 regions of the α/β receptor would disrupt recognition of Class I and II antigen presenting molecules. This reasoning may not apply to the γ/δ receptors, as these do not require MHC for the recognition of proteins and non-proteins (168, 169). Even so, hypermutation has not been reported to occur in these receptors either. I thought it possible that it may occur in the so-called γ/δ high species such as sheep, pigs and cows since γ/δ T-cells are 60% of peripheral cells. Despite numerous recent publications characterizing in detail genomic and cDNA sequences in these species (170-172), I could find no reference to hypermutation in TCR γ chains. A recent paper noted point mutations in salmon TCR γ chains (173). These results are difficult to interpret since many mutations were insertions or deletions resulting in frame shifts, and the number of mutations in the constant regions was approximately the same as for the V regions, while in our results the ratio between V region and C regions is about 200:1. I believe ours is the first report documenting hypermutation in TCR γ V regions. Interestingly, analysis of the mutation patterns shows that they are very similar to hypermutation in the immunoglobulin V regions of the nurse
shark. Clear distinguishing features are the tandem mutations, which apparently only occur in sharks (125, 158). Thus, it appears that the mechanisms for hypermutation in shark B-cells and γ/δ T-cells are similar. The high level of hypermutation in gamma V regions emphasizes to us that many γ/δ T-cells in the shark recognize and respond to antigen independently of antigen processing and MHC presentation. Considering this in conjunction with the expression of a highly diverse repertoire, it is clear that γ/δ cells play a major role in the shark immune system.

The multi-cluster arrangement raised the specter that immunoglobulin genes may not be clonally expressed, since it was difficult to envisage how rearrangement and transcription were regulated in light of what was known of mechanisms in higher vertebrates. However, it has now been unequivocally (158, 174) demonstrated that clonality, an essential requirement of the adaptive immune system, is a feature of immunoglobulin expression in B cells of the nurse shark. Hsu and colleagues (158, 174) have proposed a stochastic model for isotype and allelic exclusion in which limiting amounts of nuclear factors allow the formation of very few complexes able to catalyze transcription and rearrangement. Thus, only one to three clusters can be targeted for rearrangement in a single B-cell. In contrast to the Ig clusters, the sandbar shark TCR γ locus is a single copy, simple translocon. In common with the Ig clusters, but in marked contrast to the translocons of higher vertebrates, it is small, encompassing approximately 30 kb. Thus, for both the Ig clusters and the TCR translocons, part of the processes that are considered to be important for allelic exclusion in higher vertebrates, namely chromatin contraction and rearrangement over long distances, are not a factor here (174). However, our data
suggest that allelic exclusion is occurring at the TCRγ locus. All of the cDNA sequences I isolated were functional, i.e., they contained no frame shifts or stop codons. In contrast, about half of the genomic rearrangements I examined were nonproductive, i.e., frame shifts were present in the joining regions. These results indicate that non-productively rearranged and non-rearranged alleles are not expressed, as would be expected if allelic exclusion were occurring. Although not conclusive, we believe that the TCR γ genes are allelically excluded and clonally expressed. This suggests that the exclusion model proposed by Hsu et al (158, 174) may also apply to the TCR γ locus. Another possible explanation of this phenomenon is Nonsense-Mediated Decay of untranslatable mRNA (175). It is possible that I did not identify the mRNAs of non-productively rearranged DNA due to the decay of these mRNA.

The nature of the primordial immune receptor is subject to considerable speculation. In its simplest form, I would expect that this protein was a cell surface receptor similar to TCR, but also similar to antibody in its ability to directly bind antigen. Since a) these are properties of γ/δ receptors, b) gamma chains appear phylogenetically primitive (57) and, c) γ/δ receptors have structural features more in common with Ig Fab than TCR α/β (171, 176, 177), some have speculated that the primordial receptor was a primitive γ/δ molecule (57). Another primitive feature must have been the ability to generate a diverse binding repertoire. Was hypermutation one of these mechanisms? I believe that several considerations indicate it was. Activation-Induced Cytidine Deaminase (AID), the enzyme responsible for hypermutation and isotype switch rearrangement, is conserved in
all vertebrates. AID plays a role in the generation of diversity in agnathans, even though these animals have a completely different genetic system as the basis of their immune system (17, 18, 178). Thus, AID is ancient and was present at the evolutionary origins of the immune system. Our results for sandbar shark γ chain V regions supports this model since they strongly suggests that hypermutation occurred in the ancestral γ/δ receptor. If so, then it almost certainly occurred in the proposed γ/δ like primordial receptor. Lee et al (125) have also proposed that hypermutation is an ancient mechanism for generating diversity, perhaps preceding somatic recombination.

The characterization of the shark immune system continues to yield surprises. The finding that the TCR γ locus in the sandbar shark is a translocon was expected (142), but it is interesting that it is relatively simple and small. It is a major surprise that the V regions undergo hypermutation. This suggests to us an expanded role for γ/δ T cells in the shark immune system.

The adaptive immune system uses V(D)J recombination to generate an almost unlimited antigen repertoire with a limited number of genes. How this rather complicated and economic system evolved is always a hot topic in research of the evolution of the adaptive immune system. In higher vertebrates, this system (including both Ig and TCRs) has been described as translocon with features of clusters. Thus, the question is, which one comes first, translocon or cluster? The focus is on the genomic arrangement of the adaptive immune system of elasmobranches, the most distant living relatives of mammals with an adaptive immune system. A surprising feature of the organization of shark
immunoglobulin genes was the cluster arrangement of both light chains (135-137) and heavy chains (138-140). These findings favor the cluster-origin models. However, as reported in this dissertation and other’s reports in the literature, shark TCR genes are arranged in a classic single copy translocon. The co-existence of both translocon(TCR) and cluster(Ig) in a single animal provided no definitive answer to the previous question. However, our parallel comparison of known genomic maps of TCR γ gene of different species may provide some insight to the evolution of TCR genes, or if the immunoglobulin genes and TCR genes co-evolved, the evolution of immunoglobulin genes. Up to now, TCR γ gene of 7 species (sandbar shark, Atlantic salmon, puff fish, cattle, sheep, mice and human) has been fully or partially genomic mapped (Table 1). The sandbar shark TCR γ gene is arranged as a typical translocon(5 V genes -3 J segments -1 C gene), with a relatively small size of 40 kb. A striking feature of TCR γ gene of other species is the presence of Vs-Js-C cluster type arrangements, the exception is human which contains 2 J-C clusters, instead of V-J-C clusters. The size and the organization of this gene are relatively conserved from teleost to human. This uniform cluster type of gene arrangement is very different from that of immunoglobulin, which is mainly a translocon. Interestingly, the small translocon of shark seemingly can serve as a prototype motif for the cluster arrangement of TCR γ gene of other species. Here, I propose a theoretical model for the evolution of TCR γ gene(Figure 11). Four steps are included in this model. The ancestor gene for TCRγ is likely to be a simple V-J-C “arrangement pattern”. Then, duplication of V and J segments occurred, which leads to a similar organization as the shark TCRγ gene. Next, duplication of Vs-Js-C cluster took
place, resulting in the cluster type organization of TCR $\gamma$ genes I can see in other species. Finally, due to the selection pressure, more V or J genes were generated or lost (pseudo-genes). DNA transposon, containing the RAG gene, has been proposed and generally accepted as the cause for the duplication of V/J genes and clusters (179-183).

Considering the dominant cluster arrangement of shark immunology heavy and light chain genes, this duplication of clusters in TCR $\gamma$ is indeed not an isolated event. The RAG transposon induced duplication of clusters took place in a rather drastic way around the time chondricthyes diverged from the mainstream of vertebrate evolution, which resulted in the cluster appearance of TCR $\gamma$ and shark immunoglobulin genes. An interesting finding is that there is no duplication of the C gene inside each cluster, which is probably the reason that no class switch developed in TCR $\gamma$ during evolution.

Overall, the comparison of gene arrangement pattern between Igs and TCRs suggested that the evolution of Ig and TCR genes were on a very different track, or at least at a different speed. This difference may result from, or in the difference of their functions.
Table 1: Known Genome of TCR $\gamma$ gene of different species

<table>
<thead>
<tr>
<th>Species</th>
<th>Locus</th>
<th>Size (~kb)</th>
<th>V</th>
<th>Pseudo-V</th>
<th>J</th>
<th>C</th>
<th>notes</th>
<th>ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sandbar shark</td>
<td>1</td>
<td>40</td>
<td>5</td>
<td>none</td>
<td>3</td>
<td>1</td>
<td>Pure translocon</td>
<td></td>
</tr>
<tr>
<td>Atlantic salmon</td>
<td>2</td>
<td>$\gamma_1$:26</td>
<td>10</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>$\gamma_1$ contain 4 tandemly repeated Vs-J-C clusters.</td>
<td>(173)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_2$:9.3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$\gamma_2$ is a single non-expressed V–J–C cluster.</td>
<td></td>
</tr>
<tr>
<td>Zebrafish Pufferfish</td>
<td>1</td>
<td>n/a</td>
<td>8</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>partial information at least two linked VJC clusters.</td>
<td>(184)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>partial information at least two linked VJC clusters.</td>
<td>(185)</td>
</tr>
<tr>
<td>Chicken</td>
<td>1</td>
<td>n/a</td>
<td>~3</td>
<td>n/a</td>
<td>3</td>
<td>1</td>
<td>Unsequenced/unannotated</td>
<td>(186)</td>
</tr>
<tr>
<td>Human</td>
<td>1</td>
<td>200</td>
<td>14</td>
<td>8</td>
<td>5</td>
<td>2</td>
<td>Mainly translocon Cluster character: contain Js-C repeat.</td>
<td>IMGT database</td>
</tr>
<tr>
<td>Mice</td>
<td>1</td>
<td>200</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>Contain 3 tandemly repeated Vs-J-C clusters</td>
<td>(171)</td>
</tr>
<tr>
<td>sheep</td>
<td>2</td>
<td>$\gamma_1$:14</td>
<td>10</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>Contain 3 tandemly repeated Vs-Js-C clusters</td>
<td>(171)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_2$:90</td>
<td>3</td>
<td>0</td>
<td>6</td>
<td>3</td>
<td>Contain 3 tandemly repeated Vs-Js-C clusters</td>
<td>(187)</td>
</tr>
<tr>
<td>cow</td>
<td>2</td>
<td>$\gamma_1$:21</td>
<td>13</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>Contain 3 tandemly repeated Vs-Js-C clusters</td>
<td>(188)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\gamma_2$:92</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>Contain 3 tandemly repeated Vs-Js-C clusters</td>
<td></td>
</tr>
</tbody>
</table>
Figure 11: Scheme of the evolution of the TCR γ gene.
CHAPTER III: ANALYSIS OF SOMATIC HYPERMUTATION IN SANDBAR SHARK
TCR GAMMA LOCUS

3.1 Summary

It is an accepted dogma that T cell receptors rarely undergo the process of somatic hypermutation. However, in this chapter I demonstrate that somatic hypermutation occurs in the V region of the TCR γ gene of the sandbar shark and perform extensive analysis of the mutations. Two distinct patterns of mutation, point mutations and tandem mutations, are identified. These two types of mutations are different in many aspects. The data indicates that point mutations occur by a mechanism similar to that of the somatic hypermutation process of immunoglobulin genes of mammals. However, tandem mutations are unique to sharks and the data suggests that an error-prone DNA polymerase, with terminal deoxynucleotidyl transferase (TdT)-like function may be responsible for the generation of this unique mutation. My analysis suggests that the purpose of somatic mutations in shark TCR γ V-regions is to generate a more diverse repertoire in γ/δ receptors, rather than receptors with higher affinity. These findings have an impact on our understanding of the development of SHM during evolution, as well as the origins of diversity in the immune system.
3.2 Introduction

In mammals, Somatic Hypermutationalong with subsequent Ag-based selection, is responsible for generating antibody populations that have higher affinity to antigens, a process called affinity maturation. Activation induced cytidine deaminase (AID) has been shown to be the key player for SHM, as well as for class switch recombination (CSR) and gene conversion (GC) (60, 61). According to the current model (Figure 2)(189), SHM is a two-step process. The first step is AID mediated conversion of cytidine (C) to uridine (U). The second step is the transformation of the U:G lesion to a mutation in the V region. For C:G base pairs, a mutation is generated via base excision repair (BER) in which uracil DNA glycosylase (UNG) removes the U base. UNG mediated cleavage of the U base from the DNA backbone results in a non-instructive abasic site. Next, error prone polymerases are recruited to the abasic site to replace the C in the original C:G base pair with another base. Also, mismatch repair (MMR) can generate mutations at an A:T base pair. The MMR heterodimer MSH2/MSH6 can bind to a mismatched base pair as well as to a U:G base pair. Exonuclease1, a known partner of MSH2 (83), is responsible for the cleavage and removal of the section of DNA containing one of the mispaired bases. The recruitment of an error–prone polymerase to the DNA lesion leads to mutations in the excised region. So far, we know very little about the targeting and regulatory mechanisms underlying SHM. Although T cell receptors are structurally similar to immunoglobulins (Igs) and utilize similar mechanisms to generate receptor diversity, they are believed to be incapable of SHM (190), because mutations in the CDR 1 and 2 regions of TCR αβ would disrupt recognition of MHC molecules, and mutation of any of the three CDR
could yield autoreactive T cells. Although TCR $\gamma/\delta$ does not need MHC for antigen presentation (30, 32, 191), SHM in TCR $\gamma/\delta$ V genes has not been reported. In the previous chapter, I fully sequenced the TCR $\gamma$ translocon locus and 73 cDNA clones containing V regions of sandbar shark. The sequencing of both germline and cDNA enabled me to undertake a comprehensive study of the mutation patterns in TCR $\gamma$ V regions. Here I conclude that SHMs in sandbar shark TCR $\gamma$ V regions have a similar pattern to that of Ig genes of nurse shark, suggesting that both shark B cells and $\gamma/\delta$ T cells utilize similar mechanisms. Analysis of the pattern of tandem mutations in shark suggests that an error-prone DNA polymerase with terminal deoxynucleotidyl transferase (TdT)-like function may be responsible for the generation of this unique mutation. Surprisingly, database mining suggests that low levels of SHM may occur in TCR $\gamma$ genes of some higher vertebrates.
3.3 Materials and Methods

3.3.1 Animal

Sandbar shark spleen was provided by Dr. Carl Luer (Mote Marine Laboratories, Sarasota, FL). All the genomic DNA and mRNA are collected from a single animal.

3.3.2 Genomic and cDNA sequence of sandbar shark TCR γ gene

cDNA sequences were obtained using the 5’ Rapid amplification of cDNA ends (RACE) technique (192). Briefly, 5’ RACE was performed using the SMART™ RACE cDNA Amplification Kit (Clontech) according to the manufacturer’s instructions. Three C region specific primers were used to rule out the potential PCR bias for certain V regions.

Primer 1: 5’TGGTTACCTTGCCCTGTCCGACTGC3’,
Primer 2: 5’CGTACCCCCGATTTTCCAGATAACC3’
Primer 3: 5’GTTACCTTGCCCTGTCCGACTG3’.

PCR reaction conditions: 50 μl reaction total, 29.5 μl sterile H₂O, 10μl 5X iProof 2 PCR buffer, 1 μl dNTP mix, 1 μl 50X iProof polymerase mix (BIO-RAD), 2.5 μl 5’RACE ready cDNA, 5 μl UPM (universal primer), 1μl GSP (gene specific primer). PCR settings: 5 cycles at 94°C 30 sec, 72°C 3 min, 5 cycles: 94°C 30 sec, 70°C 30 sec, 72°C 3 min.

Agarose gel electrophoresis of PCR reactions was performed on a 1% agarose gel, and the band of proper size carefully excised. The PCR products were purified using the QIA quick gel purification kit (QIAGEN) and cloned using the TOPO4 TA vector cloning
system (Invitrogen). Mach1-T1 chemically competent cells (Invitrogen) and a QIA prep miniprep kit (QIAGEN) were used for transformation & plasmid purification.

3.3.3 Germline DNA and cDNA sequences alignment

Alignments were performed using Mac Vector 10 (MacVector, Inc.).

3.3.4 Calculation of mutability indices

The mutability index is the observed number of mutations of a specific nucleotide divided by the expected number of mutations of that nucleotide. The expected number of mutations was derived by determining the frequency of the nucleotide within the sequenced TCRGV database multiplied by the total number of observed mutations within the database. A mutability index value of 1.00 would be assumed to represent the effects of random mutations.

3.3.5 Statistical analysis

\( \chi^2 \) analysis was used to compare observed mutation frequencies to corresponding expected mutation frequencies. A p value <0.05 was considered statistically significant. \( \chi^2 \) analysis was also used for the analysis of antigen selection pressure on TCR \( \gamma \) V genes. A p value <0.05 was considered statistically significant. Fisher exact test was used for analysis of purine transition preference. A p value <0.05 was considered statistically significant. t test was used for the comparison of the mutation frequency among individual domains. A p value <0.05 was considered statistically significant. All the statistical analyses were performed by SAS program.
3.4 Results

3.4.1 Somatic hypermutation in sandbar shark TCR \( \gamma \) V genes

I identified, performed the genomic mapping, and sequenced the TCR \( \gamma \) locus of sandbar shark (accession number FJ854492) using standard genomic PCR and chromosome walking technique {Universal Genome Walker kit (Clontech)}. I sequenced a total of 73 TCR \( \gamma \) V region cDNA clones (accession numbers FG854417-FG854491, Figure 12). Despite the low identity between human and shark TCR \( \gamma \) V genes, eleven amino acid sites are conserved throughout evolution (Figure 13A), and these enabled us to accurately identify the CDRs and FRs in the shark V genes. The amino acid sequences of individual V genes share less than 50% identity with each other, and there is a major difference in the size of CDR1 (5 to 9 amino acids) and CDR2 (6 to 8 amino acids) among individual V genes. Alignment of the cDNA sequences with their parental genomic DNA sequence allowed me to analyze the mutation profile of sandbar shark TCR \( \gamma \) V genes. Almost half (33) of the 73 cDNA clones contain at least one mutation in their V regions (14 GV1 clones, 5 GV2 clones, 6 GV3 clones, 7 GV4 clones, and only 1 GV5 clone). Twenty-two clones contain two or more mutations and twelve clones contain tandem mutations. Although somatic mutations may occur in the CDR3 region, these are extremely hard to analyze due to extensive deletions and additions at the VJ junctions, and are not included in this report. Only clones containing mutations are included in the following analysis, and because of the limited number, GV5 clones are excluded. The alignment of amino acid sequences with their parental genomic sequence
Figure 12: The cDNA sequences of the GV families.

A total of 73 different cDNA clones were obtained using 5'RACE and C region specific primers. The sequences are grouped into 22 GV1 clones, 18 GV2 clones, 17 GV4 clones, and 4 GV5 clones. The parent genomic sequences are shown in uppercase above each group. Mutations in individual sequences are indicated, while identical sequences are represented as dots.
Figure 13: Somatic hypermutation in sandbar shark TCR γ V genes. A) The amino acid alignment of sandbar shark TCR γV genes. The IMGT unique numbering system was used for numbering the amino acid positions of the V regions (193). Framework (FRs) and CDRs are indicated. Conserved amino acids are highlighted. H: human TCRGV10; M: mice TCRGV1. B) Amino acid comparison of the parental sequence with the GV1 clones are shown. Only clones containing at least one mutation are shown. Individual domains are indicated by lines above the diagram. Amino acids identical to the parental genomic sequence are shown as dots. Amino acid changes caused by mutations are shown. Amino acid changes caused by tandem mutations are underlined. The number of single nucleotide mutations, tandem groups, and nucleotides in tandem mutation groups of individual clones are shown at the end of the Figure. A total of 39 single nucleotide mutations lead to 20 amino acid substitutions. A total of 14 tandem mutations lead to 14 amino acid substitutions.
Figure 14: Comparison of the parental genomic sequence with the GV2,3 and 4 clones. (A) The amino acid sequences of GV2 clones containing at least one mutation. A total of 19 single nucleotide mutations lead to 10 amino acid substitutions. A total of 3 tandem mutations lead to 2 amino acid substitutions. (B) The amino acid sequences of GV3 clones containing at least one mutation. A total of 16 single nucleotide mutations lead to 12 amino acid substitutions. A total of 9 tandem mutations lead to 12 amino acid substitutions. (C) The amino acid sequences of GV4 clones containing at least one mutation.
is shown in Figure 13B (for GV1) and Figures 14 (for GV2-4). The amino acid changes resulting from tandem mutations are underlined. I identified 191 mutations in a total of 10305 nucleotides. (the nucleotide composition of V genes is shown in Table 2) The total mutation rate is 0.018/bp, which is comparable to the mutation rate in Ig genes of mice (0.016/bp) and sharks (0.015/bp) (125). As seen in shark Ig and the new antigen receptor (NAR) genes, two different patterns of mutation were identified in TCR γ; namely, single nucleotide mutations and tandem mutations.

3.4.2 Distribution of mutations

The mutation rates of the different domains (CDRs and FRs) of the TCRγ V regions are shown in Figure 15. In the analysis of single nucleotide mutations, two domains (CDR1 and FR2) have a higher mutation rate than the other domains. CDR1 has the highest mutation rate in two V genes and the second highest rate in one, with an overall mutation rate of 0.017/bp. Similarly, FR2 has the highest mutation rate in one V gene, the second highest mutation rate in two, with an overall mutation rate of 0.009/bp. The mutation rate of CDR1 is significantly higher than that of FR1 (p=0.011), CDR2 (p<0.0001), and FR3 (p=0.024). The mutation rate of FR2 is significantly higher than that of FR1 (p=0.035) and CDR2 (p<0.0001). However, it barely failed the χ2 test in comparison with FR3 (p=0.054). There is no significant difference in the mutation rate between CDR1 and FR2 (p=0.112). One “cold” domain (CDR2) of mutation was identified when compared with
Figure 15: The mutation rates in individual domains of V genes.

(A) Mutation rates of single nucleotide mutations in individual domains of V genes. A “hot” domain for mutations is indicated by hollow arrows. A “cold” domain for mutations is indicated by solid arrows. A statistical analysis (t test) of the significance of the differences between mutation rates of different domains is shown in the table below the Figure. Significant differences are underlined.

(B) Mutation rates (including both single nucleotide and tandem mutations) in individual domains of V genes. The solid (black) curves stand for the mutation frequency of single nucleotide mutations. The dashed (gray) curves stand for overall mutation frequency. The area between these two curves shows tandem mutation frequency.
Table 2. Summary of nucleotides distribution in V genes.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>GV1</td>
<td>1246</td>
<td>938</td>
<td>1176</td>
<td>1218</td>
<td>4578</td>
</tr>
<tr>
<td>GV2</td>
<td>475</td>
<td>340</td>
<td>335</td>
<td>395</td>
<td>1615</td>
</tr>
<tr>
<td>GV3</td>
<td>441</td>
<td>419</td>
<td>425</td>
<td>475</td>
<td>1760</td>
</tr>
<tr>
<td>GV4</td>
<td>658</td>
<td>546</td>
<td>574</td>
<td>574</td>
<td>2352</td>
</tr>
<tr>
<td>Total</td>
<td>3087</td>
<td>2444</td>
<td>2762</td>
<td>2923</td>
<td>10305</td>
</tr>
</tbody>
</table>

Table 3. The single nucleotide mutation profiles of TCR γ V genes

<table>
<thead>
<tr>
<th></th>
<th>Leader seq</th>
<th>FR1:</th>
<th>CDR1</th>
<th>FR2:</th>
<th>CDR2</th>
<th>FR3:</th>
</tr>
</thead>
<tbody>
<tr>
<td>GV1</td>
<td>MR(^3): 0</td>
<td>0.45×10(^{-2})</td>
<td>0.15×10(^{-1})</td>
<td>0.12×10(^{-1})</td>
<td>0.99×10(^{-2})</td>
<td>0.72×10(^{-2})</td>
</tr>
<tr>
<td>GV2</td>
<td>MR: 0.37×10(^{-2})</td>
<td>0.13×10(^{-1})</td>
<td>0.83×10(^{-2})</td>
<td>0.15×10(^{-1})</td>
<td>NA</td>
<td>0.13×10(^{-1})</td>
</tr>
<tr>
<td>GV3</td>
<td>MR: 0.31×10(^{-2})</td>
<td>0.88×10(^{-2})</td>
<td>0.16×10(^{-1})</td>
<td>0.20×10(^{-1})</td>
<td>NA</td>
<td>0.76×10(^{-2})</td>
</tr>
<tr>
<td>GV4</td>
<td>MR: 0.13×10(^{-1})</td>
<td>0.17×10(^{-1})</td>
<td>0.26×10(^{-1})</td>
<td>NA</td>
<td>NA</td>
<td>0.78×10(^{-2})</td>
</tr>
<tr>
<td>overall bp</td>
<td>1785</td>
<td>2695</td>
<td>792</td>
<td>1785</td>
<td>696</td>
<td>3834</td>
</tr>
<tr>
<td>MR: 0.39×10(^{-2})</td>
<td>0.92×10(^{-2})</td>
<td>0.17×10(^{-1})</td>
<td>0.12×10(^{-1})</td>
<td>0.44×10(^{-2})</td>
<td>0.85×10(^{-2})</td>
<td></td>
</tr>
</tbody>
</table>

\(^3\)Mutation rate (MR) is calculated by the number of single nucleotide substitutions divided by the total nucleotide number.

Table 4. The overall mutation profile of TCR γ V genes (including both single nucleotide and tandem mutations)

<table>
<thead>
<tr>
<th></th>
<th>Leader seq</th>
<th>FR1:</th>
<th>CDR1</th>
<th>FR2:</th>
<th>CDR2</th>
<th>FR3:</th>
</tr>
</thead>
<tbody>
<tr>
<td>GV1</td>
<td>MR(^3): 0</td>
<td>0.45×10(^{-2})</td>
<td>0.21×10(^{-1})</td>
<td>0.23×10(^{-1})</td>
<td>0.23×10(^{-1})</td>
<td>0.13×10(^{-1})</td>
</tr>
<tr>
<td>GV2</td>
<td>MR: 0.37×10(^{-2})</td>
<td>0.13×10(^{-1})</td>
<td>0.83×10(^{-2})</td>
<td>0.39×10(^{-1})</td>
<td>NA</td>
<td>0.17×10(^{-1})</td>
</tr>
<tr>
<td>GV3</td>
<td>MR: 0.31×10(^{-2})</td>
<td>0.11×10(^{-1})</td>
<td>0.32×10(^{-1})</td>
<td>0.33×10(^{-1})</td>
<td>0.42×10(^{-1})</td>
<td>0.20×10(^{-1})</td>
</tr>
<tr>
<td>GV4</td>
<td>MR: 0.13×10(^{-1})</td>
<td>0.33×10(^{-1})</td>
<td>0.74×10(^{-1})</td>
<td>NA</td>
<td>NA</td>
<td>0.17×10(^{-1})</td>
</tr>
<tr>
<td>overall bp</td>
<td>1785</td>
<td>2695</td>
<td>792</td>
<td>1785</td>
<td>696</td>
<td>3834</td>
</tr>
<tr>
<td>MR: 0.39×10(^{-2})</td>
<td>0.13×10(^{-1})</td>
<td>0.33×10(^{-1})</td>
<td>0.22×10(^{-1})</td>
<td>0.19×10(^{-1})</td>
<td>0.16×10(^{-1})</td>
<td></td>
</tr>
</tbody>
</table>
other domains (p<0.0001). In fact, among all the single nucleotide mutations, only three mutations were found in this domain. Although AID deaminates substrates preferentially at WRCY or RGYW hotspots (where W= A or T, R= purine, Y= pyrimidine) (194), a detailed analysis of the hotspot motifs revealed that there are none in CDR2 of GV1-3 and only one motif in CDR2 of GV4. However, all three single nucleotide mutations took place in GV1 rather than GV4. The CDR2 region may be intrinsically resistant to single nucleotide mutations. A detailed statistical analysis is shown in Figure 15A.

No regularity in the distribution of tandem mutations (a total of 33 tandems, including 85 nucleotides) is apparent (Figure 15B, area between solid and dash curves). The overall distribution pattern of mutations (including single nucleotide and tandem mutations) is different from that of single nucleotide mutations along. CDR2 is no longer a cold domain for mutations (with a mutation rate of 0.019/bp). The mutation rate of CDR1 remains the highest (0.033/bp) among all domains; while the mutation rates of other domains are also relatively high (detailed analysis of individual V genes is shown in Tables 3 and 4).

### 3.4.3 Single nucleotide mutations

A total of 106 single nucleotide mutations were identified, with a bias towards G and C bases. Overall, 64.2% of all single nucleotide mutations are GC mutations (p=0.011, \( \chi^2 \) test), varying from 51.2% to 84.2% among individual V genes. Mutation indices (MI) of A, C, G, T are 0.76, 1.70, 1.16 and 0.59 respectively, with mutations of C bases significantly higher and mutations of T bases significantly lower than the expected
number of mutations if there were no bias. Although the MI of A is higher than that of T, statistically there is no significant difference between A and T. In addition, a bias towards transition mutations was found. If the mutations occurred randomly, the transitions should be about one third of all mutations (theoretically 33.3%). However, 53.8% of all single nucleotide mutations are transitions, with a variation from 46.1% to 70% among individual V genes. The observed frequency of transitions is significantly higher (p=0.002, \( \chi^2 \) test) than the expected frequency. Another interesting finding is that there is a relatively stable replacement mutation (R) frequency among V genes. Of the 106 single nucleotide mutations, 56 were replacement mutations (52.8%). In individual V genes, the R frequency is 51.3% (20 Rs out of 39 mutations) in GV1, 52.6% (10 Rs out of 19 mutations) in GV2, 75% (12Rs out of 16 mutations) in GV3, and 50% (14Rs out of 28 mutations) in GV4. This comparatively high replacement mutation frequency indicates that single nucleotide mutations in the V genes of TCR \( \gamma \) of shark might contribute to the generation of a larger repertoire. Detailed analyses of single nucleotide mutations are shown in Table 5 and Table 6A and Table 7-10 (for individual V regions).

3.4.4 Tandem mutation

Tandem mutation is a pattern of mutations in which two or more consecutive nucleotides are mutated. This unique pattern has been identified in Ig genes and the NAR genes of shark. In our analysis of TCR \( \gamma \) V genes, a total of 33 tandem mutations (involving 85 nucleotides) were found in 12 of the 33 cDNA clones that contained mutations. Tandem
<table>
<thead>
<tr>
<th></th>
<th>All single base substitutions</th>
<th>All tandem substitutions</th>
<th>All substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>from</td>
<td>To</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>-</td>
<td>13</td>
</tr>
<tr>
<td>G</td>
<td>19</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>106</td>
</tr>
</tbody>
</table>

*GC mutation: 64.2% (p=0.011), transition: 53.8%

*GC mutation: 45.9%, transition: 34.1%

*GC mutation: 56.0% (p=0.052), transition: 45.0% (overall)

---

*a The mutability index is the observed number of mutations in a specific nucleotide divided by the expected number of mutations in that nucleotide. The expected number of mutations was derived by determining the frequency of the nucleotide within the sequenced TCRGV database multiplied by the total number of observed mutations within the database. A mutability index value of 1.00 would be assumed to represent the effects of random mutations. The observed and the expected number of mutations were compared by X2 analysis, and the significant differences are indicated in the footnotes to this table. The significantly mutable nucleotides are shown in bold type.

b statistically significant by X2 test (p<0.05)

c GC mutations and transition mutations in single base substitutions.

d GC mutations and transition mutations in tandem substitutions.
Table 6: Statistical analysis of mutability index of single nucleotide mutations (A) and tandem mutations (B).

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Expected</th>
<th>MI</th>
<th>P=</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>22</td>
<td>33</td>
<td>0.76</td>
<td>0.085</td>
</tr>
<tr>
<td>C</td>
<td>39</td>
<td>25</td>
<td>1.7</td>
<td>0.036</td>
</tr>
<tr>
<td>G</td>
<td>29</td>
<td>28</td>
<td>1.16</td>
<td>0.877</td>
</tr>
<tr>
<td>T</td>
<td>16</td>
<td>30</td>
<td>0.59</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Expected</th>
<th>MI</th>
<th>P=</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>24</td>
<td>26</td>
<td>0.92</td>
<td>0.736</td>
</tr>
<tr>
<td>C</td>
<td>13</td>
<td>20</td>
<td>0.65</td>
<td>0.175</td>
</tr>
<tr>
<td>G</td>
<td>26</td>
<td>23</td>
<td>1.13</td>
<td>0.611</td>
</tr>
<tr>
<td>T</td>
<td>22</td>
<td>24</td>
<td>0.92</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 7: Nature of base substitutions in TCRGV1

<table>
<thead>
<tr>
<th></th>
<th>All single base substitutions</th>
<th>All tandem substitutions</th>
<th>All substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>from A</td>
<td>-</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>G</td>
<td>7</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>7</td>
<td>11</td>
</tr>
</tbody>
</table>

GC mutation: 51.2%, transition: 46.1%

Table 8: Nature of base substitutions in TCRGV2

<table>
<thead>
<tr>
<th></th>
<th>All single base substitutions</th>
<th>All tandem substitutions</th>
<th>All substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>from A</td>
<td>-</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>G</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>6</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

GC mutation: 60%, transition: 70%

GC mutation: 42.9%, transition: 57.1%
Table 9: Nature of base substitutions in TCRGV3.

<table>
<thead>
<tr>
<th></th>
<th>All single base substitutions</th>
<th>All tandem substitutions</th>
<th>All substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>to A from A</td>
<td>A</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>G</td>
<td>5</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

\(^{a}\)GC mutation: 84.2\% , transition: 52.6\%

\(^{b}\)GC mutation: 43.5\% , transition: 26.1\%

Table 10: Nature of base substitutions in TCRGV4

<table>
<thead>
<tr>
<th></th>
<th>All single base substitutions</th>
<th>All tandem substitutions</th>
<th>All substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>to A from A</td>
<td>A</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

\(^{a}\)GC mutation: 71.4\% , transition: 53.6\%

\(^{b}\)GC mutation: 43.5\% , transition: 30.4\%
mutations comprised 44.5% of all mutations. The size of tandem mutations varies from 2 to 5 nucleotides. Tandem mutations are different from single nucleotide mutations in several ways. First, unlike single nucleotide mutations, no GC bias is present in tandem mutations. Mutations at GC bases comprise only 45.9% of all the nucleotides in tandem mutations, varying from 42.9% to 50% among V genes (detailed information is shown in Table 5 and Table 6B and Table7-10). Second, no transition preference was found (34.1% in tandem mutations, compared with 64.2% in single nucleotide mutations). In individual V regions, the frequency of the transitions in tandem mutations varied from 26.1% to 57.1%. Overall, the transition to transversion ratio of tandem mutations is significantly lower than that of single nucleotide mutations (p=0.008, in χ² test). However, due to the limited number of clones in each V gene, statistical analysis of each V gene failed to show a difference between single nucleotide and tandem mutations (detailed information is shown in Table 11). In this sample, we failed to find a significant purine transition preference (p=0.143) (detailed information is shown in Table 12), as has been identified in tandem mutations of Ig light chain genes of nurse shark (125). It is reasonable to expect that tandem mutation would have a higher efficiency in the replacement of amino acids than single nucleotide mutation, since any tandem mutation is highly likely to generate one or more replacements. A total of 33 tandem mutations led to 43 replacement of amino acids, or 1.3 replacements per tandem on average. Therefore, the replacement efficiency of tandem mutations is significantly higher than that of single nucleotide mutations (p=0.001, χ² test). Our data also suggests that the RGYW/WRCY hotspots defined in mammalian systems for somatic mutations
Table 11: statistical analysis ($\chi^2$ test) of the transition to transversion ratio in sandbar shark TCR $\gamma$ V genes.

(A)  

<table>
<thead>
<tr>
<th></th>
<th>Transition</th>
<th>Transversion</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gv1</td>
<td>Single nucleotides</td>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Tandem</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>GV2</td>
<td>Single nucleotides</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>tandem</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>GV3</td>
<td>Single nucleotides</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>tandem</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td>GV4</td>
<td>Single nucleotides</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>tandem</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>Overall</td>
<td>Single nucleotides</td>
<td>57</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>tandem</td>
<td>29</td>
<td>56</td>
</tr>
</tbody>
</table>

The overall transition to transversion ratio is significantly higher in single nucleotide mutations and is highlighted in bold.
Table 12: Comparison of purine and pyrimidine mutations in transition and transversions in single nucleotide and tandem mutations \(^{a}\) normalized after correction for base composition.

<table>
<thead>
<tr>
<th></th>
<th>Single nucleotide</th>
<th>Tandem</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Transition</td>
<td>Purine (R)/ pyrimidine (Y)</td>
</tr>
<tr>
<td>GV1</td>
<td>R</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>7</td>
</tr>
<tr>
<td>GV2</td>
<td>R</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>6</td>
</tr>
<tr>
<td>GV3</td>
<td>R</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>4</td>
</tr>
<tr>
<td>GV4</td>
<td>R</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>10</td>
</tr>
<tr>
<td>total</td>
<td>R</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Single nucleotide</th>
<th>Tandem</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Transversion</td>
<td>R/Y</td>
</tr>
<tr>
<td>GV1</td>
<td>R</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>7</td>
</tr>
<tr>
<td>GV2</td>
<td>R</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>5</td>
</tr>
<tr>
<td>GV3</td>
<td>R</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>4</td>
</tr>
<tr>
<td>GV4</td>
<td>R</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>7</td>
</tr>
<tr>
<td>total</td>
<td>R</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>23</td>
</tr>
</tbody>
</table>

(i) Transition ratios comparing purine and pyrimidine mutations in individual V genes; (ii) Transversion ratios comparing purine and pyrimidine mutations in individual V genes. \(^b\) p value for the normalized R/Y ratio of total transition mutations is 0.143 (\(\chi^2\) test), p value for the normalized R/Y ratio of total transversion mutations is 0.562 (\(\chi^2\) test).
are not responsible for tandem mutations. While RGYW/WRCY motifs covered about 19.4% of the overall sequence, only 9 out of 33 tandem mutations (27.2%) lie within or partly within a RGYW/WRCY motif. Statistical analysis failed to establish an association between hotspot motifs and tandem mutations (p=0.447, χ2 test). Close examination of individual tandem mutations also yielded some interesting information. In the tandem mutations, 87.9% were found to have purine–purine, or pyrimidine-pyrimidine repeats. Eighteen tandem mutations (54.5%) contain di- and tri-nucleotide repeats (underlined in Table 13). A similar di- and tri-nucleotide repeat pattern was also observed in shark Ig genes (both light chain and heavy chain)(125, 126), but the frequencies of the repeats vary (Table 14 A and B).

3.4.5 Selection pressure

The comparison of the replacement to synonymous substitution ratio (R/S) between the CDR and FR areas is conventionally used to evaluate the selection pressure on Ig genes. The R/S ratio of CDRs is generally expected to be significantly higher than that of FRs, since CDRs form the antigen binding site and should be subjected to the most selection pressure. However, this is not the case in the V genes of sandbar shark TCR γ chain. No appreciable difference in the R/S ratios was found between CDRs and FRs of TCR γ V genes. The R/S ratios are 1.83 in FR1, 0.86 in CDR1, 1.0 in FR2, 1.17 in CDR2, and 1.0 in FR3 (including both tandem and single nucleotide mutations). Overall, the R/S ratio in CDRs is 0.95 (19R/20S), and in FRs is 1.16 (73R/63S). Statistical analysis failed to show a difference between the CDRs and FRs (p=0.591, χ2 test). Detailed information is
Table 13: Tandem mutations in the V genes of TCRγ of sandbar shark.

<table>
<thead>
<tr>
<th>Clones(TCR)</th>
<th>mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCRg18</td>
<td>TTCT</td>
</tr>
<tr>
<td>V324a21</td>
<td>TC AA CC</td>
</tr>
<tr>
<td>TCRg21</td>
<td>AG GG TC</td>
</tr>
<tr>
<td>TCRgV8</td>
<td>GG AC GA</td>
</tr>
<tr>
<td>TgVcx14</td>
<td>TC GG AA</td>
</tr>
<tr>
<td>V324a1</td>
<td>CG</td>
</tr>
<tr>
<td>V324a8</td>
<td>AG</td>
</tr>
<tr>
<td>V324a2</td>
<td>CC TAG AAA TC AAT</td>
</tr>
<tr>
<td>GV5C21</td>
<td>CG GAG</td>
</tr>
<tr>
<td>GV5C312</td>
<td>TTT GACCA</td>
</tr>
<tr>
<td>V324a22</td>
<td>TAAC GAA CA AAA</td>
</tr>
<tr>
<td>GV5C311</td>
<td>GT</td>
</tr>
</tbody>
</table>

The “duplicate” or “triplicate” tandems (AA, TT, CC, and GG) are underlined.

Table 14: Tandem mutations in the V genes of Ig H(A) and Ig L(B) of nurse shark.

<table>
<thead>
<tr>
<th>ClonesV18(IgH)</th>
<th>mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>J3</td>
<td>CG</td>
</tr>
<tr>
<td>Q10</td>
<td>CT CT TT AA</td>
</tr>
<tr>
<td>T16</td>
<td>CAC CT AAG</td>
</tr>
<tr>
<td>CH12</td>
<td>TT AT</td>
</tr>
<tr>
<td>08</td>
<td>CC CA CT CAGTACG</td>
</tr>
<tr>
<td>A5</td>
<td>TTC TC AA CTTC CG GG</td>
</tr>
<tr>
<td>N40</td>
<td>TT</td>
</tr>
<tr>
<td>H3</td>
<td>GAG</td>
</tr>
<tr>
<td>E4</td>
<td>GTT CG GC</td>
</tr>
<tr>
<td>CH7</td>
<td>GG GC AC AA</td>
</tr>
<tr>
<td>N27</td>
<td>GCA</td>
</tr>
<tr>
<td>Clones(lg l)</td>
<td>mutations</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>C5-03</td>
<td>TCC</td>
</tr>
<tr>
<td>C5-04</td>
<td>CG</td>
</tr>
<tr>
<td>C5-07</td>
<td>CGTC</td>
</tr>
<tr>
<td>C5-012,2</td>
<td>GA</td>
</tr>
<tr>
<td>C5-A</td>
<td>TC</td>
</tr>
<tr>
<td>C5-B</td>
<td>CG</td>
</tr>
<tr>
<td>C5-1</td>
<td>CGC</td>
</tr>
<tr>
<td>C5-5</td>
<td>AAT</td>
</tr>
<tr>
<td>C5-7</td>
<td>CG</td>
</tr>
<tr>
<td>C5-8</td>
<td>AG</td>
</tr>
<tr>
<td>C5-11</td>
<td>GC</td>
</tr>
<tr>
<td>C5-12</td>
<td>GGG</td>
</tr>
<tr>
<td>C5-15</td>
<td>CG</td>
</tr>
<tr>
<td>C5-16</td>
<td>TT</td>
</tr>
<tr>
<td>C5-18</td>
<td>GC</td>
</tr>
<tr>
<td>C5-05</td>
<td>GC</td>
</tr>
<tr>
<td>C5A-3</td>
<td>AG</td>
</tr>
<tr>
<td>C5A-5</td>
<td>AC</td>
</tr>
<tr>
<td>C5A-11</td>
<td>TG</td>
</tr>
<tr>
<td>C5A-14</td>
<td>TC</td>
</tr>
<tr>
<td>C5A-15</td>
<td>GC</td>
</tr>
<tr>
<td>C5A-19</td>
<td>CG</td>
</tr>
<tr>
<td>C5A-25</td>
<td>AA</td>
</tr>
<tr>
<td>C5A-29</td>
<td>TT</td>
</tr>
<tr>
<td>C5A-37</td>
<td>GT</td>
</tr>
<tr>
<td>C5B-1</td>
<td>GG</td>
</tr>
<tr>
<td>C5B-2</td>
<td>GC</td>
</tr>
<tr>
<td>C5B-3</td>
<td>TT</td>
</tr>
<tr>
<td>C5B-8</td>
<td>AA</td>
</tr>
<tr>
<td>C5B-19</td>
<td>CG</td>
</tr>
</tbody>
</table>

The “duplicate” or “triplicate” tandems (AA, TT, CC, and GG) are underlined.
Table 15: Ratio of replacement to synonymous mutations in V genes of TCR γ of sandbar shark (including both tandem and single nucleotide mutations)

<table>
<thead>
<tr>
<th>Leader seq</th>
<th>FR1</th>
<th>CDR1</th>
<th>FR2</th>
<th>CDR2</th>
<th>FR3</th>
</tr>
</thead>
<tbody>
<tr>
<td>GV1</td>
<td>R:0</td>
<td>S:0</td>
<td>R:5</td>
<td>S:1</td>
<td>R:10</td>
</tr>
<tr>
<td></td>
<td>R:5</td>
<td>S:1</td>
<td>R:2</td>
<td>S:5</td>
<td>R:5</td>
</tr>
<tr>
<td></td>
<td>R:10</td>
<td>S:10</td>
<td>R:5</td>
<td>S:2</td>
<td>R:14</td>
</tr>
<tr>
<td>GV2</td>
<td>R:1</td>
<td>S:0</td>
<td>R:3</td>
<td>S:2</td>
<td>R:1</td>
</tr>
<tr>
<td></td>
<td>R:1</td>
<td>S:0</td>
<td>R:3</td>
<td>S:7</td>
<td>R:0</td>
</tr>
<tr>
<td></td>
<td>R:1</td>
<td>S:0</td>
<td>R:3</td>
<td>S:7</td>
<td>R:0</td>
</tr>
<tr>
<td>GV3</td>
<td>R:1</td>
<td>S:0</td>
<td>R:3</td>
<td>S:2</td>
<td>R:2</td>
</tr>
<tr>
<td></td>
<td>R:2</td>
<td>S:2</td>
<td>R:7</td>
<td>S:3</td>
<td>R:2</td>
</tr>
<tr>
<td></td>
<td>R:2</td>
<td>S:4</td>
<td>R:9</td>
<td>S:4</td>
<td>R:9</td>
</tr>
<tr>
<td>GV4</td>
<td>R:2</td>
<td>S:3</td>
<td>R:11</td>
<td>S:7</td>
<td>R:7</td>
</tr>
<tr>
<td></td>
<td>R:11</td>
<td>S:7</td>
<td>R:0</td>
<td>S:0</td>
<td>R:0</td>
</tr>
<tr>
<td></td>
<td>R:11</td>
<td>S:7</td>
<td>R:0</td>
<td>S:0</td>
<td>R:0</td>
</tr>
<tr>
<td>Overall</td>
<td>R:1</td>
<td>S:3</td>
<td>R:22</td>
<td>S:12</td>
<td>R:12</td>
</tr>
<tr>
<td></td>
<td>R:12</td>
<td>S:14</td>
<td>R:20</td>
<td>S:20</td>
<td>R:7</td>
</tr>
<tr>
<td></td>
<td>R:12</td>
<td>S:14</td>
<td>R:20</td>
<td>S:20</td>
<td>R:7</td>
</tr>
</tbody>
</table>

Table 16: Ratio of replacement to synonymous single nucleotide mutations in V genes.

<table>
<thead>
<tr>
<th>Leader seq</th>
<th>FR1</th>
<th>CDR1</th>
<th>FR2</th>
<th>CDR2</th>
<th>FR3</th>
</tr>
</thead>
<tbody>
<tr>
<td>GV1</td>
<td>R:0</td>
<td>S:0</td>
<td>R:5</td>
<td>S:1</td>
<td>R:10</td>
</tr>
<tr>
<td></td>
<td>R:5</td>
<td>S:1</td>
<td>R:2</td>
<td>S:5</td>
<td>R:5</td>
</tr>
<tr>
<td></td>
<td>R:10</td>
<td>S:10</td>
<td>R:5</td>
<td>S:2</td>
<td>R:14</td>
</tr>
<tr>
<td>GV2</td>
<td>R:1</td>
<td>S:0</td>
<td>R:3</td>
<td>S:2</td>
<td>R:1</td>
</tr>
<tr>
<td></td>
<td>R:1</td>
<td>S:0</td>
<td>R:3</td>
<td>S:7</td>
<td>R:0</td>
</tr>
<tr>
<td></td>
<td>R:1</td>
<td>S:0</td>
<td>R:3</td>
<td>S:7</td>
<td>R:0</td>
</tr>
<tr>
<td>GV3</td>
<td>R:1</td>
<td>S:0</td>
<td>R:3</td>
<td>S:2</td>
<td>R:2</td>
</tr>
<tr>
<td></td>
<td>R:2</td>
<td>S:2</td>
<td>R:7</td>
<td>S:3</td>
<td>R:2</td>
</tr>
<tr>
<td></td>
<td>R:2</td>
<td>S:4</td>
<td>R:9</td>
<td>S:4</td>
<td>R:9</td>
</tr>
<tr>
<td>GV4</td>
<td>R:2</td>
<td>S:3</td>
<td>R:11</td>
<td>S:7</td>
<td>R:7</td>
</tr>
<tr>
<td></td>
<td>R:11</td>
<td>S:7</td>
<td>R:0</td>
<td>S:0</td>
<td>R:0</td>
</tr>
<tr>
<td></td>
<td>R:11</td>
<td>S:7</td>
<td>R:0</td>
<td>S:0</td>
<td>R:0</td>
</tr>
<tr>
<td>Overall</td>
<td>R:1</td>
<td>S:3</td>
<td>R:22</td>
<td>S:12</td>
<td>R:12</td>
</tr>
<tr>
<td></td>
<td>R:12</td>
<td>S:14</td>
<td>R:20</td>
<td>S:20</td>
<td>R:7</td>
</tr>
<tr>
<td></td>
<td>R:12</td>
<td>S:14</td>
<td>R:20</td>
<td>S:20</td>
<td>R:7</td>
</tr>
</tbody>
</table>

Table 17: Statistic analysis of the ratio of replacement to synonymous mutation in V genes.

<table>
<thead>
<tr>
<th>Leader seq</th>
<th>FR</th>
<th>R</th>
<th>S</th>
<th>P value(X2 test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GV1</td>
<td></td>
<td>29</td>
<td>24</td>
<td>0.753</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>GV2</td>
<td></td>
<td>9</td>
<td>15</td>
<td>0.211</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>GV3</td>
<td></td>
<td>19</td>
<td>9</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>GV4</td>
<td></td>
<td>16</td>
<td>15</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>73</td>
<td>63</td>
<td>0.591</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>
available in Table 15 and Table 17. Analysis of single nucleotide mutations alone yielded similar results, in which R/S ratios are 1.4 in FR1, 0.86 in CDR1, 1.0 in FR2, NA in CDR2, and 0.65 in FR3 (detailed information is shown in Table 16). An interesting finding is that although single nucleotide mutations rarely occurred in CDR2 (only 3 out of 106 mutations), all three single nucleotide mutations led to replacements.

A parallel comparison of the R/S ratios among different antigen recognition systems of shark (Ig heavy chain and light chain, NAR, and TCR) gives us some insight about selection pressure on the shark immune system. The R/S ratio of Ig is much higher than what we observed in TCRγ of shark. For example, in the Ig light chain, the R/S ratios are CDR1 =3.68, FR2=5.14, CDR2=2.71, FR3=3.57, CDR3=2.68 (125). In addition, the R/S ratios are 3.6 in FR1, 5 in CDR1, 1.5 in FR2, 4 in CDR2, and 2.3 in FR3 for Ig heavy chain (126). Although the data from the Ig heavy chain of shark fail to show a statistically significant difference in the R/S ratios between CDRs and FRs, the pattern is similar to that observed in mammalian Ig heavy chain mutations. On the other hand, the R/S ratios for NAR are 1.57 in FR1, 6.1 in CDR1, 5.3 in CDR2, and 3.33 in FR3 (130). Not only is there a statistically significant difference in R/S ratios between CDRs and FRs, there is also a similarity to that of human Ig genes. Overall, the low R/S ratios observed in TCRγ suggests that there may be little or no selection pressure on the TCRγ gene of shark. Our analysis suggests that Ag-driven selection may not be occurring in shark g/d T cells and gd T cells in shark may not need a survival singla in order to live after SHM, whereas mammalian B cells clearly do (and Ag is providing signal). Thus, shark gd T cells will not be susceptible to apoptosis after SHM in order to provide a bigger receptor repertoire.
3.5 Discussion

3.5.1 TCR hypermutation

My data showed that diversity in sharks is generated by mutations in TCR $\gamma$ V regions. T lymphocytes are believed to be incapable of SHM (190). It is generally accepted that mutations in the CDR 1 and 2 regions of the $\alpha\beta$ receptor might disrupt or overaly enhance (to the point of autoimmunity) recognition of MHC Class I and II antigen presenting molecules. There are only two reports that show mutations in T lymphocytes in humans and mice (145, 146). In mice, a very small population of T lymphocytes recruited into the splenic germinal centers, were found to acquire mutations in the T cell receptor $\alpha$ locus, but not in the beta-locus (146). A similar mutation pattern was found in the TCR$\beta$ locus of T cells isolated from human splenic germinal centers (145). However, the mutation rates reported are far less than those of Igs, but are similar to those of some oncogenes associated with lymphomagenesis (92) and a proviral reporter gene integrated into the B cell genome (195). It is likely that the SHM observed in T cells is due to genomic instability caused by the overexpression of AID in the germinal centers, inadequate DNA repair machinery, or recruitment of error prone polymerases. So far, SHM has not been reported in TCR $\gamma\delta$ genes of any other species. On the other hand, the mutation rate of the shark T cell receptor $\gamma$ gene (0.018/bp) is comparable to that of Ig genes of mice (0.016/bp) and shark (0.015/bp) (125). In addition, unique tandem mutations are seen in the T cell receptor $\gamma$ gene of shark. It appears that in sharks, TCR and Ig share the mutation machinery for the generation of SHM. Although the mutation profiles are similar in shark Ig and TCR genes, detailed analysis reveals subtle
differences between these two systems, suggesting that the regulatory mechanisms may be different.

### 3.5.2 Single nucleotide mutation vs tandem mutations

In mice and humans, somatic mutations in Igs are evenly distributed among A, T, C, G nucleotides (196, 197). This unbiased distribution is also present in Ig light chain, Ig heavy chain and NAR genes of nurse shark. However, this is not always the case. A GC preference was found in the Ig heavy chain of frog (129), horned shark (198), teleost fish (127), and some B cell lines (116, 117). Different hypotheses have been proposed for these diverse observations: 1) A mutator system may consist of several mutators, and some have been silenced or become subdominant in mammals later in evolution (125, 199); 2) The primary specificity of the system is biased towards mutation of GC base pairs, but this is obscured by antigenic selection (199). However, we consider it more likely that a regulatory system was developed during evolution of the immune system. The regulatory system of SHM in sharks may be different from that of humans and mice. Furthermore, the apparent lack of germinal centers in sharks suggest that SHM machinery may be less localized or more loosely regulated than what is observed in mice and humans. In addition to the single nucleotide mutations, a unique pattern of tandem mutations is present in all three systems (Igs, NARs and TCRs). Because of the existence of a tandem mutation pattern that is different from the pattern of single nucleotide mutations (125), the mutations were separately analyzed. The ratio of mutations in GC compared to AT is similar in both tandem and single nucleotide mutations in nurse shark Ig light chain gene. However, in sandbar shark TCRγ, single nucleotide mutations show a strong GC
preference (64.2%, p=0.011), whereas there is no GC bias in tandem mutations (45.9%) (Table 1). When both types of mutations are combined, the difference between GC and TA mutations is not significant (56.0%, p=0.052).

There is a universal transition preference in SHM of Igs in all species studied to date. This transition preference also applies to the Ig genes, NAR, and TCR single nucleotide mutations of sharks. The tandem mutation pattern of TCR in sharks is an exception. There are fewer transitions in tandem mutations than in single nucleotide mutations. Although it is lower than what is found in single nucleotide mutation, the transition frequency in tandem mutations of Ig light chain is still significantly higher than the random expectation of 33%. However, the transition frequency (34%) in tandem mutations of TCR is close to random (33%), suggesting that there might be a different mutator responsible for tandem mutations. The coexistence of different mutation patterns in a single system supports the hypothesis that there may be more than one mutator at the primitive stage of the adaptive immune system (125, 199).

3.5.3 What is responsible for tandem hypermutation?

Tandem mutation is a unique feature of SHM in sharks. So far, this pattern of mutation has been observed in Ig heavy chain of horn shark, Ig light chain and Ig heavy chain of nurse shark, NAR of nurse shark, and TCRγ of sandbar shark. A close look at the published Ig gene sequences of skate (131) reveals that tandem mutations also occur in skate Igs. Previous investigations to explore the mechanism of tandem mutations show that it is highly unlikely to be the results of sequential, independent point mutations or
classic gene conversion (125). In the tandem mutations of TCR γ V genes, eighteen tandem mutations (54.5%) contain di- and tri-nucleotide repeats. A similar di-and tri-nucleotide repeat pattern was also observed in shark Ig light chain and heavy chain genes (125, 126), but the frequencies of the repeats vary. In the shark, it appears that the DNA polymerase employed in repair may be highly inaccurate and can generate tandem mutations by largely or completely disregarding the sequence information in the template strand. Di-tri nucleotide repeats and untemplated addition of mutations are known characteristics of terminal deoxynucleotidyl transferase (TdT). This suggests that an unknown polymerase with TdT-like functions may be responsible for generating the tandem mutations in sharks. Among all the proposed polymerases, only polymerase μ has been shown to be capable for untemplated addition of nucleotide in the presence of a template DNA and Mn^{2+} in an in vitro study (200). Furthermore, this template-independent synthesis depends on a specific loop formation of this enzyme (201, 202). Therefore, the function and structure of shark polymerase μ and its role in SHM are worth further investigation.

3.5.4 Polymerase μ based model for generation of tandem mutations.

Here I propose a pol μ based model for the generation of tandem mutations during DNA repair syntheses in the V region, in an attempt to explain the underlining mechanism for generating tandem mutations. As illustrated in Figure 16A, this model is built on the current model of SHM and only involves the mismatch pathway. First, Us induced by AID are recognized by mismatch repair complex MSH2/MSH6. Then, Exonuclease 1
starts to excise a stretch of nucleotides, which in turn results in a single strand gap. Next, two different polymerases, Pol μ and another error prone polymerase are recruited to the DNA lesion area and compete with each other for adding nucleotides at the gap area. The open 3’ end of the DNA strand will provide a platform for Pol μ to add randomly several nucleotides, which in turn form tandem mutations found in immunoglobulin and TCR of shark. In this model, error-prone polymerase (very likely pol η as indicated in human and mice) is still responsible for the point mutations at A:T base pair as indicated in current model. As a result of competition between Pol μ and error prone polymerase, the size of tandems varies.

The untemplated addition of nucleotides by Pol μ will generate “babbles” in double strand DNA, which if too big, in a normal situation, are targets for the DNA repair machinery. However, the predominant feature of tandem mutations suggests that the DNA repair machinery in shark T/B cells may not be as efficient. Although the size of tandems varied (2-7 nucleotides), most of tandem mutations involve only two nucleotides. What count for the dominance of small tandems and scarcity of large tandems? In the first model (as indicated in Figure 16A), this is due to competition between Pol μ and error prone polymerase. Error prone polymerase tend to have higher efficiency in adding nucleotides to the DNA lesion sites than Pol μ, which in turn results in most of the gap being repaired by error prone polymerase at most times. Only the nucleotides of tandem mutations are added by Pol μ. However there are other possibilities. Here I also propose an alternative model for the generation of tandem mutations (as shown in Figure16B).
Figure 16: pol μ based tandem mutation model
During the DNA repair process, Pol μ has similar, if not higher, efficiency as that of error prone polymerase and can generate larger tandem than are actually observed. However, since the mismatch repair complex MSH2/MSH6 exon1 remain active, the newly synthesized tandems (or “babbles”) will be constantly trimmed off and the remaining single strand gap will be filled by the error prone polymerase. The third possibility is that large tandems tend to be recognized more that small tandems by the normal DNA repair machinery and are repaired more often. On the other hand, small tandem mutations tend to escape the surveillance of DNA repair system.

3.5.5 Somatic diversification (SHM and allelic difference) in TCR γ of higher vertebrates.

Antigen processing and presentation by MHC is not required for recognition by γ/δ T cells (32, 203). I wondered whether γ/δ T-cells in higher vertebrates really did lose the capacity for SHM. Therefore, I did a database search (IMGT and Genbank) in an attempt to answer this question. Several criteria were used for collecting genomic and cDNA sequence data: 1) The genomic sequence of the TCR γ locus was known. 2) Greater than 15 cDNA clones were available for a particular V gene. 3) The cDNA sequences were greater than 100 bp. 4) The cDNA samples were from an single animal. Data from only 5 species (sandbar shark, Atlantic salmon, cattle, mice, and human) satisfied these conditions. The clones used in the analysis are listed in Table 18, and sequence alignments showing possible mutations are shown in Figure17. A summary of the results is shown in Table 19. Besides sandbar shark, a quite high frequency of
Table 18: Clones used in analysis of SHM in TCRγ of different species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Clones (GenBank accession number)</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sandbar shark</td>
<td>FJ854417–38 (TCRGV1), FJ854439–57 (TCRGV2), FJ854458–70 (TCRGV3), FJ854471–87 (TCRGV4), FJ854492 (TCR γ genomic locus)</td>
<td>Including clones of 4 out 5 functional V genes. all clones are from one animal</td>
</tr>
<tr>
<td>Atlantic salmon</td>
<td>EU221103-12, EU221139-40, EU221172(TCRGV1-1); EU221113-32, EU221173-74, EU221142-44(TCRGV1-3); EU221155-71(TCRGV1-5); EU221197-102, EU221135-39, EU221141(TCRGV2-1); EU221145-54, EU221175(TCRGV2-3)</td>
<td>Including clones of 5 out of 7 functional V genes.</td>
</tr>
<tr>
<td>Zebra fish</td>
<td>AY973880,AY973895-6,AY973904-12(TCRGV1),AY973882-3,AY973897-8,AY973913-18(TCRGV2),AY973884,AY973899,AY973919-22(TCRGV3),AY973885-87,AY973900,AY973923(TCRGV4),AY973888-9,AY973901-2,AY973928-33(TCRGV5),AY973890-1,AY973934-38(TCRGV6),AY973892-4,AY973903,AY973939-43(TCRGV7)</td>
<td>Including clones from all 7 functional V genes. direct submission to genbank.</td>
</tr>
<tr>
<td>Cow</td>
<td>D16117-18, D16120-25, D16127, D16129, D16132-33, D16149-54</td>
<td>Cattle TCRGV5-2 clones only. TCRGV5 appears to be the predominant population of periphery. All clones are from an 2 year old health helstein cow.</td>
</tr>
<tr>
<td>human</td>
<td>Z11169– Z11410(TCRGV9)</td>
<td>Vg9/Vd2 T cells are dominant g/d T cells in human periphery. The cDNA sequences were taken from more than 20 individuals with or without sarcodiosis.</td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
<td>Description</td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td>-------------</td>
</tr>
<tr>
<td>00:00</td>
<td>Start</td>
<td>Session begins</td>
</tr>
<tr>
<td>01:00</td>
<td>Activity 1</td>
<td>Group discussion</td>
</tr>
<tr>
<td>02:00</td>
<td>Activity 2</td>
<td>Hands-on exercise</td>
</tr>
<tr>
<td>03:00</td>
<td>Activity 3</td>
<td>Case study analysis</td>
</tr>
<tr>
<td>04:00</td>
<td>Activity 4</td>
<td>Group presentation</td>
</tr>
<tr>
<td>05:00</td>
<td>Summary</td>
<td>Recap of key points</td>
</tr>
<tr>
<td>06:00</td>
<td>End</td>
<td>Session concludes</td>
</tr>
</tbody>
</table>

**Agenda Session**

**Date:** 2023-04-01

**Location:** Conference Room A

**Participants:**
- [Names and roles]
<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Allelic Difference</th>
<th>Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCRA</td>
<td>TCRB</td>
<td>TCRG</td>
</tr>
<tr>
<td>TCRD</td>
<td>TCRE</td>
<td>TCRF</td>
</tr>
<tr>
<td>TCRG</td>
<td>TCRH</td>
<td>TCRJ</td>
</tr>
<tr>
<td>TCRK</td>
<td>TCRL</td>
<td>TCRM</td>
</tr>
<tr>
<td>TCRN</td>
<td>TCRP</td>
<td>TCRQ</td>
</tr>
<tr>
<td>TCRR</td>
<td>TCRS</td>
<td>TCRT</td>
</tr>
<tr>
<td>TCRU</td>
<td>TCRV</td>
<td>TCRW</td>
</tr>
<tr>
<td>TCRX</td>
<td>TCRY</td>
<td>TCRZ</td>
</tr>
</tbody>
</table>

Figure 17: Comparison of the TCR V gene cDNAs of different species with corresponding parental genomic sequence. The nucleotides that are identical to the parental genomic sequence are shown as dots. Nucleotide changes caused by allelic difference are indicated in lower case. Nucleotide changes caused by mutation are underlined.
Table 19. Database search for somatic diversification (SHM and allelic difference) of TCR γ gene in other species.

<table>
<thead>
<tr>
<th></th>
<th>γ/δ T cells high species</th>
<th>γ/δ T cells low</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sandbar shark</td>
<td>Atlantic salmon</td>
</tr>
<tr>
<td>Clones included</td>
<td>73</td>
<td>63</td>
</tr>
<tr>
<td>Possible Mutation</td>
<td>yes</td>
<td>likely</td>
</tr>
<tr>
<td>Clones containing mutations</td>
<td>33</td>
<td>14</td>
</tr>
<tr>
<td>Number of mutations</td>
<td>191</td>
<td>19</td>
</tr>
<tr>
<td>Tandem mutation</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Number of bases included</td>
<td>10305</td>
<td>9552</td>
</tr>
<tr>
<td>Mutation frequency</td>
<td>0.018/bp</td>
<td>0.002/bp</td>
</tr>
<tr>
<td>Allelic difference</td>
<td>yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
mutation in TCR γ V genes was found in both atlantic salmon (173) and cattle (204). The mutation frequencies are about 0.002, which appear too high for PCR error, but are about 10 fold lower than in normal SHM. Unfortunately, I cannot make a firm conclusion because of limitations in the data set. However, our analysis suggests that high mutation rates in TCR γ V genes may be more widespread than currently thought. In contrast, not only are there no examples of SHM in TCR γ of humans and zebra fish, there is no diversity generated in the population by allelic differences. I analyzed 214 human Vγ9 sequences, which were collected from about 20 different individuals with or without sarcoidosis (205). The V sequences are completely identical. It appears that there is strong selection pressure is on this gene to keep it from mutations and polymorphism in population. However, this must not be the case in γ/δ T cell high species, such as cattle and fish, in which γ/δ T cells probably play a much larger role in their immune system. Zebra fish are highly inbred and this is probably the reason for the homogeneity I observed in its TCR γ genes.

3.5.6 Selection pressure on the immune system and evolution

Genetic analysis of the human Ig heavy chain V region sequences showed that CDR and FR sequences could differ significantly in their inherent susceptibility to amino acid replacement given any single nucleotide change (115). Selection pressure is inferred from the highamino acid replacement mutation rate in CDRs compared to the scarcity of replacement in FRs. Furthermore, genetic analysis shows that the R/S ratio bias may be a consequence of the intrinsic mutability of CDR regions even in the absence of antigen
selection (93, 121). Differences in the inherent susceptibility to replacement mutations between CDR and FR regions was proposed to be the result of selection pressure imposed on these regions during evolution. However, research on the nurse shark shows that there is no significant difference in R/S ratio between CDR and FR in both Ig light chain and Ig heavy chain (125)(126). Similar results can also be found in the Ig heavy chain of teleost fish (127) and the TCRγ of sandbar shark. All of the above findings suggest that there has been little selection pressure towards affinity maturation during the evolution of Ig genes in teleosts and sharks. Thus, SHM in Ig genes of teleosts and shark appears to be mainly for increasing repertoire diversity rather than for increasing affinity. These findings support the concept that SHM may have initially evolved in the ancestral immune system as a mechanism to principally increase repertoire diversity (128, 129)(125, 126, 130). Interestingly, although evidence of selection cannot be found in shark Igs and TCRγ, it can be found in the NAR system in which the R/S ratio of CDR is significantly higher than that of the FR (130). The authors speculate that NAR might serve as the secondary immune defense system in which higher affinity to antigens is generated, while Ig serve only as the primary immune system, similar to the role of IgM in mammals. The R/S ratio of Immunoglobulin is significantly higher than that of TCRγ, indicating that Ig might be under some selection pressure. If this is the case, this finding may provide insight regarding the development of affinity maturation in Ig. It is likely that SHM was used for generating receptor diversity in both Igs and TCRs in the ancestral adaptive immune system. However, due to selection pressure, the function of SHM in Igs
shifted from generating repertoire to generating higher affinity antibodies and the critical event was to make cell undergoing SHM dependent on Ag-mediated survival signal.

It appears that during evolution, the descendants of sharks lost the capacity for SHM in TCRγ. Lack of selection pressure may have been one of the reasons. However, there are other possibilities. 1) Decreased demand: During evolution, the size and complexity of TCR γδ genes increased, which may have diminished the need for SHM to increase the receptor repertoire. 2) Development of an antigen presenting system: TCR γδ has been speculated to be the common ancestor of both TCR αβ and Igs (55-57). Structurally, it is more like an Ig molecule than TCR αβ, and recognizes antigens in a Ig-like manner (30). Although antigen presentation by the MHC molecule is not required (32, 191), cell-cell contact may be necessary for the activation of human γδ T cells (34), indicating that other non-MHC antigen presenting molecules might be involved. The nature of antigen recognition by TCR γδ in shark is still unknown. However, it is likely that an antigen presenting system was developed during evolution. As a consequence, this non-MHC molecule/TCR interaction may have led to the restriction of SHM in TCR γδ genes. 3) Reduced antigen repertoire: the actual antigen spectrum of TCR γδ isn’t yet fully known. So far, only a few antigens have been identified as potential candidates recognized by TCR γδ in humans and mice. Most of these antigens are phosphoantigens (33, 35, 37, 206) released during microbial infection, and stress proteins (39-41, 43, 44). Furthermore, unlike αβ T cells, γδ T cells in humans and mice are restricted in location and TCR usage. Overall, the antigen spectrum of TCR γδ is much smaller than that of TCR αβ in humans.
and mice. This may not be the case in sharks, in which the antigen spectrum of TCR $\gamma\delta$ may be much larger. It is reasonable to speculate that the shrinkage of the antigen repertoire might have led to the degeneration of SHM in TCR $\gamma\delta$ during evolution. However, this cannot explain the apparent absence of SHM in $\gamma\delta$ T cell dominant species, such as chicken and fish. 4) The development of the TCR $\alpha\beta$ system during evolution may have reduced the burden of the TCR $\gamma\delta$ system for antigen recognition.

**Future directions**

Finally, I have to point out the fact that all the DNA sequences involved in this research came from one single animal. The conclusions draw from this research may not be sufficient enough to apply to all Elasmobranchii species. In future, it is critical for us that speciement from other shark or skate species should be tested to further confirm our observations.

As for the pol mu base models proposed here, I believe that overexpression of shark AID and pol mu genes in mammalian or fish cell lines may provide definitive answer to the validity of these models.
CHAPTER IV CHARACTERIZATION OF THE ARRANGEMENT AND EXPRESSION OF THE BETA-2 MICROGLOBULIN LOCUS IN THE SANDBAR AND NURSE SHARK

4.1 Summary

Beta 2 microglobulin (β2m) is an essential subunit of major histocompatibility complex (MHC) type I molecules. In this chapter, β2m cDNAs were identified and sequenced from sandbar shark spleen cDNA library. Sandbar shark β2m gene encodes one amino acid less than most teleost β2m genes, and 3 amino acids fewer than mammal β2m genes. Although sandbar shark β2m protein contains one β sheet less than that of human in the predicted protein structure, the overall structure of β2m proteins is conserved during evolution. The germline gene for the β2m in sandbar and nurse shark is present as a single locus. It contains 3 three exons and two introns. CpG sites are evenly distributed in the shark β2m loci. Several DNA repeat elements were also identified in the shark β2m loci. Sequence analysis suggests that the β2m locus is not linked to the MHC I loci in the shark genome.
4.2 Introduction:

Beta2 microglobulin(β2m) is a key component of the adaptive immune system. It is a subunit of the major histocompatibility complex (MHC) type Ia molecules, which are essential for peptide presentation to CD8+ T cells to initiate cellular mediated cytotoxicity. It is non-covalently associated with the heavy chain (composed of α1,α2,and α3 domains ) of MHC Ia molecules to stabilize the heavy chain complex(207). β2m is essential for peptide-MHC complex assembly(208), and can influence the ability of TCRs to engage the MHC-peptide complex(209). Unlike MHC molecules, β2m doesn’t contain a trans-membrane domain. Therefore, free β2m can be detected in serum. Although not a part of MHC-peptide-TCR interaction complex, β2m is involved in the activation of T cells by antigen presenting cells (APCs) via direct contact with the CD8 molecule (210, 211). β2m is also associated with non-classical MHC type Ib molecules, including HLA-E, -F, -G, and -H, CD1, and neonatal Fc receptor (FcRn). Compared with MHC Ia molecules, class Ib molecules display limited polymorphism and restricted tissue distribution, and are involved in a wide range of activities from the activation of NK cells and γ/δ T cells to placental IgG transport(212-214). Besides its role in the structure of MHC class Ia and Ib, β2m has been identified as the thymotaxin(215), which is responsible for the thymus colonization of T cell precursors(216). Furthermore, β2m regulates NK cell function by direct contact with NK cell inhibitory receptors(217-221). A role in the anti-tumor immune response (222, 223) and viral resistance (224) has also been suggested.
β2m is closely related to the membrane proximal domains of MHC molecules. It has been speculated that β2m and MHC genes share a common ancestor (225, 226). Sharks have a unique position in research on the molecular evolution of the adaptive immune system, since they are the most distantly related gnathostome relatives to mammals. All defining elements of the adaptive immune system, such as antibodies, T cell receptors (TCRs), MHCs and recombination activator genes (RAG), are present in elasmobranches (4, 5). Although β2m gene expression has been reported in the skate (227), the β2m locus hasn’t been characterized in any elasmobranches. Hence, research on the genomic arrangement of the β2m gene in sharks may provide insight into the evolution of β2m and MHC genes.
4.3 Material and method

4.3.1 Sequence

The sequences reported in this paper have been deposited in the GenBank database.

Accession numbers: GQ865620 (sandbar shark \( \beta 2 \)m allele 1 cDNA sequence), GQ865621 (sandbar shark \( \beta 2 \)m allele 2 cDNA sequence), GQ865622 (sandbar shark \( \beta 2 \)m genomic locus), GQ865623 (nurse shark \( \beta 2 \)m genomic locus).

4.3.2 Preparation of DNA and mRNA

Sandbar shark spleen was provided by Dr. Carl Luer (Mote Marine Laboratories, Sarasota, FL). Sandbar shark spleen genomic DNA was prepared using the QIAGEN Blood and Cell Culture DNA Kit. Sandbar shark spleen mRNA was prepared using the Invitrogen Micro-FastTrack 2.0 mRNA Isolation Kit.

4.3.3 Cloning of sandbar shark \( \beta 2 \)m cDNA

The finding of \( \beta 2 \)m gene was unanticipated. A 3’ RACE PCR was performed using a degenerate primer based on the skate and horned shark TCR beta constant region, in an attempt to clone sandbar shark TCR beta chain cDNA. Among all the cDNA clones obtained from this PCR reaction, one clone showed high identity to \( \beta 2 \)m of teleosts and mammals in Genbank BLAST search. Next, using this sequence information, a new primer was designed (5’CCAAGTTTGTGAAAGTCACCCCTGAG3’) for 5’ Rapid Amplification of cDNA Ends (RACE) to amplify the 5’ part of the gene. 5’ RACE was performed using the SMARTTM RACE cDNA Amplification Kit (Clontech) according to the manufacturer’s instructions. Gene specific primers as well as adaptor specific primers
were used for PCR reactions employing the following conditions: 50 μl reaction total, 34.5 μl sterile H₂O, 5 μl 10X Advantage 2 PCR buffer, 1 μl dNTP mix, 1 μl 50X Advantage 2 polymerase mix, 2.5 μl 5’RACE ready cDNA (1), 5 μl UPM (universal primer), 1 μl GSP. PCR settings: 5 cycles at 94°C 30 sec, 72°C 3 min, 5 cycles: 94°C 30 sec, 70°C 30 sec, 72°C 3 min. Then, forward primer: 5’TGCAAGAGGAGAGTCTGTGAGTG3’; reverse primer: 5’CCATTGGATGATGCTA3’ were used to amplify full length cDNA. PCR was performed under the same conditions as described above.

4.3.4 Cloning of sandbar shark beta-2 microglobulin genomic DNA

A genomic PCR was performed, using sandbar shark spleen genomic DNA and the same primers used for amplifying full length cDNA. PCR reaction was performed using the Expand 20 kb Plus PCR system (Roche) under the following conditions: 50 μl reaction total, 31.5 μl sterile H₂O, 5 μl 10X Advantage 2 PCR buffer, 1 μl dNTP mix, 1 μl Expand 20 kb Plus enzyme mix. 1 μl sandbar shark spleen genomic DNA, 5 μl forward primer, 5 μl reverse primer. PCR settings: 10 cycles at 92°C 10 sec, 62°C 10 sec, 20 cycles at 68°C 18 min, 10 cycles at 92°C 10 sec, 62°C 10 sec, 68°C 18 min+ 10s cycle elongation for each successive cycle. Elongation cycle: 68°C 10 min. A PCR product of ~3500 bp was purified and cloned for sequencing.

4.3.5 Shotgun sequencing of large DNA fragment
For genome sequencing, a shotgun sequencing strategy was applied. The GPS-1 genome primer system (New England Biolab, Inc.) was used for this purpose. Individual sequences were assembled using the MacVector 9.0 (MacVector, Inc.) program.

4.3.6 Screening the nurse shark BAC library

The nurse shark library was purchased from Arizona Genomic Institute (AGI). The PCR product of the full length β2m cDNA was used as a probe. Standard hybridization conditions were used as described by AGI(www.genome.arizona.edu/information/protocols).

4.3.7 Computer Analysis

The following software packages were used: MAFFT (148, 149) for sequence alignment, MacVector 10.5 (MacVector) for assembly of sequencing projects, sequence alignment and phylogenetic analyses, and ClustalW2 (150, 151) for sequence and phylogenetic analyses. DNA repeats are found by on line program tandem repeat finder (228) and repeatmasker (229). Putative sandbar shark leader segment was identified by using the signalP3.0 program (3).

4.3.8 DNA sequencing

Sandbar shark DNA sequencing was done by University of Arizona Genetics Core service. Nurse shark genomic BAC clone was sequenced by AGI.

4.3.9 Southern blots

Sandbar shark spleen DNA was digested by Bam HI, Eco RI, and Hind III independently. Digested DNA was transferred to Immobilon-Ny+ transfer membrane (Millipore) by
capillary flow under alkaline conditions (1.5 M NaCl/0.5 M NaOH) and fixed to the membrane with UV cross-linking (5000 micro joules). A $^{32}$P labeled PCR product of full length $\beta$2m cDNA was used as the probe. After an overnight hybridization, high stringency washes were performed in 0.1× standard saline citrate (SSC)/0.1% SDS solution at 65°C.
4.4 Results

4.4.1 Cloning and sequence analysis of sandbar shark β2m cDNA

The finding of the sandbar shark β2m gene was unanticipated. A 3’ RACE PCR was performed using degenerate primers based on the skate and the horned shark TCR β gene constant region. Among all the cDNA clones from this PCR reaction, one clone showed high identity to β2m genes of teleost and mammals in a Genbank BLAST search. Next, using this sequence information, a new primer was designed and used in 5’ RACE to amplify the 5’ end of the gene. Then, based on the sequence information of both ends, a new set of primers was designed to amplify full length cDNA clones. The full length cDNA contains 691 base pairs (bp), including the coding region (from 103 to 441 bp) and untranslated regions (UTRs) at both ends(102 bp at the 5’ end and 250 bp at the 3’ end).

The organization of the cDNA sequence is shown in Figure 18A. The mature β2m protein is predicted to be 96 amino acids (aa) in length, and the leader sequence contains 15 aa. The alignment of inferred β2maa sequences of sandbar shark and other species are illustrated in Figure 18B and 18C. As shown in Figure 18, sandbar shark β2m shows a high level of similarity to β2m of other species (from 66% overall similarity to skate, to 52% overall similarity to human). The cysteines which form the intra-domain disulfide bond are evolutionarily conserved at residue 25 and 80(230). Compared with human β2m, there are two amino acid deletions at residue 83 and 84, which are also missing in
Figure 18. Amino acid sequence alignment of sandbar shark β2m with those of other vertebrate species. (A) Scheme of sandbar shark β2m cDNA sequence. (B) Amino acid sequence alignment of sandbar shark β2m leader sequence with those of other species. Putative sandbar shark β2m leader segment was identified by using the signalP3.0 program (3). (C) Amino acid sequence alignment of sandbar shark β2m mature protein with those of other species. Identical residues are indicated by using a bold capital case and dark shaded box; similar residues are indicated by using capital case and light shaded box; mismatched residues are indicated by using lower case. Spaces for amino acids absent in shark and other species, but present in human, are indicated by dots. Percentage identities and similarities are shown on the right.
most teleost (231-236) except sturgeon (237), and one more deletion at residue 99, which is unique to elasmobranches. Furthermore, the leader sequences of both sandbar shark and skate contain only 15 aa, which is much shorter than that of most teleost (19 aa, except sturgeon which contains 25 aa) and human (20 aa). Alignment of cDNA clones show the presence of two alleles, which differ from each other at nucleotide 201 (A in allele 1, T in allele 2) and at nucleotide 207 (A in allele 1, G in allele 2), resulting in two aa difference between the two alleles. Roughly equal frequencies of the four nucleotides were found in the sandbar shark β2m cDNA sequence.

4.4.2 Phylogenetic analysis

A phylogenetic tree was generated using the UPGMA method based on the full length β2m protein sequences (Figure 19). β2m of shark and skate are grouped together, while all other teleost are located in one group and mammals are located into one another group. Chicken β2m falls between teleost and mammals. The taxonomical status of teleost is consistent with the accepted evolution pathway, in which sturgeon is believed to diverge before other teleost. A similar tree can also be generated using the neighbor joining method (data not shown).

4.4.3 Protein structure prediction

The protein structures of the sequences were predicted using the Phyre (Protein Homology Fold Recognition Engine) server created by the Structural Bioinformatics Group, Imperial College, London (1, 2). This server uses algorithms to generate models based on known sequences. Known coordinates of sequences alignments from similar
Figure19 Phylogenetic tree, derived from full length β2m protein comparisons, showing the relation between sandbar shark β2m and that of other species. Sequences were aligned by using MAFFT, and the tree was constructed with MacVector using the UPGMA method (238). Percentage bootstrap values (1000 replicates) are shown at the interior branch point. The horizontal length is proportional to the distance score generated by the computer program. The shark β2m sequence is from this work. All other sequences were obtained from GenBank: skate (AAN62852), sturgeon (CAB61322), salmon (NP_001117171), rainbow trout (AAB04663), common carp (AAA49203), zebra fish (NP_001153240), chicken (AAA64915), platypus (NP_001121090), mouse (NP_033865), domestic cat (NP001009876), human (AAA51811), horse (NP_001075971), pig (NP_999143), sheep (NP_001009284), cattle (NP_776318).
structures are changed to the residues of the sequence. Using Swiss-Pdb Viewer, an application from the Swiss Institute of Bioinformatics, the model generated from Phyre is loaded and the 3D structure is visualized. Swiss-Pdb Viewer provides an interface for analysis of proteins.

As shown in Figure20A, the predicted human β2m protein contains 9 β sheets and 8 loops. the predicted structure is well correlated with published human β2m structure based on X-ray crystallography(239), with a slight difference in the size of the first β sheet. As shown in Figure 3B, the predicted sandbar shark β2m protein contains 8 β sheets, with the G strand missing. Since the location of the G strand is outside of the Ig C domain (including from strand G to strand F), the overall structure of the β2m proteins was conserved during evolution, including the location of cysteines which form the disulfide bond and the Asp 53 β bulge in the D β sheet. In Figure20B, the eleven residues that are predicted to be involved in hydrogen binding of MHC I heavy chain in human β2m are highlighted(239). Among these hydrogen binding residues, Tyr11, Asp53, Leu54, and Trp61 are conserved between shark and human, while Ser11, Arg12, Asn24, Ser 28, His 31, Asp98 and Met99 are different.

4.4.4 Southern blot analysis

The finding that the cDNA clones isolated above shared more than 99% identity indicated the presence of only a single β2m locus. To confirm this conclusion, southern blot analyses were performed with sandbar shark spleen DNA by using the PCR product of the full length β2m cDNA as the probe. As shown in Figure21, the BamHI DNA
Figure 20. Predicted three dimensional structure of β2m protein of sandbar shark and human. (A) Predicted three dimensional structure of β2m protein of sandbar shark and human. (B) Predicted three dimensional structure of β2m protein of sandbar shark. The protein structures were predicted using the Phyre (Protein Homology Fold Recognition) (1, 2). The location of the Asp53 β-bulge in the D strand is highlighted in red (C). Using Swiss-Pdb Viewer, the model generated from Phyre is loaded and the 3D structure is visualized. The overall chain trace of β2m proteins are displayed to highlight the overall β- sandwich fold, the secondary structure elements (β sheets labelled from A to G), the loop regions (dark line), and the intra-chain disulfide bond (yellow).
Figure 21. Southern blotting of the sandbar shark β2m locus. Sandbar shark spleen DNA was digested by BamHI (lane 1), EcoRI (lane 2), and HindIII (lane 3) independently. The probe was generated by PCR of full-length β2m cDNA.
preparations gave rise to one strong hybridizing band, indicating the presence of only 1 β2m locus in the sandbar shark genome. This conclusion is buttressed by the finding of 2 hybridizing bands in the EcoRI DNA preparation line and 3 hybridizing bands in HindIII DNA preparation line, which can be explained by the fact that there are EcoRI and HindIII restriction sites inside the β2m locus. The locations of the restriction sites of individual restriction enzymes are indicated in Figure22. The distance between the two Hind III restriction sites is about 1800 bp, which correlates well with the size of the second band of the Hind III DNA preparation lane.

4.4.5 Sequencing of the β2m locus of sandbar and nurse shark

A genomic PCR was performed, using sandbar shark spleen genomic DNA and the same primers used for amplifying full length cDNA. A PCR product of about 3600 bp was cloned and sequenced. As shown in Figure22A, the sandbar shark β2m locus contains 3 exons and 2 introns. Exon1(from the 4th to the 158th nucleotide) encodes the leader sequence, the first two aa of β2m protein and the first base of the third codon; exon2(from the 1870th to the 2142th nucleotide) covers the remaining two bases of the third codon through the first base of codon 94 (aspartic acid). Exon3 (from the 3468th to the 3639th nucleotide) encodes the remaining two bases of codon94 and the last two amino acids.

Sandbar shark β2m introns are phase-one introns, with the intron interrupting a codon between the first and second bases. Phase one introns are common in teleost MHC and β2m genes(231).

Full length sandbar shark β2m cDNA sequence was used as a probe to screen a nurse shark BAC genomic DNA library. Two positive clones were identified as β2m gene
containing clones. One clone was fully sequenced. 2926 individual sequences were received. Using MacVector, 23 contigs were assembled, with the sizes varying from 2 kb to 60 kb. One 23 kb contig contained the β2m gene, based on a Genbank BLAST search and alignment with sandbar shark β2m gene. The locations of exons and introns are predicted by alignment with sandbar shark β2m cDNA and the genomic DNA sequence.

As shown in Figure 2, the nurse shark β2m locus contains 3 exons and 2 introns as well. Exon 1 covers from the first to the 90th nucleotides; exon 2 spans from the 1501th to the 1776th nucleotide; exon 3 covers from the 3600th to the 3833th nucleotide. As shown in Figure 5, the major differences between nurse and sandbar shark are the sizes of the introns and the locations of the repeats.

DNA repeats are found by online program Tandem Repeat Finder (228) and Repeatmasker (229). Three repeats (AT rich region, HE1_SINE, and HAT-N68) were identified inside the sandbar shark β2m locus. Two HER_LINE elements and one variable number tandem repeat (VNTR) were identified inside the nurse shark β2m locus. VNTR identified here is a 30 bp segment (5'AGAGATAGTAGGAACTGCAGATGCTGCGAGA3') which is repeated twice. One HE1_MM element was identified downstream of nurse β2m 3'UTR.

Roughly equal frequencies of the four nucleotides were found in β2m genomic sequence of both sandbar and nurse shark. CpG sites are evenly distributed in the β2m locus. In the sandbar shark β2m locus, a total of 175 CpG sites were identified, including 14 in exon 1, 73 in intron 1, 13 in exon 2, 60 in intron 2, and 15 in exon 3. In the nurse shark β2m
locus, a total of 187 CpG sites were identified, including 10 in exon 1, 59 in intron 1, 24 in exon 2, 81 in intron 2, and 13 in exon 3.
Figure 22 Schematics of shark β2m loci. (A) Schematic of the sandbar shark β2m locus. (B) Schematic of nurse shark β2m locus. β2m loci of both sandbar and nurse shark contain 3 exons and 2 introns. Exons are indicated in black boxes; UTRs are indicated in white boxes; Repeats are indicated in gray boxes; The locations of restriction sites of different restriction enzyme are shown as well. DNA repeats are found by on line program tandem repeat finder (228) and repeatmasker (229). The detailed information concerning the repeat sequences are listed in the table below.
4.5 Discussion

The appearance of MHC genes can’t be explained by gradual evolution model since no MHC precursor has been identified in jawless vertebrates. Gene duplication is believed to be an essential event for the emergence of the AIS. It is believed that large-scale gene duplication and subsequent reshuffling of exons is essential for the emergence of MHC genes. This gene duplication process may be continually occurring during evolution. Besides the origin of MHC I, it might also explain the cluster arrangement of shark immunoglobulin genes(135-138, 240, 241). The appearance of RAG and gene duplication is believed to have occurred after the divergence of Agnathans and before the divergence of Gnathostomata(8, 19).

4.5.1 The evolution of MHC and β2m

As a member of the C1 Immunoglobulin super-family (IgSF), β2m is related closely to the membrane proximal domains of MHC class I and class II molecules. It has been speculated that β2m and MHC genes share a common ancestor (225, 226). Therefore, the ancient β2m and MHC genes might be expected to be located nearby in the genome. However, the β2m gene is not linked to MHC genes in any species so-far examined. The separation of the MHCI genes and the β2m gene is believed to be the result of a gene shuffling event following the duplication of the MHC genes. In our sequence analysis, no MHC genes were identified in the neighborhood (about 200 kb) of the nurse shark β2m gene. Although not conclusive, it is very likely that the shark β2m gene is not linked to
MHC genes. Therefore, the shuffling of MHCI and β2m genes must have occurred before the divergence of Gnathostomata.

There is an increase in the length of β2m protein during evolution. Shark and skate β2m genes (96 aa) encode one amino acid less than those of teleosts (97 aa). The shark and skate leader sequences (15 aa) encoded solely by exon1 are also shorter than those of higher species (usually about 20 aa). As shown in Figure 3, the β strandG is missing in the predicted shark β2m protein structure and seven of eleven hydrogen binding residues are different between human and shark, suggesting there may be some difference in the β2m-MHCI molecule interface between shark and human. Whether such a change in interface would lead to a functional change is unclear.

4.5.2 Genomic organization

One of the major differences between the shark and higher vertebrate β2m genes is the number of exons and introns. Mammal and bird β2m loci contain four exons and three introns, while teleost and shark loci contain three exons and two introns. It is worth noting that exon 2 encodes most of β2m protein in all species, exon 1 and 3 encode only a few aa, and exon 4, is found only in higher vertebrate encodes no aa. Therefore, it appears that one intron was inserted into the 3’UTR to generate one more exon (exon 4). The insertion occurred after the divergence of teleosts and before the divergence of the birds during evolution. The length of the shark β2m loci (above 3.6 kb) is much larger than that of teleosts which are usually below 2 kb (231-235). The length difference of β2m genes results from the length difference of introns (more than 1kb in sharks, and
several hundred bp in teleosts). The lengths of the shark $\beta_2m$ loci are comparable to that of mammals. Another feature of $\beta_2m$ loci is that they contain rather long 3’UTR (548 bp in human, 674 bp in zebra fish, 566 bp in common carp and 673 bp in skate). However, sandbar shark contain a shorter 3’ UTR (about 250 bp), with a canonical polyadenylation signal (AATAAA box) identified 200 bp after the stop codon and a 30 bp poly (A) tail (13 bp after the AATAAA box). The 102 bp of sequence upstream from the start codon represent the 5’UTR, which is longer than that of zebra fish (44 bp), grass carp (60 bp), sturgeon (23 bp), and human (37 bp).

4.5.3 CpG sites

CpG dinucleotides occur at a much lower frequency in the vertebrate genomes than the expected frequency on the basis of GC in the genomes, which has been described as “CG suppression” (242). It is also found that most cytosines in the CpG sites are methlyated in mammal genomes (243). However, some regions of mammal genomes contain a higher concentration of CpG sites (CpG islands), which are often associated with the start of the genes. It was reported that CpG sites cluster at the 5’end of MHC genes of several species(244). The biological significance of CpG clustering at the 5’ end of MHC genes is not clear. It was suggested that the CpG clustering might make the MHC more prone to gene conversion-like events (245). Similarly, CpG sites also cluster at the 5’ end of the human$\beta_2m$ gene(246). However, in our analysis of shark $\beta_2m$ genes, CpG sites are distributed among exons and introns in a random manner, suggesting that during the divergent evolution of the mammal $\beta_2m$ gene there may be an accumulation of CpG sites
at the 5’ end, or CpG sites may have tended to be conserved at the 5’ end and be lost at the 3’ end.

4.5.4 SINE and LINES

Several short interspersed repetitive elements (SINE) and long interspersed repetitive elements (LINE) were identified in the shark β2m loci (as shown in Figure 5) and many are identified outside of the nurse shark β2m locus (data not shown), including members of HE1 SINE, HER1 LINE, and V SINE family. All mentioned SINE and LINE families have been reported and characterized in various elasmobranches(247, 248). SINEs and LINEs belong to the retroposon family which is characterized as being capable of amplification by reverse transcription of an RNA intermediate(249). All SINEs are derivatives of tRNAs (250), with lengths varying from 100 to 500 bp. SINEs contain an internal promoter for RNA polymerase III and no open reading frame (ORF).

One the other hand, LINEs are much longer, varying from 3 to 7 kb in length. LINEs contain one ORF encoding an RTase and an endonuclease (251). The overall significance of retroposons in evolution is unclear. Based on sequence analysis, it was reported that some poly(A) tails, enhancer segments, inhibitor signals, and even exons might be introduced by retroposon sequences (252, 253). Therefore, retroposons might play an important role in evolution by introducing regulatory elements and even protein coding elements into the genome. Furthermore, transposable elements have been proposed as one of the driving forces for speciation (254). As to the β2m locus, it would be worth
further investigating whether the extra intron3 found in higher vertebrates was introduced by SINEs or LINEs.
APPENDIX A: PUBLICATIONS


REFERENCES


144


162. Daggefeldt, A., E. Bengten, and L. Pilstrom. 1993. A cluster type organization of the loci of the immunoglobulin light chain in Atlantic cod (Gadus morhua L.) and rainbow trout (Oncorhynchus mykiss Walbaum) indicated by nucleotide sequences of cDNAs and hybridization analysis. *Immunogenetics* 38:199-209.


253. Brosius, J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238:115-134.