



# Kent Academic Repository

**Woolven, Ben (2000) *The mechanisms of antibody generation in the llama.* Doctor of Philosophy (PhD) thesis, University of Kent.**

## Downloaded from

<https://kar.kent.ac.uk/86217/> The University of Kent's Academic Repository KAR

## The version of record is available from

<https://doi.org/10.22024/UniKent/01.02.86217>

## This document version

UNSPECIFIED

## DOI for this version

## Licence for this version

CC BY-NC-ND (Attribution-NonCommercial-NoDerivatives)

## Additional information

This thesis has been digitised by EThOS, the British Library digitisation service, for purposes of preservation and dissemination. It was uploaded to KAR on 09 February 2021 in order to hold its content and record within University of Kent systems. It is available Open Access using a Creative Commons Attribution, Non-commercial, No Derivatives (<https://creativecommons.org/licenses/by-nc-nd/4.0/>) licence so that the thesis and its author, can benefit from opportunities for increased readership and citation. This was done in line with University of Kent policies (<https://www.kent.ac.uk/is/strategy/docs/Kent%20Open%20Access%20policy.pdf>). If y...

## Versions of research works

### Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

### Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in *Title of Journal*, Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

## Enquiries

If you have questions about this document contact [ResearchSupport@kent.ac.uk](mailto:ResearchSupport@kent.ac.uk). Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our [Take Down policy](https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies) (available from <https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies>).

**The Mechanisms of Antibody Generation in the  
Llama**

**Ben Woolven**

Submitted in accordance with the requirements for the degree of  
Doctor of Philosophy

The University of Kent at Canterbury  
Research School of Biosciences

2000

**BEST COPY AVAILABLE.**

**VARIABLE PRINT QUALITY**



## **IMAGING SERVICES NORTH**

Boston Spa, Wetherby  
West Yorkshire, LS23 7BQ  
[www.bl.uk](http://www.bl.uk)

**SOME PAGES BOUND  
INTO/CLOSE TO SPINE.**

For Jacqueline

## Declaration

No part of this thesis has been submitted in support of an application for any other degree or qualification of the University of Kent at Canterbury or any other university or institute of learning.

Ben Woolven

29 September 2000

## Abstract

The llama is able to generate a unique class of antibody. The *heavy chain* immunoglobulins consist only of two heavy chain polypeptides and bind antigen specifically through single protein domains. Although the mechanisms by which such an antibody interacts with antigen has been studied at some length the manner in which the heavy chain antibody is generated within the llama is unknown.

In this study a number of components of the llama immune system have been characterised. The isolation of genes encoding the variable domain of the heavy chain antibody indicates that specific genetic elements within the llama genome are responsible for the generation of the heavy chain antibody. The discovery of constant region genes that encode the heavy chain antibody provides an explanation for the absence of a major immunoglobulin domain from the final, secreted gene product. The lack of this domain within the expressed antibody is believed to be the result of a single nucleotide splice site mutation.

In order to investigate the process of llama antibody generation further additional components of the llama immune system, the *recombination activating genes (rag)* were isolated. One such llama *rag* gene (*rag-1*) was cloned, expressed and utilised in an *in vitro* assay system to investigate recombination events taking place during antibody generation. This assay involved the use of specific signal sequences derived from variable domain gene sequence data and represents, to our knowledge, the first examination of non-murine RAG activity. Through the use of this system distinct differences between llama and mouse recombination signal sequences (RSSs) were uncovered. These differences, located within a specific region of the RSS known as the coding flank, may play an important role in llama antibody generation.

These results have led to the proposal of a number of models for the mechanisms involved in llama antibody generation.

## **Acknowledgements**

I would like to thank my two supervisors, Dr Peter Nicholls at the University of Kent and Dr Paul van der Logt at Unilever Research for their support during the last three years.

Thanks also go to Drs Kevin Hiom and Donna Mallory at the MRC in Cambridge for their assistance in setting up protein expressions and recombination assays.

I am indebted to Andrew Starnes and Linda Johnson at the Ashdown Llama Farm for the generous gift of llama testicular material.

Throughout this project numerous people have assisted me through their advice and discussion, thanks must therefore also go to Karen Cromie, Gani, Leon Frenken, Hans de Haard, Martin Gellert, Pat Markham, Hans Peters, Christina Rada, Ian Tomlinson, Clive Wilson, Steve Wilson, John Windust and Roger Windsor.

Finally special thanks to Jacqueline for critical (very critical at times!) review of this thesis and support throughout my time at Canterbury and Colworth.



# Table of Contents

Abstract.....	i
Acknowledgements.....	ii
Table of Contents.....	iii
Table of Figures.....	vii
Table of Tables.....	ix
Abbreviations.....	x
Amino Acid Codes.....	xi

## Chapter 1 Introduction.....1

1.1 Overview.....	1
1.2 The Commercial Potential of Llama Immunoglobulins.....	1
1.3 Possible Roles for the Heavy Chain Antibody in the Llama.....	2
1.4 Potential Mechanisms of Heavy Chain Antibody Generation.....	3
1.5 Aims of Study.....	4
1.6 Research Strategy.....	4
1.7 Classical Immunoglobulin Gamma.....	5
1.7.1 The Structure of the Classical Immunoglobulin Gamma Protein.....	7
1.7.2 The Functional Domains of Classical IgG.....	7
1.7.3 The Structure of the Classical Immunoglobulin Gamma Heavy Chain Gene.....	10
1.7.4 Germline Components of the Heavy Chain Variable Domain.....	13
1.8 Controlling Classical Immunoglobulin Generation.....	17
1.8.1 The Levels of Regulation of Classical Antibody Generation and Expression.....	17
1.8.2 The Process of V(D)J recombination in Classical IgG Generation.....	19
1.8.3 Classical Immunoglobulin Gene Recombination Signal Sequences.....	21
1.8.4 Components of the Classical Immunoglobulin Gene RSS.....	23
1.8.5 Molecular Machinery Involved in V(D)J Recombination.....	23
1.8.6 Aberrant V(D)J Recombination During Classical Antibody Generation.....	31
1.9 Other Processes Involved In Classical Immunoglobulin Generation.....	31
1.9.1 Somatic Hypermutation within the Classical IgG Genes.....	31
1.9.2 The Process of Classical IgG Post-Transcriptional Modification.....	32
1.9.3 The Role of RNA Splicing in Classical IgG Generation.....	33
1.9.4 Specific Sequences Involved in Splice Site Selection.....	33
1.10 Classical Antibody Secretion and the Role of the Heavy Chain in B-Cell Development.....	35
1.11 The Process of Isotype Switching.....	38
1.12 Evolution of Classical IgG.....	40
1.12.1 Constant regions.....	40
1.12.2 Variable Regions.....	41
1.13 V Gene Assembly in Other Species.....	41
1.14 <i>Lama glama</i> .....	42
1.14.1 Evolution of the family <i>Camelidae</i> .....	42
1.15 Heavy Chain Immunoglobulins.....	43
1.15.1 Structure of the Heavy Chain Only Antibody Protein.....	43
1.15.2 Subclasses of Camelid IgG.....	45
1.15.3 Other Heavy Chain Immunoglobulins.....	45
1.16 Scope of Thesis.....	46
1.17 Contents of Thesis.....	46

## Chapter 2 Methods.....47

2.1 Overview.....	47
2.2 General Methods.....	47
2.2.1 Summary of Techniques Employed in this Thesis.....	47
2.2.2 Polymerase Chain Reaction.....	49
2.2.3 Other General Methods.....	53
2.3 Genomic Library Construction.....	56
2.3.1 Overview of library generation.....	56
2.3.2 Preparation of High Molecular Weight Llama DNA.....	56
2.3.3 Partial Digestion of Genomic DNA.....	57
2.3.4 Ligation of Genomic DNA into Phage Arms.....	60
2.3.5 Packaging of Phage DNA.....	61
2.3.6 Library Titreing.....	61
2.3.7 Library Amplification.....	62
2.4 Screening of a Llama Genomic Library.....	63
2.4.1 Generation of a Variable Gene Segment Probe by Polymerase Chain Reaction.....	63
2.4.2 Design of Separate Classical and Non-Classical V Region Oligonucleotide Probes.....	63
2.4.3 Design of a J Region Probe.....	64
2.4.4 Generation of a Constant Region Probe.....	64
2.4.5 Generation of a RAG-1 Specific Probe.....	64

2.4.6	Synthesis of PCR Generated Probes .....	65
2.4.7	Radiolabelling of PCR-Derived Probes for Library Screening .....	65
2.4.8	Radiolabelling of Synthetic Oligonucleotides.....	66
2.4.9	Plaque Lifting .....	66
2.4.10	Hybridisation .....	67
2.4.11	2nd and 3rd Round Clone Isolation .....	68
2.4.12	Generation of Plate Lysates.....	70
2.4.13	PEG Precipitation and DNA Extraction.....	70
2.5	Characterisation of Clone Sequences .....	70
2.5.1	An Alternative Strategy For J Region Sequencing.....	71
2.5.2	An Alternative Strategy for RAG-1 Gene Sequencing.....	71
2.6	Recombination Assays.....	73
2.6.1	Design of Oligonucleotide Substrates.....	73
2.6.2	Gel Purification of Oligonucleotides .....	73
2.6.3	Labelling of Oligonucleotides .....	74
2.6.4	Annealing of Complementary Oligonucleotides.....	74
2.6.5	Cleavage of Oligonucleotide Substrates by Recombinant RAG Proteins.....	74
2.6.6	Visualisation of Cleavage Products by PAGE .....	74
2.7	Baculovirus Expression and Protein Purification .....	75
2.7.1	Overview of the Bac-to-Bac™ Baculovirus system .....	75
2.7.2	PCR Amplification of Llama RAG-1 Sequence From Clone DNA .....	75
2.7.3	Gel Purification and Digestion of Llama RAG-1 PCR Product .....	77
2.7.5	Ligation of RAG-1 into pFastBac and Transformation into DH5α .....	77
2.7.6	Confirmation of Insert Orientation .....	77
2.7.7	Transformation into DH10Bac and Confirmation of Transposition into Bacmid .....	79
2.7.8	Preparation of Bacmid DNA .....	79
2.7.9	Transfection of Sf9 Insect Cells With RAG-1 Bacmid .....	79
2.7.10	Preparation of High Titres of Recombinant Baculovirus through Secondary and Tertiary Amplification .....	80
2.7.11	Expression and purification of Llama RAG-1/ Murine RAG-2 .....	80
2.8	Sequence Analysis .....	83

### **Chapter 3 Germline Components of the Llama Immunoglobulin Rearrangement Process.....85**

3.1	Abstract.....	85
3.2	Introduction .....	85
3.3	Aims of Library Screening .....	87
3.3.1	Screening Strategy .....	87
3.3.2	Problems of Direct sequencing.....	91
3.3.3	Isolation of Eight Llama Variable Gene Segments.....	91
3.4	Analysis of the Coding Potential of Isolated Llama Germline Variable Gene Segments.....	94
3.5	Detailed Characterisation of V <sub>H</sub> and V <sub>HH</sub> Sequences .....	94
3.6	Amino acid Composition of Variable Gene Segments .....	95
3.7	Analysis of Variable Gene Segments by Comparison to those of Other Species.....	97
3.8	Segregation of Isolated Variable Gene Segments into Specific Variable Gene Families.....	100
3.9	Analysis of Llama Variable Gene Segment Promoter Regions .....	101
3.10	Comparison of Llama and Murine Recombination Signal Sequences.....	104
3.11	Examination of Llama Variable Gene Segments for the Presence of Heptamer-Like Sequences.....	105
3.12	Analysis of Variable Gene Segment Variability.....	106
3.13	Examination of Llama Gene Segments for the Presence of Hypermutational Hotspots within V Region Sequences .....	107
3.14	Variable Gene Segments – Summary of Findings .....	109
3.15	An Alternative Strategy for J Region Sequencing.....	110
3.16	The Isolation of Diversity Gene Segments .....	113
3.18	Determination of Expressed V <sub>HH</sub> and J <sub>H</sub> gene segments .....	117
3.19	Dromedary and Llama V Gene Segments .....	122
3.20	Diversity and Joining Gene Segments – Summary of Findings .....	122
3.21	Discussion.....	123
3.22	Future Work.....	125

<b>Chapter 4 Isolation, Cloning and Expression of Llama Recombination Activating Proteins.....</b>	<b>127</b>
4.1 Abstract.....	127
4.2 Introduction .....	127
4.3 Screening Strategy .....	129
4.3.1 Generation of a RAG-1 Specific Probe .....	129
4.3.2 Genomic PCR to Isolated Partial Nucleotide Sequence of Llama RAG-2 .....	132
4.4 Analysis of sequence surrounding a Llama <i>RAG-1</i> Genomic Clone .....	132
4.5 The Isolation of the Full-length Sequence of The Llama RAG-1 Gene .....	134
4.6 The Isolation of the Partial Sequence of the Llama RAG-2 Gene.....	134
4.7 Comparison of the Llama <i>RAG</i> Genes with Those of Other Species. ....	138
4.8 General Structure and Features of Murine RAG-1 Proteins .....	143
4.9 Functional Murine RAG-1 Mutant Proteins .....	144
4.10 Analysis of the Llama <i>RAG-1</i> Gene for Regions of Potential Functional Significance.....	144
4.11 General Structure of the Murine RAG-2 Protein.....	146
4.12 Analysis of the Llama <i>RAG-2</i> Gene for Regions of Potential Functional Significance. ....	146
4.13 Features of Llama RAG-2 with Potential Functional Significance .....	146
4.14 RAG Protein Expression Strategy .....	147
4.14.1 Expression of Llama RAG-1 in a Baculovirus System .....	147
4.14.2 Expression and Purification of Llama RAG-1/ Murine RAG-2 .....	147
4.14.3 Testing the Activity of the Llama RAG-1/Murine RAG-2 proteins .....	148
4.15 RAG Gene Isolation and Expression – Summary of Findings .....	151
4.16 Discussion.....	152
4.17 Future Work.....	153
<b>Chapter 5 Examination of the Process of Camelid V(D)J Recombination .....</b>	<b>154</b>
5.1 Abstract.....	154
5.2 Introduction .....	155
5.3 Strategy .....	155
5.4 The Cell-Free Recombination Assay .....	156
5.4.1 Interpretation of Recombination Assays.....	156
5.4.2 Components of <i>In Vitro</i> Studies of RAG-1/RAG-2 Activity.....	159
5.4.3 Oligonucleotide Design.....	160
5.5 Analysis of Llama Recombination Signal Sequences .....	162
5.6 Initial Murine RAG Processing of Llama Recombination Signal Sequences .....	162
5.7 Interaction of Murine RAG Proteins with Llama RSSs in the Presence of Manganese .....	164
5.8 Substitution of Coding Flank.....	166
5.9 The Effect on Murine RAG/Llama RSSs of Coding Flank Substitution.....	166
5.10 Llama RAG-1/Mouse RAG-2.....	168
5.11 Comparison of Murine and Llama RAG Protein Activity.....	168
5.12 Processing of Llama Recombination Substrates with Llama RAG- 1/Murine RAG-2. ....	170
5.13 Recombination Assays – Summary of Findings .....	172
5.14 Discussion.....	173
5.15 Future Work.....	175
<b>Chapter 6 Llama Immunoglobulin Constant Region Genes. ....</b>	<b>176</b>
6.1 Abstract.....	176
6.2 Introduction .....	176
6.3 Screening Strategy .....	177
6.4 Sequence and General Organisation of Llama Gamma Immunoglobulin Constant Region Genes.....	177
6.5 Comparison of Llama Constant Genes with other Mammalian IgG Genes.....	185
6.6 Characteristics of the C <sub>H1</sub> exons .....	188
6.6.1 Splice Sites.....	188
6.7 Hinge Sequences .....	189
6.8 Fc Fragments .....	189
6.9 Evolution of Constant Domains.....	191
6.10 Constant Region Gene Isolation – Summary of Findings.....	192
6.11 Discussion.....	193
6.12 Future Work.....	195

<b>Chapter 7 General Discussion.....</b>	<b>197</b>
7.1 Overview .....	197
7.2 A Diverse Range of Heavy Chain Variable Gene Segments Provide the Llama with Unknown Evolutionary Benefits.....	197
7.3 Llama Immunoglobulin Heavy Chain Variable Gene Segments are Encoded in the Germline and May Be Components of a Single Llama Immunoglobulin Locus. ....	198
7.4 Controlling Heavy Chain Immunoglobulin Generation .....	199
7.5 Generation of an Extended CDR3 – Antigen Receptor Selection or Antibody Generation?.....	203
7.6 A Model for Llama Heavy Chain Antibody Secretion .....	206
7.7 Summary.....	207
7.8 Final Conclusions .....	208

<b>Chapter 8 References.....</b>	<b>209</b>
----------------------------------	------------

<b>Appendices.....</b>	<b>234</b>
Appendix I Bacterial Strains, Media, Buffers and Solutions .....	234
Appendix II DNA Quantification .....	239
Appendix III Cloning Vectors Used InThis Study.....	241
Appendix IV Full Sequence of the Llama D-J Locus .....	242
Appendix V Comparison of Llama and Murine RAG Protein Activities. ....	244
Appendix VI Origin of Bacteriophage Clones Derived from Library Screening.....	245
AppendixVII Map of Llama RAG-1/Maltose Binding Protein Baculovirus Expression Vector.....	246

# Table of Figures

## Chapter 1

Figure 1.1	Basic Structure of a Generalised IgG Molecule. ....	6
Figure 1.2	Structure of the Murine Immunoglobulin Heavy Chain Locus. ....	11
Figure 1.3	The Relationship between Variable Region Genes and Antibody Variable Domain Structure. ....	12
Figure 1.4	Basic Intron/Exon Structure of a Classical Heavy Chain Variable Gene Segment. ....	15
Figure 1.5	Processes Involved in the Regulation of Classical Immunoglobulin Gamma Generation and Secretion. ....	18
Figure 1.6	Steps Involved in the Process of V(D)J Recombination. ....	20
Figure 1.7	Cleavage of a V(D)J Recombination Signal by RAG proteins. ....	22
Figure 1.8	Major Domains of the Human RAG-1 Protein. ....	27
Figure 1.9	The Process of Palindromic Nucleotide Addition during Coding Flank Hairpin Resolution. ....	30
Figure 1.10	Key Features Involved in Nuclear Pre-mRNA Splice Processing. ....	34
Figure 1.11	Simplified Diagram Illustrating the Presumed Pathways of Both Classical and Heavy Chain Antibody Production and Secretion. ....	36
Figure 1.12	Diagram depicting events during Isotype Switching from Initial IgM to an IgG Isotype. ....	39
Figure 1.13	Basic Structure of Camelid IgG isotypes. ....	44

## Chapter 2

Figure 2.1	Agarose Gel of Partial Digest of High Molecular Weight Llama Testicular DNA Cut with Sau3A Restriction Endonuclease. ....	59
Figure 2.2	Example of Autoradiographs Generated during Progressive Rounds of Library Screening. ....	69
Figure 2.3	PCR of RAG-1 Clone DNA (a). ....	72
Figure 2.4	PCR of RAG-1 Clone DNA (b). ....	72
Figure 2.5	Overview of Baculovirus Expression. ....	76
Figure 2.6	Confirmation of Direction of Cloned Llama RAG-1 PCR Product by Digestion of pFastbac with <i>Xba</i> I. ....	78

## Chapter 3

Figure 3.1	Strategy for Variable Domain Probe Generation. ....	89
Figure 3.2	PCR for Generation of a Variable Region Probe. ....	90
Figure 3.3	Global Alignment of Germline V Region Gene Segments. ....	92
Figure 3.4	Alignments of Germline Llama V <sub>H</sub> /V <sub>HH</sub> and J Gene Segments. ....	99
Figure 3.5	Rooted Phylogenetic Tree Examining the Relationship between Isolated Variable Gene Segments. ....	102
Figure 3.6	Unrooted Phylogenetic Tree Illustrating the Evolutionary Distance between the Coding Sequence of Variable Gene Segments Isolated in this Study and Variable Domain Families of other Species. ....	103
Figure 3.7	Alignment of V Gene Segment Upstream Regions and Promoters Grouped by Promoter Similarity. ....	107
Figure 3.8	Alignment of Variable Gene Segments Isolated during this Study. ....	112
Figure 3.9	Alternative Strategy for the Isolation of J Region Gene Sequence. ....	115
Figure 3.10	Alternative PCR-based Strategy in order to Sequence a J region Bacteriophage Clone using <i>Taq</i> Polymerase. ....	116
Figure 3.11	Sequence of Llama Diversity Gene Segment D <sub>H</sub> L1. ....	118
Figure 3.12	Output from BLAST Comparison of D <sub>H</sub> L1 with Human D Gene Segment DHQ52. ....	118
Figure 3.13	Organisation of the Llama D-J Locus. ....	120
Figure 3.14	J Gene Segments and Translations. ....	120
Figure 3.15	Strategy Utilised in order to Attempt to Determine the Level of Expression of Particular Llama Germline Gene Segments. ....	120
Figure 3.16	Alignments Showing the Level of Identity between each Variable and Joining Gene Segment and its Best Match from the Unilever cDNA Library. ....	125

## Chapter 4

Figure 4.1	Genomic PCR to Generate a RAG-1 Specific Probe for Screening of the Llama .....	134
Figure 4.2	Genomic Library using <i>Pfu</i> Polymerase. ....	134
Figure 4.3	Subcloning Strategy for Llama RAG-1. ....	135
Figure 4.4	Genomic Consensus PCRs of Llama RAG-2 Gene .....	137
Figure 4.5	Presumed Overall Structure of Llama RAG-1 Bacteriophage Clone. ....	137
Figure 4.6	Nucleotide and Predicted Amino Acid Sequence of Llama RAG-1. ....	139
Figure 4.7a	Partial Nucleotide and Predicted Amino Acid Sequence of Llama RAG-2. ....	141
Figure 4.7b	Global Pairwise Alignment of Llama and Murine RAG-1 Amino Acid Sequence. ....	144
Figure 4.8	Global Pairwise Alignment of Llama and Murine RAG-2 Amino Acid Sequence .....	146
Figure 4.9	Unrooted Phylogenetic Tree Illustrating the Evolutionary Relationship between the <i>Rag-1</i> Nucleotide Sequence of all Characterised Species. ....	147
Figure 4.10	Purification of Recombinant Llama RAG-1 Protein. ....	154
	Demonstration of the Ability of Llama RAG-1/Murine RAG-2 to Cleave a Murine Control RSS in the Presence of $Mn^{2+}$ Ions. ....	155

## Chapter 5

Figure 5.1	Anatomy of a Recombination Substrate. ....	162
Figure 5.2	Schematic Describing the Typical Migration Pattern of Recombination Assay Products. ....	162
Figure 5.3	Diagram Illustrating the Stages of Processing of Recombination Signal Sequences (RSSs) by the RAG Proteins. ....	163
Figure 5.4	Comparison of the Sequence of the Llama Recombination Substrates. ....	167
Figure 5.5	Nicking of Llama RSS DNA Substrates by Murine RAG-1/-2 in the Presence of $Mg^{2+}$ . ....	169
Figure 5.6	Ability of Murine RAG-1/-2 to Cleave Llama Recombination Substrates in the Presence of $Mn^{2+}$ .....	171
Figure 5.7	Restoration of Hairpinning Through the Substitution of Coding Flanks in the Presence of $Mn^{2+}$ .....	174
Figure 5.8	Processing of Control 12-RSS with Murine RAG-1/-2 and Llama RAG-1/Murine RAG-2. ....	175
Figure 5.9	Processing of Llama Recombination Substrates with Llama RAG-1/Murine RAG-2. ....	177

## Chapter 6

Figure 6.1	PCR to Generate a Constant Region Probe Specific to the $C_H2$ Region of the Llama .....	183
Figure 6.2	Constant Gene. ....	183
Figure 6.3	Basic Layout of Llama Constant Genes Including Splice Sites and Exon/Intron Lengths. ....	184
Figure 6.4	Sequence of Four Llama Constant Region Clones Aligned using the ClustalW Method. ....	186
Figure 6.5	Alignment of Inferred Amino Acid Sequences from Llama Constant Region Clones. ....	191
Figure 6.6	Comparisons of Llama Immunoglobulin Constant Genes with those of other Species. ....	193
Figure 6.7	Rooted Phylogenetic Tree Comparing the Evolutionary Relationship between the Nucleotide Sequence of the Constant Region Domains of a Number of Species and Isotypes .....	194
	Proposed Mechanism of Splice Site Knockout in Germline IgG2b and IgG2c genes. ....	201

## Chapter 7

Figure 7.1	A Hypothetical Model Describing the Possible Ways in Which a B-cell may be able to Generate Heavy Chain Antibodies .....	209
------------	---	-----

# Table of Tables

## Chapter 1

Table 1.1	Crucial Residues within Human Recombination Signal Sequences.....	24
Table 2.1	Summary of Techniques Employed in this Thesis. ....	48

## Chapter 2

Table 2.2	Summary of Contrasting Properties of <i>Taq</i> and <i>Pfu</i> Thermostable DNA Polymerases .....	49
Table 2.3	Summary of PCR Reactions Used in this Study.....	50
Table 2.4	Components of a <i>Taq</i> -based PCR Reaction .....	51
Table 2.5	Components of a <i>Pfu</i> -based PCR Reaction .....	52
Table 2.6	Components of a Restriction Digest Reaction .....	53
Table 2.7	Components of a Vector/insert Ligation Reaction .....	54
Table 2.8	Composition of a Proteinase K Digestion Reaction. ....	57
Table 2.9	Composition of Partial Digestion Buffers for Library Generation .....	57
Table 2.10	Small Scale Partial Digest Reactions. ....	58
Table 2.11	Components of a Dephosphorylation Reaction Used in Library Construction .....	60
Table 2.12	Components of a Bacteriophage Vector Genomic Library Ligation Reaction .....	60
Table 2.13	Sequence of Oligonucleotide Probes Used during Library Screening for Variable Gene Components. ....	64
Table 2.14	Initial Components of a MegaPrime™ Radiolabelling Reaction .....	65
Table 2.15	Components of a T4 Kinase Oligonucleotide End-radiolabelling Reaction .....	66
Table 2.16	Post-Hybridisation Washing Conditions for Oligonucleotide-Based Hybridisation .....	68
Table 2.17	Post-Hybridisation Washing Conditions for PCR Probe-based Hybridisation .....	68
Table 2.18	Components of an in vitro Recombination Assay .....	75
Table 2.19	Buffers used during Affinity Purification of Recombinant RAG Proteins .....	84
Table 2.20	Outline of Sequence Analysis Tools Used Throughout this Study. ....	85

## Chapter 3

Table 3.1	Identity (%) Matrix for Variable Gene Segments using Coding Sequence only.....	102
Table 3.2	Transcription Factor Binding Motifs Located Upstream of the Variable Gene Segments isolated in this Study .....	106

## Chapter 4

-

## Chapter 5

Table 5.1	Summary of Recombination Signal Sequences derived from V,D and J Gene Segment Data. ....	167
Table 5.2	Sequence of Recombination Substrates used in this Study. ....	167

## Chapter 6

-

## Chapter 7

-

## Abbreviations

BiP	Binding protein
Bp	Base pairs.
BSA	Bovine serum albumin
cDNA	Complementary deoxyribonucleic acid
CDR	Complementarity determining region
C <sub>H</sub>	Heavy chain constant domain
CIAP	Calf intestinal alkaline phosphatase
C <sub>L</sub>	Light chain constant domain
EDTA	Ethylene diamine tetraacetic acid
Fab	Antigen-binding fragment
Fc	Crystalline fragment
FR	Framework region
Fv	Variable fragment
IgG	Immunoglobulin gamma
kb	Kilobases
kD	Kilodaltons
NAR	Nurse shark antigen receptor
OD	Optical density.
PAGE	Polyacrylamide gel electrophoresis.
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
RACE	Random amplification of cDNA ends
RAG	Recombination activating gene
RSS	Recombination signal sequence
SCC	Stable cleavage complex
SDS	Sodium dodecyl sulphate
SSC	Disodium citrate
TdT	Terminal deoxyribonucleotidyl transferase
Tris	Tris(hydroxymethyl)aminomethane
U	Unit
UTR	Untranslated region
V <sub>H</sub>	Heavy chain variable domain
V <sub>HH</sub>	Camelid heavy chain <i>only</i> heavy chain variable domain
V <sub>L</sub>	Light chain variable domain



## Amino Acid Codes

A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine

# Chapter 1 Introduction

## 1.1 Overview

The camelid family has been found to possess unique antibodies with the ability to specifically bind antigen by virtue of a single protein domain (1). The discovery of these novel heavy chain molecules has led to the intriguing possibility that such antibodies may, in the near future, be engineered so as to provide inexpensive, highly soluble antigen-binding units for use in combating a wide range of diseases (2). While a number of studies have investigated the physical and chemical properties of these novel antibodies (3), the mechanisms by which such structurally unconventional molecules are generated remain poorly defined.

In order to set this project in an appropriate framework the introduction firstly provides a broad discussion detailing the reasons for investigating the generation of immunoglobulins in the llama (section 1.2-1.6). This chapter goes on to define and describe the subclass of antibodies to which the heavy chain antibody and conventional llama immunoglobulin gamma are closely related (section 1.7). Subsequent sections (sections 1.8-1.10) discuss the established mechanisms of antibody generation utilised in species as diverse as the human and the goldfish. Particular attention is paid to novel and unusual examples that may provide clues to and precedents for, the process of camelid antibody formation. Finally the objectives of this project are set in context by provision of a brief description of the camelid family and review of the current state of heavy chain antibody research (section 1.14-1.15). Throughout this work the novel antibody type found in the camelids and lacking the light chain is referred to as the *heavy chain antibody*. The conventional heavy and light chain immunoglobulin is referred to as the *classical or conventional antibody*.

## 1.2 The Commercial Potential of Llama Immunoglobulins

The discovery of llama heavy chain antibodies comprising only a single variable domain provides a new minimum antigen-binding unit. This is the first naturally occurring example of an interface for antigen interaction that does not require the participation of a light chain variable domain. Not only is this single domain smaller, and therefore more soluble than conventional antigen binding domains, but the llama

single domain also provides the opportunity to express an antigen binding domain as a single continuous heavy chain variable polypeptide, rather than a combination of light and heavy chain sequences. This ensures that heavy chain antibody domains can be expressed simply and rapidly, and that a broad range of antigen binding specificities can be generated through a single cloning step. The commercial potential of llama immunoglobulin gamma was confirmed in the 1990's by the filing of a patent (Patent number EP584421 (4)) providing Unilever Research with sole rights for commercial exploitation of these antibodies in consumer products. This has led to intensive research that has identified both naturally and synthetically derived llama antibody molecules specific to a wide range of commercially relevant antigens. An example of this is the production of a heavy chain antibody specific to buccal bacteria that provides a targeted mechanism for the reduction of dental decay. Such an antibody was considered for inclusion within toothpaste.

Although the current emphasis in pharmaceutical research is on the use of human antibodies, the key features of the single-domain heavy chain immunoglobulin (5-7) may ultimately lead to applications in human disease therapy. Indeed, the small size of the variable domains provides the potential for excellent tissue penetration. The possibly toxic nature of such antibody therapies may also be reduced by the use of heavy chain antibody domains. The small size of such a domain would also ensure rapid clearance from the body. Clearly a greater understanding of the mechanisms by which such antibodies are generated and utilised by the llama would be beneficial before therapeutic strategies are considered.

### **1.3 Possible Roles for the Heavy Chain Antibody in the Llama**

The llama produces both classical and heavy chain antibodies. The unique structure of the llama heavy chain antibody is highly suggestive of novel roles for such immunoglobulins within the llama immune system. It is difficult to imagine how such structural modifications could have evolved without providing any advantage to the animal. The discovery of the heavy chain antibody, not only in the llama but also within close evolutionary relatives such as the dromedary, leads to the possibility that the harsh environments that make up the natural habitats of the camel and llama might somehow be responsible for the development of this unusual antibody class.

However, there is no firm evidence that this is the case and no suggestions have been made as to the nature of such benefits. The recent finding that heavy chain antibodies are superior enzyme inhibitors (8) may be physiologically relevant. Indeed it has been noted that camelids have unusually high resilience to a number of viral and bacterial infections that afflict evolutionarily close, but non-heavy chain antibody producing species (3, 9). It is possible, although by no means certain, that the heavy chain antibody may assist in the successful combat of such infections, perhaps utilising enzyme inhibitory characteristics during resolution of the infection.

#### **1.4 Potential Mechanisms of Heavy Chain Antibody Generation**

Given the current understanding of the processes involved in antibody generation within the developing human and murine B-cell, it is possible to speculate as to the mechanisms involved in heavy chain antibody formation. Any of the following could contribute to the generation of such antibodies:

- 1) The novel sequence and structure of the llama antibody could result through affinity maturation. This mechanism, involving the process of somatic hypermutation (section 1.9.1) leads to the gradual increase in the affinity of an antibody to a particular antigen after initial antigen exposure (10).
- 2) The unique composition of the heavy chain antibody may be the result of specific, heavy chain antibody-encoding genetic components within the llama genome.
- 3) The smaller size of the heavy chain antibody (by comparison to conventional antibodies) may be the result of the deletion of specific genetic information from the llama genome.
- 4) A novel mechanism (or mechanisms) of recombination may be responsible for the unusual sequence and structural features of the llama antibody.
- 5) The recombination mechanisms utilised during human and murine antibody generation may have been subtly refined by the llama in order to generate heavy chain antibodies.
- 6) Even after heavy chain antibody-encoding sequences are derived (by any or all of the processes described in points 1-5), the generation of biologically active heavy chain antibodies may be modulated at any number of levels including transcription and translation.

## **1.5 Aims of Study**

The llama heavy chain antibody represents an unusual and potentially valuable adaptation of the conventional immunoglobulin gamma antibody form. Two separate but equally compelling reasons led to the examination of the mechanisms by which such an antibody is generated. Firstly it was considered important to understand how an antibody with such unique structural characteristics (section 1.15.1) could be generated from germline immunoglobulin sequences. Does the generation of the novel heavy chain antibody involve entirely novel genetic mechanisms? Could such antibodies be generated by somatic events alone? Does a single immunoglobulin locus generate both heavy chain immunoglobulin and the conventional immunoglobulins? The reconstitution of events involved in heavy chain antibody generation make up Chapters 4 and 5 of this thesis.

A second motive behind the examination of llama immunoglobulin generation was the possibility that a greater understanding of the llama immune system could ultimately lead to the manipulation of immune responses, so that production of the heavy chain antibody, that typically comprises only 10-30% of the llama serum immunoglobulin gamma could be regulated. Higher levels of heavy chain antibody generated through immunisation would be of considerable commercial benefit. Speculation as to possible mechanisms by which llama immunoglobulin serum levels may be regulated is discussed in Chapters 3, 6 and 7.

## **1.6 Research Strategy**

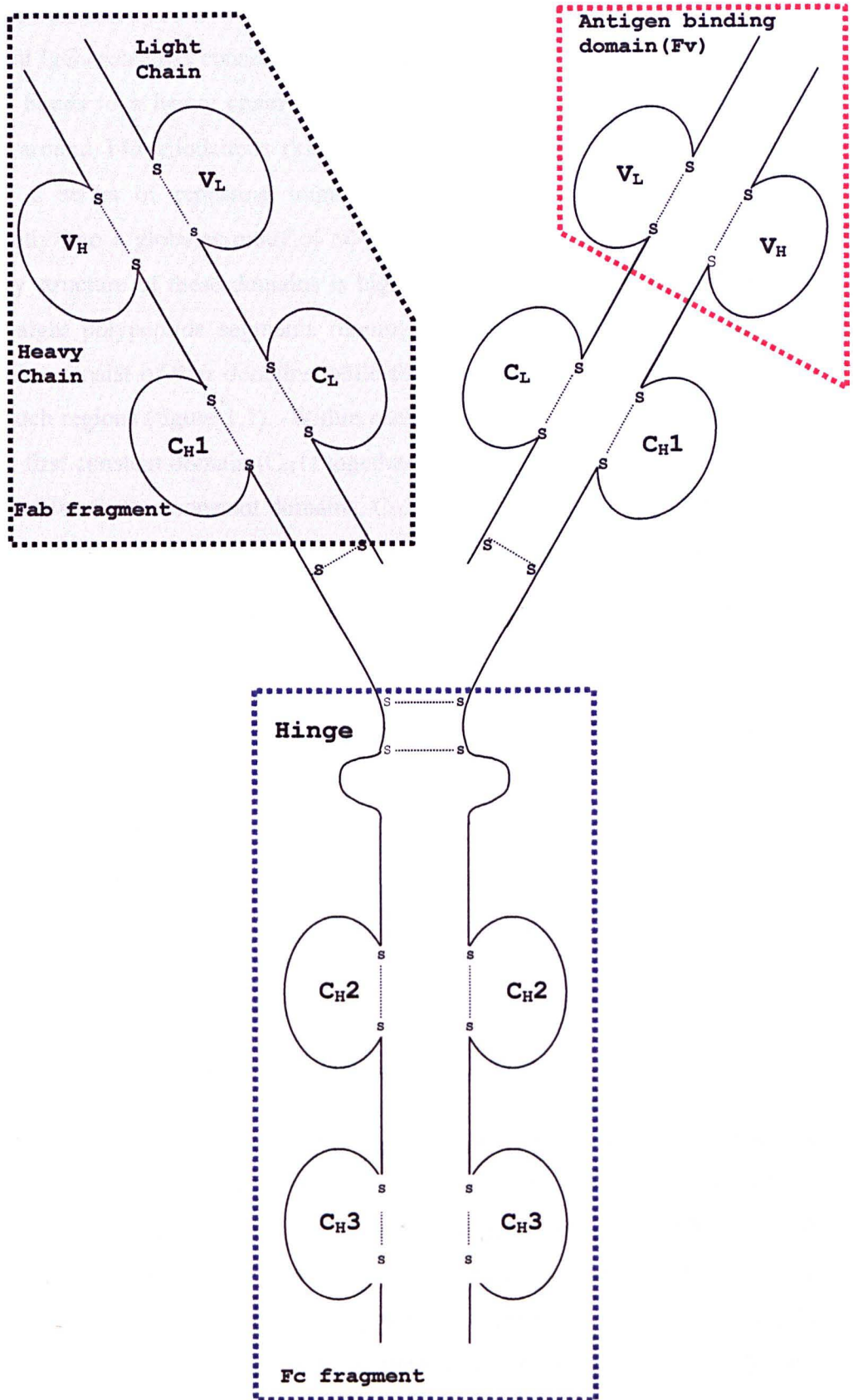
The investigation of a biological process in a little-studied species such as a llama presents a number of difficulties. Almost no biological research has been carried out on the llama other than studies into the breeding and raising the animal. Studies into human and murine antibody generation are able to take a detailed map of the antibody genetic loci and sequence of other immunological components for granted. In this study no such information was available. This meant that before the mechanisms of antibody generation could be investigated it was essential to isolate the components of the llama genome that may ultimately be responsible for generating these novel antibodies.

The research strategy was therefore as follows:

1. Isolate relevant components of the llama immune system (using a genomic library) (sections 3.3.3, 4.5-4.6 and 6.4)
2. Analyse these components for clues into possible mechanisms of antibody generation (sections 3.5-3.13, 4.7-4.13, 5.5-5-12 and 6.6-6.7)
3. Develop hypotheses to explain the manner in which heavy chain antibodies may be generated (sections 3.21, 4.16, 5.14, 6.9 and Chapter 7)
4. Test these hypotheses through reconstruction of antigen generation processes *in vitro* (sections 5.5-5.12)

### **1.7 Classical Immunoglobulin Gamma**

Immunoglobulins are a group of multifunctional serum glycoproteins essential for the prevention and resolution of infection. The highly variable structure of these molecules enables interaction with a diverse range of invading pathogens and pathogenic products. Classical IgG is the predominant immunoglobulin of the secondary immune responses that follow repeated exposure to an antigen. Once bound to an antigen the protective function of the immunoglobulin is determined by its ability to activate any number of effector mechanisms, leading to the neutralisation, and elimination of invading micro-organisms. Each immunoglobulin molecule can be structurally divided into an antigen-binding region (*fragment antigen binding* or Fab) and a portion responsible for effector functions such as interaction with cells of the immune system (Fc, so-called because it readily crystallises) (figure 1.1). Immunoglobulins are classified by virtue of their overall structure into a number of functionally distinct classes or isotypes including immunoglobulin gamma (IgG), immunoglobulin epsilon (IgE) and immunoglobulin alpha (IgA).



**Figure 1.1 Basic structure of a generalised IgG molecule.** The relative positions of heavy and light chain domains are illustrated. Also shown are the functional antibody fragments Fab, Fc and Fv. Fab fragments are responsible for antigen interaction and are comprised of an Fv and  $C_{H1}$  domain. The Fc portion is responsible for interaction with both adaptive and innate arms of the immune system

### 1.7.1 The Structure of the Classical Immunoglobulin Gamma Protein

All classical IgG molecules consist of two identical light polypeptide chains linked by disulphide bonds to a heavy chain polypeptide dimer, with a combined molecular weight of around 146 kilodaltons (kD). Both light and heavy chain polypeptides consist of a series of repeating immunoglobulin domains each of which fold independently into a globular motif of 60-70 amino acids in length. The secondary and tertiary structure of these domains is highly conserved and typically consists of several straight polypeptide segments running parallel to the axis of the domain. Heavy chains consist of four domains while the light polypeptides are composed of only two such regions (figure 1.1). Within each heavy chain the single variable ( $V_H$ ) region and first constant domain ( $C_{H1}$ ) together provide the heavy chain contribution to the Fab. Two further constant domains,  $C_{H2}$  and  $C_{H3}$  comprise the Fc portion. A hinge region of between ten and sixty amino acids separates the  $C_{H1}$  and  $C_{H2}$  domains and allows a degree of flexibility between the two antigen binding sites and the Fc region. Light chains are concerned only with Fab formation. The single variable domain ( $V_L$ ) and single constant domain ( $C_L$ ) interact with  $V_H$  and  $C_{H1}$  domains respectively to complete each antigen binding site (11). In summary, therefore,  $V_H$ ,  $V_L$ ,  $C_{H1}$  and  $C_L$  make up the antigen binding Fab while the  $C_{H2}$  and  $C_{H3}$  comprise the effector-related Fc region.

### 1.7.2 The Functional Domains of Classical IgG

As described previously the classical immunoglobulin can be divided into two functional domains, the principal characteristics of which are as follows:

#### (a) *Fab*

Within the Fab portion variable ( $V_H$  and  $V_L$ ) and constant ( $C_{H1}$  and  $C_L$ ) domains interact to generate a specific antigen-binding interface. It is the association of  $V_H$  and  $V_L$  chains that provides the surface at which antigen is encountered by classical antibodies. During association a number of loop structures within both light and heavy chain variable domains are brought together. Antibody/antigen association is typically the combined result of non-covalent electrostatic, van der Waals and hydrogen bond interactions.



- **Variable Domains**

The variable domains of both heavy and light chains ( $V_H$  and  $V_L$  respectively) provide diversity within immunoglobulin Fab regions enabling recognition of a broad range of antigen encountered by the immune system. Although exceptions have been reported (12)  $V_H$  and  $V_L$  domains together typically represent the minimal antigen-binding unit of the antibody. Each dimerised heavy and light variable domain is often referred to as an antibody variable fragment (Fv) region.

Heavy and light chain V domains are structurally similar and share a common protein subdomain organisation. Both  $V_H$  and  $V_L$  domains are approximately 110 amino acids in length and from amino to carboxyl ends respectively, comprise framework region 1 (FR1), complementarity-determining region 1 (CDR1), FR2, CDR2, FR3, CDR3 and FR4. These regions were originally determined by the analysis of variability within immunoglobulin cDNA sequences (13). Framework regions were assigned due to an overall lower level of sequence variability when compared to the CDRs. It has subsequently been shown that FRs typically correspond to regions of polypeptide  $\beta$ -chain while CDRs represent loops linking these  $\beta$ -chains. Three hypervariable heavy chain loops (H1, H2, and H3) correspond to their similarly numbered CDR (14) and interact during heavy/light chain dimerisation with corresponding light chain L1, L2 and L3 hypervariable, CDR-derived loops. Structural studies have shown the CDR3-encoded H3 loop to be situated within the centre of the antigen-binding groove formed by  $V_L$  and  $V_H$  interactions and therefore to be key to immune recognition (15, 16). The CDR3s of both light and heavy chains are generally found to possess greater sequence diversity than either CDR1 or CDR2 (17). Amino acid diversity is therefore highest within the centre of the antigen-binding groove.

Differences between  $V_H$  and  $V_L$  structure include specific motifs within their respective FR2 enabling interaction of heavy and light chains. In addition,  $V_H$  domains tend to have longer FR1 and CDR2 regions but shorter FR2 and CDR1 regions than their light chain counterparts. Overall, the composition of the Fv domain can be seen as a compromise between the requirement to maximise antibody

variability and the need for light/heavy chain association that is essential for the formation of the antigen-binding groove.

- **Constant domains**

The first constant domains of the light and heavy chains ( $C_{H1}$  and  $C_L$ ) lie immediately C-terminal of the variable domains that form the antigen-binding cleft. The constant domain of the light chain pairs with the first constant region domain of the heavy chain ( $C_{H1}$ ) through hydrophobic and disulphide interactions that stabilise heavy and light chain association. The  $C_{H1}$  domain is also important in immunoglobulin synthesis, providing a region for *Binding Protein* (BiP) association, an interaction that prevents heavy chain secretion in the absence of light chain synthesis (section 1.10 and 7.6).

**(b) Fc**

In contrast to the Fab portion of the antibody, the Fc region is comprised only of heavy chain domains ( $C_{H2}$  and  $C_{H3}$ ). While the functional significance of the Fab region has been intensively studied in recent years, the Fc domains await more comprehensive characterisation. Although not directly relevant to antigen binding, definition of the role of components of the Fc region may enable the engineering of antibodies able to interact with the immune system in more restricted and tightly regulated ways. By studying the effect of domain-shuffled antibody variants (18) the  $C_{H2}$  region has been shown to have a crucial role in Fc function. Mutagenesis has identified crucial sites responsible for binding of complement proteins and the Fc receptor within the  $C_{H2}$ . To our knowledge no other such studies have been conducted to examine the roles of the  $C_{H3}$  domain. Recently the three-dimensional structure of the interaction between the Fc portion of the antibody and an Fc receptor has been published (19). This should provide a greater understanding of the manner in which effector mechanisms are activated by the IgG molecule.

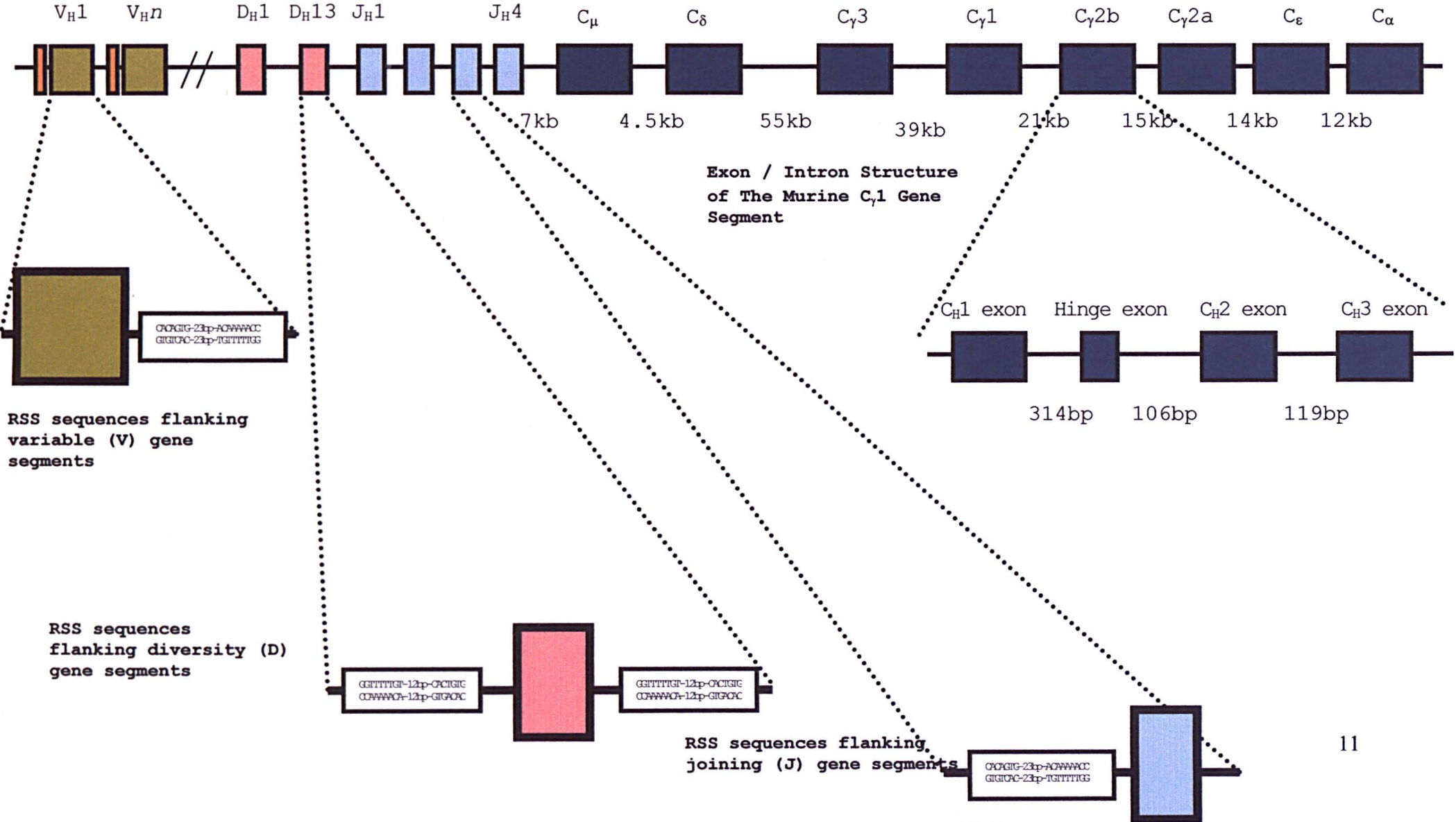
### **1.7.3 The Structure of the Classical Immunoglobulin Gamma Heavy Chain Gene**

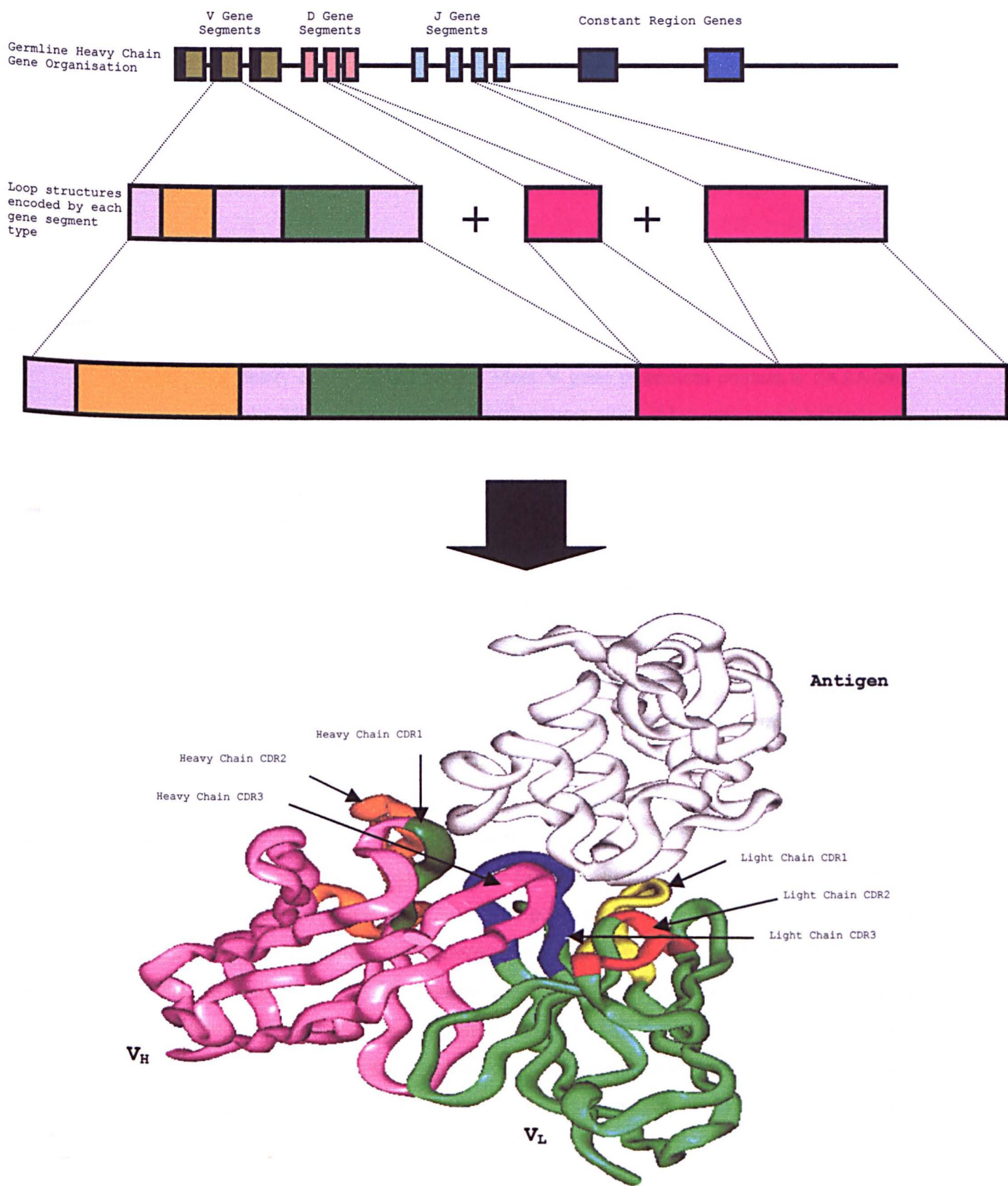
In order to set in context the study of llama heavy chain immunoglobulin generation reported in this thesis, the remainder of this introduction deals only with the genetic organisation of the heavy chain locus and the generation of heavy immunoglobulin chains. Classical immunoglobulin light chains are encoded at a separate locus of similar organisation. The classical immunoglobulin gamma genes are comprised of a number of elements that, through the process of V(D)J recombination, combine to provide separate exons encoding variable, hinge and constant domains. Separate loci containing multiple variable (V), diversity (D) and joining (J) gene segments are found within the heavy chain locus of all vertebrates. Together these provide a diverse range of coding sequences from which  $V_H$  domains can be encoded. Downstream of these genes, relatively short introns (typically 100-500bp in length) separate hinge and constant region exons. The distance between the variable domain exon and that of the first constant domain is dependent not only on the choice of variable region elements during V(D)J recombination, but also on the antibody isotype chosen through the process of isotype switching (section 1.11). This equates to a minimum of approximately 7kb in the murine locus. An example of heavy chain gene organisation is given in figure 1.2. The relationship between the position of germline sequence elements within the heavy chain locus and the components of the three dimensional variable domain structure is illustrated in figure 1.3.

**Figure 1.2**

**Structure of The Murine Immunoglobulin Heavy Chain Locus.** This consists of a large number of Variable, Diversity and Joining gene segments arranged in tandem (a reduced number are shown for illustrative purposes), downstream of which lie constant region genes corresponding to each isotype. The top of the figure shows the layout of these genes and an indication of approximate distances between the different components. Below close ups of the sequences flanking each variable, diversity and joining gene are shown and a more detailed diagram illustrating the typical exon/intron structure of a constant region gene

$n = 300-1000$





**Figure 1.3 The Relationship Between Variable Region Genes and Antibody Variable Domain Structure.** V gene segments encode CDR1 and CDR2 regions and are responsible for eventual H1 and H2 loop composition. Both D and J regions combine to encode the CDR3 that dictates H3 loop amino acid primary sequence. The antigen interacting with the V<sub>H</sub> domain is given in white.

#### 1.7.4 Germline Components of the Heavy Chain Variable Domain

Within the overall heavy chain locus organisation described above, particular sets of gene elements (or *gene segments* as they are usually referred to) make different contributions to the final variable region domain sequence. Each type of gene segment, and its contribution to the antibody, is here discussed in turn.

##### (a) The Classical Variable Gene Segment and Promoters

'Variable' or  $V_H$  gene segments are situated at the 5' most end of conventional heavy chain loci. Upstream of each V segment lie promoters and enhancers responsible for regulating heavy chain transcription. Most V gene promoters contain a TATA box approximately 25bp 5' of the transcription start site. Also conserved between all heavy chain  $V_H$  gene sequences is the presence of an octamer sequence (consensus ATGCAAAT), the inverted complement of which lies upstream of light chain V gene segments (20). This octamer region has been shown to be crucial to B-cell specificity and is required for optimal *in vitro* transcription in B-cell nuclear extracts, while having no effect on HeLa nuclear extract-based transcription (HeLa nuclear extracts are derived from a human non-B-cell line and therefore contain all ubiquitously expressed nuclear proteins but none found specifically in B-cells) (21). Two proteins, Oct-1 and Oct-2 are known to bind the octamer motif and there is some evidence that Oct-2 alone may provide B-cell-specificity (22, 23). Unlike many  $V_L$  promoters, the  $V_H$  promoters often contain additional functional elements. For example, the heptamer CTCATGA is commonly found 2-22bp upstream of the octamer motif, and is thought to play a role in co-operative Oct protein binding (24, 25). Upstream of this heptamer, a polypyrimidine tract also plays a role in optimal promoter function (26). Of possible significance in immunoglobulin heavy chain transcriptional control is the presence of an A/T rich element 125-250bp upstream of the transcription site in a particular murine  $V_H$  gene segment (VHS107) (27) a region that can mediate increased transcription in the presence of increased interleukin-5 (IL-5). This region, designated the *B-cell regulator of immunoglobulin heavy chain (IgH) transcription* (Bright), has not, however, been found in association with the majority of known  $V_H$  genes. The variations in  $V_H$  gene promoter content may be crucial in differential regulation of V region sterile transcription, a process whereby variable gene sequences are transcribed prior to V(D)J recombination and immunoglobulin expression. Sterile transcription is thought to alter the accessibility of the DNA such

that specific V regions are more readily utilised during subsequent rearrangement (28, 29).

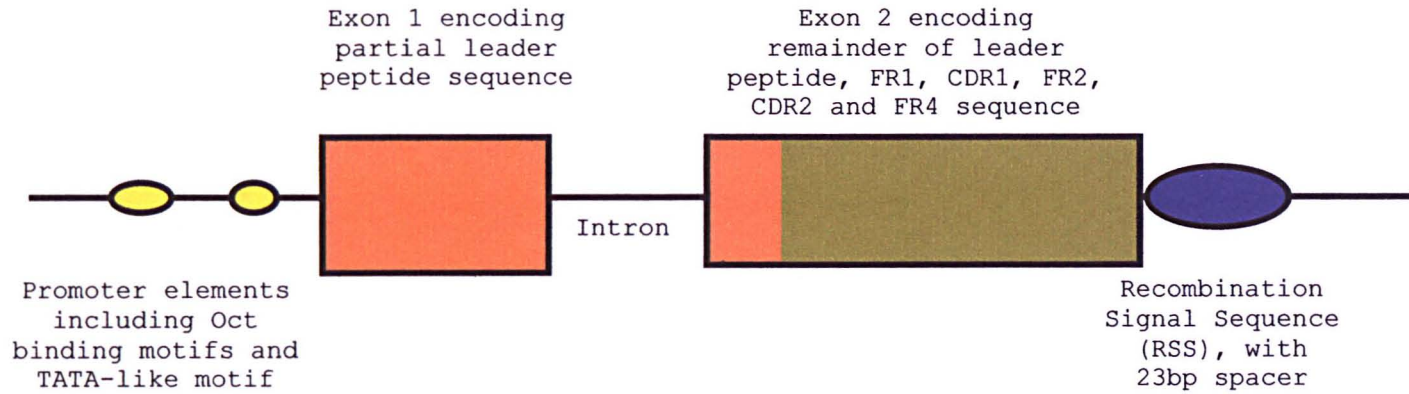
Human  $V_H$  region genes consist of a leader (L) peptide exon, a short intron of approximately 200bp and a second exon completing the leader peptide and also encoding FR1, CDR1, FR2, CDR2 and FR3 (figure 1.4). The FR1 region typically encompasses the first 90bp of coding  $V_H$  sequence downstream of the leader peptide. The CDR1 is the result of the next 15 or so bases, FR2 the subsequent 42 nucleotides, CDR2 roughly 45 bases and FR3 the final 83-87bp (30). 3' of the second exon lies a specific motif, including highly conserved heptamer and nonamer sequences separated by either 12bp or 23bp nucleotide spacers (23bp spacers are associated with V gene segments). This motif is involved in somatic rearrangement of immunoglobulin genes and known as the recombination signal sequence or RSS (section 1.8.3). The typical composition of a  $V_H$  region gene RSS is illustrated in figure 1.2.

The total size of the human  $V_H$  locus is thought to be about 1.2Mb and the locus contains 123  $V_H$  segments in total, of which 79 are pseudogenes (31). The most 3'  $V_H$  gene is approximately 77kb upstream of the nearest J segment while the most 5'  $V_H$  segment is located only a few kilobases from the telomeric sequence at the end of chromosome 14 (32).  $V_H$  segments are grouped into families within each species. For example human  $V_H$  genes can be divided into seven families based on overall homology and similarities within framework and complementarity-determining regions.

#### **(a) Classical D region gene segments**

Located downstream of the variable gene segments, the human immunoglobulin locus contains 27 'diversity' or  $D_H$  gene segments, most arranged in four clusters of approximately 9kb each (33, 34). Comparisons between human, mouse and germline immunoglobulin sequences of other species demonstrates a high level of D and J sequence conservation between species (35).  $D_H$  gene segments are typically between 11-37bp in length (36) and in general one reading frame provides hydrophilic residues, the second hydrophobic residues while the third often contains stop codons

**Figure 1.4 Basic intron/exon structure of a classical heavy chain variable gene segment.** Positions of promoter elements and recombination signal sequences are also shown.





(37).  $D_H$  gene segments are flanked, both 5' and 3', by 12bp-spacer RSSs that play a role in  $D_H$ - $J_H$  and  $V_H$ - $D_HJ_H$  joining (figure 1.2). A number of extended diversity gene-like (DIR) sequences are also found within the diversity gene locus (36) (section 1.8.6 and 7.5).

**(b) Classical J region gene segments**

The most 3' elements to make a contribution to variable domain composition are the 'joining' or  $J_H$  gene segments. The human  $J_H$  locus consists of 9  $J_H$  segments of which 3 are pseudogenes. The  $J_H$  gene segments are sandwiched between  $D_H$  gene segments and the constant region genes with the nearest 5'  $D_H$  segment approximately 100bp upstream of the first functional  $J_H$  (38) (this distance is similar in the mouse locus) (figure 1.2). Each  $J_H$  segment not only contributes to CDR3 formation during the initial  $D_H$ - $J_H$  joining event but also contains the FR4 coding sequence that adjoins the heavy chain hinge in the completed antibody (figure 1.1). The  $J_H$  segment also provides a 5' 23-spacer RSS crucial for  $D_H$ - $J_H$  joining and a 3' splice site for post-transcriptional splicing to the hinge exon (figure 1.2).

**(c) Classical Constant Region Genes**

The previous sections have discussed the regions responsible for the structure of the heavy chain variable domain. The final components of the classical heavy chain immunoglobulin locus are the constant region genes. Constant genes are responsible for encoding the invariant regions of the antibody including  $C_{H1}$  and Fc regions of the heavy chain. The murine constant heavy ( $C_H$ ) locus covers approximately 200kb of germline DNA (39). The most 5' constant gene encodes the Fc portion of IgM ( $C_\mu$ ) and is located roughly 8kb downstream of the nearest J gene segment. The constant region genes are arranged as discrete, tandem sets of exons encoding each isotype (from  $C_\mu$ ,  $C_\delta$ ,  $C_\gamma3$ ,  $C_\gamma1$ ,  $C\alpha1$ ,  $C_\gamma2$ ,  $C_\gamma4$ ,  $C_\epsilon$  to  $C\alpha2$  moving 5'-3' through the human loci). The genes encoding each isotype are separated by distances of between 6-60 kilobases (40). Within each isotype separate domains (typically of 100-110 amino acids) are generally encoded by individual exons and flanked by introns of 100-300bp (41-43) (figure 1.2). For example, the mouse IgG2b constant region consists of three domains,  $C_{H1}$ ,  $C_{H2}$  and  $C_{H3}$  as well as a short hinge domain with the following germline gene organisation:

$C_{H1}$  - intron - *hinge* - intron -  $C_{H2}$  - intron -  $C_{H3}$   
(292) (314) (64) (106) (328) (119) (322)

where the number of nucleotides in each exon/intron is given in parenthesis. A number of exceptions to this gene structure have been found and these are discussed in section 1.12.1

## **1.8 Controlling Classical Immunoglobulin Generation**

While an understanding of the protein structure and gene organisation of classical heavy chain immunoglobulins (section 1.7) is crucial in order to understand and attempt to predict the nature of the genes involved in llama heavy chain antibody generation, it is the understanding of the generation of the expressed heavy chain antibody that remains the ultimate goal of this thesis. The mechanisms involved in controlling classical immunoglobulin generation are therefore given consideration below.

### **1.8.1 The Levels of Regulation of Classical Antibody Generation and Expression.**

The processes involved in the generation and expression of immunoglobulin proteins combine a number of ubiquitous mechanisms such as transcription, translation and post-translational modification (the relationship between these mechanisms is depicted in figure 1.5) with mechanisms unique to immunoglobulin formation and consequently concerned with diversity generation. It has been proposed that the somatic recombination events that take place during the generation of combinatorial diversity are closely linked to transcription (44). Transcription may, therefore, play a role not only in the generation of RNA encoding the fully rearranged immunoglobulin, but also influence the rearrangement of the immunoglobulin genes through the process of sterile transcription (44, 45). It is thought that the generation of RNA transcripts from promoter regions upstream of the variable gene segments may alter the local DNA conformation such that proteins involved in recombination can preferentially access, and interact with those variable genes undergoing sterile transcription. It is not thought that any products of sterile transcription undergo translation. Key elements described in section 1.7.4 may therefore play roles not only in regulating levels of antibody production, but also in preference of

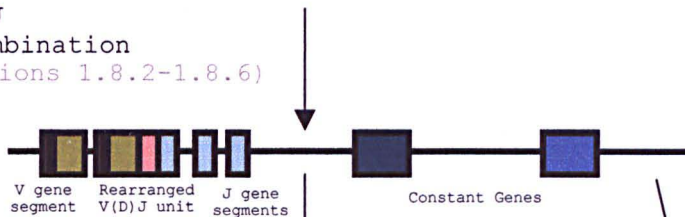
**Figure 1.5**

**Processes Involved in the Regulation of Classical Immunoglobulin Gamma Generation and Secretion.** Generation of Immunoglobulins is a complex multi-step process. Sterile transcription may prepare a particular gene element for rearrangement. V(D)J recombination then brings together V,D and J elements to form a single variable region exon. This V(D)J unit and downstream constant exons are then transcribed as a single RNA and spliced to form a complete immunoglobulin heavy chain transcript. Alternatively V(D)J recombination may be followed by isotype switching in which constant region genes encoding particular antibody isotypes are removed from the DNA to enable expression of differing immunoglobulin classes.

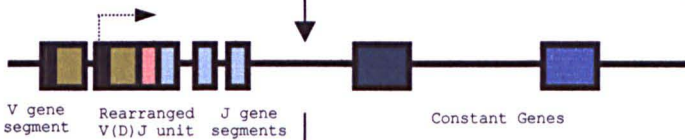
- 1) Sterile Transcription (section 1.7.4 (a))



- 2) V(D)J Recombination (sections 1.8.2-1.8.6)



- 3) Transcription (section 1.7.4)



- 4) Post-Transcriptional Modification (including splicing) (sections 1.9.2-1.9.4)



- 5) Isotype Switching (optional) (section 1.11)



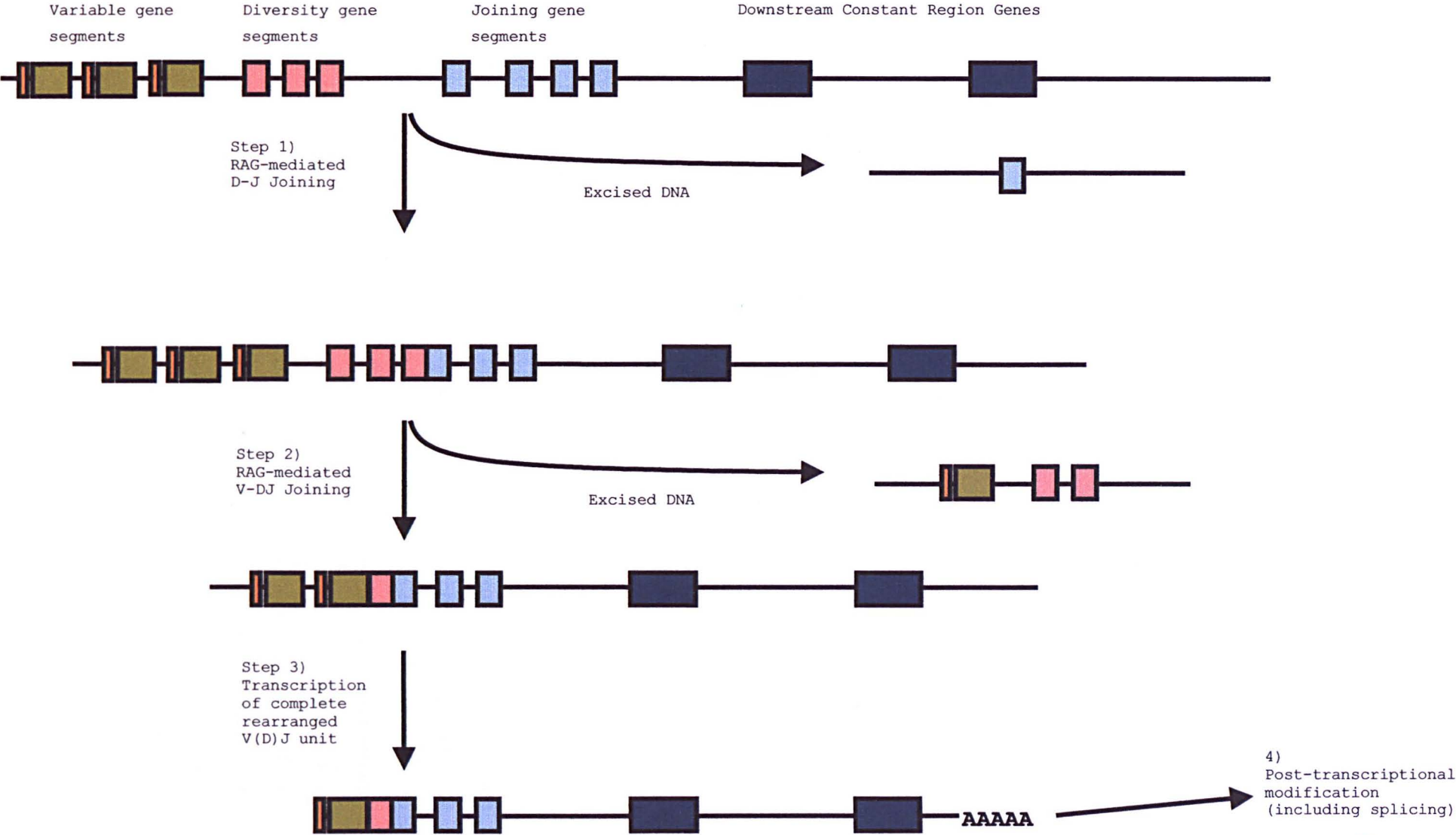
V gene segment undergoing recombination. Any or all of these levels of control may play a role in heavy chain antibody generation.

### **1.8.2 The Process of V(D)J recombination in Classical IgG Generation**

The camelid family, while exhibiting novel and functionally significant heavy chain antibody production, is closely related to other vertebrates (section 1.14.1) and it is therefore reasonable to assume a degree of similarity, if not considerable overlap, between the mechanisms of generation of heavy chain and classical antibodies. Indeed the existence of classical antibodies within camelid sera provides confirmation of the importance of 'traditional' recombination mechanisms within the *Camelidae*. Any study of heavy chain antibody generation must therefore consider the systems characterised in, and common to a wide range of organisms such as the mouse and human. It is estimated that the mammalian immune system is able to generate in excess of  $10^8$  different immunoglobulin specificities. The task of generating such a diverse range of antibody proteins falls to a set of genetic mechanisms, the most significant of which is thought to be the process of V(D)J recombination (figure 1.6).

In 1965 a theoretical model was put forward (46) whereby two separate genes encode single polypeptide chains of each immunoglobulin molecule, one gene encoding the V region, and the other the C region. It was proposed that these two genes come together by some unknown mechanism to form a continuous stretch of DNA from which the mRNA required to encode a light or heavy chain can be transcribed. This model went on to predict the existence of multiple V-region genes within the germline from which the required diversity could be generated. In addition, the presence of single germline copies of C-region class genes was predicted. It was not until almost a decade later that this hypothesis was verified directly. Verification came when the DNA of embryonic cells and that of lymphoid myelomas was compared by restriction endonuclease digestion and hybridisation through the use of immunoglobulin mRNA probes. This showed the transposition of DNA specific to antibody genes from two separate loci within the embryonic material to a single locus in myelomic material (47).

**Figure 1.6 Steps involved in the process of V(D)J recombination.** V(D)J recombination is a somatic process restricted to T and B-cell progenitors, and occurring only at T-cell receptor (TCR) and immunoglobulin loci. The first step in immunoglobulin heavy chain V(D)J recombination is the removal of DNA lying between specific diversity and joining gene segments to form a DJ recombination unit. During this process N and P nucleotide addition may also occur (not shown, for P nucleotide addition see figure 1.9). The next step (Step 2) is the excision of DNA between this DJ unit and a specific variable gene segment to form a complete VDJ unit. Again N and P nucleotide addition may occur at this step. Together these two discreet steps lead to the generation of a complete complementarity determining region 3 (CDR3) prior to transcription (Step 3)



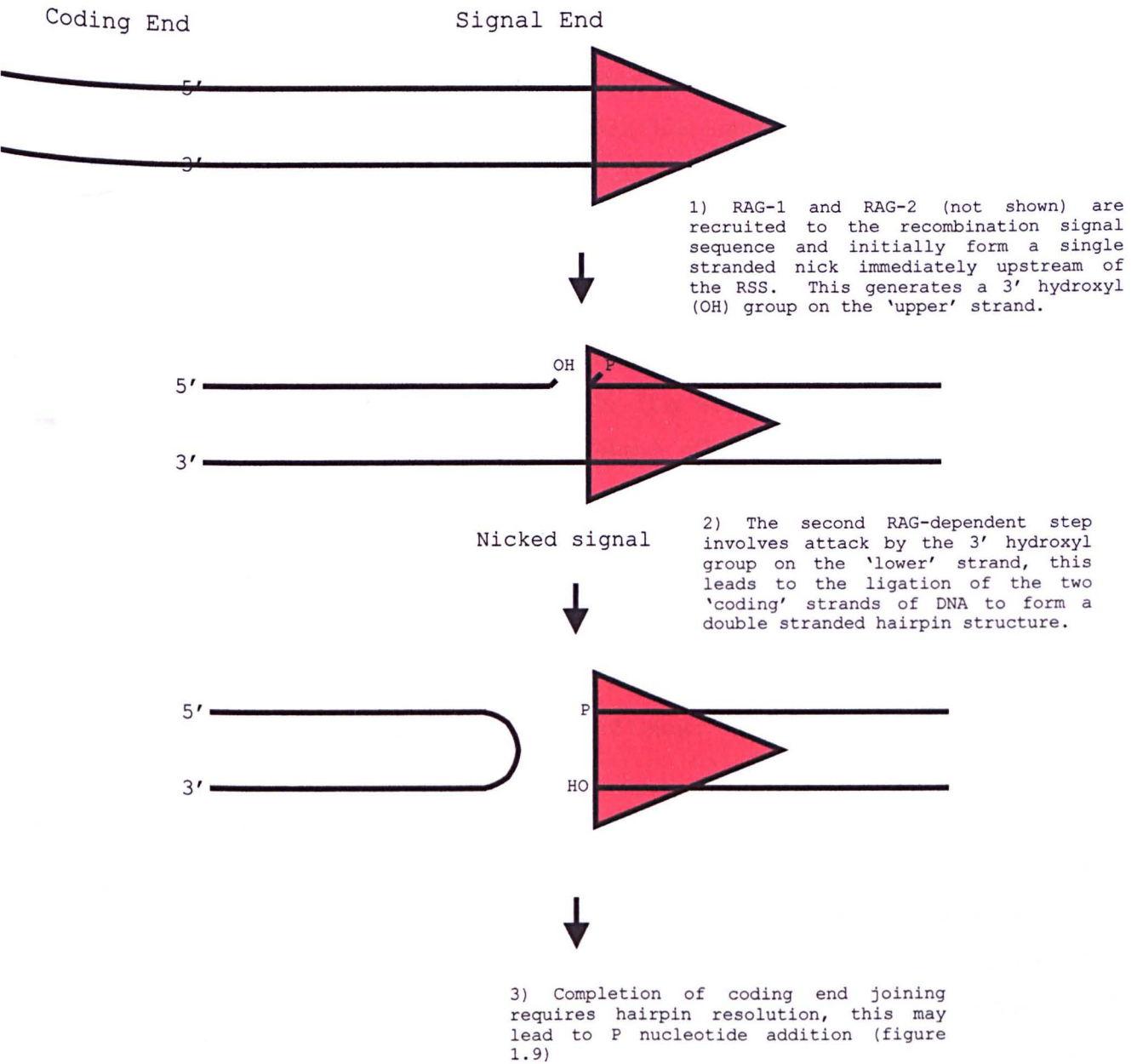
The recombination events leading to the assembly of functional genes encoding immunoglobulin polypeptide chains are the only known example of site-specific DNA rearrangement in vertebrates. Generation of complete, functional heavy-chain immunoglobulin genes requires the contribution of two separate rearrangement events. Classical V(D)J recombination involves the joining of a  $D_H$  gene segment to a  $J_H$  segment, the product of which is then rearranged to lie adjacent to a  $V_H$  segment thereby generating a full  $V_H D_H J_H$  unit encoding the complete variable region. This unit consists of a short leader peptide (L) exon, an intron, the completed VDJ segment, a second intron and finally a series of C exons (figure 1.6)

### 1.8.3 Classical Immunoglobulin Gene Recombination Signal Sequences

Two closely related and highly conserved sequences within variable region germline DNA provide the key to V(D)J recombination specificity. Recombination signal sequences or RSSs flank each germline V, D and J gene segment such that a single RSS lies 5' of each V segment and 3' of every J region while each D gene segment is sandwiched between two such sequences. Each RSS contains a conserved palindromic heptamer (consensus CACAGTG) and a conserved nonamer sequence (consensus ACAAAAACC) separated by 12 or 23 intervening base pairs corresponding to a single or double turn in the DNA helix respectively (48) (figure 1.2). This suggests that protein interactions with these signal sequences may be DNA conformation dependent. The signal sequences associated with the  $V_H$  and  $J_H$  segments contain a 23bp spacer while those flanking the diversity gene segments are of a single turn (12bp) only. Recombination, catalysed largely by two proteins, or recombination-activating gene products (RAG-1 and RAG-2), leads not only to joining of coding sequences (a coding joint) but also the fusion of the two heptamer regions of each signal sequence to form a signal joint and circular excision product (figure 1.7). This process normally occurs between gene segment RSSs of differing lengths (a phenomenon known as the 12/23 rule) (49) and is actively targeted by the RSSs even when the two signal sites are separated by distances in excess of a megabase (50). It is possible to examine the biochemistry of such events by generation of artificial recombination substrates (section 1.8.5).

**Figure 1.7**

**Cleavage of a V(D)J recombination signal by RAG proteins.** Lines represent single strands of DNA while triangles are recombination signal sequences. Note that these are only the initial steps in V-DJ or D-J recombination. Full recombination requires the synapsis of two such recombination signal sequences. See figure 5.3 for more specific details of *in vitro* cleavage.



#### **1.8.4 Components of the Classical Immunoglobulin Gene RSS.**

Mutational analysis of the 12bp-spacer and 23bp-spacer RSSs in human and mouse has identified crucial sequence elements. These analyses involved the testing of plasmid and retroviral substrates containing mutated RSSs in pre-B-cell lines actively undergoing V(D)J recombination (51-53). It is important to distinguish these studies from previous work, which examined the frequency of particular RSSs found in germline loci. While the earlier studies enabled determination of a consensus RSS, they did not directly determine the ability of such sequences to recombine (52). The most crucial nucleotides in the heptamer and nonamer sequences are shown in table 1.1. In general the heptamer is found to be the most important region, in particular the first three nucleotides (typically CAC). This work has demonstrated that a diverse range of signal sequences can be utilised, often deviating significantly from the consensus, and that these may be relevant to biased gene segment usage (54-56).

#### **1.8.5 Molecular Machinery Involved in V(D)J Recombination**

The signal sequences described previously act as targets to direct a number of proteins to carry out the complex biochemical events involved in V(D)J recombination. The key proteins in this process are described below:

##### **(a) The RAG Proteins**

The RAG proteins were identified through the design of a retroviral construct, capable of conferring antibiotic resistance on a cell only after recombination of V and J gene segments engineered into the viral genome. Transfection of cells containing this construct with random fragments of pre-B-cell DNA led, through restoration of antibiotic resistance, to the isolation of two such genes, recombination-activating gene 1 and 2 (*rag-1* and *rag-2*) (57, 58). During lymphoid development the expression of the RAG-1 and RAG-2 proteins parallels the process of gene rearrangement. RAG-1 and RAG-2, like the adaptive immune system they assist in generating, are found only in jawed vertebrates possessing a thymus (an evolutionary group known as the Gnathostomata). In all species examined the *rag* genes are closely linked on the chromosome and are convergently transcribed (that is to say transcription of the *rag-1* gene takes place 5'-3' on one strand, while transcription of the *rag-2* gene occurs 5'-



**Table 1.1 Crucial residues within human recombination signal sequences, (From (52, 53)).** Note: Recombination levels shown are for 12-spacer RSSs, however similar results were obtained for 23-spacer sequences. N is any other nucleotide

Substitution(s) from consensus sequence																Recombination frequency (100 is unmutated level)
Heptamer							Nonamer									
C	A	C	A	G	T	G	A	C	A	A	A	A	A	C	C	
N																<2
	N															<2
		N														<3
			C													6
			T													29-33
			G													77
				C												6
				A												85
					A											40
					C											87
						C										52
						A										74
				A	C	A										26
T	G	G	C	G	A	T										<1
							G									120
								G								61
								T								13
								A								3
									G							44
										G						27
											C					10
											T					87
											G					80
												G				25
													G			14-17
							T	G	T	C	T	C	T	G	A	5

3' on the opposite strand). The two proteins possess a transposase-like ability that allows *in vitro* excision and insertion of RSS-containing DNA into unrelated target DNA (59, 60), an ability that not only implicates the proteins in immune diversity generation, but may also have been significant in evolution of the immune system and the generation of subsets of V gene segments within the heavy chain locus. Indeed the RAG-1 protein shows sequence homology to bacteriophage lambda integrase (61). It is thought that the RAG proteins originated as a component of a retrovirus capable of integrating into the host genome as a transposable element (59). This theory is strengthened by the similarity of the RAG protein mode of action to that of HIV integrase.

RAG-1 and RAG-2 act in concert to introduce a double-stranded break (DSB) at the interface between RSS and coding DNA (57). These DSBs have been detected in cells actively undergoing V(D)J recombination (62-64). This two-step process begins with the introduction of a nick at the 5' end of the heptamer element, followed by nucleophilic attack on a phosphodiester of the opposite strand by the ensuing 3' hydroxyl (65) (figure 1.7). The initial nicking event is thought to play a significant role in generation of junctional diversity (66). This reaction generates two products, a signal end terminating in a blunt 5' phosphorylated DSB (67) and a coding end terminating in a DNA hairpin via a transesterification mechanism similar to that used in strand transfer by HIV integrase (63, 67-69). The remainder of the recombination reaction is less well understood but is believed to involve not only RAG proteins, but also a number of proteins previously identified as members of the double stranded break repair protein family (70).

- *In Vitro* Studies of Murine RAG Protein Activity

*In vitro* studies have assisted in the definition of the combined function of RAG-1 and RAG-2. However, the insolubility of full-length RAG-1 initially hindered development of an *in vitro* model for V(D)J recombination. Deletion analysis eventually led to the isolation of smaller, truncated, and therefore more soluble RAG protein derivatives or 'core' proteins (71). It was shown that purified, truncated RAG-1 and RAG-2 proteins are able to specifically cleave single RSSs present on an artificial oligonucleotide substrate (65). The proteins form a protein-DNA complex (or stable cleavage complex (SCC)) on this single RSS (72). The extent of RAG-

1/RAG-2 activity *in vitro* is dependent on the divalent cation available for cleavage such that full cleavage and generation of hairpins requires the presence of  $Mn^{2+}$  when single RSS oligonucleotides are processed, while  $Mg^{2+}$  is needed for complete processing of a paired 12/23 signal complex (72). Divalent cations associate with the RAG-1 and RAG-2 proteins during formation of the SCC. In the absence of RAG-2 the RAG-1 protein demonstrates poor binding specificity for RSSs (73, 74) although there appears to be some interaction with the nonamer (74-77). In addition, RAG-1 is known to make direct RAG-2-dependent contact with recombination signal sequences at the conserved heptamer (78). The coding sequence that flanks recombination signal sequences is also thought to affect RSS mediated recombination *in vivo* (79-82).

Recombinant purified RAG proteins show sequence specificity similar to that identified by *in vivo* mutational RSS analysis (83) (table 1.1). The first three bases of the heptamer are thought particularly crucial to *in vitro* cleavage with these purified RAG proteins. By contrast, substrates lacking a nonamer motif have been found to be competent for cleavage and full recombination (although the level of recombination is significantly reduced).

- Structure of the *Rag* Genes

The two RAG genes are unusual in that they are closely linked within the genomes of all species examined (57, 84). Approximately 7kb separate the RAG-1 and RAG-2 genes of the mouse. The two genes are also convergently transcribed in all species studied. The coding sequences of both human and murine RAG-1/-2 proteins are, unusually, encoded by single exons.

- Structure and Function of the RAG Proteins<sup>1</sup>

Several functional domains and essential 'core' regions required for recombinase activity have been identified within the RAG proteins by mutagenesis studies. The approximate position of such domains within the RAG-1 protein is indicated in figure 1.8. Most *in vitro* studies of the RAG-1 protein have been performed with a biologically active but more soluble truncated form lacking residues 1-383 and 1009-

---

<sup>1</sup> Numbering in this section is with respect to the murine RAG-1 amino acid sequence






**Figure 1.8**

**Major Domains of the Human RAG-1 Protein.** The protein is shown N- to C-terminal (left to right) and the position of the various domains shown in different colours. Note that positions of domains are only approximate



1

1040

-  Regions 'dispensable' during *in vitro* recombination
-  Nonamer binding domain
-  Region with homology to prokaryotic integrase
-  'Core' region essential to *in vitro* V(D)J recombination
-  RAG-2 interacting domain

1040. However it is thought that the amino terminal region may be of importance to catalytic activity *in vivo* (85, 86). A zinc-binding domain within the core protein is believed important to RAG-1 dimerisation (87-89) (figure 1.8). The central region of RAG-1 contains an evolutionarily conserved domain recognising the nonamer sequence of the RSS (75, 77). This domain anchors the RAG-1/RAG-2 complex to the RSS allowing subsequent cleavage of DNA. There is also a second conserved domain within RAG-1 related to a prokaryotic integrase as described above. In addition a broad RAG-2 interacting domain has been located between amino acids 504-1008 (87). Although the RAG-2 protein shows considerable evolutionary conservation along its full length the carboxyl-terminal region (at least 145 residues) is apparently dispensable so that the core active region is located within the first three-quarters of the protein (90).

In addition to the identification of broad functional domains, mutational studies, combined with the *in vitro* assays described previously have identified a number of sites crucial to murine RAG-1 function. Certain RAG-1 mutations cause sensitivity to changes in the coding sequence flanking the RSS, indirectly suggesting a role for RAG-1 in heptamer interaction. Of particular note, two mutant RAG-1 proteins have been described (91, 92) which show greater (by comparison to wild-type RAG-1) sensitivity to coding flank sequence adjacent to the heptamer of the RSS than the wild-type RAG-1. More recently several groups have undertaken more thorough examinations of potentially significant RAG-1 mutations (93, 94) and uncovered a number of acidic residues that are thought crucial to RAG-1 catalytic activity (95).

#### (b) Additional Processes Complementing V(D)J Recombination

Although V(D)J recombination alone provides a high level of variable domain diversity during B-cell development, additional processes taking place at coding ends during V(D)J recombination add to this diversity. These are described in the following sections.

- N Nucleotide Addition

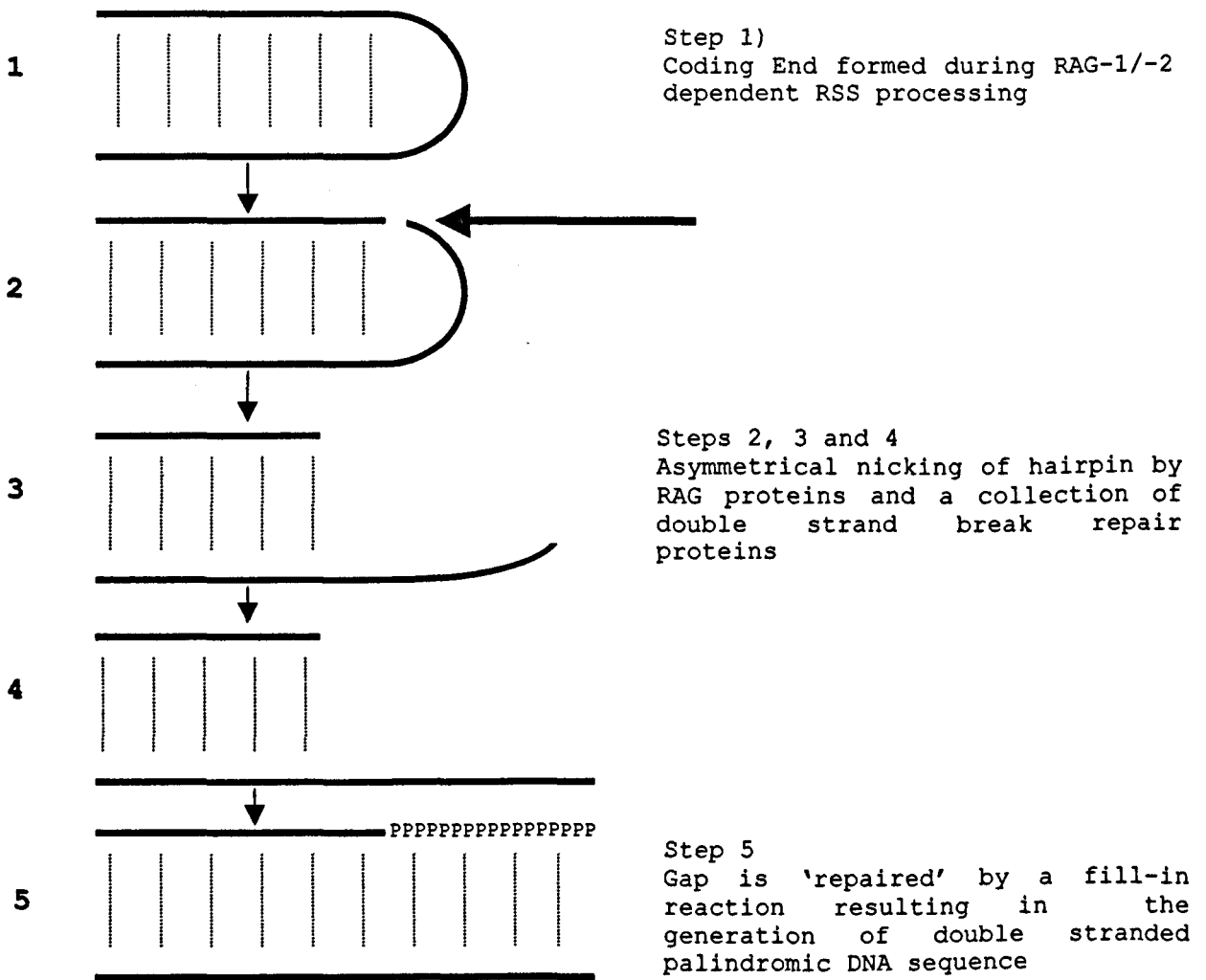
Non-germline (or N) nucleotide addition is the addition of random nucleotides to the 3' ends of coding ends during V(D)J recombination. The enzyme terminal deoxyribonucleotidyl transferase (TdT) mediates the random integration of a few

(typically less than ten) nucleotides (N diversity) to coding ends during V-(D)J and D-J joining thereby contributing to junctional diversity within the CDR3. TdT catalyses the addition of nucleotides onto the 3' end of DNA strands, preferentially adding dG residues. Lymphocytes with defects in their TdT genes produce rearranged variable exons containing virtually no N additions (96, 97), consequently reducing immunoglobulin variable domain diversity. There is limited evidence that N nucleotide addition and recombination site choice is linked (98).

- P Nucleotide Addition

P (or palindromic) nucleotides are added at the undeleted ends of V, D and J coding sequences during V(D)J recombination and as such contribute to the CDR3 sequence. These additional nucleotides are so named, as they are palindromic to the end nucleotides of the coding sequence (99). P nucleotide additions are typically only 1-2 nucleotides in length although longer (up to 15 bases) P regions have been identified in severe combined immune deficient (or *scid*, DNA-dependent protein kinase mutant) mice. Palindromic nucleotide addition is believed to result from the resolution of RAG-generated hairpins containing palindromic nucleotides derived from the two complementary strands of the coding end (figure 1.9). Completion of recombination requires the resolution of such hairpins, presumed to be the result of single or double-stranded nicking of the hairpin. Nicking of the hairpin at a position other than that at which initial RAG-mediated nicking takes place results in the insertion of palindromic nucleotides (66) (figure 1.9). This coding end processing is conducted in a coding end-specific manner such that sequence motifs within the coding end are thought to influence the degree and nature of P-region formation (100).

**Figure 1.9 The Process of Palindromic Nucleotide Addition During Coding Flank Hairpin Resolution.** Initial hairpin formation by RAG proteins must be resolved to bring DJ and V-DJ sequences together. If hairpins are nicked assymmetrically, as shown, DNA repair machinery will fill in single stranded DNA with palindromic sequence derived from the hairpin. A nick is introduced (Step 2) and the single stranded hairpin DNA opened out (Step 3). Once open the single stranded region is 'filled in' with complementary nucleotides (Step 4 and 5).



### **1.8.6 Aberrant V(D)J Recombination during Classical Antibody Generation**

Although it is clear that the majority of rearranged, functional immunoglobulin sequences can be generated by the mechanisms described above, it has been hypothesised that these mechanisms alone cannot account for the entire antibody sequence repertoire. The presence of extended germline  $D_H$  genes (or DIR genes (36)) containing irregular spacer signals interspersed between functional D gene segments has led to the suggestion that such elements, being flanked by multiple 12bp and 23bp spacer RSS may also be incorporated into the CDR3-derived H3 loop of the variable domain (36, 101) during some immunoglobulin rearrangements. It is possible that these, and conventional D regions may also occasionally use their 5' RSS in J segment joining, a process that would result in the presence of an inverted D segment within the CDR3 (102, 103). Another suggested mechanism is D-D recombination, between two 23-spacer RSSs, therefore flouting Tonegawa's 12/23 rule (104-106). Despite PCR-based evidence that inversions and D-D recombination do occur at very low frequency (103, 107), an analysis of a database of approximately nine hundred rearranged human sequences, made in an attempt to identify the presence of DIR segments, D-D joining and so forth within recombined V(D)J units has yielded little evidence that the expressed human antibody repertoire actively utilises such mechanisms (34).

## **1.9 Other Processes Involved in Classical Immunoglobulin Generation**

### **1.9.1 Somatic Hypermutation within the Classical IgG Genes**

V(D)J recombination is an ordered process introducing diversity to a discrete region of the variable domain sequence during a particular period of B-cell development. By contrast somatic hypermutation may affect the sequence encoding the entire variable domain and may occur throughout the lifetime of the B-cell. Somatic hypermutation leads to changes in DNA sequence within the rearranged V(D)J gene of the B-cell heavy chain locus. Hypermutation is restricted to approximately 1kb of sequence encoding the V, D and J segments. The C genes remain unmutated and there appears to be a sharp 5' boundary of mutation, probably within the leader intron. Within these thousand or so bases there are large variations in the frequency at which mutations are found. The CDR1 is particularly prone to hypermutation (108, 109). Mutational



'hotspots', presumed to be the result of peculiarities of DNA structure such as repeated sequences (110-112), palindromes and particular motifs such as TAA, RGYW (R=A or G, Y=C or T, W= A or T) (112), CAGCT/A and AAGTT (113) have also been described.

The process occurs primarily within B-cells located in the germinal centres of the lymphoid tissues (114, 115). Both the molecular mechanism of somatic hypermutation and the manner in which mutation is targeted to the rearranged V(D)J gene remain obscure.

Hypermutation is responsible for affinity maturation, whereby the average affinity of antibodies produced by an individual increases with the length of time during which the individual is exposed to the antigen. The rate of point mutation occurring within the V regions of immunoglobulins prior to antigenic stimulation is estimated to be between  $10^3$  and  $10^4$  times greater than the rate of spontaneous mutation elsewhere in the genome (116-118). It is estimated that for every mutation that improves antigen binding, three or four neutral mutations also occur.

### **1.9.2 The Process of Classical IgG Post-Transcriptional Modification**

Generation of heavy chain antibodies not only relies on specific mechanisms of recombination such as those described in section 1.8. The expression of immunoglobulin heavy chains also relies on conventional processes that are required prior to the expression of any eukaryotic gene product. Subsequent to V(D)J recombination and any isotype switching events (section 1.11), both of which occur at the DNA level, transcription of pre-mRNA containing completed V(D)J coding elements and the exons containing constant region domains takes place. Before translation of the heavy chain polypeptide can commence a number of post-transcriptional modifications of the pre-mRNA must take place. These include 5' capping of the RNA, 3' polyadenylation of the transcript and splicing of the intronic RNA that separates the various domains of the heavy chain. This latter process provides an additional level of possible alteration of the final gene product that may be significant to the generation of the immunoglobulin (section 1.9.3-1.9.4).

### **1.9.3 The Role of RNA Splicing in Classical IgG Generation.**

The pre-mRNA of most immunoglobulin isotypes contains four introns (non-coding RNA) which must be removed before successful translation of the sequence can be achieved. The removal of introns from the primary transcript of such a discontinuous gene occurs through splicing. These mRNA splicing events are each the result of a two step mechanism whereby specific recognition sites, initially at the 5' and later the 3' end of each intron are recognised by small nuclear ribonucleoprotein particles (snRNPs) to form a spliceosome (119, 120) (figure 1.10). Cleavage at the 5' recognition site (or splice donor) is followed by covalent linkage of the free 5' end of the intron to an intronic adenosine residue typically 20-450 nucleotides upstream of the 3' recognition site (or splice acceptor). This results in the formation of a loop (or lariat) of intronic RNA followed by cleavage at the 3' acceptor site and ligation of the two exons. A two step process is believed to be involved in exon/intron discrimination. The first step, 'exon definition', involves splice site recognition, while the second step involves binding of accessory splicing factors to additional enhancer sequences, usually located in or in the vicinity of exons (121-123). The mechanisms by which donor and acceptor splice sites are selected are not fully understood and although splice site consensi have been established there are no hard and fast rules governing usage of one splice site over any other (124, 125). The significance of splice site preference and alternative splicing are discussed in more detail in section 1.12.1

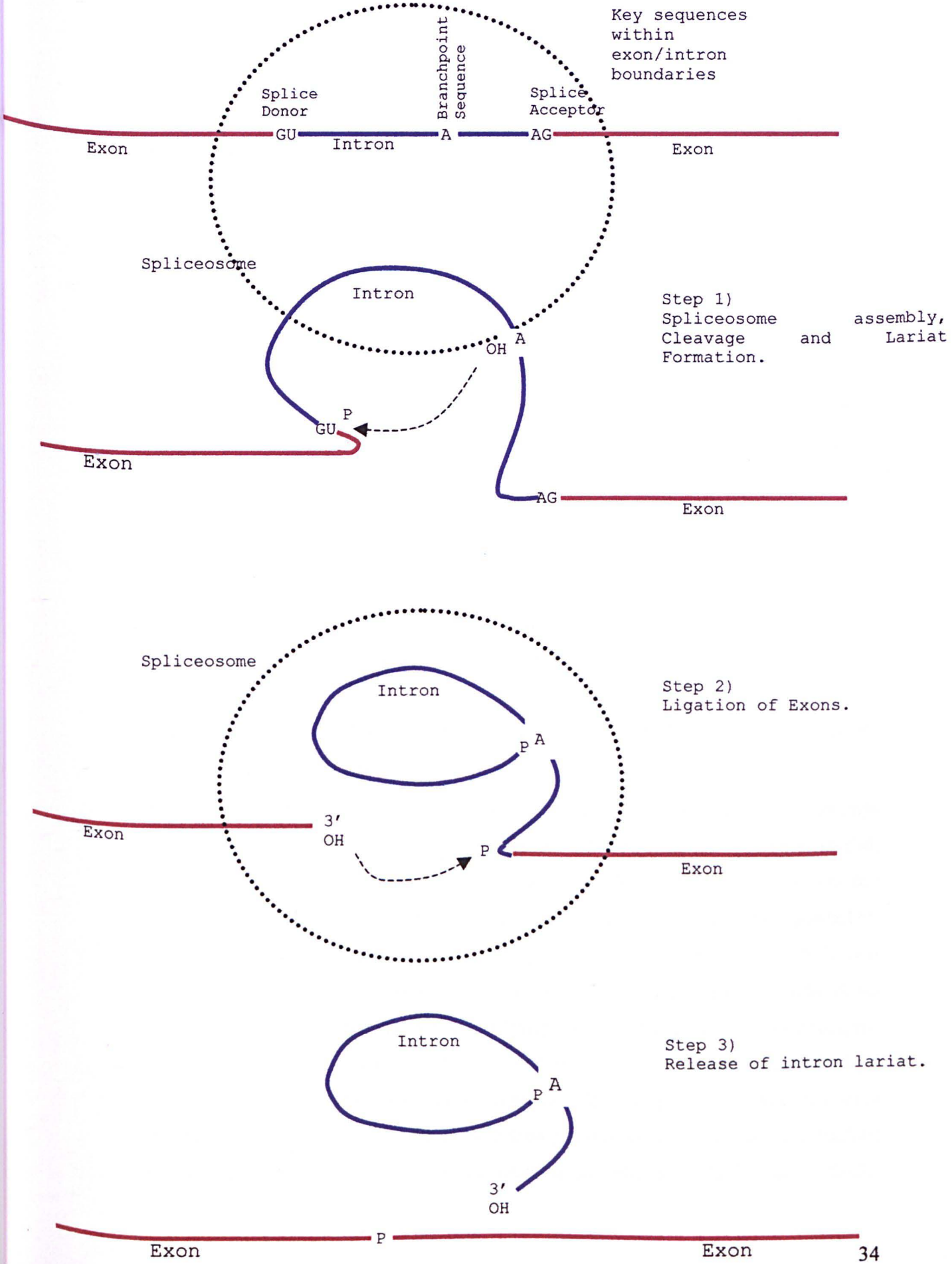
### **1.9.4 Specific Sequences Involved in Splice Site Selection**

Each sequence element involved in the splicing reaction is thought to play multiple roles. The relative positions of these sequences are indicated in figure 1.10

#### **(a) Splice Donor**

The 5' end of the intron to be spliced contains the strongly conserved AGGUAAGU sequence, where the internal GU represents the beginning of the intron. The U1 snRNP initially binds to the 5' splice site through intermolecular basepairing with the consensus sequence (126-128). Later in the process the U1 snRNP is replaced by the U6 snRNP (129).

**Figure 1.10**  
**Key features involved in Nuclear Pre-mRNA Splice Processing.** The biochemical steps involved in removal of introns prior to translation are shown. During generation of immunoglobulin heavy chains splicing of introns lying between the VDJ unit and C<sub>H</sub>1, between C<sub>H</sub>1 and hinge, hinge and C<sub>H</sub>2 and between C<sub>H</sub>2 and C<sub>H</sub>3 must be removed.



#### (b) Branch Point Sequence

The branch point sequence contains the binding site for the U2 snRNP and lies close to the polypyrimidine tract upstream of the 3' acceptor splice site (typically 18-40 nucleotides upstream of the 3' splice junction) (130-132). Recognition of the branch point is determined not only by the presence of the adenosine residue and surrounding consensus (UACU AAC in yeast) but also by the composition of the polypyrimidine tract.

#### (c) Splice Acceptor

The 3' end of each intron comprises a polypyrimidine tract (typically of 12bp length) followed by a pyrimidine residue directly preceding the AG dinucleotide at which spliceosome catalysed cleavage occurs (133, 134). Splice site cleavage generally occurs at the first AG downstream of the branch point (135, 136).

#### (d) Exon Sequences

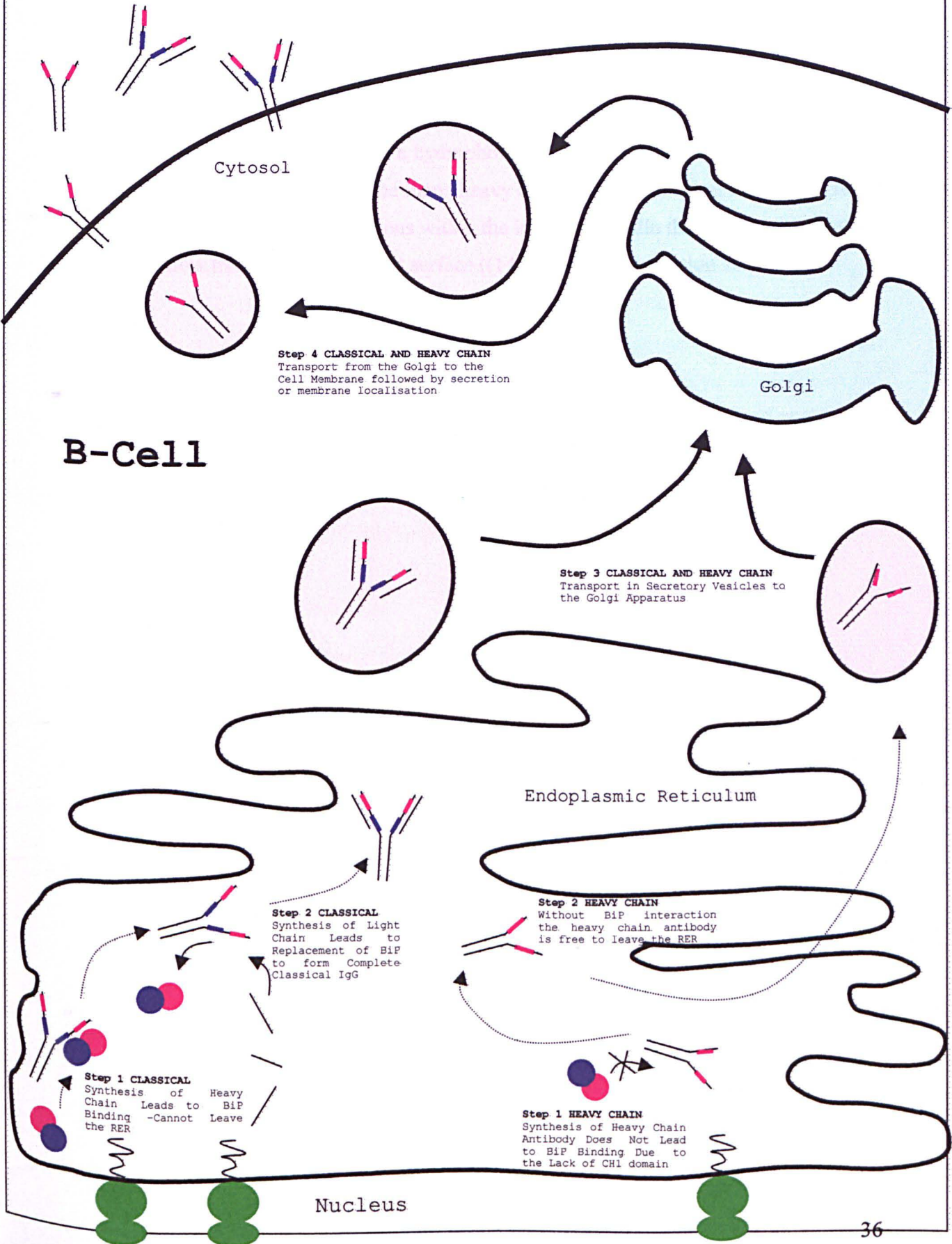
*In vitro* splicing studies using pre-mRNA substrates containing competing 5' and 3' splice sites have revealed the importance of exon sequences in splice site selection. Of these, the most studied are purine-rich exon enhancers, most of which are found in alternative exons and operate by activating weak 3' acceptor splice sites (121).

### 1.10 Classical Antibody Secretion and the Role of the Heavy Chain in B-Cell Development.

Even after successful translation of the spliced immunoglobulin heavy chain transcript further processes within the B-cell regulate the secretion of correctly folded, functional antibody. Functional IgG secretion from B-cells generally requires the association of light and heavy chains, a requirement that is believed to be regulated by a member of the heat shock protein 70 (Hsp70) molecular chaperone family known simply as *binding protein* (BiP) (137). In the absence of functional light chain synthesis BiP binds to the heavy chain polypeptide preventing the pre-B-cell receptor (pre-BCR) complex (consisting of functional heavy chains coupled to pseudo-light chains) from leaving the endoplasmic reticulum (ER) (figure 1.11). BiP has been shown to associate with immunoglobulin heavy and light chains during their folding and assembly (138-142) but the exact location and nature of the immunoglobulin

**Figure 1.11**

Simplified diagram illustrating the presumed pathways of both classical and heavy chain antibody production and secretion. Ribosomes from which heavy chain polypeptides are synthesised are shown in green. Variable domain are indicated in pink and CH1 domains in blue. Antibodies pass from rough endoplasmic reticulum (Steps 1 and 2) via vesicles to the golgi apparatus and then the cell surface where they may be secreted or remain within the cell membrane.



sequences recognised by BiP is not well characterised. Indirect evidence suggests that, while the C<sub>H</sub>1 domain is required for BiP association (139), mutant heavy chains lacking C<sub>H</sub>2 or C<sub>H</sub>3 domains remain within the ER. Studies have also shown that association with BiP is also dependent on sequences within the V<sub>H</sub> domain (143). *In vitro* studies have identified a possible heptameric BiP binding motif of HyXH<sub>2</sub>HyXH<sub>2</sub>HyXH<sub>2</sub> where Hy is a hydrophobic or bulky aromatic amino acid (144). It is thought that the ability of the llama heavy chain antibody to leave the ER is a result of the absence of specific regions within the immunoglobulin that prevent BiP binding and allow transport to the B-cell surface ((145), figure 1.11, section 7.6).

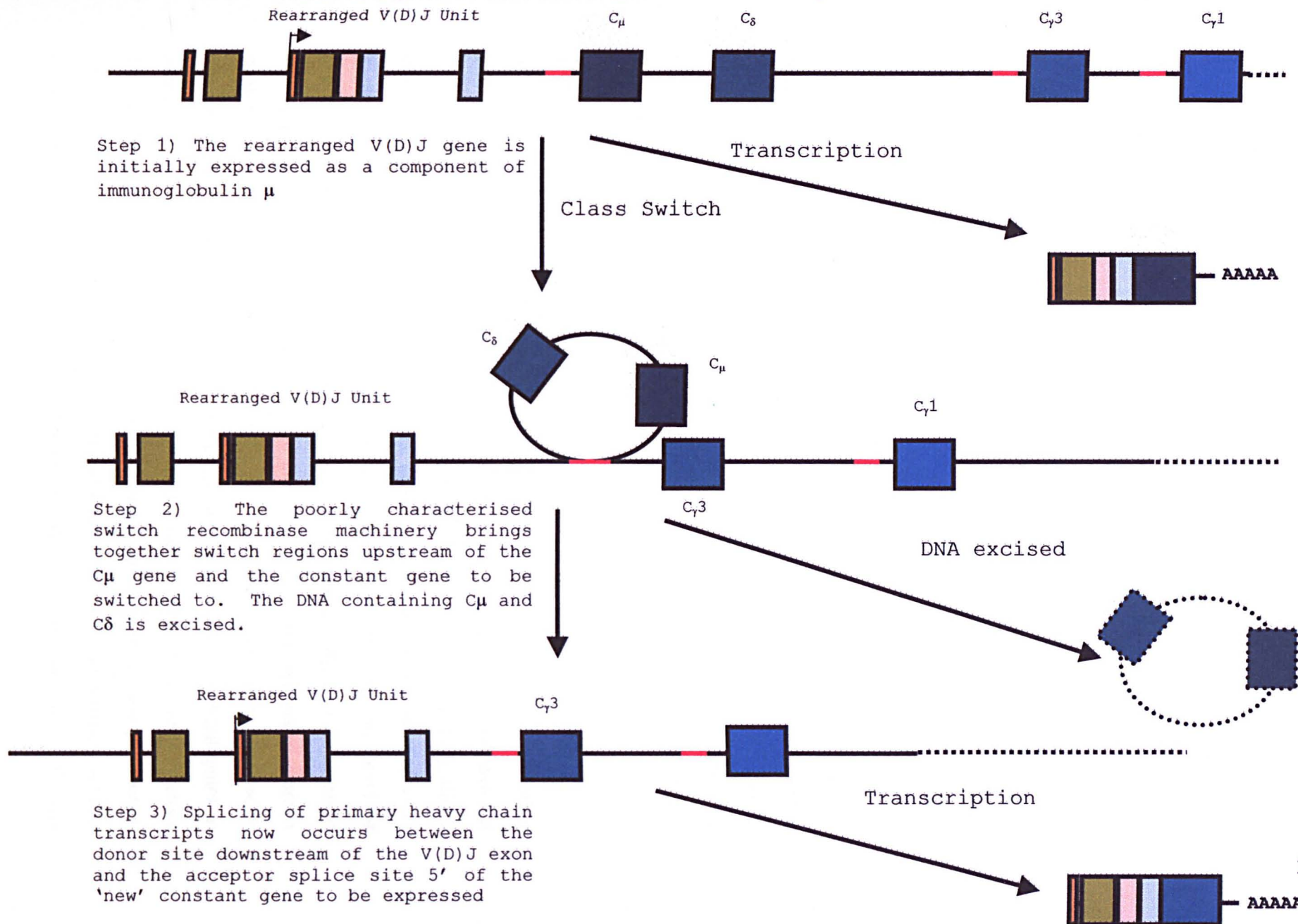
### **1.11 The Process of Isotype Switching.**

While the processes described thus far would allow functional antibodies to successfully reach the B-cell cell surface, regulation of immunoglobulin expression does not end as the B-cell reaches maturity. Indeed, another novel somatic process, isotype switching, can vary the properties of the immunoglobulin considerably.

The initial assembly of light and heavy immunoglobulin chains during B-cell development leads to the expression of an IgM molecule on the cell surface. B-cells then migrate to encounter antigen within the peripheral lymphoid organs leading to proliferation of particular B-cells and differentiation into plasma cells that actively secrete antibody. Activated B-cells are able to secrete not only IgM but also other isotypes (such as classical IgG) by changing the constant regions of the expressed heavy chain, a process known as isotype, or class switching (figure 1.12). The production of various heavy chain isotypes directs the immune system along different functional pathways. The production of heavy chain antibody of the llama may well represent a separate functional pathway available to the camelid immune system, and therefore the possibility that isotype switching plays a role in heavy chain antibody generation must be considered.

Isotype switching was first observed when the different immunoglobulin classes within serum were characterised after antigenic stimulation (146). During a primary response IgM levels are found to drastically increase, while secondary and subsequent responses are characterised by an increase in IgG secretion and reduction in IgM generation. Experiments have shown that isotype switching occurs at the antigen-dependent stage of B-cell differentiation and that a single clone of B-cells is able to switch to multiple isotypes (147) via the process of switch recombination. Furthermore, it has been shown that V-region specificity is retained during the switching process (148).

Figure 1.12 Diagram depicting events during isotype switching from initial IgM to an IgG isotype. Switch (S) regions are shown as red bars.



Step 1) The rearranged V(D)J gene is initially expressed as a component of immunoglobulin  $\mu$

Class Switch

Transcription

AAAAA

Step 2) The poorly characterised switch recombinase machinery brings together switch regions upstream of the C $\mu$  gene and the constant gene to be switched to. The DNA containing C $\mu$  and C $\delta$  is excised.

DNA excised

Step 3) Splicing of primary heavy chain transcripts now occurs between the donor site downstream of the V(D)J exon and the acceptor splice site 5' of the 'new' constant gene to be expressed

Transcription

AAAAA



Switch recombination takes place within regions composed of tandem repetitive sequences known as switch or S regions located 5' of each C<sub>H</sub> gene (other than C<sub>δ</sub>) (149-152) rather than at specific sites (153, 154). However, isotype specificity of recombination is not determined by the nucleotide sequences of the S regions themselves (155). These switch regions typically vary from between 1-10kb in length.

## 1.12 Evolution of Classical IgG

Having described both the organisation of the heavy chain immunoglobulin locus and the mechanisms of classical immunoglobulin generation in detail attention now turns to the evolutionary development of the immunoglobulin gene locus and the immunoglobulin molecule. Similar evolutionary events may well have led to the generation of the heavy chain antibody within the llama.

Members of the immunoglobulin superfamily have been characterised in organisms as primitive as the Porifera (Sponges) (156). While V domains conferring specificity to immunoglobulin-like proteins are present within insects (157) and snails (158) there are no examples of immunoglobulin-type gene rearrangement found below the Gnathostomata.

### 1.12.1 Constant regions

The immunoglobulin heavy chain constant genes are not well conserved in evolution (159). A number of exceptions to the typical constant region gene structure have led to speculation that the evolution of heavy chain genes may have included mutations that created or destroyed RNA splice sites, converting exons into introns and vice-versa (42, 160). A well-characterised example of this is the startling level of similarity between the intron 5' of the mouse  $\gamma 2b$  hinge exon and the C<sub>H1</sub> sequence of the same locus (41). This suggests that the hinge may have originated from a complete Ig domain in which a form of splice site modification took place. The use of alternative splice sites (section 1.9.3) is a common feature of IgM constant genes in many species. C<sub>μ</sub> genes typically consist of exons encoding membrane and secretory IgM variants, the processing of which is regulated by varying splice site preference. In teleost fish splicing is responsible for the absence of a complete constant domain (C<sub>H4</sub>) in the membrane form (161) and holostean fish contain motifs within their C<sub>H4</sub>

domain that may act as cryptic splice donor sites (162). These examples provide precedents for the use of splicing as a mechanism for varying Fc composition. Splice site mutation, rather than alternative splicing results in hinge deletion within the  $C\alpha$  gene locus of the pig (163).

### **1.12.2 Variable Regions**

There is considerable conservation of framework region sequences encoded by V gene segments between species. In addition all Gnathostomata variable domains contain conserved canonical CDR1 nucleotides. While the immunoglobulin variable domains of all species are assembled by joining of germline V, D and J gene segments, the number of V, D and J regions present within the germline vary considerably between species ranging from 700  $V_H$  elements in the turtle (164) to only four associated with the nurse shark antigen receptor (section 1.15.3) (165).

### **1.13 V Gene Assembly in Other Species**

Although the overall mechanism by which variable domains are generated does not vary greatly between species, the layout of the gene elements involved and the relative usage of each element vary considerably. Although mammals have very similar heavy chain gene layouts differences exist in the manner in which these genes are utilised. For example rabbits and chickens have a single  $V_H$  family of which only a few (and sometimes only a single gene segment) are functional. In the chicken V(D)J recombination occurs at a single V gene segment. Somatic gene conversion events subsequently lead to the formation of a unique rearranged gene by recombination with a pool of otherwise non-viable pseudogenes (166). This rearranged gene then encodes all expressed immunoglobulin heavy chains. Although the chicken demonstrates no combinatorial joining diversity it is still able to mount a highly diverse immune response due to multiple rounds of this gene conversion within different regions of the V gene segment. In the rabbit the most D-proximal  $V_H$  gene segment is preferentially rearranged and accounts for the majority of heavy chains expressed. Variability is brought about largely as a result of gene conversion events between this and other V gene segments (167). Another unconventional example of variable gene usage occurs within the heavy chain locus of the shark which consists of

many gene clusters (typically of ~10kb in length) each containing V, D, J and C elements. Rearrangement occurs exclusively within a single cluster (168).

### **1.14 *Lama glama***

The full understanding of any unique biological mechanism or phenomenon such as the generation of heavy chain antibody requires an appreciation of the physiological context in which the phenomenon has developed. It is important therefore to provide an introduction to the species in which heavy chain antibodies have developed.

The llama (*Lama glama*) is a long-haired ruminant native to South America. Although principally domesticated as a beast of burden the llama has more recently been farmed for its fur and ability to guard sheep against predators.

#### **1.14.1 Evolution of the family *Camelidae***

The llama is a member of the camelid family, which comprises four South American llama-like species: llama, alpaca, guanaco and vicuna, in addition to Asian dromedaries and bactrian camels. The *Camelidae* family is a ruminating member of the suborder Tylopoda or cud-chewing Artiodactyls (169). The order Artiodactyl includes hooved mammals such as the pig, giraffe and hippopotamus (none of which are known to express heavy chain antibodies). Camelids are thought to have originated in harsh glacial conditions in North America approximately 9-11 million years ago. Although they gradually migrated apart, llamas and camels were not completely separated until the beginning of the glacial Pleistocene period (or epoch) approximately 1.6 million years ago when llamas migrated into South America and camels moved across the Bering Strait land bridge into Asia (169, 170).

Given that both Asian and South American camelids express heavy chain antibodies it is likely that the mechanism of their generation therefore evolved at least 1.6 million years ago (assuming that the two species did not develop the novel antibodies independently).

As a ruminating mammal the llama may have an immune system similar to that characterised in other ruminants such as the sheep. Initial sheep B-cell production is the role of gut-associated lymphoid tissue (GALT) where immunoglobulin rearrangement is thought to originate, in an antigen-independent manner, from a relatively small pool of variable gene segments in such mammals. Such

immunoglobulin generation may utilise different mechanisms from those described above for human and murine antibody production.

### **1.15 Heavy Chain Immunoglobulins**

The remainder of this introduction considers the current state of research into the properties of the heavy chain antibody and related proteins.

The heavy chain antibody that has evolved within the camelid family was discovered in 1993 when Hamers-Casterman and co-workers discovered protein A and protein G-binding molecules of around 100 kilodaltons within the serum of the camel (*Camelus dromedarius*) (1). The group went on to demonstrate the presence of heavy chain antibodies within all members of the camelid family including the llama (*Lama glama*) and to show that the novel immunoglobulins have an extensive antigen-binding repertoire.

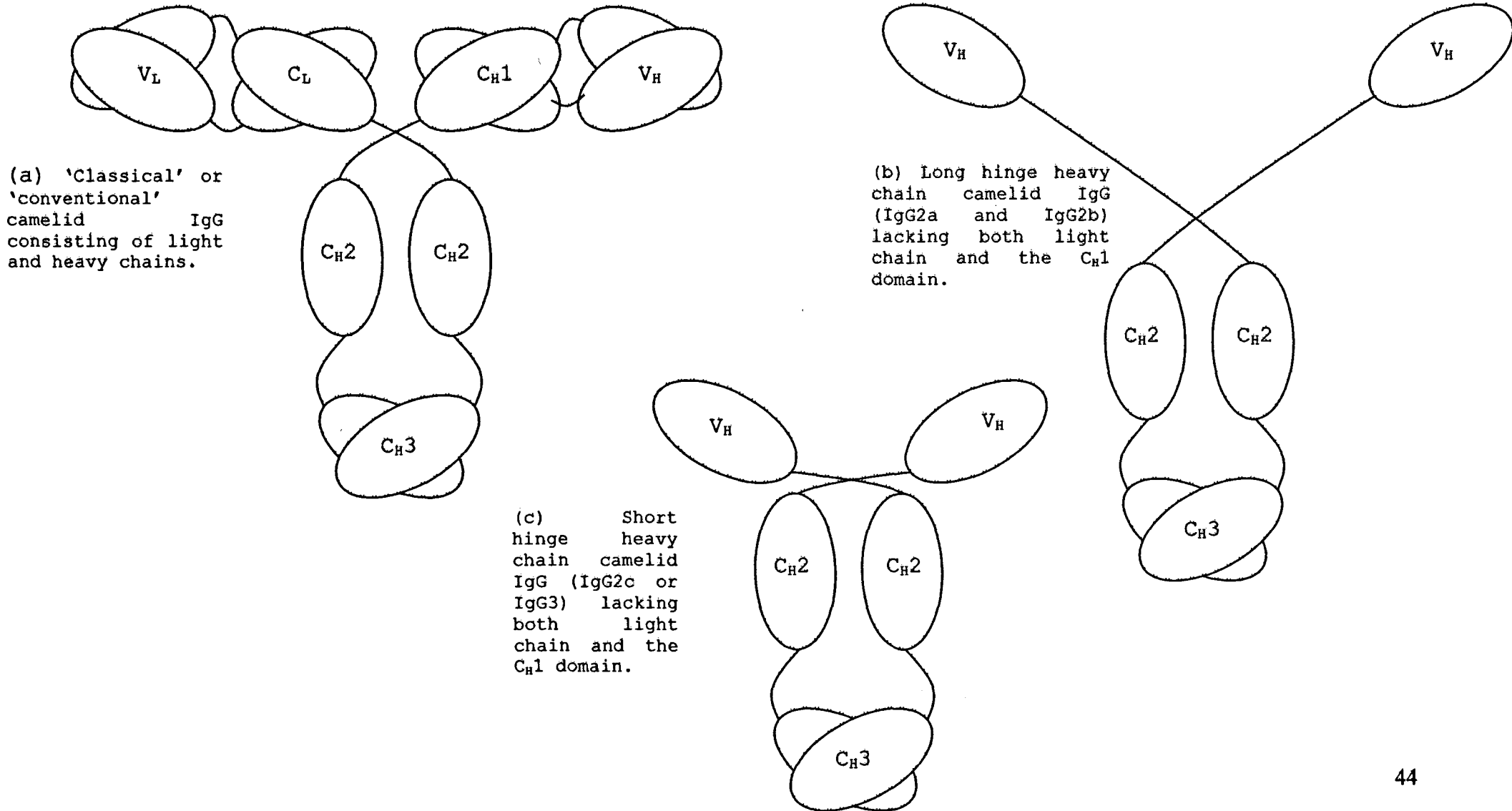
#### **1.15.1 Structure of the Heavy Chain Only Antibody Protein**

The two principle structural characteristics of camelid heavy chain antibodies that set them apart from classical IgG are the lack of associated light chains, so that the antibody unit consists only of a heavy chain dimer, and the complete absence of a C<sub>H</sub>1 domain (1, 145). Three types of heavy chain are classified according to their differing hinge lengths (section 1.15.2) (figure 1.13). Heavy chain immunoglobulin-derived variable domains (V<sub>HH</sub>) isolated from camelids have been found to contain relatively long CDR3 regions, by comparison to both camelid and human/murine antibodies (171). This extended CDR3 is believed to have a crucial effect on antibody/antigen interaction (2). The H3 loop encoded by the CDR3 has been shown through X-ray crystallography (172) to loop out of the immunoglobulin variable domain rather than form an antigen-binding cleft typical of conventional antibodies. This may lead to the inhibitory effects of the heavy chain antibody on enzymes (section 1.3 and (8)). The CDR3 of the camelid heavy chain antibody frequently contains a cysteine residue while a further cysteine is commonly found within the CDR1, second framework region (FR2) or CDR2. As a result of these additional cysteine residues heavy chain antibodies are thought to contain additional disulphide bridges that may stabilise the longer CDR3 loop.

The heavy chain immunoglobulin CDR1 also differs from that observed in conventional antibodies (172-174). Four of the seven highly conserved residues vital

**Figure 1.13 Basic structure of Camelid IgG isotypes (After (1)).**

Two major hinge types are found. Both heavy chain antibodies are characterised by the absence of the C<sub>H1</sub> domain



to the hydrophobic association of classical heavy chain variable domains with their corresponding light chains (at positions 37, 44, 45 and 47 by Kabat numbering) are changed to more hydrophilic residues. In addition camels (but not llamas) have a leucine rather than serine residue at position 11 (145). Serine11 is a residue that conventionally interacts with the C<sub>H1</sub> domain (2, 3). The camelid heavy chain immunoglobulin variable domain is referred to as the V<sub>HH</sub> (in order to distinguish the domain from the classical V<sub>H</sub>).

### 1.15.2 Subclasses of Camelid IgG

Three subclasses of camelid IgG have been defined, IgG1, IgG2 and IgG3. IgG1 comprises the conventional antibody of two light and two heavy chains. In llamas this class can be further subdivided into IgG1a and IgG1b corresponding to long hinge (19 residues) and short hinge (12 residues) variants respectively. The IgG2 class comprises two hinge variants of the heavy chain antibody form. IgG2a contains a long hinge of 29 amino acids while IgG2b has an even longer 35 residue hinge. IgG3 includes heavy chain antibodies with a short hinge of 12 amino acids (175). In the llama the IgG3 form is classified as IgG2c.

### 1.15.3 Other Heavy Chain Immunoglobulins

Antibodies lacking light chains have also been described in human serum. These occur during the lymphoproliferative heavy chain disease. The antibodies are not functional due to extensive deletions within the V<sub>H</sub> – hinge region (176). The deletion of this region prevents interaction with light chain. It is this deletion that leads to exclusion of the C<sub>H1</sub> domain and failure to associate with light chains. Human heavy chain antibodies do not include the amino acid differences found in camelid antibodies that allow binding of antigen in the absence of a light chain.

Since the discovery of camelid heavy chain antibodies further “antigen receptors” also composed of a single heavy chain dimer and thought to bind antigen in a similar manner have been characterised in the shark (known as the nurse shark antigen receptor (NAR)) (177, 178). In contrast to the camelid heavy chain antibody, such molecules are composed of a dimer of two single variable domains followed by five constant domains. The NAR is therefore the only other known example of a heavy chain only antigen-binding immunoglobulin-like molecule.

### **1.16 Scope of Thesis**

Through the study of the mechanisms involved in the generation of the heavy chain antibody in the llama it was hoped to discover not only the nature and context of the sequences responsible for encoding the novel antibody type, but also to dissect the mechanisms by which such antibodies are generated. If heavy chain antibodies are to fulfil their commercial potential in the future, an understanding of the mechanisms by which they are generated in the developing B-cell may provide important clues as to the optimum manner in which commercially and therapeutically relevant llama immunoglobulins can be produced.

### **1.17 Contents of Thesis**

This thesis describes firstly the isolation of a number of components of a llama heavy chain locus (or loci) responsible for encoding the variable domain of the antibody. (Chapter 3). Analysis of these sequences suggests the active utilisation of a number of levels of control during heavy chain immunoglobulin generation by the developing llama B-cell.

In subsequent sections attempts to reconstitute the process of llama V(D)J recombination *in vitro* are described. Firstly through the isolation and expression of components of the llama recombination machinery (Chapter 4) and secondly through the development of an assay system to test the ability of recombination proteins to interact with specific signal sequences derived from isolated germline variable gene data (Chapter 5).

Chapter 6 reports the isolation of a number of the constant region genes responsible for encoding both classical and heavy chain immunoglobulins in the llama. This chapter goes on to provide an explanation for crucial absence of the C<sub>H</sub>1 domain within heavy chain antibodies.

## **Chapter 2 Methods**

### **2.1 Overview**

The generation of a llama genomic library was crucial to the isolation of components of the llama immune system described in Chapters 3, 4 and 6. This chapter considers both the strategy and methodology underlying the generation of this library, baculovirus expression and recombination assays in addition to other techniques used throughout the thesis, such as polymerase chain reaction (PCR) and probe generation.

### **2.2 General Methods**

A number of methods were utilised repeatedly throughout this project. These methods include basic molecular biological techniques involving manipulation of DNA to generate probes and PCR products for cloning and characterisation of components of the llama immune system. This chapter begins with the description of these supporting protocols, the use of which is referred to not only in subsequent strategy-specific sections, but also in table 2.1. Where any significant deviations are made from the general methods described here, the nature of the deviation is described in the relevant section. Supporting protocols are followed in this chapter by description of library generation techniques (section 2.3), library screening protocols (section 2.4) and sequencing strategies (2.5). The description of sequence isolation techniques leads on to the discussion of recombination assay protocols (section 2.6), baculovirus expression techniques (section 2.7.1-2.7.9) and protein purification methods (section 2.7.10-2.7.11). The chapter concludes with a description of sequence analysis techniques used within the thesis (section 2.8). Reagents used in Chapter 2 were from Gibco-BRL, Paisley, UK unless otherwise stated

#### **2.2.1 Summary of Techniques Employed in this Thesis.**

Before describing the various techniques employed in this report, the steps involved in the molecular biological strategies used are summarised in table 1.1.



**Table 2.1 Summary of Techniques Employed in this Thesis.** Major Molecular Biology Strategies are shown in the left column while the techniques employed are described in columns to the right.

Method	Pfu-based PCR <sup>*</sup>	Taq-based PCR	Agarose Gel Visualisation	Agarose Gel Purification	Restriction Digestion	Ligation	Heat Shock Transformation	Mini/Midi Prep of DNA	Generation of Multiple Alignment <sup>^</sup>	Primers Used for Sequencing <sup>#</sup>
Library Generation	No	No	Yes Integrity of Purified DNA and Partial Digestion Integrity using 0.6% Agarose/TBE gel at 15V overnight	No	Yes Partial	Yes Into Bacteriophage Lambda DASH vector arms High Concentration Ligase (Roche Biochemicals, Lewis UK)	No	Preparation of DNA by PEG precipitation	No	No
Generation of Variable and Constant Region Probes	No	Yes	Yes 1% Agarose/TBE	Yes	No	No	No	No	No <sup>1</sup>	No
Generation of RAG-1 Probe	Yes	No	Yes 1% Agarose/TBE	Yes	No	Yes Blunt ended pBluescript cut with <i>Sma</i> I	Yes Chemically competent XL-1 Blue. Plated onto LB Tet/Amp plates	Yes	Yes	RAG-1 A+B
J region subcloning and sequencing	Yes	No	Yes 1% Agarose/TBE	Yes	No	No - Cloned using TOPO - Blunt Vector (Invitrogen, Carlsbad, US)	Yes - TOP 10 competent cells (Invitrogen). Plated onto LB Kan plates	Yes	No	J Sequencing 1+2
RAG-1 Sequence Generation	Yes	No	Yes 1% Agarose/TBE	Yes	No	Yes - Blunt pCR - Blunt	Yes TOP 10 competent cells. Plated onto LB Kan plates	Yes	Yes	RAG-1 SeqGen A+B
RAG-2 Sequence Generation	Yes	No	Yes 1% Agarose/TBE	Yes	No	No Cloned using TOPO - Blunt Vector	Yes TOP 10 competent cells. Plated onto LB Kan plates	Yes	Yes	RAG-2 A-F
Generation of truncated RAG-1 sequence for baculovirus expression	Yes	No	Yes 1% Agarose/TBE	Yes	Yes <i>Xba</i> I	Yes Rapid Ligation Kit (Roche Biochemicals, UK)	Yes into DH10Bac, plated onto selective plates	Yes - Midiprep prior to transfection with Cellfectin reagent	Yes - in order to identify points of truncation	Trunc RAG-1 A+B

<sup>\*</sup>PCR conditions are described in table 1.2 below

### 2.2.2 Polymerase Chain Reaction

The amplification of specific regions of DNA by polymerase chain reaction was a technique essential to the successful sequencing and cloning of isolated llama immune system components described in Chapters 3-6 of this thesis. All PCRs conducted in this work utilised either *Taq* or *Pfu* (Stratagene, California, USA) polymerase. Differences between these enzymes are shown in table 2.2:

Characteristic	<i>Taq</i> Polymerase	<i>Pfu</i> Polymerase
Fidelity	Low	High
Speed of DNA Synthesis	High	Low
Nature of PCR product	Poly-T ended	Blunt-ended
Cost	Low	High

Table 2.2 Summary of contrasting properties of *Taq* and *Pfu* thermostable DNA polymerases (Data derived from manufacturer literature)

#### (a) *Taq*-based PCR

The lower fidelity of *Taq* polymerase makes *Taq*-based PCR ideal for generating products where a degree of sequence degeneracy is beneficial (for example when generating a variable region probe, section 2.4.1). For the duration of this project *Taq*-based PCR was typically better able to generate specific PCR products than other commercially available enzymes such as *Pfu*. Sequences reported within this thesis derived from *Taq*-based PCR are the result of the generation of a consensus sequence from at least three separate PCR clones (in order to account for sequence errors resulting from the low fidelity of the *Taq* enzyme). *Taq*-based PCR was used in the experiments indicated in table 2.1. Conditions for PCR including annealing temperature and number of cycles are indicated in table 2.3.

*Taq* PCR reactions were set up in 500µl thin-walled microfuge tubes as described in table 2.3



Reagent	Volume
10x <i>Taq</i> Buffer (including MgCl <sub>2</sub> ) (Gibco-BRL, UK)	10µl
25mM dNTPs (25mM each of dATP, dGTP, dCTP and dTTP) (all Amersham-Pharmacia, UK)	2µl
10µM 5' primer (all synthesised by MWG-Biotech, Milton Keynes, UK)	5µl
10µM 3' primer (all synthesised by MWG-Biotech, Milton Keynes, UK)	5µl
<i>Taq</i> Polymerase (5U/µl) (Gibco-BRL, Paisley UK)	0.5µl
Template DNA (100ng/µl)	1µl
MilliQ deionised water to total volume	100µl

**Table 2.4 Components of a *Taq*-based PCR reaction**

Tubes were then placed in a heated lid 'Touchdown' thermocycler (Hybaid, Ashford, UK) and incubated at 94°C for three minutes to denature the template DNA. X cycles of PCR amplification were performed as follows:

Denature	94°C	45 seconds
Anneal*	Y°C	30 seconds
Synthesis	72°C	90 seconds

\*Consult table 2.3 for value of X and Y in particular experiment

A further 10 minute incubation at 72°C allows completion of PCR product synthesis. PCR products were generally visualised as necessary on a 1.0% agarose/TBE gel at 15V/cm on horizontal gel apparatus.

### (b) *Pfu*-based PCR

By contrast with *Taq* polymerase the fidelity of *Pfu* polymerase is extremely high (particularly over low numbers of cycles). *Pfu*-derived PCR product sequences of less than 1kb in length resulting from PCRs of 25 cycles and less are published within this thesis from single clones. Larger PCR products, or those resulting from greater numbers of amplification cycles are shown after confirmation of the sequence from two separate clones. *Pfu*-based PCR reactions proved less robust than *Taq*-based PCR throughout this report with non-specific PCR products a common problem. For this reason *Pfu*-based PCR was used when suitable (as indicated in table 2.1), but not throughout this study. *Pfu*-based PCR reactions were set up in 500µl thin-walled microfuge tubes as detailed in table 2.4.

Reagent	Volume
10x Pfu Polymerase Buffer (Stratagene, UK)	10µl
25mM dNTPs (25mM each of dATP, dGTP, dCTP and dTTP) (all Amersham-Pharmacia, UK)	0.8µl
10µM 5' primer (all synthesised by MWG-Biotech, Milton Keynes, UK)	5µl
10µM 3' primer (all synthesised by MWG-Biotech, Milton Keynes, UK)	5µl
Pfu Polymerase (5U/µl) (Stratagene, UK)	2µl
Template DNA (100ng/µl)	1µl
MilliQ deionised water to total volume	100µl

Table 2.5 Components of a Pfu-based PCR reaction

Tubes were then placed in a heated lid 'Touchdown' thermocycler and incubated at 94°C for three minutes to denature template DNA. X cycles of PCR amplification were performed as follows:

Denature	94°C	45 seconds
Anneal *	Y°C	30 seconds
Synthesis	72°C	600 seconds

\*Consult table 2.3 for value of X and Y in particular experiment

A further 10 minute incubation at 72°C was performed to allow completion of PCR product synthesis. PCR products were visualised, when necessary on a 1.0% agarose/TBE gel at 15V/cm using horizontal electrophoresis apparatus.

## 2.2.3 Other General Methods

### (a) Restriction Digest Preparation

Restriction digestion was used throughout this study to prepare both plasmid and bacteriophage vectors for acceptance of insert DNA and to prepare insert DNA when blunt or TA ligation (section 2.2.4 (i)) and cloning techniques were not utilised (more specifically during library generation (section 2.3) and pre-protein expression cloning (section 2.7.3)). Restriction digestion was also used to check for the presence of cloned DNA inserts after plasmid DNA preparation. Restriction digests were set up in a total volume of 20 $\mu$ l as described in table 2.6 and incubated in a waterbath at 37°C for 2 hours unless otherwise stated.

Reagent	Volume
React Restriction Buffer (10X) )	2 $\mu$ l
Restriction Enzyme (10U/ $\mu$ l)	0.5 $\mu$ l
DNA sample (1 $\mu$ g)	1-4 $\mu$ l
dH <sub>2</sub> O to final volume	20 $\mu$ l

Table 2.6 Components of a Restriction Digest Reaction

### (b) Agarose Gel Electrophoresis

Resolution of discrete bands of DNA was achieved using horizontal agarose gel electrophoresis. Unless otherwise stated agarose gels were prepared at 1% (appendix I) containing ethidium bromide (500ng/ml) and run in 1x Tris Borate EDTA (TBE) Buffer (appendix I) for 45 minutes at 80V and visualised using a GelDoc 1000 (BioRad, Hemel Hempstead, UK) imaging system.

### (c) Gel Purification of DNA

The cloning of a number of DNA sequences within this thesis relied on the removal of DNA bands of known size from agarose gels. This was achieved using a QIAquick gel purification kit (Qiagen Ltd, Crawley, UK) as per manufacturers instructions.

### (d) Purification of DNA by Phenol/Chloroform Extraction

Phenol/Chloroform extraction and ethanol precipitation were used to purify DNA from the components of various enzymatic reactions such as restriction digests. An equal volume of Tris-EDTA (TE)-buffered phenol/chloroform (Sigma-Aldrich, Poole, UK) was added to the DNA to be purified and the sample vortexed thoroughly. The sample was then spun at 13,000xg for 10 minutes and the upper aqueous phase

transferred to a clean tube. The DNA was then precipitated by the addition of 0.5 volumes of ammonium acetate solution (7.5M) and 2 volumes of absolute ethanol. After gently mixing and a 30 minute incubation at  $-20^{\circ}\text{C}$  the DNA was pelleted by centrifugation at 13,000xg for 30 minutes. The supernatant was discarded and the DNA pellet rehydrated by washing with 70% ethanol before a further 5-minute spin. After removal of the 70% ethanol the pellet was allowed to air dry before resuspension in distilled water and storage at  $4/-20^{\circ}\text{C}$  until required.

#### (e) Vector/Insert Ligation

The crucial step in cloning of novel llama sequences was the ligation of llama-derived DNA into plasmid and bacteriophage vectors. DNA ligation reactions were set up as described in table 2.7 unless otherwise stated. As suggested in the literature (179) a molar excess of insert to vector, giving a molar ratio of between 3:1 and 1:1, was utilised.

Reagent	Volume
Vector DNA	100ng
Insert DNA	200-500ng
T4 DNA Ligase (1U/ $\mu\text{l}$ )	1 $\mu\text{l}$
5x Ligase Buffer	2 $\mu\text{l}$
dH <sub>2</sub> O to final volume	10 $\mu\text{l}$

Table 2.7 Components of a vector/insert ligation reaction

Reactions were incubated at  $4^{\circ}\text{C}$  overnight

#### (f) Transformation of Chemically-Competent *E.coli*

The uptake of successfully ligated insert/vector DNA by *E. coli*-derived bacterial cells, for subsequent propagation and large-scale DNA preparation was achieved by heat-shock at  $42^{\circ}\text{C}$ . A volume of a vector/insert ligation reaction (3 $\mu\text{l}$ ) was added to chemically ( $\text{CaCl}_2$ ) competent *E.coli* cells (100 $\mu\text{l}$ ) (for strains used see appendix I) and incubated on ice for 30 minutes. The transformation mix was then incubated at  $42^{\circ}\text{C}$  for 45 seconds and returned to ice for a further 2 minutes. Pre-warmed ( $37^{\circ}\text{C}$ ) 2xTY (0.8ml) was added to the cells which were then incubated at  $37^{\circ}\text{C}$  with shaking for 1 hour. Cells were then plated out at a range of dilutions onto 2xTY selective plates (appendix I) (for details of selection see table 2.1) and incubated overnight at  $37^{\circ}\text{C}$ .

#### **(g) Preparation of Plasmid DNA from overnight cultures**

Preparation of sufficient quantities of cloned DNA for sequencing or subsequent downstream manipulation was performed by a modified alkaline lysis method. Positive transformants were picked from a 2xTY selective plate and used to inoculate volumes of 2xTY liquid media (5ml). The culture was then grown up overnight at 30°C with shaking (180rpm). Cells were then centrifuged at 4000xg for 10 minutes and the supernatant discarded. Subsequent plasmid DNA purification was performed using the QIAprep mini spin kit (Qiagen Ltd, Crawley, UK) as per manufacturers instructions. Briefly cells underwent alkaline lysis and neutralisation before treatment with RNase, and binding and washing on a spin column. DNA was eluted into 10mM Tris-Cl (pH7). The quantity of DNA recovered was calculated by UV spectrophotometry (Ultraspec 2000, Amersham-Pharmacia, Uppsala, Sweden) at  $OD_{260}$  using the formula described in the appendix I.

#### **(h) Preparation of DNA and Oligonucleotide Primers for Sequencing**

DNA template was diluted to a concentration of 0.5 $\mu$ g/ $\mu$ l. 5 $\mu$ l of template was used per reaction. Sequencing primers (see appendix I) were diluted to a concentration of 1pmol/ $\mu$ l. Sequencing was conducted using an automated sequencer 373A DNA Sequencer (Applied Biosystems, Warrington UK) and the 'BigDye Terminator' DNA sequencing kit (Applied Biosystems, Warrington, UK).

#### **(i) TA/Blunt-End Cloning**

To clone PCR products two cloning kits were used throughout this study. The first, the TOPO TA cloning kit (Invitrogen, Carlsbad, US). PCR products generated with *Taq* polymerase (section 2.2) have a single overhanging deoxyadenosine (A) on each strand at the 3' end. The linearised TA cloning vector contains single overhanging 3' deoxythymidine (T) residues at each end, to which topoisomerase I molecules are bound. Topoisomerase I allows efficient ligation of the insert DNA without the need for addition of a separate DNA ligase. The Zero Blunt cloning kit (Invitrogen, Carlsbad, US) utilises a similar linearised vector, again with topoisomerase I bound to the vector ends. The vector ends are, however, blunt-ended in order to ligate PCR products that lack 3' deoxyadenosine overhangs (for example, PCR products generated with *Pfu* polymerase).



## **2.3 Genomic Library Construction.**

### **2.3.1 Overview of library generation**

The aim of the generation of a genomic DNA library was to provide a collection of clones sufficient in number to include all the genes of a particular organism. Such libraries are prepared through the purification of total cell DNA followed by partial restriction digest resulting in DNA fragments for cloning into a vector that can, not only maintain large, stable inserts, but can also be propagated easily and provide easy access to the genomic sequence through some form of screening. Bacteriophage libraries fulfil all of these criteria through their ability to maintain DNA inserts in excess of 15kb and to readily infect specific *E.coli* strains. Screening of such libraries is achieved by infection of bacteria with large numbers of phage prior to plating out. Discrete regions of bacteriophage-induced lysis (plaques) result, allowing the transfer of viral DNA onto solid nylon membranes (the process of plaque lifting). The nylon membranes can then be treated to remove all contaminating material leaving only library DNA. Large numbers of discrete library DNA molecules can then be screened by DNA hybridisation. DNA hybridisation relies on the ability of radiolabelled single-stranded DNA probes (derived from sequences of interest) to anneal to viral DNA containing llama genomic sequence on the membrane. The position of the hybridisation signal, and consequently the relevant clone, can then be determined by autoradiography.

During this work two llama genomic libraries were constructed using either testicular or leukocyte-derived llama DNA. These libraries were therefore intended to represent both non- and partially rearranged DNA respectively. The same genomic DNA was also used to obtain a custom-made non-rearranged genomic library (Stratagene, La Jolla, USA). Both non-rearranged libraries were utilised during the screening process. Initial screening of the partially rearranged library did not indicate the presence of partially recombined llama immunoglobulin sequences and therefore was not considered further in this study.

### **2.3.2 Preparation of High Molecular Weight Llama DNA**

The preparation of clean, high molecular weight non-rearranged llama DNA was essential to the generation of a high quality genomic library. DNA was obtained from testicular material (kindly donated by The Ashdown Llama Farm, Surrey, UK) which was snap frozen on liquid nitrogen before crushing by pestle and mortar and overnight

proteinase K digestion. Proteinase K digestion and cell lysis was set up as described in table 2.8 and incubated at 37°C for 12 hours with gentle agitation.

Reagent	Mass/Volume
Proteinase K (Sigma-Aldrich, UK)	1mg
Ground testicular tissue	1g
10% SDS (0.5%) (Sigma-Aldrich, UK)	0.5ml
5M NaCl (100mM)	0.2ml
0.5M EDTA pH8 (25mM)	0.5ml
dH <sub>2</sub> O to final volume	10ml

Table 2.8 Composition of a proteinase K digestion reaction.

Digestion was followed by phenol/chloroform extraction and ammonium acetate/isopropanol precipitation using the method described by Maniatis *et al* (179) (section 2.2.3) with the following modifications to prevent DNA shearing. Rather than vortexing after phenol-chloroform addition the digested tissue sample was mixed gently for 30 minutes using an orbital shaker at low speed. After gently mixing and a 30-minute incubation at -20°C, DNA was removed by gentle spinning and attachment to a glass rod. After washing with 70% ethanol the DNA was then allowed to dry and resuspended in dH<sub>2</sub>O. DNA quality was determined by agarose gel electrophoresis.

### 2.3.3 Partial Digestion of Genomic DNA

#### a) Small-scale reactions

In order to determine the ideal digestion conditions to generate pseudo-random genomic DNA fragments of the correct size for library generation small-scale reactions were set up as described in table 2.9

Two initial buffers were prepared as follows:

Dilution Buffer	Volume
10x React buffer 4 (Gibco-BRL, Paisley, UK)	150µl
dH <sub>2</sub> O to final volume	1500µl
<b>Assay Buffer</b>	
10µg high molecular weight DNA	10µl
10x React buffer 4 (Gibco-BRL, Paisley, UK)	45µl
dH <sub>2</sub> O to final volume	450µl

Table 2.9 Composition of Partial Digestion Buffers for Library Generation

Restriction enzyme dilutions were then set up on ice as described in table 2.10

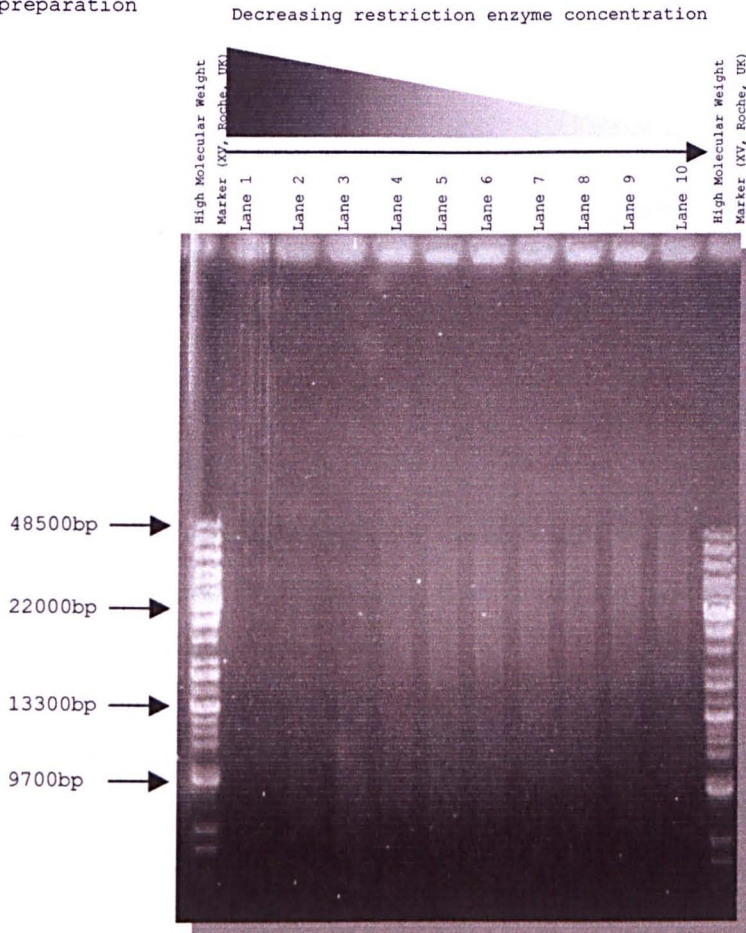
Tube	Preparation			Overall Dilution
1	10µl <i>Sau3A I</i>	+	140µl dilution	1/15
2	10µl 1/15 dilution	+	90µl dilution	1/150
3	10µl 1/150 dilution	+	10µl dilution	1/300
4	10µl 1/150 dilution	+	30µl dilution	1/600
5	10µl 1/150 dilution	+	50µl dilution	1/900
6	10µl 1/150 dilution	+	70µl dilution	1/1200
7	10µl 1/150 dilution	+	90µl dilution	1/1500
8	10µl 1/150 dilution	+	110µl dilution	1/1800
9	10µl 1/150 dilution	+	190µl dilution	1/3000
10	10µl 1/150 dilution	+	290µl dilution	1/5000

Table 2.10 Small Scale Partial digest reactions.

Small-scale restriction digests were then set up by addition of assay buffer (45µl) to a small volume of each dilution (5µl) and incubation at 37°C for 30 minutes. Reactions were stopped by addition of 0.5M EDTA (1µl) (Sigma-Aldrich, Poole, UK).

Digests were then analysed by agarose gel electrophoresis on a 0.6% gel, running at 1.4V/cm overnight with high molecular weight markers (High Molecular Weight Marker 'XV', Roche Biochemicals, Lewis, UK). The optimum amount of enzyme was then estimated by eye from the agarose gel (figure 2.1) as half of that used to produce maximum UV fluorescence on the gel within the required size range (9-16kb).

**Figure 2.1 Agarose Gel of partial digest of high molecular weight llama testicular DNA cut with *Sau3A* restriction endonuclease.** A range of enzyme concentrations were used in order to ascertain the optimum DNA size for bacteriophage library generation. The intensity of DNA fluorescence by ethidium bromide staining is related to mass distribution of DNA so that half the concentration of enzyme required to generate maximum fluorescence was used for large-scale partial digestion. Inserts of approximately 15kb were required and so half the concentration used in Lane 4 (Refer to table 2.10) was used in large-scale preparation



### b) Large-scale partial digestion

Using the optimum enzyme concentration as determined above (in this case 0.0125 units of *Sau3A* I per  $\mu\text{g}$  DNA) 100 $\mu\text{g}$  of DNA was partially digested as described above. The DNA was then purified by phenol/chloroform extraction (section 2.2.3) and concentration determined by ultraviolet spectrometry at  $\text{OD}_{260}$  using the formula described in the appendix II.

### c) Dephosphorylation of Insert DNA

To prevent insertion of multiple genomic DNA fragments into the bacteriophage vector during the subsequent ligation reaction, partially digested DNA was treated with calf intestinal alkaline phosphatase (CIAP, Roche Biochemicals, Lewis, UK) to remove 5' phosphate groups. A reaction was set up as described in table 2.11.

Reagent	Volume
CIAP 10X reaction buffer	10 $\mu\text{l}$
CIAP 1U/ $\mu\text{l}$ (Roche, UK)	2 $\mu\text{l}$
Partially Digested Llama Genomic DNA	2 $\mu\text{g}$
dH <sub>2</sub> O to final volume	100 $\mu\text{l}$

Table 2.11 Components of a dephosphorylation reaction used in library construction

The reaction was incubated for 60 minutes at 37°C and stopped by addition of 0.5M EDTA (2 $\mu\text{l}$ ) (Sigma-Aldrich, Poole, UK).

### 2.3.4 Ligation of Genomic DNA into Phage Arms

Library construction was achieved using the Lambda DASH II/*Bam*HI vector kit (Stratagene, La Jolla, US) as per manufacturer instructions. The optimal ligation conditions were as set out in table 2.12

Reagent	Volume
Lambda DASH II pre-digested with <i>Bam</i> HI (1 $\mu\text{g}$ )	1.0 $\mu\text{l}$
Partially digested genomic DNA (0.4 $\mu\text{g}$ )	1.0 $\mu\text{l}$
10 $\times$ Ligase buffer (Stratagene, UK)	0.5 $\mu\text{l}$
10mM rATP (pH 7.5)	0.5 $\mu\text{l}$
T4 DNA Ligase (Stratagene, UK)	2 U
dH <sub>2</sub> O to final volume	5 $\mu\text{l}$

Table 2.12 Components of a bacteriophage vector genomic library ligation reaction

The ligation was incubated at 4°C overnight.

### **2.3.5 Packaging of Phage DNA**

Packaging of phage DNA involves the incubation of DNA with components of the bacteriophage viral coat, allowing synchronised assembly of viable virus particles containing the DNA of interest. The Lambda DASHII ligation was packaged using Gigapack XL packaging extracts (Stratagene, La Jolla, USA) optimised to maximise insert size as per manufacturers instructions. Extracts were thawed rapidly before addition of the ligation reaction (1-4 $\mu$ l). The packaging reaction was then mixed briefly before incubation at room temperature (22°C) for 2 hours. The completed packaging reaction was diluted in SM Buffer (500 $\mu$ l) containing chloroform (20 $\mu$ l) and stored at 4°C.

### **2.3.6 Library Titreing**

The packaged library was pooled, titred and plated out as quickly as possible to prevent the inactivation of phage particles which occurs gradually with time during storage. A culture of XL1-Blue MRA (P2) (Stratagene, La Jolla, USA), the host strain for Lambda DASH II infection was grown up in appropriate media (appendix I) to an OD<sub>600</sub> of approximately 0.8 and pelleted by centrifugation at 500xg for 10 minutes. The cells were then gently resuspended in sterile MgSO<sub>4</sub> (10mM) to a spectrophotometer absorbance at OD<sub>600</sub> of 0.5. Dilutions of the packaging reaction were prepared so that 1 $\mu$ l of the reaction and 1 $\mu$ l of a 1:10 dilution of the packaged phage were each added to 200 $\mu$ l of cells. The phage/bacteria mix was then incubated at 37°C for 15 minutes to allow phage attachment. Molten, pre-warmed (48°C) LB top agarose (3ml) was then mixed with the cells and plated immediately onto pre-warmed (37°C) 90mm diameter LB agar plates. Plates were then incubated overnight at 37°C and the number of plaque forming units (pfus) determined by counting the number of plaques formed in the bacterial lawn. The titre of the pooled genomic testicular library (containing three packaging reactions) was 1.4 x 10<sup>6</sup> prior to amplification. The leukocyte derived genomic library had a significantly lower titre of 1 x 10<sup>5</sup>. The lower titre of the leukocyte derived library may have been due to any number of factors including reduced efficiency of ligation and packaging perhaps resulting from lower quantities of pure DNA of the appropriate size within the reactions.

### **2.3.7 Library Amplification**

Library amplification involves the multiplication of the specific bacteriophage clones within the original packaged library. Excessive amplification can lead to over and under-representation of particular clones within the library. For this reason libraries were amplified only once during library generation described here. The libraries were amplified in order to generate a stable, high-titre stock of library. A fresh stock of host cells (OD<sub>600</sub> of 0.5) were prepared in MgSO<sub>4</sub> solution and approximately  $5 \times 10^4$  pfu (infecting 600 $\mu$ l of host cells) were plated onto large (150mm) LB agarose plates (section 2.3.6) with molten LB top agarose (6.5ml). Phage plaques were allowed to grow until plaques made contact with each other at which point each plate was overlaid with sterile SM buffer (10ml) and incubated overnight at 4°C with gentle rocking. The bacteriophage suspension was recovered from each plate and chloroform added to a 5% (v/v) final concentration. The phage solution was then spun for 10 minutes at 500xg to remove cell debris and the library stored in a sterile polypropylene container with 0.3% (v/v) chloroform. Amplified titres of approximately  $5 \times 10^9$  pfu/ml were achieved from both libraries.

## **2.4 Screening of a Llama Genomic Library**

Attempts to isolate genomic variable domain sequence components were based on proprietary sequence information generated at Unilever Research, Colworth. This comprised a large (209 member) database of heavy chain immunoglobulin cDNA sequences derived from non-immunised llama leukocyte RNA (180). Determination of a consensus sequence from the variable domain of the cDNA sequences using multiple alignment software (table 2.20) allowed the synthesis of specific oligonucleotides to generate a mixed (containing many different variable sequences) species variable probe during PCR of llama leukocyte cDNA (sections 2.4.1-2.4.3, 3.3.1 and figure 3.1). While this technique enabled the isolation of a small number of variable gene segment clones, it did not allow isolation of the significantly smaller diversity (D) or joining (J) mini-genes, nor did it allow for differentiation between classical and heavy chain variable sequences at the level of screening. Smaller oligonucleotide probes were generated for these purposes (sections 3.4.2-3.4.3 and figure 3.1).

### **2.4.1 Generation of a Variable Gene Segment Probe by Polymerase Chain Reaction**

Variable gene segments are, by virtue of their nature, heterologous in sequence. Degenerate primers were therefore designed using consensi derived from a database of cDNA  $V_{HH}$  sequences (180). These oligonucleotides (designated variable 1 and variable 2, table 2.3) were designed to span the framework 1 (FR1) to framework 3 (FR3) regions of the variable gene segments.

### **2.4.2 Design of Separate Classical and Non-Classical V Region Oligonucleotide Probes**

While initial rounds of library screening enabled isolation of a small number of variable region-containing clones, these gene segments were found to be exclusively of a classical nature. That is, all contained sequence that would lead to formation of classical  $V_H$  domains. In order to screen the genomic library specifically for non-classical and classical variable gene segments two oligonucleotide probes were derived from analysis of 22 classical and 179 non-classical variable domain cDNA sequences. The sequence of these probes is given in table 2.13.



### 2.4.3 Design of a J Region Probe

A single oligonucleotide probe was designed by reference to a database of llama immunoglobulin gamma cDNA sequences (corresponding to position 102-113 (30, 180)). Heavy chain locus J segments are typically very short (between 49-63bp) (38) and encode approximately 12 amino acids within the CDR3 and FR4 of variable cDNA sequences. In order to minimise the potential problems associated with somatic hypermutation and possible additional recombination-related mechanisms, the probe was designed with modest degeneracy exclusively to the FR4 coding sequence (table 2.13).

Oligonucleotide Probe Name	Probe Generated	Oligonucleotide Sequence
VCLASS	Variable ( $V_H$ ) Region Oligonucleotide Probe	TGGGTGCGCCAGGCTCCAGGGAAGGGGCTCGAGTGG
VNCLASS	Variable ( $V_{NH}$ ) Region Oligonucleotide Probe	TGGTACCGCCAGGCTCCAGGGAAGCAGCGCGAGTTG
JOINING	J Region Oligonucleotide Probe	TACTGGGGCCAGGGGACCC (A/T) GGTACCGTCTCCTCA

Table 2.13 Sequence of oligonucleotide probes used during library screening for variable gene components.

### 2.4.4 Generation of a Constant Region Probe

The constant region probe was generated by PCR of llama cDNA (kind gift of Dr K. Cromie, Unilever Research, Colworth) with primers (Constant and Constant 2, table 2.3) specific to the  $C_H2$  region of llama immunoglobulin gamma (sequence not shown). PCR was performed using *Taq* polymerase for the number of cycles indicated, and at the annealing temperature detailed in table 2.1 as per the protocol described in section 2.2.2. The relevant band was gel purified and radiolabelled by random priming (sections 2.4.7).

### 2.4.5 Generation of a RAG-1 Specific Probe

Generation of a llama RAG-1 specific probe required first the amplification of a small region of the llama RAG-1 gene from genomic DNA. This was performed by generating consensus primers derived from a multiple alignment of previously characterised RAG-1 gene sequences from other species (data not shown). The subsequent 100bp *Pfu*-derived PCR product was subsequently gel purified (section 2.2.3) and cloned into pBluescript (appendix III) by a blunt end ligation strategy employing the *Sma* I restriction endonuclease (section 2.2.3) and heat shock

transformation (section 2.2.3). A probe was then generated by redigestion of large quantities of the pBluescript DNA with *Sma* I, agarose gel purification (section 2.2.3) and radiolabelling (section 2.4.7).

#### 2.4.6 Synthesis of PCR Generated Probes

PCR was performed with *Taq* Polymerase as described in table 2.3 and section 2.2.2. Product size was checked on a 1% agarose gel against standard markers (1kb *plus* DNA marker) and the reaction products purified using a gel extraction kit (section 2.2.3).

#### 2.4.7 Radiolabelling of PCR-Derived Probes for Library Screening

PCR-derived probes were labelled using the Megaprime DNA labelling system (Amersham-Pharmacia, Uppsala, Sweden) coupled with  $\alpha$ -<sup>32</sup>P dCTP (Amersham-Pharmacia, Uppsala, Sweden). The 5'-3' primer used previously in PCR product synthesis was used in the labelling reaction, in order to provide a specific priming site for polymerase extension. Otherwise labelling was carried out as per the manufacturers instructions. Briefly the reaction was set up in a 1.5ml microtube as detailed in table 2.14.

Reagent	Volume
PCR product to be labelled (20ng/ $\mu$ l)	2 $\mu$ l
Labelling Buffer	10 $\mu$ l
PCR Primer (5'-3') (50ng/ $\mu$ l)	2 $\mu$ l
MilliQ deionised water to volume	31 $\mu$ l

Table 2.14 Initial Components of a MegaPrime™ Radiolabelling reaction

The reaction components were then boiled for 5 minutes before being placed on ice. Each unlabelled nucleotide solution (dATP, dTTP and dGTP) was added (3 x 4 $\mu$ l) followed by  $\alpha$ -<sup>32</sup>P dCTP (5 $\mu$ l of isotope at a concentration of 3000Ci/mmol) and Klenow enzyme (2 $\mu$ l of 1U/ $\mu$ l enzyme). The reaction was mixed briefly and incubated at 37°C for 3hrs.

## 2.4.8 Radiolabelling of Synthetic Oligonucleotides

Synthetic oligonucleotides were utilised during both recombination assays and library screening (in the case of  $J_H$  region and  $V_{HH}/V_H$  specific probes). Sequence details are given in table 2.13.

Kinase reactions were set up to label the 5' end of each oligonucleotide as detailed in table 2.15.

Reagent	Volume
T4 Polynucleotide Kinase 10U/ $\mu$ l (New England Biolabs, Hitchin, UK)	1 $\mu$ l
Oligonucleotide to be labelled (10 $\mu$ M)	1 $\mu$ l
10x Kinase buffer (New England Biolabs, Hitchin, US)	1 $\mu$ l
$\gamma^{32}P$ -ATP (5 pmol/ $\mu$ l)	3 $\mu$ l
dH <sub>2</sub> O to final volume	10 $\mu$ l

Table 2.15 Components of a T4 kinase oligonucleotide end-radiolabelling reaction

Reactions were incubated at 37°C for 1 hour and the reaction stopped by addition of 0.5M EDTA (1 $\mu$ l). Reactions were then diluted to a total volume of 40 $\mu$ l by addition of 30 $\mu$ l dH<sub>2</sub>O. A Sephadex G-25 spin column (Amersham-Pharmacia, Uppsala, Sweden) was prepared by brief centrifugation before addition of diluted kinase reaction. Unincorporated nucleotides were removed from the reaction by brief spinning through the G-25 column and purified, labelled oligonucleotide collected.

## 2.4.9 Plaque Lifting

To generate library filters, large LB agarose plates were prepared containing approximately  $5 \times 10^4$  pfu per plate using the phage infection and plating methods described previously (section 2.3.6). These plates were pre-cooled to harden the top agarose prior to plaque lifting. Hybond N+ filters (Amersham Pharmacia, Uppsala, Sweden) were carefully placed onto the agarose surface and both plate and membrane marked for orientation. After 30 seconds the filter was removed from the plate surface with forceps and placed plaque side up on filter paper. A second duplicate filter was then placed onto the plate for 5 minutes. The lifted filters then proceeded through a series of saturated filter paper (Whatman, Maidstone, UK) washes with the plaque side up. Initial washing with denaturation solution (appendix I) for 5 minutes liberated DNA from the bacteriophage. This was followed immediately by 5 minute

washes in neutralisation buffer (appendix I) and 2 x SSC solution (appendix I) to wash away cell debris. The DNA was then cross-linked to the filters (Spectrolinker XL-1000, Spectronics Corporation, Westbury, US) and the filter allowed to air dry.

#### **2.4.10 Hybridisation**

The hybridisation procedure used during screening consisted of four steps

- 1) Prehybridisation
- 2) Hybridisation
- 3) Stringency washing
- 4) Autoradiography

Methods vary slightly for oligonucleotide probes compared to PCR probes as the shorter length probes exhibit different hybridisation kinetics.

##### **1) Prehybridisation**

Prepared filters were incubated in prehybridisation buffer (or oligonucleotide prehybridisation buffer, appendix I) containing heat-denatured, sheared salmon sperm DNA (to a final concentration of 50µg/ml (Sigma-Aldrich, Poole, UK)) at 65°C with agitation for between 3-14 hrs in an airtight sandwich box.

##### **2) Hybridisation**

Prior to hybridisation (PCR-derived probes only) the labelled probe was denatured at 100°C for 5 minutes before cooling on ice. Labelled probe (approximately one complete oligonucleotide or PCR labelling reaction per sandwich box containing 100ml prehybridisation buffer) was then added to the prehybridisation solution and filters to be screened. Hybridisation with both oligonucleotide and PCR probes then took place at 65°C overnight with agitation.

##### **3) Stringency Washing**

Stringency washes were carried out in prewarmed SSC/SDS solutions as shown in table 2.16 for oligonucleotide hybridisation and table 2.17 for PCR probe hybridisation.

Number of Repetitions and Duration of Each Wash Step	Wash Buffer Used	Stringency of Wash (Detergent concentration and wash temperature)
2 x 15 minutes	6 x SSC	0.1% SDS at 65°C
1 x 2 minutes*	6 x SSC	0.1% SDS at 60°C

Table 2.16 Post-hybridisation washing conditions for oligonucleotide-based hybridisation

\*denotes optional wash used only if filter remains highly radioactive as determined by Geiger counter monitoring.

Number of Repetitions and Duration of Each Wash Step	Wash Buffer Used	Stringency of Wash (Detergent concentration and wash temperature)
2 x 15 minutes	2 x SSC	0.1% SDS at 65°C
1 x 30 minutes	1 x SSC	0.1% SDS at 65°C
1 x 10 minutes*	0.1 x SSC	0.1% SDS at 65°C

Table 2.17 Post-hybridisation washing conditions for PCR probe-based hybridisation

\*denotes optional wash used only if filter remains highly radioactive as determined by Geiger counter monitoring.

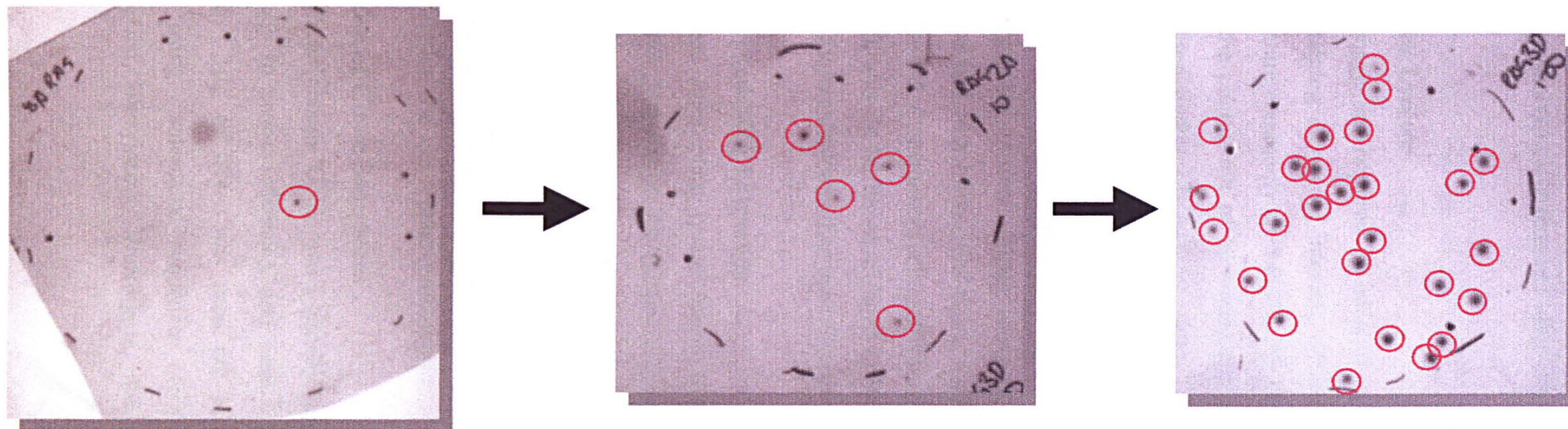
#### 4) Autoradiography

Filters were covered in Saran Wrap and placed in an autoradiograph cassette complete with intensifying screen and Biomax MS autoradiography film (Kodak, Hemel Hempstead, UK). Exposure took place overnight at  $-70^{\circ}\text{C}$  before developing in standard manual developer and fixer (Kodak, Hemel Hempstead, UK). Positive clones were displayed on the autoradiograph as small, discrete black dots (figure 2.2). The position of positive clones was determined by reference to the plate from which the plaque lifting was performed.

##### 2.4.11 2nd and 3rd Round Clone Isolation

Positive clones identified through the first round of screening were removed from LB agarose plates as an agarose plug using the narrow end of a glass pasteur pipette. The plug containing the relevant clone was subsequently suspended in SM buffer (500 $\mu\text{l}$ ). The plug was then vortexed and stored at  $4^{\circ}\text{C}$  overnight to allow diffusion of bacteriophage particles from the plug into the buffer. SM Buffer containing positive bacteriophage was then plated out on small (90mm diameter) LB agarose plates after infection of host bacteria. Approximately 50 pfu were plated per plate. The screening process was then repeated from plaque lifting, hybridisation and autoradiography in order to confirm the positive, uncontaminated nature of the original positive clones.

Figure 2.2 Example of autoradiographs generated during progressive rounds of library screening



**Round 1)** Initial screening of large (140mm diameter) filters containing approximately 50 000 clones per plate. Small numbers of candidate positive clones are isolated, plaques eluted in SM buffer and replated

**Round 2)** Replating of the initial candidate clones on a smaller plate (approximately 50 plaques per 90mm diameter plate). Filters are generated and rehybridised. Approximately 20-80% of 2<sup>nd</sup> round plaques represent the candidate clone. Single candidate clones (which are now a greater distance from possibly contaminating clones) are then replated

**Round 3)** Replating of a 2<sup>nd</sup> round candidate clone (approximately 30 plaques per 90mm diameter plate). All plaques at this stage should represent the clone of interest. Clones isolated after this round of hybridisation are used for generation of plate lysates (section 2.3.6)

Two rounds of screening were typically required after the initial screen in order to isolate pure clones.

#### **2.4.12 Generation of Plate Lysates**

Large titres of bacteriophage are required in order to purify large quantities of bacteriophage DNA containing clones of interest. Two established methods for bacteriophage propagation are commonly used (179). The method used in this thesis (plate lysates) involved the plating out of approximately  $5 \times 10^4$  pfu infecting host strain bacteria per 140mm-diameter LB agarose plate. Typically 10 plates were required per phage, and consequently, DNA preparation. Bacteriophage plaques were grown at 37°C until they reached confluence. Plates were then overlaid with SM buffer (10ml) and incubated overnight with gentle agitation to allow diffusion of bacteriophage particles into the buffer. Phage-containing SM buffer was then removed and the plates washed with further SM buffer (2ml). This buffer was then combined with the initial 10ml of phage-containing SM buffer removed. Chloroform was then added (2% (w/v)) and the buffer centrifuged at 10,000xg for 10 minutes to remove residual agarose. Supernatant was then used directly for DNA preparation.

#### **2.4.13 PEG Precipitation and DNA Extraction**

Phage DNA was prepared from plate lysates using a QIAGEN Lambda Midi bacteriophage lambda DNA preparation kit (Qiagen, Crawley, UK). Briefly this involved separation of phage from media with polyethylene glycol (PEG) a long polymeric compound which in the presence of salt, absorbs water causing bacteriophage to precipitate. This was followed by binding of DNA to an anion-exchange resin under low-salt conditions. Washing of the resin-bound DNA was followed by a high-salt elution and isopropanol precipitation.

### **2.5 Characterisation of Clone Sequences**

Direct sequencing of bacteriophage DNA was typically performed using specific primers designed initially to probe sequences, and later through 'walking' along known clone sequence. 1µg of DNA was used per sequencing reaction and sequencing performed using automated sequencer (Applied Biosystems DNA Sequencer 373A, Applied Biosystems, Warrington, UK ) and the 'BigDye

Terminator' DNA sequencing kit (Applied Biosystems, Warrington, UK) in-house at Unilever Research, Colworth.

### **2.5.1 An Alternative Strategy For J Region Sequencing**

High sequence homology between multiple J gene segments in heavy chain immunoglobulin loci of most species make direct sequencing of clone DNA with primers specific to J segment cDNA sequences impossible. Successful sequencing was only possible through the use of primers specific to flanking non-coding regions surrounding J segments. However, such sequences were not present within the cDNA library and therefore not immediately available. In order to determine the nature of such flanking sequences a novel PCR strategy was employed. Phage clone DNA was amplified using a single primer (J sequencing 1) and its reverse complement (J sequencing 2) (table 2.3). Both primers were specific to the FR4 consensus from the Unilever cDNA library and therefore corresponding to the region that would typically be encoded by J gene segments. Amplification of the bacteriophage clone DNA used 25 cycles at 50°C annealing and *Pfu* polymerase in order to maintain sequence fidelity as detailed in table 2.3. This strategy is outlined in figure 3.9. Two major PCR products were generated of approximately 500bp and 300bp in length.

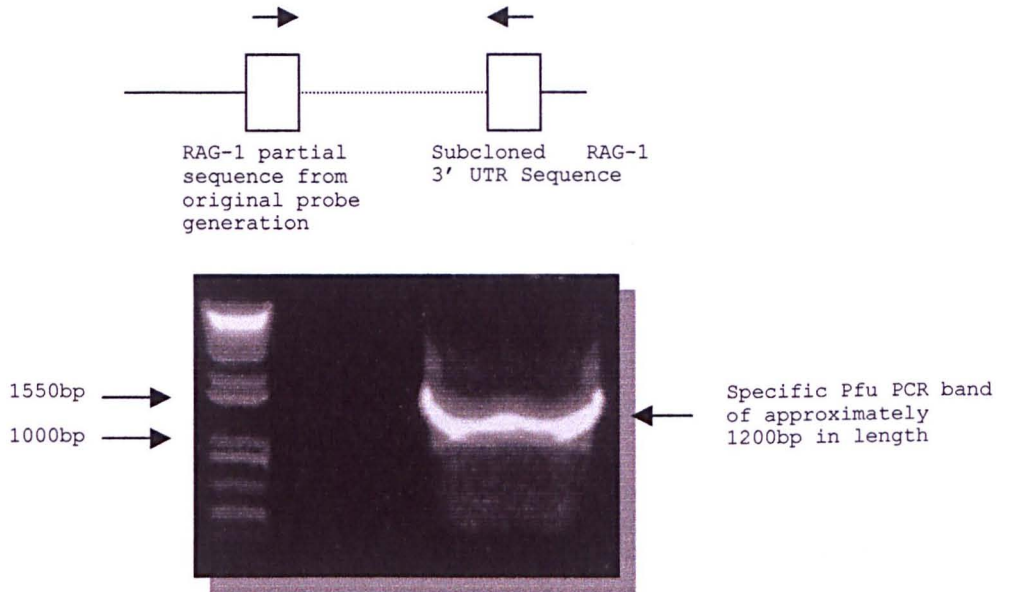
PCR products were cut out of an agarose gel and gel purified using the Qiaquick gel extraction kit (Qiagen, Crawley, UK, section 2.2.3) before cloning using the Zero Blunt™ blunt end PCR product cloning kit (Invitrogen, Carlsbad, USA) and sequencing using M13 primers. This strategy allowed the determination of non-coding sequence lying between homologous J gene segments and led to successful direct sequencing of J gene segment clone DNA. Sequencing of the D-J locus was made possible by design of overlapping sequencing primers to 'walk' across the locus.

### **2.5.2 An Alternative Strategy for RAG-1 Gene Sequencing**

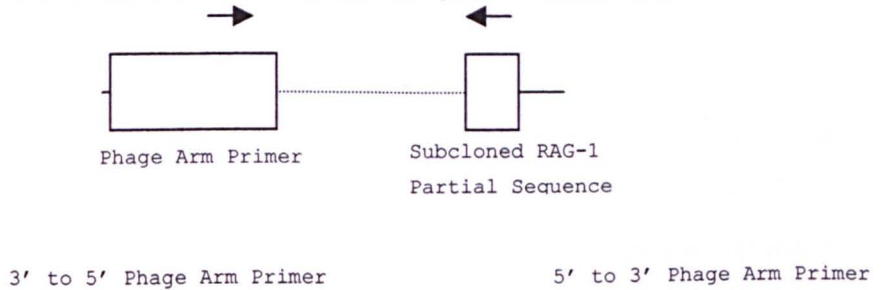
In order to obtain the full-length llama RAG-1 gene sequence. PCR primers (table 2.3) were designed to known sequence both within the gene itself and within surrounding sequences defined during sub-cloning. These primers were then used to amplify the remaining gene sequence (figure 2.3 and 2.4). PCR products were then cloned into a blunt-end cloning vector (table 2.1) and sequenced.



**Figure 2.3 PCR of RAG-1 clone DNA with Pfu polymerase in order to generate the full length RAG-1 sequence.** Subcloning of Phage DNA had previously generated partial sequence of the RAG-1 3' UTR. Primers to this sequence and derived from our original RAG-1 hybridisation probe were used to PCR phage DNA and generate a band spanning the majority of the RAG-1 sequence for subsequent cloning and sequencing



**Figure 2.4 PCR of RAG-1 clone DNA with Pfu polymerase in order to generate the full length RAG-1 sequence.** Primers specific to bacteriophage lambda arm sequence were used in conjunction with primers derived from our original RAG-1 hybridisation probe. Bands were generated using both arm primers, however the weak band seen with the 3' to 5' phage arm primer was a non-specific unrelated PCR product. Both bands were cut and gel purified prior to cloning using a TOPO Blunt cloning kit (Invitrogen, Carlsbad, USA)



## 2.6 Recombination Assays

Recombination assays provide a method of reconstructing the recombination events that take place within developing B-cells during the generation of immune diversity. A number of the individual steps involved in V(D)J recombination can be dissected *in vitro*, through the incubation of short, radio-labelled double-stranded oligonucleotide substrates in the presence of recombinant RAG proteins under the correct ionic conditions and in the presence of a suitable buffer.

### 2.6.1 Design of Oligonucleotide Substrates

Oligonucleotide substrates for utilisation in RAG cleavage assays were derived from the recombination signal sequences of the variable gene segment clones reported in Chapter 3. Sequences were chosen as those most likely expressed by the llama *in vivo* by comparison to expressed cDNA sequences (section 3.18), so as to provide the most accurate recreation of the *in vivo* recombination system. The principal oligonucleotide sequences used in these assays are summarised in table 5.2

### 2.6.2 Gel Purification of Oligonucleotides

The nature of the RAG cleavage assay is such that the presence of incomplete labelled oligonucleotides, present during visualisation, often as a ladder of bands, may obscure the cleaved products of the RAG/DNA interactions. In order to remove such bands oligonucleotides were purified on a 20% polyacrylamide sequencing gel (Sequagel, National Diagnostics, Hull, UK appendix I). On migration of the loading buffer to the base of the gel, the polyacrylamide gel was exposed to long wave UV light, allowing visualisation of the single-stranded DNA oligonucleotide. The uppermost portion of the oligonucleotide band, containing the major, correct length DNA product was excised and crushed into elution buffer (3ml, appendix I). Oligonucleotides were eluted overnight at 4°C with gentle agitation, before retrieval using Sepak C60 HPLC columns (Waters Chromatography, Milford, US). Columns were prewet with methanol (10ml) and then dH<sub>2</sub>O (10ml). Eluted oligonucleotides were then passed through a 0.45µm filter syringe prior to capture on the prewet column. The column was subsequently washed twice with dH<sub>2</sub>O (2 x 5ml) and the purified oligonucleotides removed from the column by elution in 60% methanol (2 x 1ml).

Oligonucleotides were then dried down using a RotaVap (Sartorius, Roettingen, Germany) vacuum dryer for 3-4 hours.

### 2.6.3 Labelling of Oligonucleotides

Oligonucleotides were labelled with  $^{32}\text{P}$  by a 3' kinase reaction (section 2.4.8).

### 2.6.4 Annealing of Complementary Oligonucleotides

10 $\mu\text{M}$  labelled top-strand oligonucleotide (1 $\mu\text{l}$ ) was added to 10 $\mu\text{M}$  unlabelled reverse complement oligonucleotide (3 $\mu\text{l}$  (excess)) in a final volume of 50 $\mu\text{l}$ . The oligonucleotides were incubated at 95 $^{\circ}\text{C}$  for 3 minutes, 65 $^{\circ}\text{C}$  for 10 minutes and 37 $^{\circ}\text{C}$  for 10 minutes to anneal the two complementary strands. A test gel of oligonucleotides was then run to identify any labelled truncated oligonucleotides (data not shown).

### 2.6.5 Cleavage of Oligonucleotide Substrates by Recombinant RAG Proteins

Cleavage of oligonucleotide substrates by RAG proteins recreates a number of stages of the V(D)J recombination process *in vitro* (sections 1.8.5 and 5.4). RAG cleavage reactions were set up in 1.5ml microfuge tubes as described in table 2.18

Reagent	Volume
Radiolabelled double-stranded oligonucleotide substrate	1 $\mu\text{l}$
250mM MOPS	1 $\mu\text{l}$
10mM DTT	1 $\mu\text{l}$
1mg/ml BSA (Sigma-Aldrich, UK)	1 $\mu\text{l}$
Recombinant RAG-1/2 Protein*	2 $\mu\text{l}$
Divalent Cation (either $\text{MgCl}_2$ or $\text{MnCl}_2$ ) (10 $\mu\text{M}$ )	1 $\mu\text{l}$
High Motility Group 1/2 Proteins*	1 $\mu\text{l}$
dH <sub>2</sub> O	to total volume 10 $\mu\text{l}$

Table 2.18 Components of an *in vitro* recombination assay

\*Either Mouse RAG-1/Mouse RAG-2 or Llama RAG-1/Mouse RAG-2

\*Optional

The role of each of these components is discussed in some detail in section 5.4.2.

Reactions were incubated at 37 $^{\circ}\text{C}$  for 30 minutes.

### 2.6.6 Visualisation of Cleavage Products by PAGE

Polyacrylamide sequencing gels were prepared for visualisation of RAG cleavage reactions at either 12% or 15% using the SequaGel system (National Diagnostics, Hull, UK). The volumes for preparation of a 12% gel are given in the appendix I. The gel was allowed to set for 25 minutes before washing and pre-running at 25 W for 1-2 hours prior to loading. After loading the gel was run at 25W for 2 hours before

transfer of the gel to filter paper (Whatman, Maidstone, UK) and dried at 80°C for 2 hours in a vacuum gel drier (Sartorius, Roettingen, Germany). The gel was then exposed overnight in a phosphorimager cassette before reading in a phosphorimager (Typhoon 8600) and processing using ImageQuant™ image processing software. All phosphorimager equipment and software was from Amersham-Pharmacia, Uppsala, Sweden.

## **2.7 Baculovirus Expression and Protein Purification**

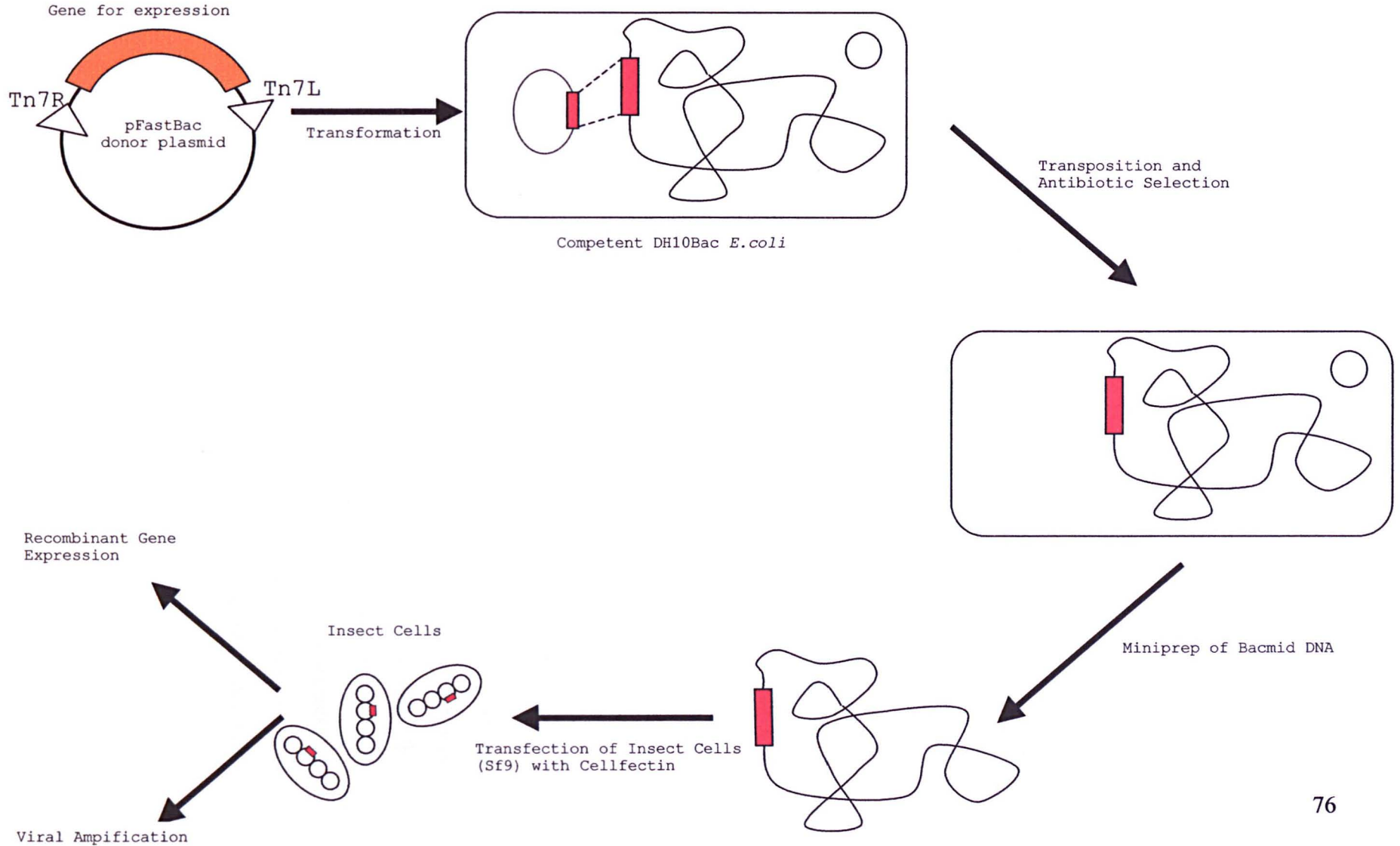
### **2.7.1 Overview of the Bac-to-Bac™ Baculovirus system**

The *in vitro* infection of insect cells by baculovirus differs from the *in vivo* infection in that the polyhedrin gene from the naturally occurring baculovirus genome is replaced by the sequence to be translated and expressed. In this study the Bac-to-Bac™ expression system (Gibco-BRL, Paisley UK) was utilised. This system is based on the site-specific transposition of an expression cassette into a bacmid or baculovirus shuttle vector which can be grown up in *E.coli*. The sequence of the protein to be expressed is first cloned into a donor (pFastBac™, appendix III) plasmid containing a baculovirus specific promoter in a region surrounded by transposition elements (Tn7). Successful transposition results in the transfer of the gene into the mini-attTn7 attachment site within the bacmid, an event that can be selected for through the disruption of the bacmid *lacZ* $\alpha$  gene through blue/white selection. The bacmid contains the complete baculoviral genome so that preparation of bacmid DNA through propagation of *E.coli* can be followed immediately by transfection of insect cells in the presence of a lipid-based transfection agent (in this case Cellfectin™). This strategy is outlined in figure 2.5

### **2.7.2 PCR Amplification of Llama RAG-1 Sequence From Clone DNA**

The 'core' region of the llama RAG-1 gene was identified by comparison to the murine form by pairwise alignment (data not shown). Primers containing suitable restriction sites and an in-frame 6xHis tag were synthesised (table 2.3). PCR was performed on purified phage clone DNA (section 2.2.2) using these primers and the *Pfu* polymerase for 25 cycles (table 2.2).

Figure 2.5 Overview of Baculovirus Expression (Modified From Gibco-BRL Technical Manual 10359). Transposition sequences are labelled (Tn7R+Tn7L)



### **2.7.3 Gel Purification and Digestion of Llama RAG-1 PCR Product**

The PCR product was run on a 1% agarose gel and visualised before gel purification (section 2.2.3). Eluted DNA was subsequently subjected to a double digest with the restriction enzyme *Xba*I in order to generate staggered ends for cloning. The digest was then cleaned up by phenol/chloroform extraction (section 2.2.3).

### **2.7.4 Preparation of pFastBac For RAG-1 Cloning**

A modified form of the pFastBac vector (Gibco-BRL, appendix III) containing the sequence of the maltose binding protein was digested with *Xba*I and cleaned up by phenol/chloroform extraction and ethanol precipitation.

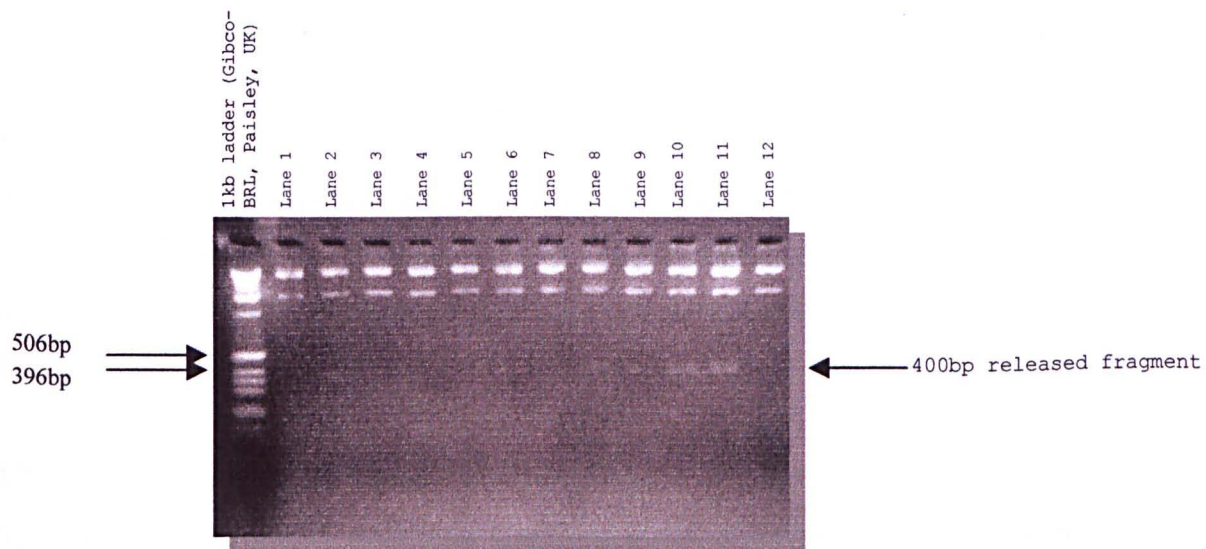
### **2.7.5 Ligation of RAG-1 into pFastBac and Transformation into DH5 $\alpha$**

The prepared pFastBac-MBP vector and llama RAG-1 insert DNA were ligated together using the rapid ligation kit (Roche Biochemicals, Lewes, UK) in the molar ratio of 3:1 insert to vector. The ligation reaction (2 $\mu$ l) was subsequently used to transform chemically-competent DH10Bac *E.coli* by heat shock at 42°C (section 2.2.3). Transformants were then selected on LB ampicillin plates (appendix I). A map of the RAG-1/MBP fusion vector is included in the appendix.

### **2.7.6 Confirmation of Insert Orientation**

In order to determine the orientation of the cloned llama RAG-1 insert DNA, positive DH10Bac clones were picked and grown up in LB ampicillin (5ml, appendix I) before plasmid DNA preparation using the QIAGEN miniprep kit (Qiagen, Crawley, UK, section 2.2.3). Plasmid DNA was then cut with the *Sac*I restriction enzyme (section 2.2.3). Clones containing inserts orientated in the suitable direction for expression generated a 170bp fragment, as confirmed by agarose (1%) gel electrophoresis (figure 2.6). The fidelity of the cloned *Pfu* PCR products was also confirmed at this point through sequencing. No PCR errors were identified.

**Figure 2.6 Confirmation of direction of cloned llama RAG-1 PCR product by digest of pFastbac with *Xba*I.** Clones releasing a 170bp fragment are in correct orientation for translation while those releasing a ~400bp fragment are in the reverse, out-of-frame orientation. Therefore clones in lanes 1, 4, 5, 7 and 12 were in the correct orientation. Note that the 170bp fragment is not visible on this 1% agarose gel. The presence of the 170bp band was confirmed by running the digestion products on a polyacrylamide mini-gel (data not shown)



### **2.7.7 Transformation into DH10Bac and Confirmation of Transposition into Bacmid**

Plasmid DNA containing correctly orientated insert was transformed into competent DH10Bac competent cells by heat-shock at 42°C (section 2.2.3). Cells were allowed to recover in SOC for 1 hour with shaking. Positive transformants were selected by plating onto selective LB agar (appendix I) plates. Colonies were allowed to grow for 48hrs at 37°C. White colonies represented successful transposition of the expression cassette into the bacmid DNA. The transposed phenotype was subsequently confirmed by restreaking of white colonies onto fresh plates and incubation at 37°C overnight.

### **2.7.8 Preparation of Bacmid DNA**

A single verified white colony was subsequently picked and grown up at 37°C in LB medium (appendix I) supplemented with kanamycin (50µg/ml), gentamycin (7µg/ml) and tetracyclin (10µg/ml) (all from Sigma-Aldrich, Poole, UK). Bacteria were then centrifuged at 4000xg for 10 minutes and the supernatant discarded. Bacmid DNA was then prepared using a Qiagen plasmid mini-prep kit (section 2.2.3) and eluted into 40µl Tris-HCl pH 7.5.

### **2.7.9 Transfection of Sf9 Insect Cells With RAG-1 Bacmid**

Approximately  $9 \times 10^5$  Sf9 cells from a mid-log phase culture were placed in a flat-bottomed 25ml flask and allowed to settle for one hour at 27°C. At the same time serum-free media (SFM) (100µl) was mixed with previously prepared bacmid DNA (approximately 5µl) while further SFM (100µl) was mixed with Cellfectin transfection agent (6µl). The two SFM mixes were then pooled and left to stand for 45 minutes at 27°C.

After allowing the insect cells to settle media was drawn off and the bacmid DNA/Cellfectin/SFM mix added to the cells with an additional 2ml of SFM. The cells were then left at 27°C for 5 hours. The SFM was then drawn off and replaced with complete media (2ml) (appendix I). Cells were then grown for 72 hours.



### **2.7.10 Preparation of High Titres of Recombinant Baculovirus through Secondary and Tertiary Amplification.**

To generate large titres of recombinant baculovirus the initial virus generated through transfection with bacmid DNA was used to reinfect larger quantities of Sf9 cells. Typically reinfection takes place twice to generate a sufficient infected cell population for protein expression.

After transfection of the cells and 72 hours of cell/viral growth the media containing primary virus was removed and 1ml of primary virus used to inoculate further Sf9 insect cells (50ml at  $1 \times 10^6$  cells/ml). The cells were then grown for a further 3-4 days to yield secondary virus. Tertiary virus was then achieved by infection of further Sf9 cells (50ml), at the same concentration with secondary virus (2ml).

### **2.7.11 Expression and purification of Llama RAG-1/ Murine RAG-2**

Purification of llama RAG-1 protein in combination with murine RAG-2 was successful after co-infection of insect cells with tertiary virus containing both llama RAG-1 and murine RAG-2. The results of the purification of insect cell lysates, first through an affinity column containing Talon™ Superflow resin (Clontech, Basingstoke, UK) for purification of the expressed RAG-1 and RAG-2 proteins by virtue of their 6xHis tags resulted in protein bands on SDS-PAGE of acceptable purity (figure 4.9). Subsequent purification of the relevant RAG-containing fractions on a column containing amylose resin purified the protein further through interaction of the amylose with the maltose binding protein elements present within the expressed proteins. The high purity of the resulting protein was demonstrated by SDS-PAGE (figure 4.9).

#### **a) Co-Infection of Sf9 with Recombinant Llama RAG-1 and Murine RAG-2 Baculovirus**

Co-expression of llama RAG-1 and murine RAG-2 proteins was achieved through co-infection of Sf9 cells with 1ml each of tertiary stocks of both llama RAG-1 and murine RAG-2 encoding virus.

**b) Preparation of Llama RAG-1/Mouse RAG-2 Protein For Downstream Purification**

Tertiary viral stock (50ml) was centrifuged at 500xg for 10 minutes and the supernatant removed. The RAG protein-containing insect cell pellet was subsequently resuspended in nickel buffer A (20ml) (table 2.19) on ice. The resuspended pellet was then sonicated on ice for 2 minutes to lyse the insect cells. The resulting supernatant containing recombinant proteins was then poured into ultracentrifuge tubes and centrifuged at 40 000xg for 40 minutes to remove insoluble protein and cell debris.

• **Initial Purification on Cobalt Resin Affinity Column**

Purification on a cobalt affinity column separated the 6xHis tagged recombinant RAG proteins, which include a 6xHis tag from other insect cell proteins. The recombinant protein demonstrated an affinity for the cobalt resin at low imadazole concentrations (i.e. during binding and washing). Once the column was washed the recombinant protein was removed and collected by passing cobalt buffer B, containing a significantly higher imidazole concentration, through the column. The higher imidazole concentration leads to disassociation of the His-tagged protein from the column.

Prior to loading the column was initially washed through with cobalt buffer A (with open loading valve) for 10 column volumes (at a flow rate of 10ml/min for 1 minute). The resultant supernatant from the ultracentrifugation step was loaded onto a 9 mm diameter 'K' affinity column (Amersham-Pharmacia, Uppsala, Sweden) containing approximately 1.5ml (corresponding to 150mm depth) Talon™ Superflow Metal Affinity Resin (Clontech, Basingstoke, UK). Loading of the proteinaceous supernatant took place at 0.5ml/min. The column was then washed with a further 10 column volumes until a baseline UV reading was restored. Cobalt buffer B (table 2.19) levels were subsequently increased from 0% to 75% over a 10ml volume to elute the His-tagged proteins.

- **Secondary Purification on an Amylose Affinity Column**

To further purify the recombinant RAG proteins a second column containing amylose resin (New England Biolabs, Hitchin, UK) was prepared using column apparatus as described previously. The maltose binding protein sequence expressed as part of the recombinant protein allowed high affinity interaction between the amylose within the column and the recombinant RAG proteins. 1ml of resin was used to prepare the column. The protein-containing fractions collected during initial purification were loaded onto this column in the presence of amylose buffer A (table 2.19). The column was then washed in 10 column volumes of amylose buffer A before elution of the protein through washing in 100% amylose buffer B containing maltose.

Buffers used during affinity purification (volumes for 500ml buffer in parenthesis) are given in table 2.19

Step	Binding and Washing on Cobalt Column	Elution from Cobalt Column	Binding and washing on Amylose Column	Elution from Amylose column	Dialysis (for 2 litres)
Buffer	Cobalt Buffer A	Cobalt Buffer B	Amylose Buffer A	Amylose Buffer B	Dialysis Buffer
Composition	Tris-HCl 20mM (10ml 1M Tris-HCl (pH 7.5))	Tris-HCl 20mM (10ml 1M Tris-HCl (pH 7.5))	K-P pH 7.2 25mM (125ml 0.1M K <sub>2</sub> HPO <sub>4</sub> / KH <sub>2</sub> PO <sub>4</sub> )	K-P pH 7.2 25mM (125ml 0.1M K <sub>2</sub> HPO <sub>4</sub> / KH <sub>2</sub> PO <sub>4</sub> )	Tris-HCl 25mM (pH7.5) (50ml of 1M)
	KCl 0.5mM (18.6g KCl)	KCl 0.5mM (18.6g KCl)	KCl 0.5M (18.6g KCl)	KCl 0.5M (18.6g KCl)	KCl 150mM (22.36g)
	Imidazole 5mM (0.17g Imidazole)	Imidazole 1M (3.4g Imidazole)		Maltose 10mM (5ml 1M Maltose Solution)	
	Glycerol 10% (50ml Glycerol)	Glycerol 10% (50ml Glycerol)	Glycerol 10% (50ml Glycerol)	Glycerol 10% (50ml Glycerol)	Glycerol 10% (200ml)
	β-mercaptoethanol 2mM*	β-mercaptoethanol 2mM*	DTT 2mM* ( xg DTT)	DTT 2mM* ( xg DTT)	DTT 2mM*
	0.1% Triton* (5ml 10% Triton)	0.1% Triton* (5ml 10% Triton)	0.1% Triton* (5ml 10% Triton)	0.1% Triton* (5ml 10% Triton)	

Table 2.19 Buffers used during affinity purification of recombinant RAG proteins  
(All reagents in this table from Sigma-Aldrich, Poole, UK)

\*Add immediately prior to use

**c) Confirmation of Collection of Purified Recombinant RAG proteins**

To confirm successful purification of the recombinant proteins, 20µl samples from 0.5ml fractions collected off the affinity columns were run on precast 6% SDS – Tris Glycine minigels using minigel vertical electrophoresis apparatus (both from Invitrogen, Carlsbad, US). Prior to loading protein samples 5x SDS loading dye (appendix I) was added and the samples boiled for 20 minutes to denature the protein.

Gels were run at 120V for approximately 1.5 hours (figure 4.9)

## **2.8 Sequence Analysis**

All sequence analysis was performed using the GCG Winsconsin package or readily available web-based sequence analysis tools such as the Multalin multiple sequence alignment interface and a Transfac, transcription factor binding motif database search tool. The sequence analysis tools used in this report are outlined in table 2.20.

Analysis Type	Function	Program Used	Settings	References
<b>Splice Site Detection</b>	Determination of boundaries of intron/exons	Splice Site Detection by Neural Network ( <a href="http://www.fruitfly.org/seq_tools/splice.html">http://www.fruitfly.org/seq_tools/splice.html</a> )	'Human' Sequences	(181)
<b>Transcription Factor Binding Site Determination</b>	Determine the nature of possible transcription factors binding upstream of promoter sequences	MatInspector™ in combination with the Transfac database ( <a href="http://genomatix.gsf.de/products/index_mat.html">http://genomatix.gsf.de/products/index_mat.html</a> )	Core and matrix similarity thresholds of 1.0 and 0.90 (a measure of similarity to the established transcription factor binding motif sequence, 1.0 being identical) were used to filter out non-significant matches	(182, 183)
<b>Global Multiple Alignment</b>	Comparison between families of related genes and equivalent genes within different species	Clustal W, a component of the GCG Winstconsin Package, accessed through the eBioinformatics server ( <a href="http://www.bionavigator.com">http://www.bionavigator.com</a> )	DNA weight matrix: IUB Protein weight matrix: Blosum 62 Gap opening penalty: 10 Gap extension penalty: 0.05 End gap separation penalty: yes Gap separation distance: 8	(184)
<b>Phylogenetic Tree Generation</b>	Provide a graphical representation of the approximate evolutionary relationship between a number of genes or proteins	Programs from the Phylogeny Interface Program (PHYLIP), in combination with the Clustal W multiple alignment program accessed through the eBioinformatics server ( <a href="http://www.bionavigator.com">http://www.bionavigator.com</a> )	'DNADIST' Distance method: Kimura Transition/transversion ratio: 2.0 Base frequencies: empirical Coefficient of variation: 1  'NEIGHBOR' Tree building method: UPGMA Outgroup: 0 Subreplicates: no Number of jumbles: 0	(185)
<b>Local Pairwise Analysis</b>	Comparison of regions of homology between two genes or proteins	Blast 2.0	Default	(186, 187)

Table 2.20 Outline of Sequence Analysis Tools used throughout this study.

## Chapter 3 Germline Components of the Llama

### Immunoglobulin Rearrangement Process

#### 3.1 Abstract

The identification of high affinity antibodies consisting only of heavy chain variable domains not only led to the possibility of small single domain antibody fragments being engineered for use in a wide range of biotechnological applications, but also posed a variety of questions regarding the process of immunoglobulin rearrangement within the camelid family. The subsequent elucidation of heavy chain antibody structure and cDNA sequence only served to make these questions more pertinent.

How are the structural differences between classical and heavy chain camelid antibodies encoded? Are heavy chain antibodies merely a result of antigen receptor selection or are they specifically encoded in the germline sequence? Does somatic hypermutation take place during camelid B-cell development? Are there specific regulators of heavy chain antibody expression? Is the same set of D and J segments utilised in the generation of both antibody types?

This chapter addresses a number of these issues, using sequence information obtained through screening of a llama genomic phage library. Data generated through this screening not only indicates that heavy chain antibodies *are* specifically encoded in the germline, but also provides evidence for the utilisation of the process of somatic hypermutation during antibody development, the presence of specific transcription binding sites that may be key to the regulation of particular antibody families and the use of a closely linked D-J gene locus during both heavy chain and classical antibody generation.

#### 3.2 Introduction

The multiple biological mechanisms involved in classical antibody generation described in Chapter 1, hint at the plethora of complex, highly regulated processes that take place within a developing B-cell prior to the production of immunoglobulin. The understanding of the generation of antibodies in well-characterised systems such as human and mouse has been highly dependent on the characterisation of the genetic

elements within the germline that are ultimately responsible for encoding the antibody. This is not merely because such sequences reveal much about the amino acid composition, size and nature of the immunoglobulins generated. Indeed it is arguable that the genetic context in which such genetic elements are found within the germline, rather than the composition of the elements themselves, provides greater clues to the mechanisms of antibody generation as a whole.

The isolation of cDNA sequences that represent the sequence that encodes the final products of antibody generation in the llama (145, 180) suggests considerable overlap between the classical recombination pathways such as those well characterised in the human and mouse, and those of the *Camelidae*. cDNA sequence analysis clearly demonstrates the presence of conventional framework and complementarity-determining regions within the llama heavy chain variable domain.

Work described within this chapter has led to the identification of the crucial sequence elements responsible for encoding the final, recombined, processed and expressed llama antibody. Such information allows not only the determination of the degree of similarity between the sequence elements of the llama and those of better-characterised species but also enables a number of questions to be addressed regarding the llama heavy chain locus, in particular, regarding the formation of the unique heavy chain antibody. Sequence data also provides a platform from which the mechanisms of antibody generation can be explored in more detail.

The benefit of isolation of such sequences is not limited to increasing the understanding of the processes involved in llama antibody formation. Germline sequence information provides a scaffold of framework region sequences onto which diverse CDR repertoires can be grafted in order to construct large libraries of llama antibody fragments for screening against a wide range of commercially relevant antigens by phage display. This technique has previously proved successful in the generation of artificial human antibody fragment libraries (188, 189).

### **3.3 Aims of Library Screening**

Llama heavy chain antibodies contain numerous amino acid sequence differences with respect to their conventional IgG counterparts (145, 175). Elucidation of the full sequence of V gene segments therefore provides the opportunity to determine whether these sequence differences are encoded in the germline or are the result of other post-transcriptional processes. The relatively large size (~300bp) of the coding sequence of the variable region segments also made these regions ideal first targets for genomic library screening, therefore allowing:

1. The determination of the degree of similarity between camelid and murine/human V segments.
2. Confirmation of the existence of separate heavy chain only and classical V segments.
3. Examination of the nature of RSSs associated with the sequences.
4. Comparison of germline sequences with known llama cDNA sequences to give an indication of the level of somatic hypermutation taking place during antibody generation.

Although the V segment sequences comprise the majority of the variable domain sequence, the major antigen-binding interface is largely the result of H3 loop composition, which in turn is due to the sequence composition of the CDR3. The formation of the CDR3 sequence is, as previously discussed (section 1.8.5), the result of the juxtaposition of D and J gene segments and associated processes. The isolation of germline D gene segments provides the complementary 12-spacer recombination signal sequences that combine with 23-spacer RSSs derived from V and J gene segments. Together these RSSs form the targets for synaptic cleavage complex formation by the RAG proteins (chapters 1, 4 and 5). The J gene segments provide, in addition to further 23-spacer sequences, the donor splice sites that interact with acceptor splice sites adjacent to the constant region genes (chapter 6).

#### **3.3.1 Screening Strategy**

The heavy chain loci of well-characterised species such as the human and mouse contain considerable information beyond coding sequence alone. The sequence of variable (V), diversity (D) and joining (J) elements that together encode the principle antigen-binding loop of the antibody is complemented by other sequence elements. The recombination signal sequences guide sequence-specific V(D)J recombination, promoter and enhancer motifs recruit various transcription factors leading to



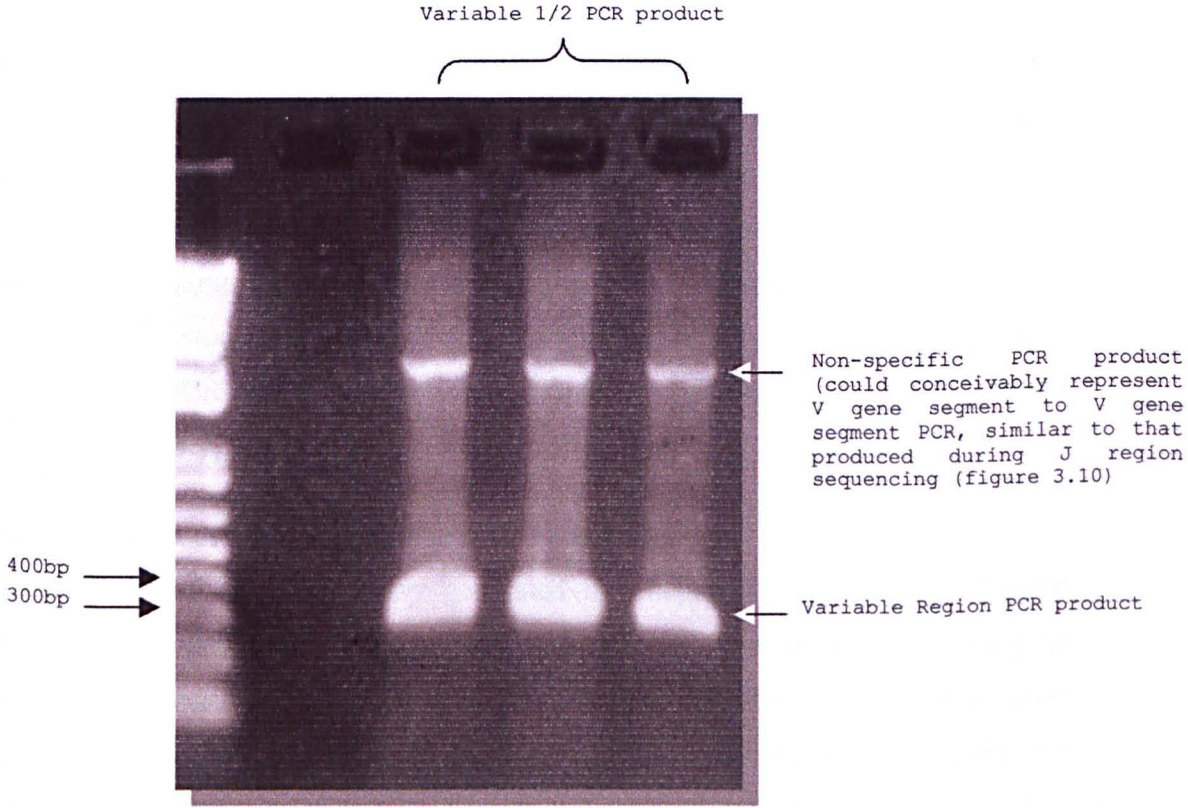
transcription of specific gene elements both prior to and post-recombination, while splice sites determine the manner in which the rearranged transcripts are processed before translation. All of these sequence elements therefore play key roles in guiding the processes that will lead to functional immunoglobulin production by the mature B-cell.

Isolation of bacteriophage clones described in this, and subsequent chapters utilised both the library constructed and described in section 2.3 and also a custom-made library constructed by Stratagene (La Jolla, California, US). A table summarising the number of clones derived from each library after each round of screening for each target gene is given in the appendix (Appendix VI). The production of the genomic library described in Chapter 2 provides a tool that enabled the isolation of immunologically relevant genetic elements from within the llama genome. The screening of genomic libraries pivots on the ability of small, labelled oligonucleotide probes to hybridise to specific clones within the library that contain this sequence. Therefore library screening provides a method for isolating large (5-15kb) regions of the llama genome provided the sequence of small regions is known.

The initial information used to design probes for library screening was obtained by careful analysis of cDNA sequences of llama classical and heavy chain immunoglobulin gamma sequences isolated previously within Unilever (180). The relationship between germline sequences and rearranged immunoglobulin sequence present within the cDNA database has been found to adhere to certain rules within other species. For example, the FR1, CDR1, FR2 and CDR2 sequence present within murine cDNA sequences is always derived from variable gene segments. The corresponding region of the llama cDNA sequence was therefore used to design a probe to isolate llama V gene segments (section 2.4 and figure 3.1). cDNA was amplified by PCR during probe generation (figure 3.2). Similarly, downstream regions of the cDNA sequence are generally derived from the D and J gene segments and these were used for probe generation in order to isolate these components. However the short length of the D and J regions (typically 10-15bp and 35-45bp respectively) (31) and the high levels of somatic hypermutation that are typical of classical immunoglobulin cDNA sequences made screening for D regions impossible, and screening for J regions challenging. The technical details of probe generation and screening methodology are described in Chapter 2.



Figure 3.2 PCR for generation of a variable region probe using primers Variable 1 and Variable 2. The band of the correct size was cut out and gel purified before radiolabelling



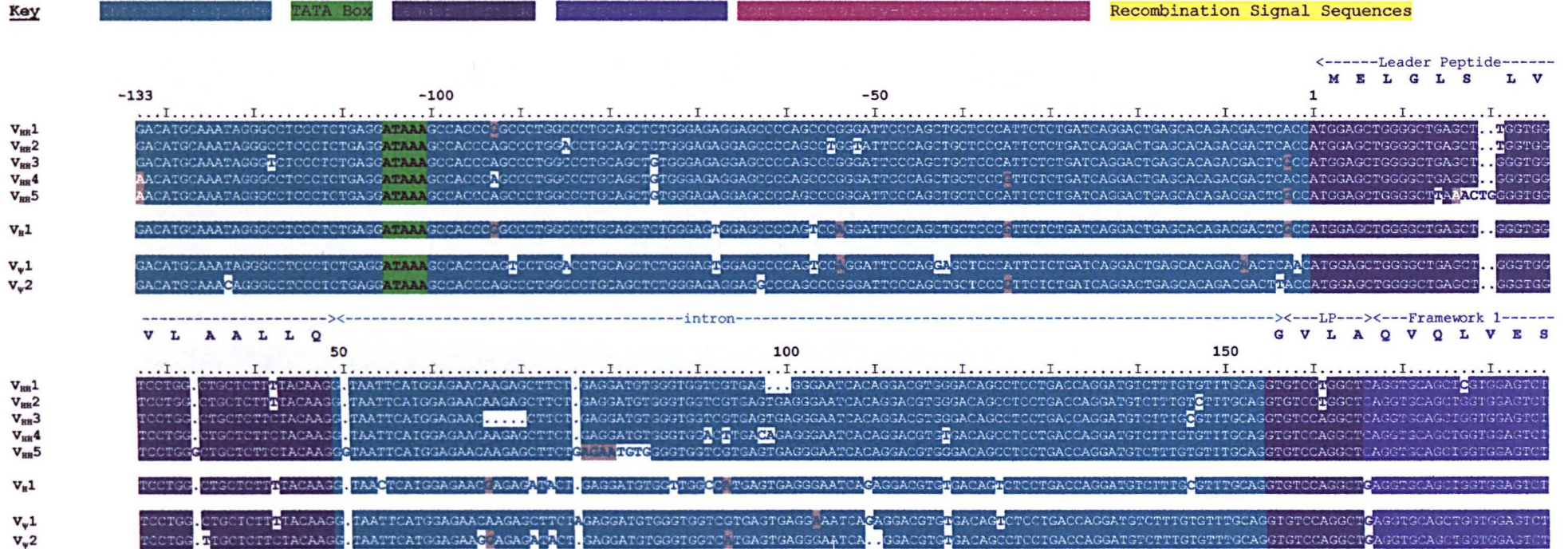
### **3.3.2 Problems of Direct sequencing**

The original strategy for the characterisation of isolated phage clones containing V, D and J elements involved the purification of phage DNA followed by direct sequencing of phage DNA using primers derived from the probe sequences (table 2.3 and table 2.13). Although this strategy proved successful in sequencing the majority of clones isolated, a number of V region clones and the single J region clone could not be directly sequenced. This is most likely due to high sequence homology between tandem gene segments within the heavy chain locus. Restriction analysis of the genomic library (data not shown) suggests an average insert size of approximately 13kb. If it is assumed that the llama heavy chain locus adopts a similar gene layout to that of the human, it is quite conceivable that multiple V (or J) gene segments may be present within the same phage clone. Indeed, human V gene segments are clustered as close as ~500bp apart within the human germline (31). Distances between J gene segments are even shorter (in the 100bp range). However, sufficient data was obtained from the eight single V gene segment clones isolated to make lengthy restriction mapping, southern hybridisation and sub-cloning unnecessary for the purposes of this study. The short distances typically separating J gene segments led to an alternative PCR-based strategy for successful sequencing of this clone (section 3.15 and figure 3.9).

### **3.3.3 Isolation of Eight Llama Variable Gene Segments**

Screening of the genomic library generated in Chapter 2 with a cDNA-derived variable gene segment probe led to the full characterisation of eight variable gene segments (Genbank accession numbers AF305944-AF305951). Variable gene segments are the most 5' elements located within human and murine heavy chain loci and encode the FR1, CDR1, FR2, CDR2 and FR3 of the variable domain. The nucleotide sequence of these clones is given in figure 3.3.

**Figure 3.3 Global Alignment of Germline V region gene segments with Amino Acid Sequence of V<sub>H</sub>1 for comparison.** Generated using the ClustalW algorithm (184). Purine-Purine differences in this and subsequent multiple alignments in this thesis are shaded grey





### **3.4 Analysis of the Coding Potential of Isolated Llama Germline Variable Gene Segments**

Specific sequence differences within the nucleotide sequences of variable gene segments isolated in this thesis indicate their potential to encode different antibody types. Five of the eight clones isolated contain sequence typically found to encode heavy chain antibodies, while only one gene segment has the potential to encode the classical immunoglobulin gamma that is also present within llama sera. This provides the first direct evidence that specific sequences within the llama genome are responsible for the generation of heavy chain antibodies (figure 3.3). These findings remove any possibility that such antibodies could have resulted merely by other immunoglobulin-related mechanisms such as somatic hypermutation and antigen receptor selection. The final two clones represent pseudo-gene segments by virtue of stop codons present within their predicted coding sequence.

### **3.5 Detailed Characterisation of $V_H$ and $V_{HH}$ Sequences**

The overall layout of each isolated variable gene segment is similar to that of human and mouse gene segments. The eight isolated clone sequences each span a region approximately 250bp upstream of the conserved Oct domain through to the 3' recombination signal sequences (figure 3.3). All eight clones contain a conserved TATA box motif upstream of the initiation (ATG) codon, from which the leader peptide and variable gene are encoded. The leader peptide is interrupted by a short intron of between 99-106bp.

As discussed previously, of the eight full sequences isolated two ( $V_{\psi 1}$  and  $V_{\psi 2}$ ) contain stop codons, within the FR2 and FR3 regions respectively and are therefore considered pseudogenes. Of the remaining six sequences five ( $V_{HH1}$ - $V_{HH5}$ ) have been assigned as heavy chain variable genes ( $V_{HH}$ ) by virtue of Phe37 or Tyr37, Glu44 and Arg45 (and Leu47 or Gly47 in four of five cases) (section 3.6, (145)). The remaining clone  $V_{H1}$  is of classical origin by virtue of Val37, Gly44, Leu45 and Trp47.

### 3.6 Amino Acid Composition of Variable Gene Segments

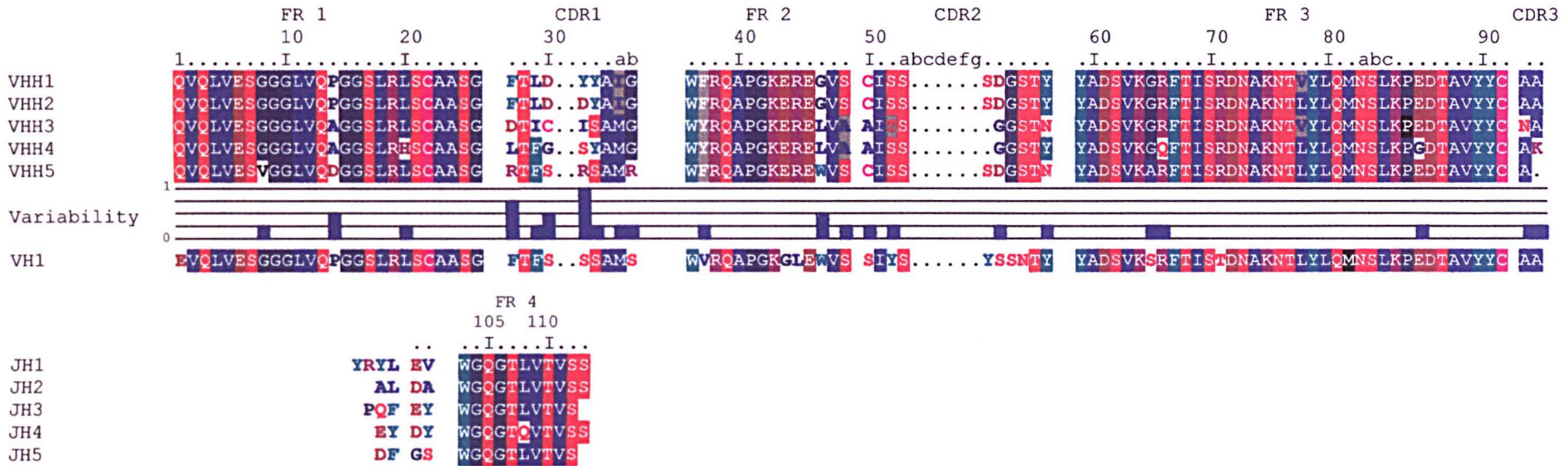
A number of differences in amino acid composition define llama heavy chain variable antibodies (145, 175). Examples of these substitutions are present within the germline gene segments isolated in this study. An alignment of V gene segment amino acid sequences is given in figure 3.4. The germline llama heavy chain sequences reported here match similar llama and dromedary cDNA and germline sequences previously published (145, 175, 190) in possessing a Pro residue at position 84 (typically Ala in conventional  $V_H$  domains) and in two cases ( $V_{HH3}$  and  $V_{HH4}$ ) an Ala residue at position 16 (typically Pro). These residues are located within the turns between  $\beta$ -sheets at opposite diagonals of the variable domain. A Leu residue at position 11 of conventional antibody heavy chains is normally responsible for interactions with Phe149 and Pro 150 of the  $C_H1$  domain. This amino acid is conserved in all six llama germline clones despite the lack of  $C_H1$  domain typically associated with the five  $V_{HH}$  sequences. The presence of Leu11 has also been noted in llama  $V_{HH}$  cDNA sequences (171, 191) and is thought to represent a species specific difference between llama and camel V domains (where Ser11 rather than Leu11 is frequently found). While the 5  $V_{HH}$  germline segments presented in this thesis are unlikely to encode the majority of expressed  $V_{HH}$  domains, the lack of the Ser/Leu substitution is in contrast to the results of cDNA analysis made by the Muyldermans group (175).

Llama and camel heavy chain sequences also exhibit a number of amino acid differences with respect to conventional IgG within the region of the variable domain responsible for contact with the  $V_L$  domain (173). Of particular note are the Val/Tyr37Phe, Gly44Glu, Leu45Arg substitutions within all five  $V_{HH}$  gene segments and Trp47Gly present in  $V_{HH1}$  and  $V_{HH2}$ . Other residues involved in  $V_L$  association (including Gln39 and Tyr91) remain unmodified within the  $V_{HH}$  gene segments reported here.

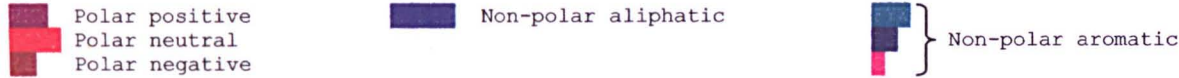
The presence of Cys50 within  $V_{HH1}$ ,  $V_{HH2}$  and  $V_{HH5}$  suggest that these gene segments may be responsible for interacting with long hinge heavy chain antibodies of the IgG2b isotype as cDNA sequences commonly associate this residue with this isotype (175). Interestingly, of the two  $V_{HH}$  clones that lack Cys50 within the CDR2,  $V_{HH3}$  contains Cys30 within the CDR1, which may provide a site for disulphide bond



**Figure 3.4 Alignments of germline llama  $V_H/V_{HH}$  and J gene segments.** Numbering is according to a modified Kabat scheme (30). An indication of  $V_{HH}$  variability at each residue is given as a blue bar chart below the  $V_{HH}$  gene segment alignment. Similar residues are shaded in the same colour as given in the key below.



**Key**



- Notes
- i) All sequences aligned using modified Kabat (30) numbering.
  - ii) Alignment using ClustalW (184,192) and the Blosum 62 scoring matrix

formation in the absence of Cys50. A cysteine is present at this position within a llama short hinge (IgG3) cDNA sequence in the work of Vu *et al* (175) (referred to as IVHH56 in this work) in the absence of the more common Cys33. Cys33 is not, however, present within the germline sequences isolated here. This therefore provides a precedent for the use of a gene segment encoding Cys30 in llama heavy chain antibody formation.

Overall the predicted amino acid sequences derived from the isolated germline sequences reported here concur with and confirm the residue differences between camelid heavy chain and classical immunoglobulins reported elsewhere (175, 193). It is clear therefore that most of these residues are germline-encoded, and not seen within cDNA populations simply as a result of the structural constraints of functional heavy chain antibody production.

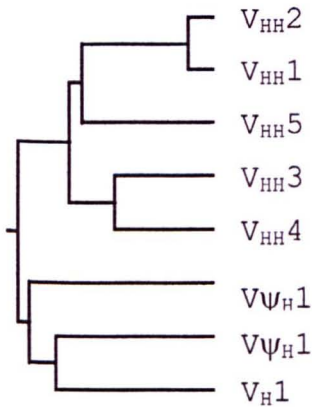
### **3.7 Analysis of Variable Gene Segments by Comparison to those of Other Species.**

If the mechanism by which heavy chain antibodies are generated is to be determined it is important to gain an appreciation of the level of similarity between the sequences that encode them and classical immunoglobulin sequences. High similarity would suggest considerable overlap in the mechanisms of generation whereas regions of complete disparity might suggest major differences in downstream recombination and expression mechanisms.

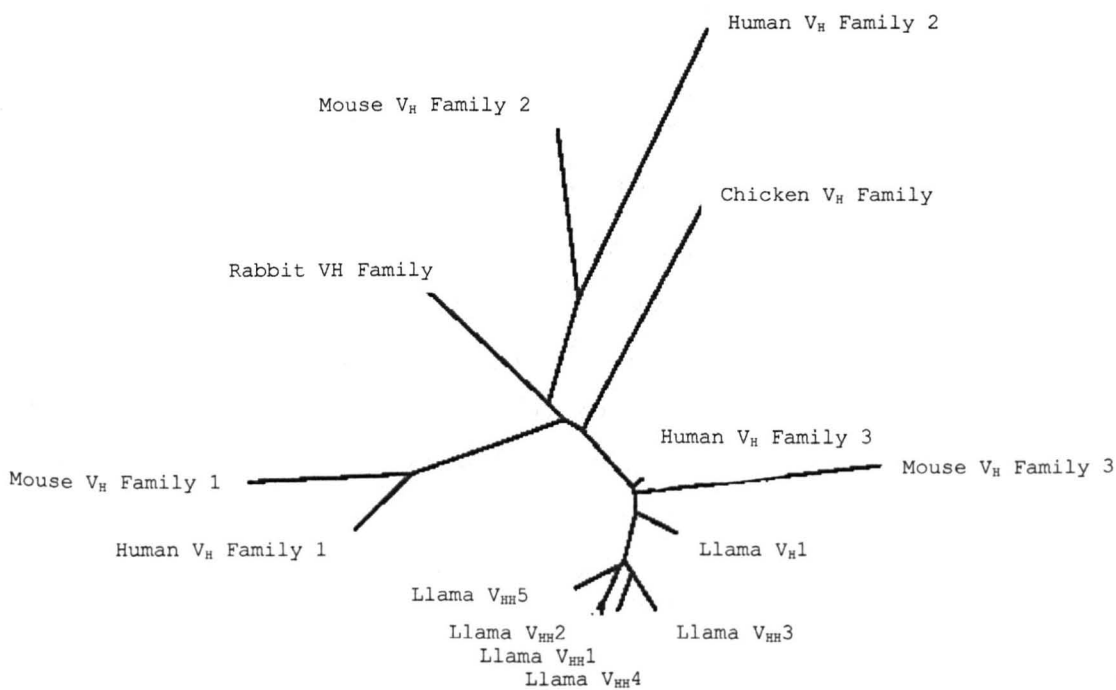
The sequences show a high level of identity at protein and nucleotide level, V<sub>HH1</sub> and V<sub>HH2</sub> demonstrating highest homology (98.3% identical at nucleotide level) (Table 3.1 and figure 3.5). The greatest disparity between heavy chain genes is between V<sub>HH4</sub> and V<sub>HH5</sub>, which share only 78.4% identity at amino acid level. The average nucleotide identity between heavy chain only variable gene segments is 91.2% while, perhaps unsurprisingly, the mean nucleotide identity between V<sub>HH</sub> and V<sub>H</sub> encoding gene segments is only 87.8%. As with previous llama V<sub>H</sub> and V<sub>HH</sub> sequences the six isolated clones show greatest similarity to human and mouse V<sub>H</sub> subgroup III members (194). The consensus sequence derived from the genomic clones shows considerable similarity to families of classical V gene segments in other species. The second phylogenetic tree generated below (figure 3.6) summarises the relationship

**Table 3.1 Identity (%) Matrix for Variable Gene segments using coding sequence only.** Blue figures represent nucleotide identity while red figures illustrate amino-acid identity.

	V <sub>HH1</sub>	V <sub>HH2</sub>	V <sub>HH3</sub>	V <sub>HH4</sub>	V <sub>HH5</sub>	V <sub>H1</sub>
V <sub>HH1</sub>	x					
V <sub>HH2</sub>	98.3 97.9	x				
V <sub>HH3</sub>	89.1 83.5	88.8 82.5	x			
V <sub>HH4</sub>	89.8 83.5	90.1 84.5	93.5 87.5	x		
V <sub>HH5</sub>	91.2 86.6	91.2 87.6	89.8 82.5	90.1 78.4	x	
V <sub>H1</sub>	88.1 79.4	87.4 80.4	87.4 77.3	88.1 78.4	88.1 81.4	x



**Figure 3.5** Rooted phylogenetic tree examining the relationship between isolated variable gene segments. Tree generated by generation of a multiple sequence alignment



**Figure 3.6 Unrooted phylogenetic tree illustrating the evolutionary distance between the coding sequence of variable gene segments isolated in this study with the corresponding sequence derived from the variable domain families of other species. Distance between ends of lines represents a measure of evolutionary distance. Genbank accession numbers clockwise from human family 1 P01743, J00530, M93173, K01569, AB019440, A33334, S31108, X01113, AF305949, AF305946, AF305944, AF305947, AF305945 and AF305948.**

between the various  $V_H$  families of human, mouse and other species with the consensus of the heavy chain antibody and classical sequences isolated. This analysis supports the notion that germline heavy chain variable segments differ only to enable the encoding of antibodies that can interact with antigen through a single variable domain. The similar exon layout and high levels of nucleotide and amino acid identity to llama classical variable gene segments is suggestive of a similar antibody secretion mechanism. High nucleotide and amino acid identity also suggests that the heavy chain variable gene segments have evolved gradually, and relatively recently from a precursor (or number of precursor) classical variable gene segments.

### **3.8 Segregation of Isolated Variable Gene Segments into Specific Variable Gene Families**

Llama heavy chain only cDNA sequences can be broadly divided into three families by virtue of their amino acid composition (191). Each family has unique characteristics and a propensity for the formation of particular loop structures within the antigen binding interface (195). Although germline sequences cannot be directly classified in this way as they are not subject to the structural constraints that an expressed cDNA must conform to, it would appear that specific germline gene segments contain the code for many of the key amino acids within this three family classification. Of these clones,  $V_{HH1}$  and  $V_{HH2}$  are the most readily classified and contain sequence that encodes six of eight key residues of Family 3 (Phe27, Asp30, Tyr/Asp31, Ile34, Ser49 and Cys50). The potential cDNA families encoded by  $V_{HH2}$ ,  $V_{HH3}$  and  $V_{HH4}$  are less clear although they demonstrate limited coding potential essential to Family 1, Family 2 and Family 1 respectively. The classification of germline gene segments in this manner may relate particular variable segments to the production of antibodies able to bind particular classes of antigen. Although the classification system is both broad and prone to exceptions the use of a family classification correlates with findings from the study of variable region promoter regions which are reported later in this chapter (section 3.9) suggesting that the B-cell may be able to control the type of antibody that is both rearranged and transcribed, perhaps to provide optimum protection against different antigen types.

### 3.9 Analysis of Llama Variable Gene Segment Promoter Regions

Classical immunoglobulin variable region promoters typically contain a number of motifs through which a variety of transcription factors can bind and regulate transcription. This section discusses the motifs found within the promoters of the germline sequences isolated in this chapter. Examination of upstream promoter sequence derived from heavy chain variable gene segment promoters demonstrates the presence of the well-characterised and highly conserved TATA promoter and Oct-1/2 binding motifs that are common to the immunoglobulin variable gene promoters of many species (section 1.7.4). In addition, further motifs corresponding to transcription factor binding sites are present upstream of all the variable gene segments (table 3.2). These include an Ikaros family binding site. The Ikaros family of transcription factors is believed to play a key role in the early lineage determination as cells differentiate from haematopoietic stem cells to lymphocytes, and particularly at the stage of B-cell/T-cell/NK-cell differentiation (196). Overall sequence conservation is high from the Oct binding site through to the transcription start site (-170 to -1). It is clear from the alignment given in figure 3.7 that the region upstream of the Oct motif (-282 to -171) lacks this high conservation and divides the  $V_{HH}$  promoters into two groups. The first of these groups includes heavy chain family 3-related (section 3.8) gene segments ( $V_{HH1}$  and  $V_{HH2}$ ) in addition to the classical-derived  $V_{H1}$  sequence.

Analysis of the promoter regions was performed using a web-based analysis tool Matinspector Professional™ (Genomatix, München, Germany) in concert with the Transfac database of transcription factor motifs (table 3.2 (182, 183)) and indicates a number of other possible transcription factor binding sites specific to individual V gene segments. It is important to remember that transcription factor binding site analysis is prone to the identification of false positives, and that results are typically suggestive rather than indicative of a particular binding site. However a number of different binding sites upstream of the Oct domain suggest differential regulation of transcription of  $V_{HH}$  gene segments.  $V_{HH1}$  contains two motifs of interest, CCAAT/enhancer binding protein beta (C/EBP  $\beta$ , position -250 to -226 of figure



<u>Putative Transcription Factor Binding Motif</u>	<u>Ref</u>	<u>V<sub>HH</sub>1, V<sub>HH</sub>2 and V<sub>H</sub>1</u>	<u>V<sub>HH</sub>3-V<sub>HH</sub>5</u>
Octamer Binding Factor 1 CATGCAAATA	(20)	Present Nucleotides -122 to -129	Present Nucleotides -122 to -129
Transcriptional Repressor CDP/CLOX GAAATTATCGATTTT	(197)	Not Present	Present Reverse Strand Nucleotides - 240 to -254
Activator Protein 1 AGTGACCCAAT	(198)	Not Present	Present Reverse Strand V <sub>HH</sub> 4 only -180 to -200
TATA box-like domain GGATAAAGCCA	(199)	Present -96 to -106	Present -96 to -106
Activator Protein 4 CACAGCTGCT or AGCAGCTGGG	(200)	Present Reverse Strand -38 to -47  Present Forward Strand -189 to -199	Present Reverse Strand -38 to -47  Not Present
Ikaros 2 AGCTGGAATCC	(201)	Present Reverse Strand -41 to -52	Present Reverse Strand -41 to -52
CCAAT/enhancer binding protein beta ACTTCCAGAAATTA or AATTCTGGAAGTT	(202)	Present Reverse Strand -226 to -247 (-250 for V <sub>HH</sub> 1)	Not Present
Fork head related activator-2 ANNGTAAACAA	(203)	Present V <sub>HH</sub> 1 only -243 to -261	Not Present
E47 CCTGCAGCTGTGGGA	(204)	Present -69 to -83	Present -69 to -83
v-Myb CCCAGCTGCTCCCGTTCTC	(205)	Not Present	Present Reverse Strand V <sub>HH</sub> 4 only -29 to -48

Table 3.2 Potential transcription Factor Binding Motifs within the Promoter Regions of Isolated Llama Germline Variable Gene Segments





3.7)) and Related Activator 2 (FREAC 2, position -261 to -246 of figure 3.7)). C/EBP  $\beta$  is characterised by its role in interleukin-1 (IL-1) stimulated IL-6 expression during the acute-phase of inflammatory reactions lying between nucleotides -122 and -129 of the germline clones, and the related activator 2 motif (FREAC 2) that is involved in DNA bending (203) (position -261 to -246).  $V_{HH3-5}$  contain a transcriptional repressor motif, CLOX/CDP 2, not found within the other upstream regions (197). The full results of this analysis are shown in table 3.2, and the corresponding alignment of promoter sequences in figure 3.7. In conclusion it is clear that llama variable gene segments are associated with at least two distinct patterns of transcription factor binding motif pattern. However, all the gene segments isolated in this study also contain well-characterised and highly conserved elements found within the promoters of variable gene segments in other species. Crucially no major differences were found between the promoters of the isolated classical gene segment and the heavy chain gene segment promoters corresponding to heavy chain variable family 3. This suggests that differences in transcription factor binding sites are not responsible for simple distinction between heavy chain and classical antibody transcription. Overall this analysis strengthens the findings of the evolutionary analysis (section 3.7) concluding that heavy chain gene segments are highly similar and closely related.

### **3.10 Comparison of Llama and Murine Recombination Signal Sequences**

As discussed in Chapter 1 (section 1.8.3) classical immunoglobulin generation is reliant on specific targeting of the cellular recombination machinery to short signal sequences known as recombination signal sequences or RSSs. The variable region gene segments isolated in this study are all associated with RSS similar to those found associated with human and mouse variable gene segments (52). In all cases a standard heptamer (CAC(A/T)GTG) and nonamer are separated by a 23 nucleotide spacer sequence. The RSSs of the heavy chain gene segments ( $V_{HH1}$  to  $V_{HH5}$ ) show greater than 90% identity to one another while identity to the classical gene segment ( $V_{H1}$ ) RSS is only slightly lower. Previous work, reported whilst this study was in progress, examining camel variable gene segments has described the presence of heptamer-like sequences within the coding sequence of the variable gene segments

(206) and has led to the suggestion that such cryptic recombination signal-like sequences may have a role in the generation of short codon deletions found in variable cDNA but not germline gene segments (section 3.11). As shown in the alignment given in figure 3.3 no significant differences are present between the recombination signal sequences associated with classical and heavy chain germline gene segments. However the first few nucleotides of the coding flanks associated with all of the variable segments isolated in this chapter are different to those used to demonstrate recombination in other species (Chapter 5). The similarity between classical and heavy chain RSSs and of llama RSSs to those found in other species further confirms the overall similarity of the llama heavy chain variable segment to classical variable gene segments.

### **3.11 Examination of Llama Variable Gene Segments for the Presence of Heptamer-Like Sequences**

Recent work from Serge Muyldermans and co-workers has led to the proposal that an insertion/deletion mechanism involving recombination signal sequence heptamer-like sequences acts during heavy chain antibody generation (190). This group noticed that germline variable gene segments isolated from the dromedary (another camelid species expressing heavy chain antibodies) frequently contained sequences similar to the CACAGTG recombination signal heptamer consensus (section 1.8.4). Alignment of germline sequences containing these heptamer-like signals with similar cDNA sequences by this group showed that the nucleotides immediately upstream of the heptamer-like sequence within germline sequences were typically deleted at corresponding positions of similar cDNA sequences. Furthermore they noted that the frequency of the occurrence of these heptamer-like sequences was considerably higher within heavy chain gene segments than classical segments (most heptamer-like sequences were found within the framework 3 region). These authors went on to suggest that a RAG-mediated insertion/deletion mechanism may be at play leading to increased diversity within heavy chain antibodies. Heptamer-like sequences may predispose sequence immediately upstream to a RAG-mediated gene replacement mechanism similar to a heptamer-mediated mechanism previously reported in both the mouse and human (207, 208).

An analysis of the germline sequences reported in this chapter was made to search for the presence of heptamer-like sequences similar to those reported in the dromedary. This was accomplished through use of the GCG-based 'FindPattern' program (GCG Group, Wisconsin, USA) (using the search string CAC(A,T)GTG) and revealed only one heptamer-like site. This site was present in all of the llama  $V_{HH}$  and  $V_H$  germline gene segments at the same position (within the FR3). Higher numbers of such motifs were, however, located on the reverse strand where it is also conceivable that RAG proteins could act. Alignment of isolated germline sequences with those from the llama cDNA library (180) similar to that described in section 3.18 has revealed no evidence of an insertion/deletion mechanism (data not shown).

Were such a RAG-based mechanism to play an active role in llama heavy chain immunoglobulin generation it must do so without either interaction with a spacer or nonamer element as these are not present either in camel or llama gene segments. The most probable explanation for the observations of the Belgian group is that the presence of the heptamer-like element predisposes to some ill-defined nicking event (perhaps involving RAG proteins, that are recruited to specific RSSs sequences in the vicinity). RAG mediated nicking may then lead to insertions and deletions as part of the normal DNA repair mechanism. As discussed in section 1.8.5 there is considerable overlap between recombination mechanisms and those responsible for DNA repair.

In the absence of a greater number of llama variable gene segments it is impossible to confirm the finding that greater numbers of heptamer-like sequences are found within heavy chain gene segments. No further investigation of this mechanism is made in this thesis.

### **3.12 Analysis of Variable Gene Segment Variability.**

To generate a large repertoire of immunoglobulin variable domains the llama requires a diverse range of variable gene segments. An estimate of  $V_{HH}$  gene segment amino acid variability accompanies the V gene segment amino acid alignment given in figure 3.4. Even within the limited number of sequences reported in this thesis variability is notably higher within the CDR regions previously defined from llama cDNA sequences, suggesting that a considerable proportion of the overall antibody variability is encoded in the germline.

### **3.13 Examination of Llama Gene Segments for the Presence of Hypermutational Hotspots within V Region Sequences**

As discussed previously (section 1.9.1) a number of key nucleotide motifs have been described within germline variable region genes that predispose to somatic hypermutation in their vicinity. Of particular significance are the triplets AGY and TAY (where Y= C or T) (209, 210). To ascertain the importance of these hypermutational hotspots on llama cDNA sequence the occurrence of such hotspots within germline clones can be superimposed on a cDNA variability plot derived from the Unilever cDNA library (figure 3.8). This clearly demonstrates a positive correlation between the occurrence of hotspots within germline sequence and cDNA variability at corresponding positions. Both hypermutational hotspots and cDNA variability are concentrated within the variable domain CDRs. It seems likely therefore that both classical and heavy chain gene segments are targets for a similar hypermutational process to that undergone by human and murine variable gene segments.



### 3.14 Variable Gene Segments – Summary of Findings

Subsequent sections within this chapter present further sequence data derived from additional variable gene components. Prior to the presentation of further sequence data it is useful to summarise the results of the analysis of variable gene segments performed above:

- Separate germline gene segments encode heavy chain and classical variable domains.
- Heavy chain and classical gene segments share near-identical promoter/exon/intron/RSS layouts, which are also highly similar to previously characterised variable gene segments in other species.
- While the classical variable gene segment shares the least identity with the other isolated germline sequences, sequence identity between variable gene segments is generally high at both nucleotide and amino acid level.
- There are no discernible differences between the recombination signal sequences or promoters of classical and heavy chain gene segments isolated.
- Heavy chain gene segments loosely fall into families previously described to categorise heavy chain cDNA sequences.
- Although classical and heavy chain promoters exhibit no major differences, promoter sequences can be divided into two distinct families that correspond to the heavy chain family classification.
- Somatic hypermutation catalysed by ‘hotspot’ triplets may well take place at llama variable gene segments.
- Heavy chain variable gene segments are highly variable, particularly within the CDR regions.
- No evidence for a previously reported heptamer-like insertion/deletion mechanism could be found.

### **3.15 An Alternative Strategy for J Region Sequencing**

Isolation of joining and diversity gene segments was not as straightforward as the isolation of variable gene segments. As discussed previously, the typically high sequence homology between multiple J gene segments in heavy chain immunoglobulin loci of most species make direct sequencing of clone DNA with primers specific to J segment cDNA sequences near impossible. Successful sequencing was only possible through the use of primers specific to flanking non-coding regions surrounding J segments (figure 3.9 and 3.10). However, such sequences were not present within the cDNA library and therefore not immediately available. To determine the nature of such flanking sequences a novel PCR strategy was employed. Phage clone DNA was amplified using a single primer (J sequencing 1) and its reverse complement (J sequencing 2) (table 2.3). Both primers were specific to the FR4 consensus sequence derived from the Unilever cDNA library and therefore correspond to the region that would typically be encoded by J gene segments during conventional recombination processes. This led to the generation of PCR products spanning introns between tandem J gene segments. Using this strategy it was possible to determine the nature of non-coding sequence lying between homologous J gene segments. By determining the nature of the sequence between adjacent J gene segments successful direct sequencing of J gene segment clone DNA was made possible. Full sequencing of the D-J locus was made possible by design of overlapping sequencing primers to 'walk' across the locus. The full D-J locus sequence and corresponding exon translations are shown in appendix IV.

Figure 3.9 Alternative Strategy for the isolation of J region gene sequence

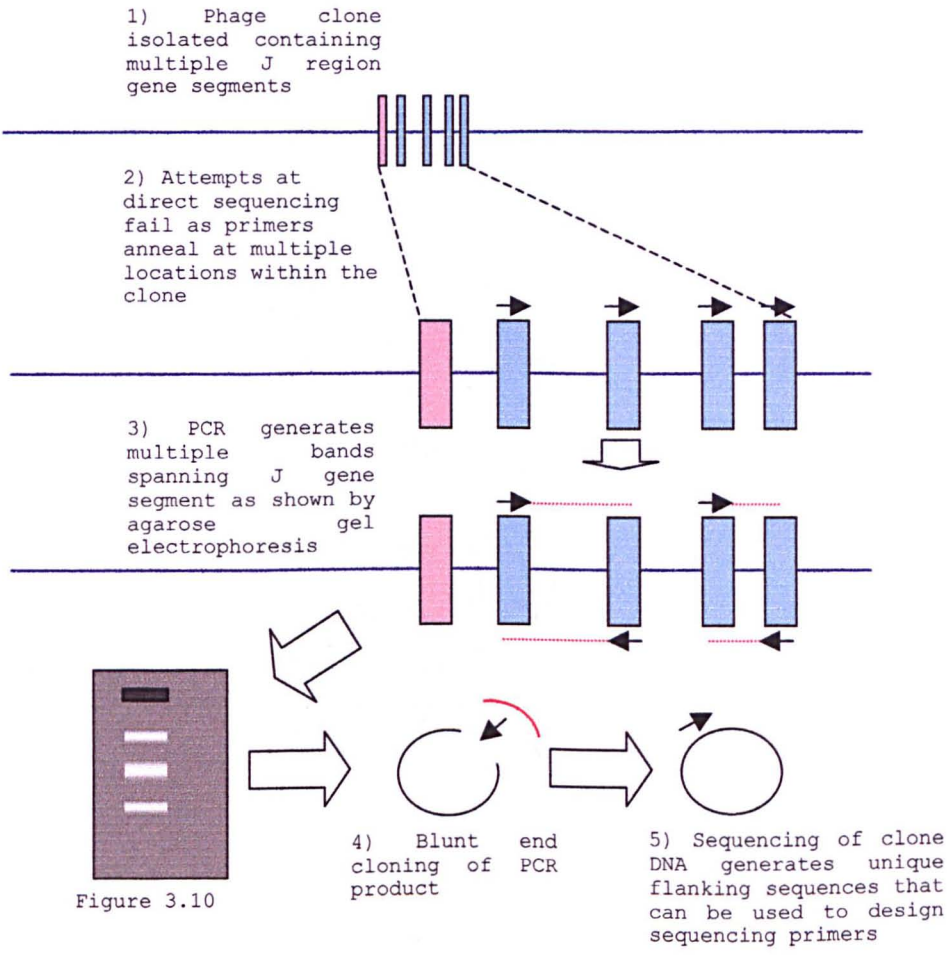
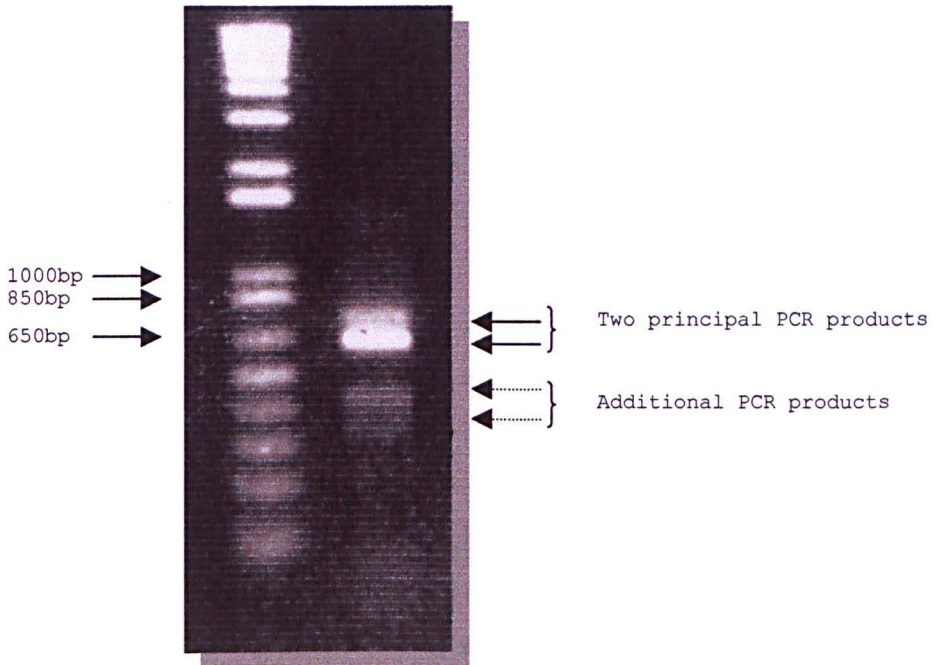


Figure 3.10



**Figure 3.10** Alternative PCR-based strategy in order to sequence a J region bacteriophage clone using *Taq* polymerase. Primers were designed to amplify regions between J region gene segments, resulting in a ladder of PCR products. PCR products were cut-out, gel purified and cloned using a TA-cloning kit (Invitrogen, Carlsbad, US).



### 3.16 The Isolation of Diversity Gene Segments

Two further genetic elements responsible in part for encoding the llama  $V_H$  domain were isolated through the screening strategies described earlier (section 3.3.1).

Downstream of the variable gene segments within classical immunoglobulin heavy chain loci, lie a number of diversity or D gene segments. A single diversity gene segment ( $D_{HL1}$ ) has been isolated (figure 3.11). This sequence includes a reading frame of 11bp.  $D_{HL1}$  shows considerable nucleotide sequence homology to DHQ52, the human sequence most 5' proximal to the human J gene segment locus (figure 3.12). Indeed,  $D_{HL1}$  was derived from a J region genomic library clone (section 3.1.7). Interestingly DHQ52 is the only D region segment to have a close homologue in other species. It would seem that conservation of this sequence therefore extends to the llama, providing further evidence for the high level of conservation between llama and murine/human variable gene elements. The RSSs associated with  $D_{HL1}$ , like those of the variable gene segment, conform to known consensi and take the heptamer, 12-spacer, nonamer conformation typical of heavy chain D gene segments in other species.

Figure 3.11 Sequence of Llama Diversity (D) gene segment D<sub>B</sub>L1 (Genbank Accession No AF305951). Translations of the three possible reading frames are given below.

	nonamer	12-spacer	heptamer	coding	heptamer	12-spacer	nonamer
D <sub>B</sub> L1	GGTTTGGC	TGAGCTGGGAAC	CGCAGTG	CTAACTGGAGC	CACAGTG	ACTGACAACCTCT	ACAAAAACT
Frame 1				N W S			
Frame 2				* L E			
Frame 3				L T G			

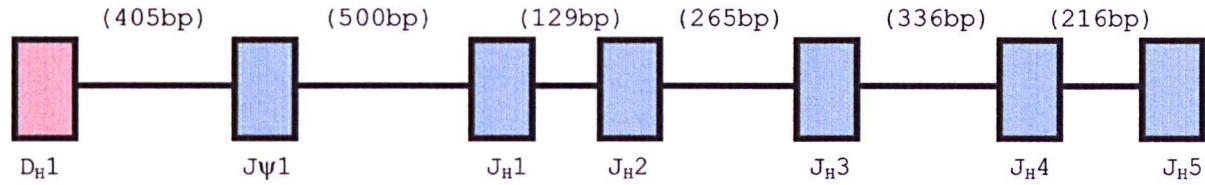
Figure 3.12 Output from BLAST Comparison (186, 187) of D<sub>B</sub>L1 with human D gene segment DHQ52 (Accession No X97051). Vertical lines '|' indicate identical nucleotides

	nonamer	12-spacer	heptamer	coding	heptamer	12-spacer	nonamer				
D <sub>B</sub> L1:	ggttttg	tgagctg	ggaac	cgcagtg	ctaactg	ggagc	cacagtg	actgaca	actct	acaaaaa	act
D <sub>H</sub> Q52:	ggttttg	tgagctg	gagaac	cactgtg	ctaactg	gggga	cacagtg	attggc	agctct	acaaaaa	ac

### 3.17 The Isolation of Joining Gene Segments

The most 3' set of variable domain components present within the heavy chain immunoglobulin locus is the joining gene segments, the final gene segments to be reported in this chapter. Screening of the genomic phage library led to the isolation of a clone containing a number of J gene segments ( $J_{HL1-5}$ ) and an associated D gene segment ( $D_{HL}$ , section 3.16). The organisation of the D-J locus is given in figure 3.13 while the full nucleotide sequence can be found in appendix IV (Genbank accession no AF305951). J gene segments were identified not only by coding sequence and 5' 23-spacer RSS signal but also by the presence of a strong splice donor sequence at the 3' end. The presence of RSS and splice donor sequences also identified a J segment-like pseudogene ( $J_{\psi 1}$  is located upstream of  $J_{H1}$ ). The five functional J segments are shown in figure 3.14.

**Figure 3.13 Organisation of the llama D-J locus** (full sequence in appendix IV)  
Distances separating gene segments given in parenthesis



**Figure 3.14 J Gene Segments and Translations**

	Nonamer	12-Spacer	Heptamer	Coding	Splice
JH1	GGGTTTGTG	CACTGGGGCCAGGCAGGCAGAC	CAGTGTG	GCTACAGGTATCTCGAAGTTTGGGGCCAGGGCACCCCTGGTCACTGTCTCCTCAG Y R Y L E V W G Q G T L V T V S S	GT
JH2	GGTTTATGT	CTGGGGGAGAGCCGGGACTATGT	CCCTGTG	CA...ATGCTTTGGACGCATGGGGCCAGGGACCCCTGGTCACTGTCTCAG A L D A W G Q G T L V T V S S	GT
JH3	GGTTTTTGC	ACAGCACCTAACGGGGCCCGTGG	CGCTGTG	AT.....GAGTATGACTACTGGGGCCAGGGACCCAGGTCACCGTCTCCTCAG E Y D Y W G Q G T Q V T V S S	GT
JH4	AGCATTTC	CTGGGTCTTGACACAGTTGTCA	CAATGTG	AC...CCCCAGTTTGAATACTGGGGCCAGGGCACCCCTGGTCACTGTCTCAG... P Q F E Y W G Q G T L V T V S	GT
JH5	GGTTTTTGC	ACACCACCTAACGGGGCCCGTGG	CGTTGTG	CT.....GACTTTGGTTCCTGGGGCCAGGGACCCCTGGTCACTGTCTCCTC D F G S W G Q G T L V T V S	

### **3.18 Determination of Expressed $V_{HH}$ and $J_H$ gene segments**

Sequences of variable (V), diversity (D) and joining (J) gene segments are described in previous sections (3.4-3.17). Although these sequences show high levels of similarity to known gene segments in other species, and are associated with conventional immunoglobulin promoters and signal sequences, it is not possible to determine beyond any doubt, whether any of these particular sequences result in the production of secreted antibodies or even whether such sequences are actively transcribed by the llama.

It is often difficult to identify the precise germline gene segments responsible for encoding specific cDNA sequences even in well-characterised species such as the mouse or human. The reason for this is that the human and murine heavy chain loci contain multiple, highly similar V, D and J gene segments. Several of these highly similar gene segments may be sufficiently similar to the cDNA sequence to be considered as the potential originator of the sequence. Transcribed, rearranged gene segment sequences have undergone junctional diversity in the form of N and P nucleotide addition and deletion which leads to changes in sequence. More significantly somatic hypermutation introduces near-random changes within the sequence. If the error-prone nature of the PCR process used to generate cDNA sequences is also considered, it is clear that the original germline sequence may be mutated beyond all recognition in cDNA form. In some cases, however, a particular cDNA can be ascribed to a gene segment when the full gene repertoire is known.

The small number of  $V_{HH}$  gene segments isolated in this report make it difficult to determine with any certainty the expression of a given gene segment. However, examination of the llama heavy chain immunoglobulin cDNA library is sufficient to gain an insight into the proportion of llama cDNA sequences that utilise llama germline gene segments similar to, if not identical to those that have been isolated. It is important to obtain an indication of the likely usage of llama gene segments such as those that have been isolated in this thesis, especially if the sequences are to be used in tests of llama antibody generation mechanisms (as in Chapter 6). Clearly analysis of the recombination of sequences associated with a redundant gene segment is of little value.

To compare germline and expressed variable sequences the llama cDNA library of  $V_{HH}$  clones was searched for sequences similar to the germline gene segments isolated in this report, through the creation of a GCG-based database (figure 3.15). The search was made through the secure eBioinformatics ([www.ebioinformatics.com](http://www.ebioinformatics.com)) internet server using the Blast 2.0 algorithm (186). This generated a number of local alignments between previously characterised cDNA sequences within the Unilever database and the germline gene segments. The Blast search also produces a score that allows for a comparison of sequence similarity between different alignments, based not only on percentage identity but also length of the local alignment. The best match to each variable and joining gene segment are shown in figure 3.16. The single diversity gene segment was too short to generate any significant alignment.

In an attempt to give a measure of the number of sequences within the cDNA database with similarity to the coding sequence of each germline gene segment the average Blast score of the top ten matches to each segment is included in figure 3.16 as well as the Blast score for the twentieth best match (the twentieth being an arbitrarily chosen match). This data gives an indication of the depth of representation within the llama cDNA database.

Comparison of variable gene segments with the cDNA library show best matches with identities varying from 93.7% ( $V_{HH3}$ ) to 89.6% ( $V_{HH5}$ ). The depth of representation within the library also varied considerably, with a large number of library sequences showing high identity to  $V_{HH3}$  in particular. (20<sup>th</sup> Blast match score of 329). While the low best-match identity to  $V_{HH5}$  is paralleled by a rapid drop off in the identity of next best matches to this sequence (i.e. very few database sequences match this sequence), the relatively high average top ten match blast score for  $V_{HH2}$  (Blast average of 325) is not coupled with a high twentieth best match. This can be interpreted as a very limited number of sequences having reasonable identity to the  $V_{HH2}$  gene segment sequence. Overall the cDNA library contained more sequences with higher identity to  $V_{HH3}$  than any other gene segments and for this reason this sequence was used throughout further studies of llama antibody generation.

The overwhelming majority of cDNA sequences within the Unilever database include FR4 sequence with very high homology to at least one of the five J<sub>H</sub> segments isolated. Interestingly the amino acid sequence of one of the germline J region sequences reported here (section 3.16) matches exactly the predicted J gene segment responsible for over half the clones examined in the work of Vu *et al* (175) and provides evidence that J and therefore also the D gene segments isolated here may represent a locus utilised both by V<sub>H</sub> and V<sub>HH</sub> gene segments. The coding sequence of the J<sub>H</sub> gene segments was also compared to the sequence of a limited number of classical llama cDNA sequences (data not shown). The FR4 of classical cDNA sequences showed high levels of identity to these J gene segments suggesting that the same J gene segments are utilised during generation of both heavy chain and classical immunoglobulins.

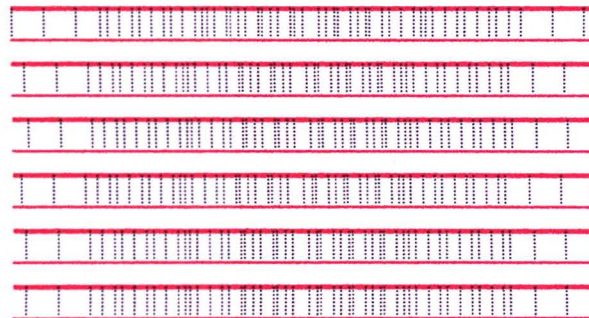
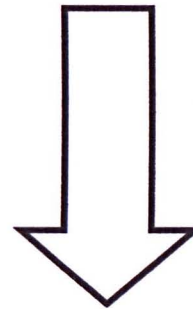
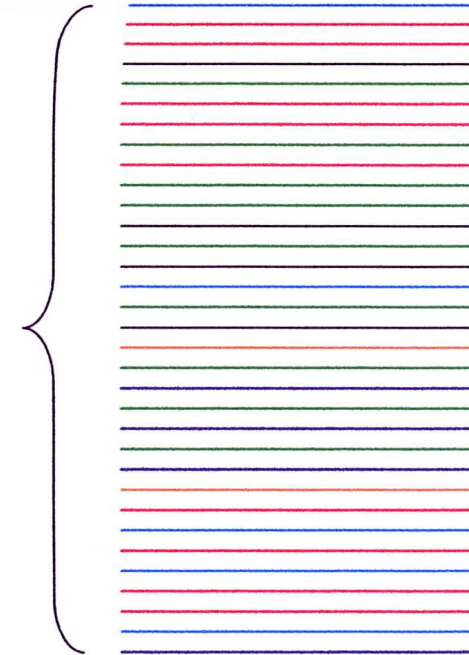


Germline Variable (or Joining) Gene Segment

---



Library of sequences is searched for best 'matches' (i.e. sequences most similar to the germline sequence)



Both the number of matches and level of identity within each match give an indication of the likely level of expression of a particular germline gene segment

Figure 3.15 Strategy utilised to attempt to determine the level of expression of particular llama germline gene segments.

**Figure 3.16 Alignments showing the level of identity between each variable and joining gene segment and its best match from the Unilever cDNA library. Colour indications correspond to framework and complementarity determining regions. It is clear from these alignments either that these gene segments are expressed by the llama and therefore found within the cDNA library, or that highly similar gene segments are expressed. The gene segments reported here are therefore considered representative of at least a proportion of the expressed repertoire of heavy chain variable gene segments. Associated signal sequences are therefore likely to be representative of the full variable segment complement. Note that the incomplete nature of the J<sub>H</sub>5 sequence excludes this gene segment from the analysis.**

Variable vs cDNA Nucleotide Identity	Blast Score For This Match	Average Blast Score	Blast Score For 20 <sup>th</sup> Best Match	Match Count
<b>V<sub>H</sub>1 vs cDNA 101517 Nucleotide Identity 92.3%</b>	<b>389</b>	<b>338</b>	<b>260</b>	1
<pre> GGAGGCTTGGTCAGCTCGGGGGTCTCTGAGACTCTCTGTGACGCTCTGGATCACTTGGATTATTATGCCATAGCTGGTTCCGCCAGGCCCAAGGGAAGGAGCGGAGGCGTCTCATGTATTAGTATGATGGTAGCAGACTATGCGAGCTCCGTGAAGGCCGATTACCACTCTCCAGAGCAACGCCAGGAACAGGTGTATCTGCAATGAAACAGCTGAAACCTGAGGACACGGCCGTTTATTACTGTGCAGCA GGAGGCTTGGTCAGCTCGGGGGTCTCTGAGATTTCTCTGTGACGCTCTGGATCACTTGGCTTATTATGCGATAGCTGGTTCCGCCAGGCCCAAGGGAAGGAGCGGAGGCGTCTCATGTATTAGTCAAGTTGATGGTAAATACACTATGCGAGCTCCGTGAAGGCCGATTACCACTCTCCAGAGCAATGCCAGGAACAGGTGTATTTGCAATGAAACAACTGAAACCTGAGGACACGGCCGTTTATTACTGTGCAGCA </pre>				
<b>V<sub>H</sub>3 vs cDNA 105617 Nucleotide Identity 93.7%</b>	<b>412</b>	<b>369</b>	<b>329</b>	1
<pre> GGAGGCTTGGTCAGCTCGGGGGTCTCTGAGACTCTCTGTGACGCTCTGGAGACCACTGTATCTCTGCCATAGGCTGGTTCGCCAGGCCCAAGGGAAGGAGCGGAGGCGTCTCATGTATTAGTATGATGGTAGCAGACTATGCGAGCTCCGTGAAGGCCGATTACCACTCTCCAGAGCAACGCCAGGAACAGGTGTATCTGCAATGAAACAGCTGAAACCTGAGGACACGGCCGTTTATTACTGTGAAATGCA GGAGGCTTGGTCAGCTCGGGGGTCTCTGGATTTCTCTGTGACGCTCTGGATCACTTGGCTTATTATGCGATAGCTGGTTCGCCAGGCCCAAGGGAAGGAGCGGAGGCGTCTCATGTATTAGTCAAGTTGATGGTAAATACACTATGCGAGCTCCGTGAAGGCCGATTACCACTCTCCAGAGCAACGCCAGGAACAGGTGTATTTGCAATGAAACAACTGAAACCTGAGGACACGGCCGTTTATTACTGTGAAATGCG </pre>				
<b>V<sub>H</sub>2 vs cDNA 101517 Nucleotide Identity 91.5%</b>	<b>373</b>	<b>325</b>	<b>238</b>	1
<pre> GGAGGCTTGGTCAGCTCGGGGGTCTCTGAGACTCTCTGTGACGCTCTGGATCACTTGGATTATTATGCCATAGCTGGTTCCGCCAGGCCCAAGGGAAGGAGCGGAGGCGTCTCATGTATTAGTATGATGGTAGCAGACTATGCGAGCTCCGTGAAGGCCGATTACCACTCTCCAGAGCAACGCCAGGAACAGGTGTATCTGCAATGAAACAGCTGAAACCTGAGGACACGGCCGTTTATTACTGTGCAGCA GGAGGCTTGGTCAGCTCGGGGGTCTCTGAGATTTCTCTGTGACGCTCTGGATCACTTGGCTTATTATGCGATAGCTGGTTCGCCAGGCCCAAGGGAAGGAGCGGAGGCGTCTCATGTATTAGTCAAGTTGATGGTAAATACACTATGCGAGCTCCGTGAAGGCCGATTACCACTCTCCAGAGCAACGCCAGGAACAGGTGTATTTGCAATGAAACAACTGAAACCTGAGGACACGGCCGTTTATTACTGTGCAGCA </pre>				
<b>V<sub>H</sub>4 vs cDNA 402217 Nucleotide Identity 91.4%</b>	<b>365</b>	<b>329</b>	<b>283</b>	1
<pre> GGAGGCTTGGTCAGCTCGGGGGTCTCTGAGACTCTCTGTGACGCTCTGGATCACTTGGATTATTATGCCATAGGCTGGTTCGCCAGGCCCAAGGGAAGGAGCGGAGGCGTCTCATGTATTAGTATGATGGTAGCAGACTATGCGAGCTCCGTGAAGGCCGATTACCACTCTCCAGAGCAACGCCAGGAACAGGTGTATCTGCAATGAAACAGCTGAAACCTGAGGACACGGCCGTTTATTACTGTGCAAAA GGAGGCTTGGTCAGCTCGGGGGTCTCTGAGATTTCTCTGTGACGCTCTGGATCACTTGGCTTATTATGCGATAGCTGGTTCGCCAGGCCCAAGGGAAGGAGCGGAGGCGTCTCATGTATTAGTCAAGTTGATGGTAAATACACTATGCGAGCTCCGTGAAGGCCGATTACCACTCTCCAGAGCAACGCCAGGAACAGGTGTATTTGCAATGAAACAACTGAAACCTGAGGACACGGCCGTTTATTACTGTGCAGCA GGAGGCTTGGTCAGCTCGGGGGTCTCTGAGATTTCTCTGTGACGCTCTGGATCACTTGGCTTATTATGCGATAGCTGGTTCGCCAGGCCCAAGGGAAGGAGCGGAGGCGTCTCATGTATTAGTCAAGTTGATGGTAAATACACTATGCGAGCTCCGTGAAGGCCGATTACCACTCTCCAGAGCAACGCCAGGAACAGGTGTATTTGCAATGAAACAACTGAAACCTGAGGACACGGCCGTTTATTACTGTGCAGCA </pre>				
<b>V<sub>H</sub>5 vs cDNA 213317 Nucleotide Identity 89.6%</b>	<b>313</b>	<b>279</b>	<b>230</b>	1
<pre> GGAGGCTTGGTCAGCTCGGGGGTCTCTGAGACTCTCTGTGACGCTCTGGATCACTTGGATTATTATGCCATAGCTGGTTCGCCAGGCCCAAGGGAAGGAGCGGAGGCGTCTCATGTATTAGTATGATGGTAGCAGACTATGCGAGCTCCGTGAAGGCCGATTACCACTCTCCAGAGCAACGCCAGGAACAGGTGTATCTGCAATGAAACAGCTGAAACCTGAGGACACGGCCGTTTATTACTGTGCAGCA GGAGGCTTGGTCAGCTCGGGGGTCTCTGAGATTTCTCTGTGACGCTCTGGATCACTTGGCTTATTATGCGATAGCTGGTTCGCCAGGCCCAAGGGAAGGAGCGGAGGCGTCTCATGTATTAGTCAAGTTGATGGTAAATACACTATGCGAGCTCCGTGAAGGCCGATTACCACTCTCCAGAGCAACGCCAGGAACAGGTGTATTTGCAATGAAACAACTGAAACCTGAGGACACGGCCGTTTATTACTGTGCAGCA GGAGGCTTGGTCAGCTCGGGGGTCTCTGAGATTTCTCTGTGACGCTCTGGATCACTTGGCTTATTATGCGATAGCTGGTTCGCCAGGCCCAAGGGAAGGAGCGGAGGCGTCTCATGTATTAGTCAAGTTGATGGTAAATACACTATGCGAGCTCCGTGAAGGCCGATTACCACTCTCCAGAGCAACGCCAGGAACAGGTGTATTTGCAATGAAACAACTGAAACCTGAGGACACGGCCGTTTATTACTGTGCAGCA </pre>				
<b>J<sub>H</sub>1 vs cDNA 405017 Nucleotide Identity 90.7%</b>	<b>86</b>	<b>61.8</b>	<b>42</b>	1
<pre> GCTACAGGTATCTCGAAGTTTGGGGCCAGGGCACCGTGTCTCTCTCAG ::: ::::: : ::::: CGTACAGGTATTTCGAAGTTTGGGGCCAGGGCACCGTGTCTCTCTCAG- </pre>				
<b>J<sub>H</sub>2 vs cDNA 202917 Nucleotide Identity 84%</b>	<b>50</b>	<b>45.6</b>	<b>42</b>	1
<pre> CAATGCTTTGGACGCATGGGGCCAGGGCACCGTGTCTCTCATCAG ::: ::::: : ::::: CAA-CCTTTTCG-CGCAATGGGGCCAGGGCACCGTGTCTCTCTCAG- </pre>				
<b>J<sub>H</sub>3 vs cDNA 202117 Nucleotide Identity 97.9%</b>	<b>94</b>	<b>87</b>	<b>78</b>	1
<pre> ATGAGTATGACTACTGGGGCCAGGGGACCCAGGTCACTCTCTCAG ::: ::::: : ::::: ATGAGTATGACTACTGGGGCCAGGGGACCCAGGTCACTCTCTCA- </pre>				
<b>J<sub>H</sub>4 vs cDNA 204817 Nucleotide Identity 72%</b>	<b>48</b>	<b>48.4</b>	<b>40</b>	1
<pre> ACCCCAAGTTTGAATCTGGGGCCAGGGCACCGTGTCTCTCAG-- ::: ::::: : ::::: GATATGAGTTTGAATCTGGGGCCAGGGGACCCAGGTCACTCTCTCA </pre>				

### **3.19 Dromedary and Llama V Gene Segments**

While this chapter presents the results of the isolation of six coding variable gene segments another study conducted in parallel by the Muyldermans laboratory in Belgium reported the isolation of a larger number of dromedary derived variable gene segments of both classical and heavy chain immunoglobulin coding potential (190, 206). They also reported the isolation of a single diversity gene segment (with no identity to that reported here). No joining gene segments were isolated during the Muyldermans study. This study used a PCR-based strategy to provide a more comprehensive database of germline camelid variable gene segments (190).

### **3.20 Diversity and Joining Gene Segments – Summary of Findings**

In addition to the findings summarised earlier regarding variable gene segments, the following additional results are reported in this chapter:

- Sequences of a single diversity and five joining gene segments have been isolated
- Diversity and joining gene segments are closely linked.
- Joining and variable gene segments, or sequences highly similar to them are expressed during generation of heavy chain antibodies by the llama.
- Joining gene segments show homology to FR4 sequences derived from both classical and heavy chain immunoglobulins. It is therefore likely that the same J gene locus is utilised during generation of classical and heavy chain genes.

### 3.21 Discussion

Attempts at assigning particular gene segments isolated here to expressed cDNA sequences (sections 3.18) suggest that the sequences reported here are representative of those used by the llama immune system.

In section 1.4 a number of potential mechanisms of heavy chain antibody generation were discussed. In this chapter a variety of the genetic components involved in the generation of llama heavy chain antibodies have been isolated and characterised. More specifically, representatives of each of the three gene segment types that together encode the variable domain of the antibody are reported.

The first conclusion from this study confirms point (2) of section 1.4. That is to say, specific heavy chain antibody-encoding genetic components are shown to be present within the llama genome. Five such components ( $V_{HH1-5}$ ) are reported here, the sequence composition of each confirming their roles in encoding heavy chain immunoglobulins (sections 3.4-3.6). However it is not possible to discount the role of affinity maturation (point (1), section 1.4) completely. Evidence is provided in this chapter (section 3.13) that supports a role for somatic hypermutation during llama heavy chain antibody generation.

Through the comparison of germline and cDNA sequences (section 3.11) no evidence was found to support a mechanism of deletion of specific sequence information (section 1.4 point 3))

The comparison of the recombination signals associated with the variable gene segments isolated in this chapter with those characterised previously in human and mouse (section 3.10) is highly suggestive of considerable conservation between the mechanisms of recombination in the llama and other species. The high degree of sequence identity between the heavy chain and classical germline gene segments isolated here (section 3.7) indicates that heavy chain gene segments are only relatively recently evolved from classical gene segments. The development of a completely new recombination mechanism over such a short evolutionary period seems unlikely. These two findings suggest that a completely new recombination mechanism (section

1.4 point (4)) does not play a major role in heavy chain antibody generation. Nevertheless the possibility that subtle differences between the recombination mechanisms of the llama and other species exist (section 1.4, point (5)) cannot be ruled out.

The analysis of the promoter regions of the germline gene segments isolated in this report has identified two patterns of binding motifs. Although the number of promoter regions examined is small there would appear to be a correlation between the functional class of heavy chain antibody generated by a particular gene segment (section 3.8-3.9) and the motifs present upstream of the gene. It is possible therefore that the generation of both heavy chain and classical llama antibodies is controlled at the level of transcription (section 1.4 point (6))

All evidence from analysis of the V, D and J gene segments isolated in this study suggests that those involved in heavy chain antibody generation are of similar size, composition and exon/intron layout to those found at conventional heavy chain immunoglobulin loci. This concurs with the relatively short (~2 million years) evolutionary time-scale over which the heavy chain antibodies are thought to have developed (section 1.14.1).

### **3.22 Future Work**

The functional analysis of recombination signal sequences isolated in this chapter is described in Chapter 5 of this thesis.

The first objective of any future work investigating the germline gene segments of the llama will be to confirm the presence of a single immunoglobulin heavy chain locus within the llama genome. This could be achieved by the construction and screening of either a yeast artificial chromosome (YAC) or cosmid library. Both of these would allow cloning and isolation of significantly larger fragments of the llama genome than the bacteriophage library described and used in this thesis (sections 2.3-2.4). The linkage of the genetic elements involved in classical and heavy chain antibody generation will allow the examination of two levels of antibody generation control. Firstly specific enhancer regions such as those isolated in the human heavy chain locus (213) may be identified that regulate transcription of heavy chain gene segments. Secondly heavy chain and classical gene segments may be found to be located within specific, discrete locations within the heavy chain locus. If this were the case factors such as chromatin accessibility may play a role in heavy chain/classical gene segment choice. The possibility that more than one locus is responsible for llama heavy chain immunoglobulin production cannot yet be ruled out.

A second objective would be the isolation of larger numbers of variable, diversity and joining gene segments. This could be achieved during the mapping of the heavy chain locus using the library described above, or through a PCR-based strategy such as that used to derive dromedary gene segments (190).

Once a larger number of gene segments are available the correlation between variable gene promoter composition and variable domain family can be confirmed. To test the significance of particular motifs present upstream of the variable gene segments, constructs containing various combinations of the promoter motifs linked to a reporter gene could be transfected into a pre-B-cell line. This would allow the examination of differential transcription that may result from the presence of particular motifs upstream of variable gene segments during B-cell development.

To better understand the levels of control of antibody generation within the llama it would be useful to examine the levels of heavy chain and classical immunoglobulin within the llama sera. Although the level of heavy chain antibody has been shown to vary, typically comprising between 5-20% of serum IgG no study has yet examined the possible changes in these levels during llama development or during the course of an infection. Such studies could provide an insight not only into the type of antibody favoured by the immune system during these times but may also give an indication of how well the llama is able to modulate antibody levels in response to the immune system requirements.

# **Chapter 4 Isolation, Cloning and Expression of Llama**

## **Recombination Activating Proteins**

### **4.1 Abstract**

Chapter 3 of this thesis reports the isolation of a range of sequences from the heavy chain immunoglobulin locus of the llama. This data provides answers to a number of questions regarding llama antibody formation as summarised in section 3.20. Sequence data alone can only provide limited clues to the manner in which heavy chain antibodies are generated. The formation of immunoglobulin molecules is the result of a number of complex protein-protein, protein-DNA and DNA-DNA interactions (Chapter 1), the full nature of which cannot be understood simply through an appreciation of the sequences taking part in mechanisms such as V(D)J recombination and somatic hypermutation. It is for this reason that Chapters 4 and 5 of this thesis investigate the most crucial biochemical process involved in the generation of antibody diversity, V(D)J recombination. It is V(D)J recombination that ultimately leads to the production of high-affinity antibodies to combat a vast range of invading pathogens. In this chapter the successful isolation of sequences encoding llama recombination activating proteins (RAG-1 and RAG-2) is described. Additionally the successful cloning, expression and purification of biologically active llama RAG-1 protein is also reported.

### **4.2 Introduction**

Antibody generation is controlled at numerous levels. Control of llama immunoglobulin generation at the level of transcription has already been described, through the examination of promoter binding sites that may lead to preferential recombination of particular V region gene segments (section 1.7.4 and section 3.9). Control at the level of somatic mutation has also been discussed (sections 1.9.1 and 3.13). Control at the level of RNA processing, whereby splicing events are responsible for the removal of large regions of the antibody primary sequence will be described in chapter 6.



## IMAGING SERVICES NORTH

Boston Spa, Wetherby

West Yorkshire, LS23 7BQ

[www.bl.uk](http://www.bl.uk)

**PAGE MISSING IN  
ORIGINAL**

*rag* genes of all species. The isolation of the equivalent genes in the llama was therefore considered a priority.

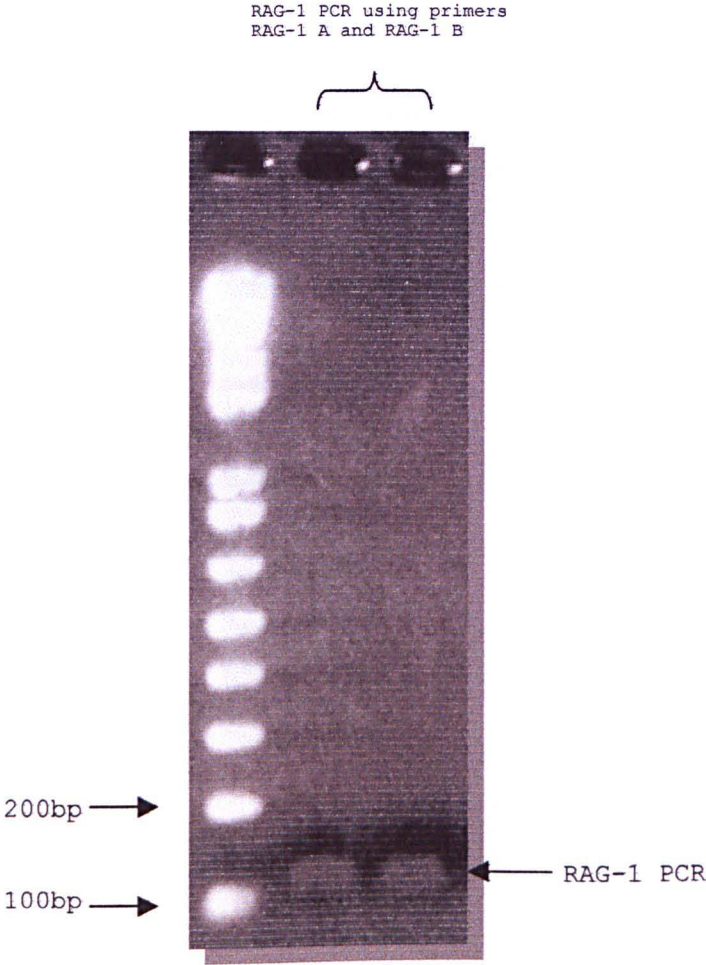
### 4.3 Screening Strategy

In order to isolate llama *rag* genes the genomic library generated and described in section 2.3 was again utilised. A screening strategy based on the generation of llama RAG-specific probes through PCR was devised (section 2.4.5). Whereas sequence data and material for the generation of probes used in the library screening of previous chapters was readily available from analysis of cDNA sequences obtained at Unilever, no llama *rag* gene sequences have previously been reported. Upon successful screening *rag* genes were cloned and llama RAG proteins expressed for optimisation in a cell-free system where individual llama recombination signal sequences could be tested (Chapter 6).

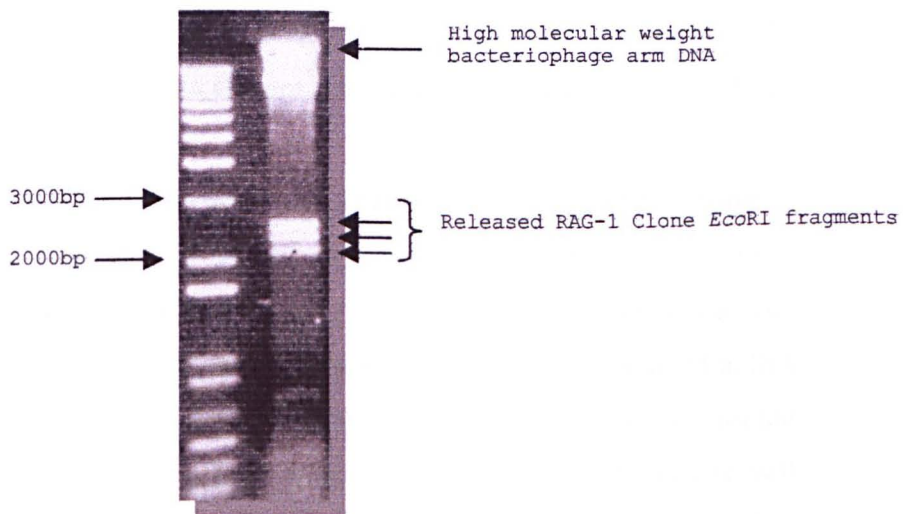
#### 4.3.1 Generation of a RAG-1 Specific Probe

The genomic library screening strategy for *rag* gene sequence isolation was based largely on two findings. Firstly the *rag* genes are highly conserved in evolution (219-221) and secondly the coding sequence of the *rag-1* gene is restricted to a single exon in all species examined (58). In order to generate a probe specific to the llama *rag-1* gene a multiple alignment was constructed to illustrate areas of strong *rag-1* sequence conservation between other characterised species (data not shown). The steps involved in the generation of this probe including genomic PCR (figure 4.1) are described fully in section 2.4.5. A single *rag-1* bacteriophage clone was isolated from the germline genomic library using this strategy. However, problems in obtaining high quality bacteriophage clone DNA by direct sequencing were encountered and a high fidelity PCR/sub-cloning strategy was subsequently adopted to obtain the full *rag-1* sequence (section 2.5.2 and figures 2.3-2.4 and 4.2). Given the close (~8kb) linkage of the *rag-1* and *rag-2* genes in a range of other species, it was hoped that it might be possible to isolate the llama *rag-2* gene from this same bacteriophage clone. However, sequencing of both ends of this clone, coupled with consensus PCR gave no indication that the *rag-2* sequence was present within the clone. For this reason an alternative strategy was employed to obtain the *rag-2* sequence (section 4.3.2).

**Figure 4.1 Genomic PCR to generate a RAG-1 specific probe for screening of the llama genomic library using Pfu Polymerase. A band of approximately 130bp was generated as predicted by sequence alignment. The band was cut out and gel purified before cloning into the pBluescript vector.**



**Figure 4.2 Subcloning strategy for llama RAG-1.** Subcloning of the llama RAG-1 clone. Direct sequencing of the RAG-1 genomic clone was not possible. This gel shows the digestion of RAG-1 bacteriophage clone DNA with restriction enzyme *EcoRI* prior to subcloning in pUC19. Three bands, all of approximately 2-3000bp were cut out and gel purified before ligation into *EcoRI*-cut pUC19. Subsequent sequencing of clones revealed sequence specific to the 3' UTR of llama RAG-1 and RAG-2 and sequence with homology to the human RAG-1/-2 intergenic region.

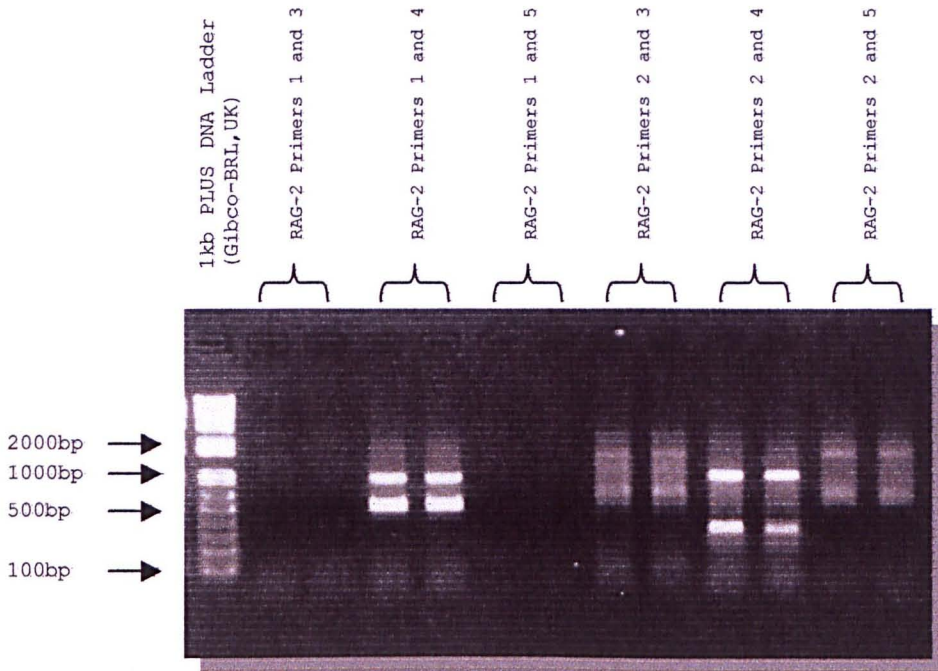


#### **4.3.2 Genomic PCR to Isolated Partial Nucleotide Sequence of Llama RAG-2**

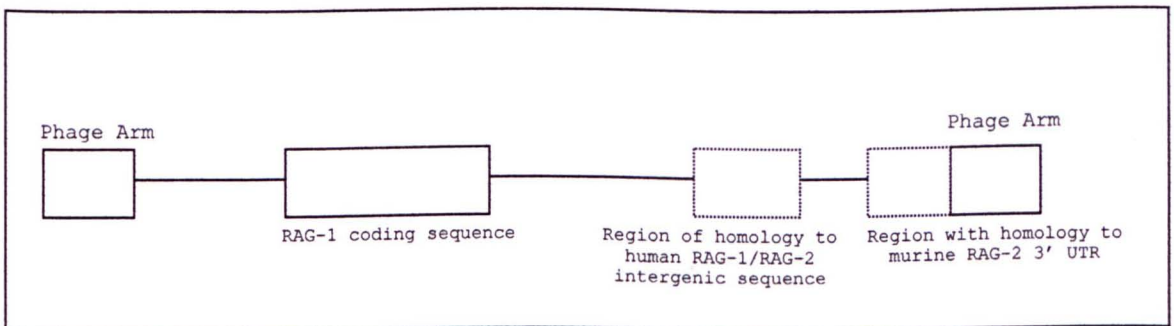
As discussed in section 4.3.1, it was not possible to isolate *rag-2* germline nucleotide sequence data from the *rag-1* bacteriophage clone. In order to obtain *rag-2* gene sequence data a second strategy was employed. This involved the construction of a multiple-nucleotide sequence alignment of the *rag-2* nucleotide sequences from a variety of species (data not shown). This was used to generate a number of consensus PCR primers (table 2.3). By using overlapping primers the sequence obscured by the primers themselves was also elucidated. This method made the assumption that conservation of sequence between other species was indicative of conservation between these sequences and the unknown sequence of llama *rag-2*. This assumption was shown to be correct after isolation of the majority of the llama *rag-2* sequence. By including degeneracy within the primers and using relatively low annealing temperatures (these modifications allow a degree of primer/template mismatch) (section 2.2.2) a number of DNA products were cloned and characterised (figure 4.3).

#### **4.4 Analysis of sequence surrounding a Llama RAG-1 Genomic Clone**

As discussed previously (Chapter 1.8.5) the *rag* genes of all species previously characterised are closely linked and convergently transcribed. Although it was not possible to isolate the *rag-2* gene from within the single genomic clone isolated in this report, the subcloning of the *rag-1* clone DNA fragments did reveal regions within the clone with homology both to the intergenic region of the human *rag-1/2* locus as well as sequence with some similarity to the 3' untranslated region of murine *rag-2* (data not shown). It is therefore assumed that the llama *rag-2* gene lies just beyond the range of this clone. The presumed layout of the *rag-1* bacteriophage clone is given in figure 4.4.



**Figure 4.3 Genomic Consensus PCRs of Llama RAG-2 gene** using Pfu polymerase (6 minute extension). A number of primer sets were used in different combinations and two lanes of each reaction run. Specific bands were cut out and cloned using the TOPO-Blunt cloning kit (Invitrogen, Carlsbad, US)



**Figure 4.4 Presumed Overall Structure of Llama RAG-1 Bacteriophage Clone.** Subcloning revealed regions of DNA with some homology to both human RAG-1/-2 intergenic regions and the 3' downstream untranslated region of the human RAG-2 gene. Solid boxes represent known sequences. Dotted lines represent regions of homology to the human and murine RAG-1 loci, as derived from BLAST analysis (data not shown)

#### **4.5 The Isolation of the Full-length Sequence of The Llama RAG-1 Gene**

The full-length sequence of the llama *rag-1* gene and corresponding translation is shown in figure 4.5. No promoter region is present upstream of this sequence as the llama *rag-1* coding sequence, like the coding sequence characterised in other species, is thought to lie within the second exon of the *rag-1* gene, with the promoter upstream of the first, non-coding exon. This non-coding exon was not isolated in this study.

#### **4.6 The Isolation of the Partial Sequence of the Llama RAG-2 Gene**

Approximately 80% of the llama *rag-2* gene sequence was isolated through the use of genomic PCR (table 2.3 and figure 4.6). The nature of this sequence is discussed in subsequent sections (section 4.11-4.13).

Figure 4.5 Nucleotide and predicted amino acid sequence of Llama RAG-1 (Genbank accession no AF305953). Colour indications illustrate regions of interest and homology.

**Key** Colour indications are as for figure 1.8. Sequence in green lies outside the 'core' domain of the protein (Residues 1-385 and 1011-1042). The zinc dimerisation domain (268-382) is indicated in *italics* The amino acid sequence of the RAG-2 interacting domain is shown underlined while the nonamer binding domain (392-462) is depicted in blue and the region with homology to a prokaryotic integrase family (503-525) is given in purple. Note: all regions are described in section 4.10 and are defined by analogy with the murine protein

```

1
M A V S L P P T L G L S S A P D E I Q H P H I K F S E W K F K L F R V R S F E K A P E N A Q K E K Q
ATGGCTGTCTCTTTGCCCCCGACCTGGGACTCAGTTCGCCCCCGATGAAATCCAGCATCCACATATTAATTTTCAGAGTGGAAATTTAAGCTATTGAGGGTGAATCCTTTGAAAAGGCCCTGAAAATGCTCAAAAAGAAAAGCAA
50
D S S E G K P F L E Q S P A V L D K G G G Q K P A L T Q P A L K P H P K F L K K P P D D G K A R D K
GATTCCTCCGAGGGGAAGCCCTTCTCGAGCAATCCTCAGCAGTCCGGACAAGGGTGGTGGTCAAGAGCCAGCCCTGACTCAACCAGCATTAAAGCCTCACCCAAAGTTTGAAGAAAACCCCTGATGATGGGAAAGCGAGAGACAAA
100
A I H Q A N L R H L C R I C G N S F N S D G H N R R Y P V H G P V D G K T Q V L L R K K E K R A T S
GCCATCCACCAAGCCAACCTGAGACACCTCTGCCGATCTGTGGGAATTCCTTCAACAGCGATGGGCAACAGGAGATATCCGGTCCACGGGCCTGTGGACGGGAAAACCAAGTCTTTTGGCAAAGAAAAGAGAGACCACTTCC
150
W P D L I A K V F R I D V K A D V D S I H P T E F C H N C W S F M H R K F S S A P C E V Y C P K S V
TGGCCAGACTCATCGCAAGGTTTTCGGATCGATGTGAAGGAGATGTGGACTCGATCCACCCACTGAGTTCCTGTGATAATGCTGGAGCTTATGACACAGGAAGTTTAGCAGTGCCCATGTGAGGTTTATTGCCCAAAGAGTGA
200
T M E W H P H T P S C D I C H A A R R G L K R K S P Q P N L Q L S K K L K T V I D R A K Q A R R H K
ACCATGGAGTGGCATCCCCACACCCCTCTGCGACATCTGCCATGCTGCCGCTGGTGGACTCAAGAGGAAGAGTCCCGAGCCAACCTGCAGCTCAGCAAAAACCTCAAACTGTGATTGACCGAGCAAAAACAGCCCGTCCGACAAAG
250
R R A Q A R I S S K E L M K K I A S C S K I H L S T K L W Q W T S G T L V K S I S C Q I C E H I L A
AGGAGAGCTCAGGCAAGGATCAGCAGCAAGGAAGTGAAGAGATGCCAGCTGCAGTAAAGATACATCTTAgCACCAAgCTCTGGCAGTGGACTTCCGGCACACTGTGAAATCTATCTCTGCCAGATTGTGAGCAGACTTCTGGCT
300
D P V E T S C K H V F C R I C I L M C L K V M G S Y C P S C Q Y P C F P T D L E S P V R S F L S I L
GACCTGTGGAGACCAGCTGTAAGCATGATTTTGCAGGATCTGCATCTGATGTGCCCTCAAAGTCATGGCAGCTACTGTCCCTCTTGCCAGTATCCCTGcTTCCCTACTGACCTGGAgAGTCCGGTGGGCTTTTCTGAGCATCTG
350
N S L T V K C P A Q E C N E E V S L E K Y N H H V S S H K E S K E T F V H I N K G G R P R Q H L L S
AATTCCTGACTGTGAAATGTCCAGCACAGAGTGAATGAGGAGTCAAGCTTGGAAAATAACAATCACCATGTCTCAAGCCATAGGAATCAAAGAGACTTTGTAACATATCAATAAAGGAGGCCGGCCCGCCAGCACCTCTGTCC
400
L T R R A Q K H R L R E L K L Q V K A F A D K E E G G D V K S V C L T L F L L A L R A R N E H R Q A
CTGACCCGGAGGGCTCAGAAGCACCGTCTGAGGGAGCTCAAGCTGCAAGTCAAGGCTTTTGTGCAAAAAGAAGAGTGGGATGTGAAGTCCGGTGTCCCTGACCTTGTCTGCTGGCACTGAGGGCGAGGAATGAGCACAGACAAGCA
450
D E L E A I M Q G R G S G L Q P A V C L A I R V N T F L S C S Q Y H K M Y R T V K A I T G R Q I F Q
GACGAGCTGGAGGCCATCATGACGGGACGGGGCTCTGGTCTACAGCCAGCTGTTTGTCTGGCCATCCCGCTCAACACTTCTCAGCTGCAGTACACCAAGATGTACAGGACTGTGAAAGCCATCAGGGGAGGCAGATTTTCCAG
500

```



550  
P L H A L R N A E K V L L P G Y H P F E W Q P P L K N V A S S T D V G I I D G L S G L S S S V D D Y  
CCTTTGACGCCCTTCGGAACGCTGAGAAGGTCCTCTGCGGGCTACCACCCTTCGAGTGGCAGCCACTCTGAAGAAGCGTGGCTTCCAGCACCGAGTGGCATTATTGACGGGCTGTCCGGACTGTCTCCTCTGTGGACGATTAC

600  
P V D T I A K R F R Y D S A L V S A L M D M E E D I L E G M R G Q D L D D Y L N G P F T V V V K E S  
CCGGTGGACACCATTCGAAGCGCTTCGCTATGATTCAGCTTTGGTGTCTGCTCTGATGGACATGGAAGAAGACATCCTGGAAGGCATGAGAGGCCAGGACCTTGATGACTACCTGAATGGCCCTTCACCCTGGTGGTGAAGGAGTCT

650  
C D G M G D V S E K H G S G P A V P E K A V R F S F T I M K I T I A H G S Q N V K V F E E A K P N S  
TGTGATGGGATGGGAGCGTGGAGGAGAAGCATGGGAGCGGGCCGCGAGTTCGGGAGAAGCGGTTGGTTTTCTTTCACGATCATGAAGATTACCATCGCGATGGGTACAGAACGTGAAGTGTTCGAGGAAGCCAGCCGAATCT

700  
E L C C K P L C L M L A D E S D H E T L T A I L S P L I A E R E A M K S S E L M L E M G G I L R T F  
GAGCTGTGTGCAAGCGTGTGCCTGATGCTGGCTGATGAGTCTGACCACGAGACCTGACGGCCATCTGAGCCCTCATCGCCGAGAGGGAGCCATGAAGAGCAGGAATTAATGCTGGAGATGGGAGGCATCTCCGGACCTTC

750  
K F I F R G T G Y D E K L V R E V E G L E A S G S V Y I C T L C D A T R L E A S Q N L V L H S I T R  
AAGTTCATCTTCAGGGCACCGGATACGACGAGAACTTGTCCGGGAGTGAAGGCCTCGAGGCTTCTGGCTCAGTCTACATTTGTACTCTGTGTGATGCTACCCGCTGGAAGCCTCTCAGAATCTTGCTCCACTCCATAACAGG

800  
S H A E N L E R Y E V W R S N P Y H E T V E E L R D R V K G V S A K P F I E T V P S I D A L H C D I  
AGCCACGCCGAGAACCTGGAGCGTACGAGGCTGGCGGTCGAACCCCTTACCACGAGACGGTGGaGAGCTGCGGGATCGGGTGAAGGGGCTCTCGCCAAACCCCTCATCGAGACGGTCCCGTCCATCGACGCGCTCCACTGCGACATT

850  
G N A A E F Y K I F Q L E I G E V Y K N P N A S R E E R K R W Q A T L D K H L R K K M N L K P I M R  
GGCAACGGCGCTGAGTTTACAAGATCTCCAGCTAGAGATAGGGGAGGTGTATAAGAATCCCAACGCCTCCAGGAGGAAAGAAAGATGGCAGCCGACCTGGACAAGCACCTCCGGAAGAAGATGAACCTGAAGCCCATCATGAGG

900  
M N G N F A R K L M T K E T V E A V C E L V P S E E R H E A L R E L M D L Y L K M K P V W R S S C P  
ATGAACGGCACTTCGCCAGGAAGCTCATGACCAAGAGACGGTTGAAGCAGTCTGTGAATTAGTTCCTCCGAGGAGAGGCACGAAGCTCTGAGGGAAGCTGATGGACCTTTATTTGAAGATGAAACCCGCTGCGGATCGTCATGCCCT

950  
A K E C P E S L C Q Y S F N S Q R F A E L L S T K F K Y R Y E G K I T N Y F H K T L A H V P E I I E  
GCTAAGGAGTGCCAGAATCCCTCTGCCAGTACAGTTCAATTCACAGCGTTTTGCTGAGCTCCTCTCCACCAAGTTCAAGTATAGATAGAGGGCAAGATCACCAATATTTTACAAGACCCTGGCCACGTCCTGAAATATTAGG

1000  
R D G S I G A W A S E G N E S G N K L F R R F R K M N A R Q S K Y Y E M E D V L K H H W L Y T S K Y  
AGGGATGGCTCCATTGGGGCTGGGCAAGCGAGGAAATGAGTCTGGCAACAACTGTTTCAGGCGTTTTCCGGAAGATGAATGCCAGGAGTCCAGTACTACGAAATGGAAGACGCTTGAAGCATCATGGTTGTACACCTCCAATAC

1042  
L Q K F M N A H N A L K N S G F T L N S Q G S L G D L L D L E D S P D S Q D V M E F \*  
CTGCAGAAGTTTATGAATGCTATAATGCGTTAAAAAACTCGGGTTCACC-TAAACTCACAGGGAAGCTTAGGGGACTTGTGTAGACTTAGAGGACTCTCCAGACTCTCAAGATGTAATGGAATTTAA

**Figure 4.6 Partial nucleotide and predicted amino acid sequence of Llama RAG-2** (Genbank accession no AF305954). Numbering is for the murine protein. Residues shown in black comprise the active core of the protein while the dispensable portion is given in green. Regions with similarity to HimA and HimD are italicised.

```

36                               50
. . . . . I . . . . . I . . . . . I . . . . . I . . . . .
W P K R S C P T G V F H F D V K H N H L K L K P A V F S K D S C Y L P P L R Y P A T C T L K G S L E
TGGCCGAAAAGATCCTGCCCCACTGGAGTTTTCCATTTTGATGTAAAGCATAATCATCTCAAAGTGAAGCCTGCAGTTTTCTCTAAGGATTCTGTACCTTCTCTCTTCGCTACCCAGCTACTTGCACACTCAAAGGCAGCCTAGAG

                               100
. . . . . I . . . . . I . . . . . I . . . . . I . . . . . I . . . . .
S E K H Q Y I I H G G K T P N N E L S D K I Y V M S V V C K N N K K V T F R C I E K D L V G D V P E
TCTGAAAAGCATCAGTACATCCATCCATGGAGGGAAAACACCTAATAATGAGCTTTCAGATAAGATTTATGTCATGCTGTTGTTTCAAGAAATAACAAAAAGTTACTTTTCGCTGCATAGAGAAAGACTTGGTAGGTGATGTTCCTGAA

                               150
. . . . . I . . . . . I . . . . . I . . . . . I . . . . . I . . . . .
G R Y G H S I D V V Y S R G K S M G V L F G G R S Y I P S A Q R T T E K W N S V A D C L P H A F L V
GGCAGATATGGTCATTCCATTGATGTGGTGTATAGTAGAGGGAAAAGTATGGGTGTTCTCTTTGGAGGACGTTTCATACATACCTTCTGCCCAAAGAACCCAGAAAAATGGAATAGTGTAGCTGACTGCCTGCCCATGCTTTCTTGGTG

                               200
. . . . . I . . . . . I . . . . . I . . . . . I . . . . . I . . . . .
D F E F G C S T S Y I L P E L Q D G L S F H V S I A R N D T I Y I L G G H S L A N N I R P A N L Y R
GATTTTGAATTTGGGTGCTCTACATCATACTCTCCAGAAGTTCAGGATGGGCTATCTTTTCATGTCTCCATTGCCAGAAATGATACCATTATATTTTAGGAGGACATTCACTTGCCAATAACATCCGCCCTGCCAATCTATACAGA

                               250
. . . . . I . . . . . I . . . . . I . . . . . I . . . . . I . . . . .
I R V D L P L G S P A V N C T V L P G G I S V S S A I L T Q T S N D E F V I V G G Y Q L E N Q K R M
ATAAGGGTTGATCTCCCCCTGGGTAGTCCAGCTGTGAATTGCACAGTCTTGCCAGGAGGAATCTCTGTATCCAGTGAATCCTGACGCAAACAAGCAATGATGAGTTTGTATTGTTGGTGGCTATCAGCTTGAAAAATCAAAAAAGAAATG

                               300
. . . . . I . . . . . I . . . . . I . . . . . I . . . . . I . . . . .
I C N I I S F E D N K I E I R E M E T P D W T P D I K H S K I W F G S N M G N G T I F L G I P G D N
ATCTGCAACATCATCTCTTTTGGAGACAACAAGATAGAAAATCCGTGAGATGGAAACCCAGATTGGACTCCAGATATTAACACAGCAAGATATGGTTTGGAAAGCAACATGGGAAATGGAATATTTTCTTGGCATACCCGGGAGACAAC

                               350
. . . . . I . . . . . I . . . . . I . . . . . I . . . . . I . . . . .
K Q A V S E A F Y F Y M L K C A E D D V N E D Q K T F T N S Q M S T E D P G D S T P F E D S E E F C
AAACAGGCTGTTTCAAGCATCTATTTCTATATGTTGAAATGTGCTGAAGACGATGTGAATGAAGATCAGAAAACATTCACAAATAGTCAGATGTCAACAGAAGACCCAGGGGACTCTACTCCCTTTGAAGACTCGGAAGAATTTTGT

                               400
. . . . . I . . . . . I . . . . . I . . . . . I . . . . . I . . . . .
F S A E A N S F D G D D E F D T Y N E D D E E D E S E T G Y W I T C C P T C D V D I N T W V P F Y S
TTCAGTGGCGAGGCAAAATAGTTTTGATGGCGATGATGAATTTGACACCTATAATGAAGATGATGAGGAAGATGAGTCTGAGACAGGCTACTGGATTACATGTCTGCCCTACATGTGATGTGGATATCAATACTTGGGTACCATTTTATTC

                               450
. . . . . I . . . . . I
T E L N K P A M I Y C S H G D G H W V H A Q C K
ACTGAGCTCAACAAACCTGCCATGATCTACTGCTCTCATGGAGATGGGCACTGGGTACATGCTCAGTGCAAG

```

#### 4.7 Comparison of the Llama RAG Genes with Those of Other Species.

The necessary and sufficient nature of the RAG proteins in directing the first stages of V(D)J recombination means that even subtle residue differences with respect to the murine form may lead to major differences in the recombination mechanism. It is important, therefore, to analyse the similarity of the sequences isolated in this report with previously published data from other species. Broadly speaking the greater the difference between llama and other *rag* sequences, the more likely it is that recombination mechanisms may differ.

The llama RAG-1 protein shows considerable homology to other mammalian recombination proteins characterised including the well-studied murine form (figure 4.7a, a comparison of murine and llama RAG-2 is given in figure 4.7b). The global pairwise alignment in figure 4.7a illustrates the high level of amino acid identity (88.5%) between murine and llama proteins. An approximate unrooted phylogenetic tree comparing the llama *rag-1* nucleotide sequence to all other known full length *rag-1* nucleotide sequences (figure 4.8) is in agreement with previous evidence that the *rag-1* gene may mimic an evolutionary clock (217) (that is to say that the overall evolutionary distance between species is accurately reflected by sequence differences within this gene). Within the llama protein the area of least homology is located between residues 240-286, where two single residue gaps are required for alignment and 22 amino acids differ. The core domain of the protein is considerably more conserved than the 'dispensable' ('dispensable' in so much as *in vitro* cleavage of substrates can be performed in the absence of this region (215, 222)) portion of RAG-1 ('core' region amino acid identity is 96.8%), a finding that is echoed throughout a number of *rag-1* genes in other species.

While it is impossible to predict the activity of proteins based entirely on knowledge of primary amino acid sequence, the high level of identity between llama and other mammalian RAG proteins is suggestive of considerable overlap in their modes of action.



```

                    500                                550
Llama RAG-1 I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....
CSQYHKMYRTVKAITGRQIFQPLHALRNAEKVLLPGYHPFEWQPPLKNVASSTDVGIIDGLSGLSSVDDYPVDTIAKRFRYDSALVSALMDEEDILEGMRGQDLDDYLNQPFPTVVVKE
Mouse RAG-1 ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
CSQYHKMYRTVKAITGRQIFQPLHALRNAEKVLLPGYHPFEWQPPLKNVSSRTDVGIIDGLSGLASSVDEYPVDTIAKRFRYDSALVSALMDEEDILEGMRSQDLDDYLNQPFPTVVVKE
..I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....
                    500                                550

                    600                                650                                700
Llama RAG-1 I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....
SCDGMGDVSEKHGSGPAVPEKAVRFSFTIMKITIAHGSONVKVFEEPKPNSLCCCKPLCLMLADESDHETLTAILSPLIAEREAMKSSSELMLEMGGITRTFKFIFRGTGYDEKLVREVEG
Mouse RAG-1 ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SCDGMGDVSEKHGSGPAVPEKAVRFSFTVMRITIEHGSONVKVFEEPKPNSLCCCKPLCLMLADESDHETLTAILSPLIAEREAMKSSSELTLEMGGIPRTFKFIFRGTGYDEKLVREVEG
..I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....
                    600                                650                                700

                    750                                800
Llama RAG-1 I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....
LEASGSVYICTLCDATRLEASQNLVLSHSITRSHAENLERYEVWRSNPYHETVEELRDRVKGVSAPKF IETVPSIDALHCDIGNAAEFYKIFQLEIGEVYKHPNASREERKRWQATLDKHL
Mouse RAG-1 ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
LEASGSVYICTLCDTRLEASQNLVLSHSITRSHAENLQRYEVWRSNPYHESVEELRDRVKGVSAPKF IETVPSIDALHCDIGNAAEFYKIFQLEIGEVYKHPNASKEERKRWQATLDKHL
..I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....
                    750                                800

                    850                                900                                950
Llama RAG-1 I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....
RKRKMNLPIMRMNGNFARKLMTKETVEAVCELVPSEERHEALRELDLYLKMKPVWRSSCPAKECPESLCOYSFNSQRF AELLSTKFKYRYEGKITNYFHKT LAHVPEI IERDGSIGAWA
Mouse RAG-1 ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
RKRKMNLPIMRMNGNFARKLMTQETVDAVCELIPSEERHEALRELDLYLKMKPVWRSSCPAKECPESLCOYSFNSQRF AELLSTKFKYRYEGKITNYFHKT LAHVPEI IERDGSIGAWA
..I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....
                    850                                900                                950

                    1000                                1042
Llama RAG-1 I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....
SEGNESGNKLFRRFRKMNARQSKYEMEDVLKHHWLYTSKYLQKFMNAHNALKNSSGFTLNSQGSLGDLDDLEDSPDSQDVMEF
Mouse RAG-1 ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SEGNESGNKLFRRFRKMNARQSKCYEMEDVLKHHWLYTSKYLQKFMNAHNALKNSSGFTMNSKETLGDPLGIEDSLESQDSMEF
..I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....I.....
                    1000                                1040

```



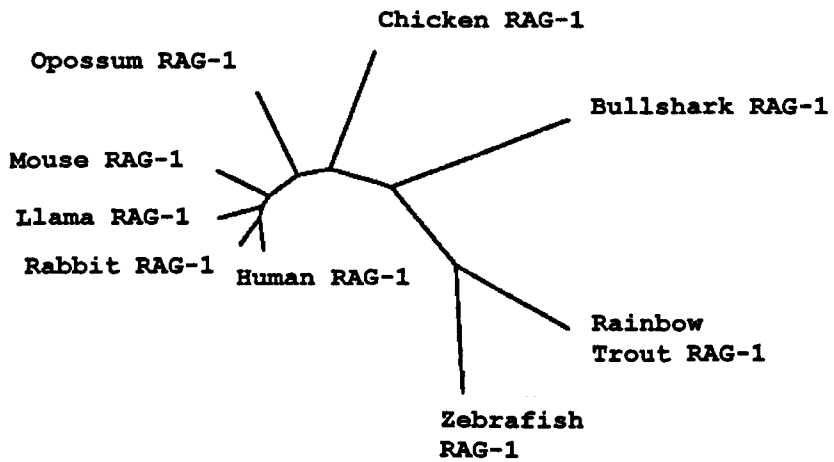


Figure 4.8 Unrooted phylogenetic tree illustrating the evolutionary relationship between the *rag-1* nucleotide sequence of all characterised species (Genbank accession numbers anticlockwise from bottom U71093, U15663, U62645, M58530, U51897, M29475, AF305953, M77666 and M29474). The evolutionary distance between *rag-1* genes of each species is indicated by the distance between the ends of each line.

## 4.8 General Structure and Features of Murine RAG-1 Proteins<sup>2</sup>

While the evolutionary analysis conducted in the previous section finds llama RAG-1 to be highly similar to other RAG proteins, further clues as to the possible characteristics of the protein can be gleaned through comparison of a number of specific functionally-related regions of the protein that have been described in the mouse.

A number of detailed mutagenesis studies have been carried out on murine RAG proteins to define minimum core regions capable of initiating recombination (71, 90, 222, 223). Both the N-terminal third of the RAG-1 protein and the C-terminal quarter of RAG-2 have been shown to be dispensable without reducing the ability of recombinant RAG proteins to process isolated oligonucleotide substrates *in vitro* (chapter 6). This strongly implicates the remaining 'core' portions of the murine RAG amino acid sequences in RSS processing. A fairly large region of murine RAG-1, within the functional core, has been shown to be responsible for interaction with RAG-2 (amino acids 504-1008 (*506-1010*<sup>2</sup>)) (figure 4.5) (87). The crystal structure of a zinc-binding domain that acts as a specific dimerisation domain has been published (amino acids 265-380 (*268-382*)) (88, 89). This domain is directly adjacent to the catalytic core region (amino acids 384-1008 (*386-1010*), which itself includes a putative DNA binding domain (384-477 (*386-479*)). This DNA binding region has subsequently been more accurately defined as a site of interaction between RAG-1 and the nonamer region of the RSS, consisting of residues 390-460 (*392-462*) (214). In addition the catalytic core contains a region with high homology to a prokaryotic integrase (501-523 (*503-525*)) (section 1.8.5) and three acidic residues (D600, D708 and E709 (*D602, D710 and E711*)) recently shown to be essential for the two initial steps of V(D)J cleavage (93-95). At least two of these three residues are thought to be involved in co-ordination of a divalent metal ion. Many biological enzymes require the presence of a divalent cation associated with their biochemical structure in order to exhibit catalytic activity. The finding that these RAG-1 acidic residues may be

---

<sup>2</sup> Numbering in this section is given with respect to the murine RAG-1. Italicised numbering in parenthesis is with respect to llama RAG-1



involved in cation association therefore supports the hypothesis that the active site of the V(D)J recombinase lies within RAG-1 rather than RAG-2.

#### **4.9 Functional Murine RAG-1 Mutant Proteins**

A number of mutant RAG-1 proteins have been examined to further dissect the action of this protein (71, 91). Two such mutations have been shown to alter the specificity of the RAG-1 protein for its target recombination sequence. The first mutation involves an H609L (H611L) substitution, and represents a naturally occurring mutant mouse RAG-1 allele (92) while the second (known as the D32 mutation) is an artificially generated mutation involving the substitution of 4 amino acid residues and the deletion of a further two all within the same region (91). Both mutations, located within similar regions of murine RAG-1 lead to an increase in sensitivity of the protein to the region immediately upstream of the recombination signal sequence, known as the coding flank. Mutant proteins require particular nucleotides immediately upstream of the heptamer sequence to catalyse the initial stages of recombination. These mutations demonstrate that murine RAG-1 interacts not only with the RSS but with adjacent nucleotides as well.

#### **4.10 Analysis of the Llama *RAG-1* Gene for Regions of Potential Functional Significance.**

Major residue differences within the 'core' regions of the llama and murine RAG-1 proteins are rare. Only six non-similar residue differences are located within this region in contrast to the 'dispensable' region where a total of 36 such differences may be found. Possible functional differences between the RAG-1 proteins of the two species can be predicted through the examination of differences between the two proteins within each of the function domains described in section 4.9.

The large region of RAG-1 known to interact with RAG-2 contains four non-similar residue differences between llama and murine forms and demonstrates overall amino acid identity of 95.6%. By contrast the zinc dimerisation domain that lies within the 'dispensable' portion of the protein contains major differences including the residues 'WQWTSGL' in the llama sequence that are not present in the murine protein. The DNA binding domain between residues 390 and 460 (392-462) of the murine and

llama RAG-1 proteins are highly similar (95.7% identity) containing only a single non-similar residue difference (E417 (E419)), while the short region with homology to prokaryotic integrases is completely conserved. The three essential acidic residues described in section 4.8 are also found in the llama RAG-1 proteins.

The region of the llama RAG-1 protein corresponding to that mutated in the H609L and D32 forms (section 4.9) contains a number of differences with respect to the wild-type murine form. These include an alanine rather than glutamic acid residue at murine position 632 (E634A) and other differences such as V626I, R628K and P644A (V628I, R630K and P646A). V626 is conserved amongst RAG-1 proteins from species as evolutionarily distant as salmon and *Xenopus* (95) so replacement of the valine residue with isoleucine may have considerable implications for the action of llama RAG-1.

Although any of the amino acid differences within key regions of the RAG-1 primary sequence described above may affect the process of V(D)J recombination in the llama, the lack of a full understanding of the three-dimensional interaction between RAG proteins and RSS in either mouse or llama makes any prediction of effects difficult. It seems possible from the residue differences described in this section that the llama RAG-1 protein may possess differences in the way it interacts both with other RAG-1 molecules, during dimerisation, and with llama RAG-2. The conservation of the DNA binding domain and integrase-like region suggests that the interaction between llama RAG-1 and llama RSSs may be similar to the better characterised murine RAG/DNA cleavage mechanism (section 1.8.5). At the same time however, the differences in the vicinity of the regions key for D32 and H609L related coding flank sensitivity suggest that llama RAG-1 may possess subtle differences in the manner in which it interacts with RSSs.

#### **4.11 General Structure of the Murine RAG-2 Protein**

By contrast to murine RAG-1 the murine RAG-2 protein shares relatively little homology to other protein domains. The only homology to other proteins lies within two regions (14-141 and 321-446 respectively) which share similarity with two integration host factor proteins known as HimA and HimD (61). These proteins form a heterodimer required during prokaryotic integration. It is likely therefore that the regions of RAG-2 with homology to these proteins play a similar essential role during V(D)J recombination (the RAG-1 protein has a region of similarity to prokaryotic integrases, section 1.8.5). The active core of the RAG-2 protein is located within the first three-quarters of the protein (amino acids 1-382 of the 517 residue protein). This corresponds approximately to the 'core' region of the RAG-2 protein used during *in vitro* cleavage experiments (section 5.4). Within the RAG-2 'core' a six-fold motif of 50 amino acids is found that is related to the kelch/mipp motif (224) that may play a role in protein/protein interactions such as RAG-1/RAG-2 dimerisation.

#### **4.12 Analysis of the Llama RAG-2 Gene for Regions of Potential Functional Significance.**

Without the complete sequence of the llama *rag-2* gene it is not possible to perform an evolutionary analysis similar to that performed on the *rag-1* gene. Regions of the *rag-2* gene that have not been isolated in this study may not show similar levels of homology to the murine form demonstrated by the characterised region. The finding that the region isolated in this study has 89% identity at nucleotide level to the corresponding murine region (llama *rag-1* shares 88.5% identity) strengthens the suggestion that the RAG proteins may act as an evolutionary clock (section 4.7)

#### **4.13 Features of Llama RAG-2 with Potential Functional Significance**

Approximately 80% (assuming an identical protein length to murine RAG-2) of the llama RAG-2 amino acid sequence has been isolated in this study (91% of the 'core' domain). Within this region a number of similar and non-similar residue differences are present with relation to the murine sequence. Of particular note are a total of eight differences between residues 286 and 309 within the 'core' domain of the sequence, N-terminal of the HimD motif. Overall the structure of the RAG-2 protein in the mouse and other species has been less well studied than that of RAG-1 and it is

therefore difficult to draw any conclusions as to the possible significance of residue differences.

#### **4.14 RAG Protein Expression Strategy**

The isolation of the sequence of the llama RAG-1 and RAG-2 proteins is described in section 4.3. After isolation of these genes it was possible to begin the process of recreating llama V(D)J recombination *in vitro*. The first step in developing a suitable assay for examination of the llama recombination process involved the expression and purification of biologically active RAG proteins. Given the better-defined role of RAG-1 (by comparison with RAG-2) in initial nonamer and heptamer binding (77, 225) and the more complete nature of the available sequence data regarding RAG-1, expression of llama RAG-1 alone was undertaken. Although the dual expression of both proteins could be argued to provide a more physiologically-relevant recreation of the *in vivo* recombination process, RAG-1 expression alone provided the prospect of isolating any unusual features of recombination that were specific to the RAG-1 protein. Given the considerable homology between llama and murine RAG-1, llama RAG-1 expression also provided a test of compatibility with murine RAG-2. The assay used to investigate llama V(D)J recombination is described further in Chapter 5.

##### **4.14.1 Expression of Llama RAG-1 in a Baculovirus System**

To examine the activity of the llama RAG-1 protein within the context of a cell-free assay system, the protein was expressed within an insect cell culture system by infecting insect cells with recombinant baculovirus containing the llama *rag-1* coding sequence. An overview of the expression system is given in figure 2.5. To produce soluble, high purity virus, the region corresponding to the murine 'core' protein (amino acids 384-1008 by murine numbering or 386-1010 by llama numbering) was cloned in-frame with both maltose binding protein (which provided not only a means of purification but also increased the RAG-1 protein solubility) (226) and a 6x histidine (6xHis) tag for secondary purification (227) (section 2.7).

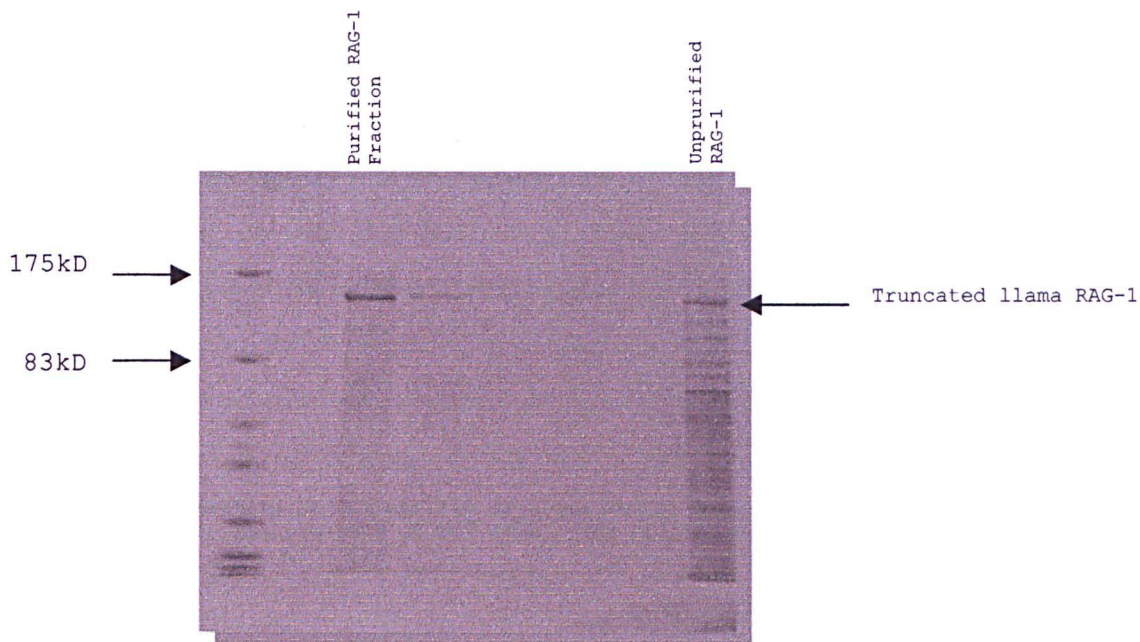
##### **4.14.2 Expression and Purification of Llama RAG-1/Murine RAG-2**

Llama RAG-1 protein in combination with murine RAG-2 was successfully expressed after co-infection of insect cells with bacmid DNAs containing both truncated murine

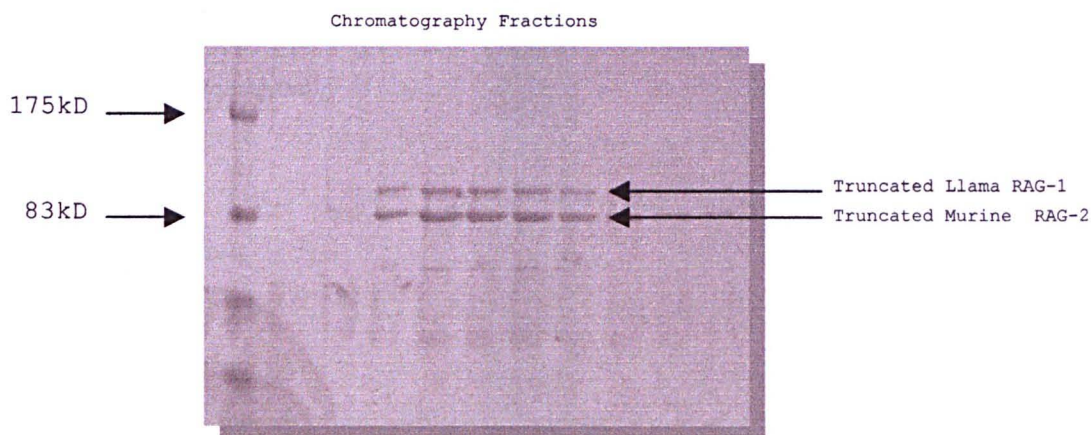
RAG-2 and the truncated llama RAG-1 insert. The results of first and second stages of RAG protein purification are shown in figure 4.9.

#### **4.14.3 Testing the Activity of the Llama RAG-1/Murine RAG-2 proteins**

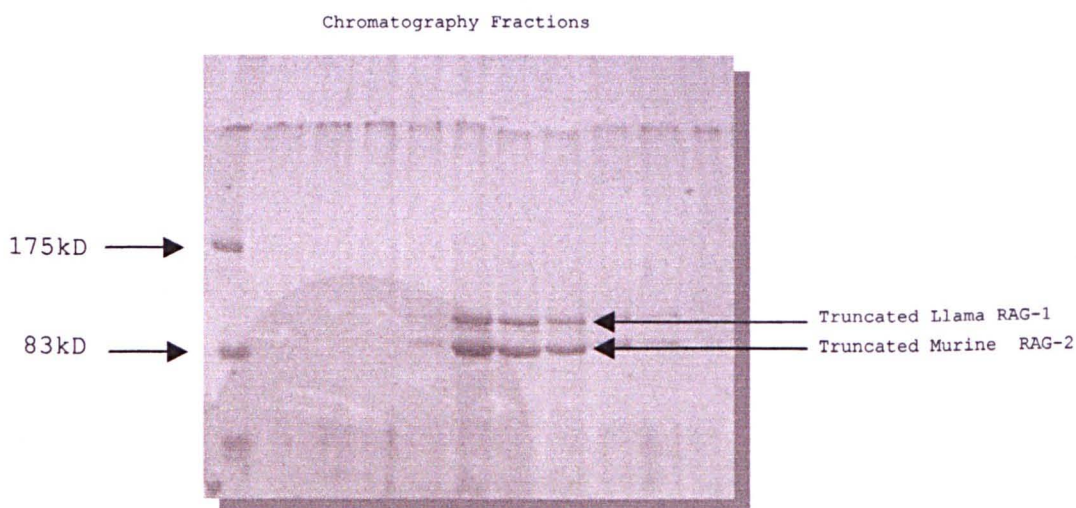
To test the biological activity of the purified proteins, a test assay was performed whereby a DNA oligonucleotide substrate was cleaved with the llama/mouse proteins (figure 4.1) (further details of the assay system are given in Chapter 6).



**Gel 1** Test purification of llama RAG-1 comparing protein before and after affinity chromatography



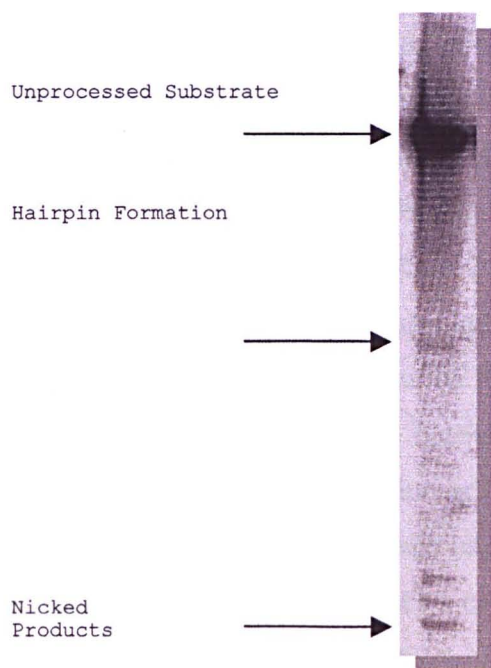
**Gel 2** Initial cobalt resin/ His tag purification of co-expressed llama RAG-1



**Gel 3** Secondary MBP/Amylose purification of co-expressed llama RAG-1 and murine RAG-2

**Figure 4.9 Purification of recombinant llama RAG-1 protein.** Gel 1 compares llama RAG-1 protein pre and post-cobalt resin affinity purification. Gel 2 shows co-purification of llama RAG-1 and murine RAG-2 by cobalt resin/ His tag affinity purification. Gel 3 shows purified proteins after secondary purification on MBP/Amylose column. All proteins were run on precast 6% SDS -Tris glycine polyacrylamide mini-gels. Elution conditions are described in the text. Markers were prestained broad range protein markers (New England Biolabs, Hitchin, UK)

Figure 4.10 Demonstration of the ability of llama RAG-1/ murine RAG-2 to cleave a murine control RSS in the presence of  $Mn^{2+}$  ions. The principle of the assay is described in detail in section 5.4.



#### 4.15 RAG Gene Isolation and Expression – Summary of Findings

In this chapter the successful isolation of the llama recombination activating gene sequences and the expression of a truncated form of the llama RAG-1 protein are described. In this chapter the following findings are reported:

- That llamas possess both recombination activating genes 1 and 2.
- Identification of the full coding sequence of the llama RAG-1 protein
- Identification of ~80% of the coding sequence of the llama RAG-2 protein
- Evidence that llama RAG-1 and RAG-2 may closely linked and convergently transcribed
- Evidence that llama *rag* genes are similar to those isolated from other species, and that llama *rag-1* shares highest nucleotide identity with the rabbit *rag-1* gene.
- Phylogenetic analysis of a range of RAG-1 proteins, including that of the llama that confirms the previously reported finding that RAG-1 sequence may act as an ‘evolutionary clock’. Partial llama RAG-2 sequence supports this.
- That the llama RAG-1 protein contains amino acid sequence corresponding to all the major domains reported in the murine RAG-1 protein.
- Identification of subtle differences between the function of llama and murine RAG-1 which may result from amino acid differences, particularly surrounding the domains involved in protein-protein interaction and coding flank recognition.
- That a truncated form of llama RAG-1 can be expressed as soluble biologically active protein in a baculovirus system.



#### 4.16 Discussion

This chapter reports the successful isolation and cloning of the complete sequence of the llama RAG-1 gene in addition to 80% of the coding sequence of the llama RAG-2 gene. The finding that the llama possesses both *rag-1* and *rag-2* genes suggests that the mechanisms of recombination utilised in the generation of llama antibodies are similar to those characterised in other species, rather than a completely novel mechanism (section 1.4 point (4)). This complements the evidence from the analysis of recombination signal sequences reported in chapter 3 (section 3.10). The differences between murine and llama RAG protein sequences are suggestive of subtle differences in the mechanisms of recombination (section 1.4 point (5)). The nature of the differences between murine and llama RAG proteins indicates that the mechanism of recombination may differ through contrasting mechanisms of protein-protein interaction, or perhaps differences in the recognition of coding flank sequences (section 4.10).

The successful expression of a truncated llama RAG-1 protein provides the opportunity to establish an *in vitro* assay to test the ability of llama recombination proteins to act on both heavy chain and classical antibody recombination signals. The results of these assays are described in chapter 5.

#### 4.17 Future Work

To further examine the recombination process that leads to the generation of llama antibodies the isolation of other components of the recombination machinery may be of value. In particular the isolation of the full llama RAG-2 nucleotide sequence, either through rescreening of the genomic library using the partial sequence reported in this chapter, or through a technique such as *random amplification of cDNA ends* (RACE) PCR could lead to the dual expression of llama RAG-1 and RAG-2 proteins. This would provide a more physiologically relevant assay to recreate the recombination processes undertaken within the llama.

A further natural extension of this project would be an examination of other proteins implicated in V(D)J recombination including perhaps terminal deoxynucleotidyl transferase, that is responsible for N addition and the DNA repair proteins known to act in the later stages of recombination (section 1.8.5). However, suitable assays to allow the investigation of the effects of these proteins on the latter stages of recombination have not, to our knowledge, yet been reported and so subsequent studies could not rely on an existing assay system.

## Chapter 5 Examination of the Process of Camelid V(D)J

### Recombination

#### 5.1 Abstract

The *in vitro* reconstitution of events taking place during llama antibody generation provides the opportunity to examine a range of phenomena related to V(D)J recombination that cannot be predicted by sequence data alone. This includes the ability of RAG proteins to interact both with each other and with llama DNA sequences, as they must if successfully recombined antibodies are to be generated *in vivo*.

Understanding of the intricacies of V(D)J recombination has been greatly advanced in recent years through the reconstitution of the process within an *in vitro* environment (214). The development of such a cell-free assay has led to a flurry of research activity that has dissected the detailed biochemical events that lead to the generation of antibody diversity. While the early stages of the murine recombination system have been investigated exhaustively through these studies, demonstrating for example, the ability of the RAG proteins to initiate recombination and the role of divalent cations within the reaction, relatively little work has examined potential abnormalities within recombination systems. To our knowledge, no published study has yet examined the process of V(D)J recombination within other species. Here the results of the first study to investigate non-murine recombination are reported. These demonstrate the interchangeable nature of components of the cell-free system, both through studies of murine RAG protein activity on llama derived RSSs and llama RAG protein activity on murine RSS. After the optimisation of these assays murine RAG proteins were found to be unable to fully process llama recombination substrates. In this chapter a method is provided that allows full restoration of this recombination activity.

Despite a number of residue differences between the expressed llama RAG-1 protein and its murine counterpart, this chapter demonstrates that both proteins are able to act in concert with murine RAG-2 and fully cleave mouse recombination substrates.

## 5.2 Introduction

The events that take place during human and murine V(D)J recombination *in vivo* are complex and highly regulated. RSSs may be more than a megabase apart (31) and factors such as nucleosomal positioning and higher order chromatin structure are thought to play a crucial role (228, 229). Given the limited understanding of such factors it is not possible to fully reconstitute the process of llama V(D)J recombination as it occurs within the B-cell. However, given the sequence data obtained from V, D and J segments (Chapter 3) the development of an assay allowing the demonstration of the role of these sequences in targeting recombination was considered a priority.

After assessment of a variety of plasmid-based recombination systems (230-233) a simple, yet relatively robust assay system was chosen, originally developed within the laboratory of Martin Gellert at the National Institutes of Health in Bethesda (65). To provide a more physiologically-relevant reconstruction of the events taking place during camelid B-cell development both murine RAG-1/2 proteins (typically used in such studies) and llama RAG-1 protein were used to cleave the RSSs isolated from the llama heavy chain loci. It is to be noted that throughout these studies a range of assay conditions best suited to the activity of murine RAG1/2 proteins were used. It is possible therefore that the llama protein may be physiologically active outside this range of conditions

## 5.3 Strategy

The aims of this study were twofold. Firstly the ability of murine RAG proteins to process RSSs derived from llama sequences was examined, thereby demonstrating interspecies conservation of the initial stages of V(D)J recombination. The second aim was to examine the ability of llama RAG-1 to interact with murine RAG-2 and murine and llama RSSs.

Demonstration of

- 1) Llama RAG-1/Murine RAG-2 interaction
- 2) Llama RAG-1/ Llama RSS interaction, and
- 3) Llama RAG-1/ Murine RSS interaction

would illustrate the considerable degree of conservation between methods of recombination, and also highlight any potential differences between the species caused, for example, by amino acid differences within the llama RAG-1 sequence.

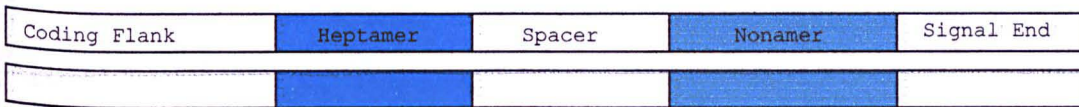
## **5.4 The Cell-Free Recombination Assay**

The basic principle of the cell-free recombination system utilised in this study is described in a number of recent papers (65, 72, 215, 234, 235). Briefly, short oligonucleotides, containing approximately 16 nucleotides of coding end, heptamer, 12 or 23 nucleotide spacer, nonamer and 6 nucleotides of the signal end were synthesised together with their reverse complements. The oligonucleotides were then radiolabelled at their 5' ends before annealing to their reverse complements (figure 5.1). Biologically active RAG-1/2 proteins expressed within a baculovirus system (sections 2.7 and 4.14) were purified by affinity chromatography and added to a buffer containing divalent cations and in some cases high motility group proteins 1 and 2 (HMG1/2) (sections 1.8.5, 2.6 and 5.4.2(a) and (b)). The oligonucleotide substrates were then incubated with the recombinant RAG proteins. After incubation the reaction was run on a polyacrylamide sequencing gel to allow visualisation of the oligonucleotide substrates and any processing that may have taken place.

### **5.4.1 Interpretation of Recombination Assays**

The products of cell-free recombination assays using the Gellert system described above (section 5.4) fall into two categories. Initial nicking of the double stranded DNA substrate by the RAG protein complex generates a short (typically 16 base), labelled single-stranded DNA product that migrates towards the base of polyacrylamide gels during electrophoresis. The second stage of recombination involves hairpin formation between the two annealed DNA strands at the coding end (figure 1.7). This product migrates as a higher band approximately midway between any nicked DNA products and the dense non-processed substrate that runs at the top of the gel. This migration pattern is illustrated in figure 5.2 and the events that take place during RAG interaction with the recombination substrate(s) are summarised in figure 5.3.

**Figure 5.1 Anatomy of a Recombination Substrate.** Substrates typically consisted of 16 nucleotides of coding flank including the 3' end of the variable gene segment coding sequence, the heptamer, 23 or 12 nucleotide spacer, nonamer and a further five nucleotides at the signal end. A full recombination substrate consisted of a double strand DNA molecule including the recombination substrate annealed to its reverse complement:



**Figure 5.2 Schematic describing the typical migration pattern of recombination assay products.** Line diagrams illustrate processing of recombination substrates indicated by these bands. Red triangles indicate RSSs, the significance of which is illustrated in figures 1.7 and 5.3

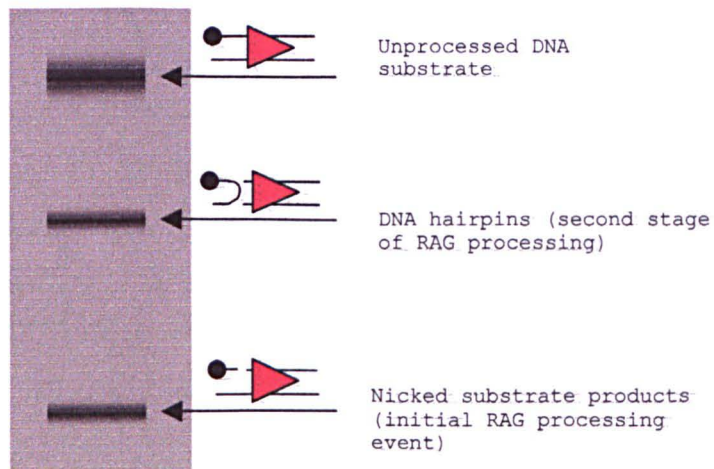
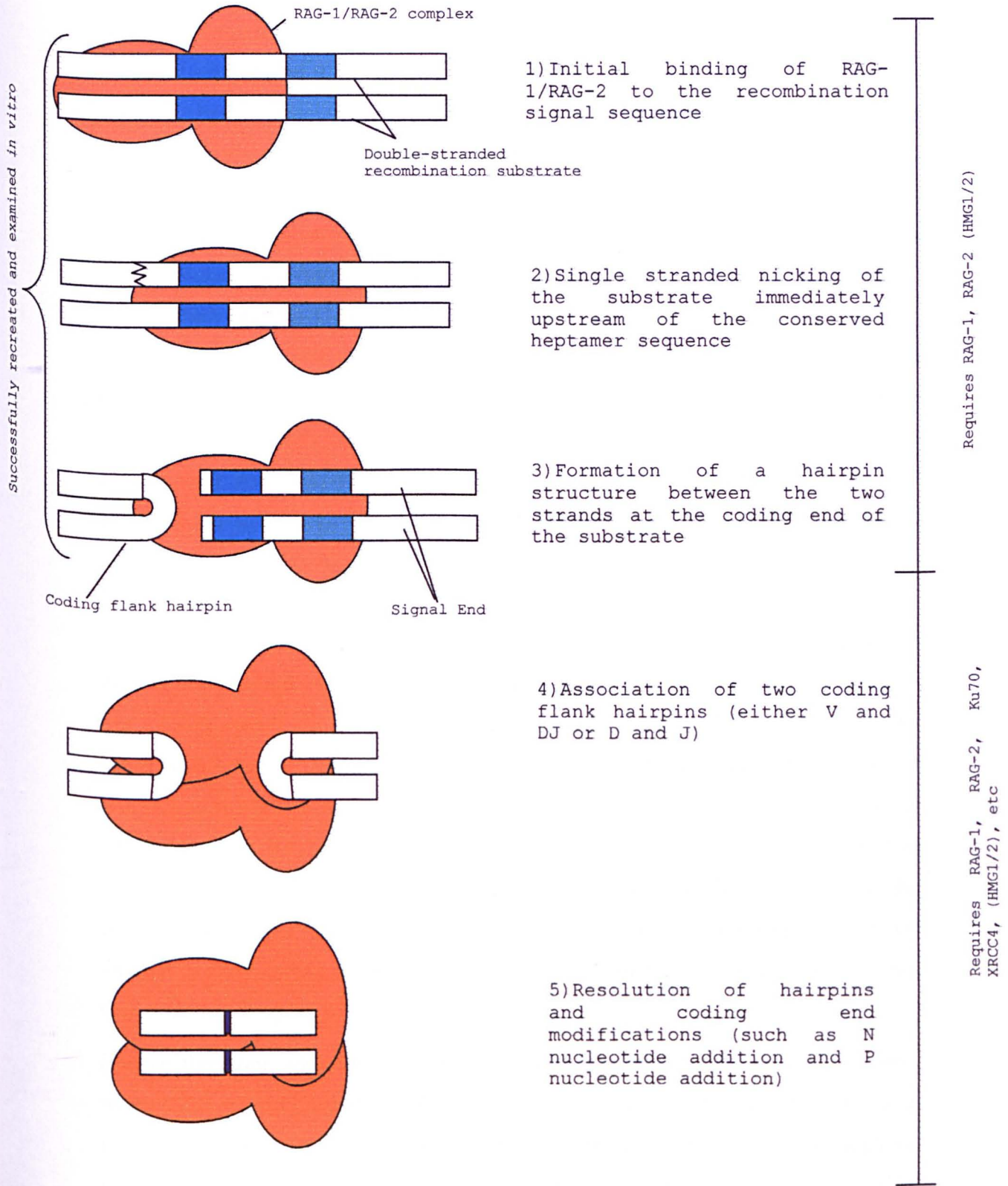


Figure 5.3 Diagram illustrating the stages of processing of recombination signal sequences (RSSs) by the RAG proteins. Stages 1-3 can be completed *in vitro* by RAG proteins alone. However resolution of hairpins and rejoining of coding ends require a number of other factors described below. This chapter describes only the analysis of Steps 1-3 of the recombination process.



#### 5.4.2 Components of *In Vitro* Studies of RAG-1/RAG-2 Activity

Although under optimal conditions murine 'core' RAG-1/RAG-2 proteins are both necessary and sufficient to promote the early stages of V(D)J recombination, a number of the components of the cell-free reaction can affect the results obtained during artificial RAG/RSS interaction.

##### a) Recombinant RAG-1/RAG-2 Protein

Batch to batch variation in protein purity between preparations of the RAG-1/2 proteins are common and can have considerable effects on the assay (236). For this reason the assay must be fully optimised for each batch of protein. In some cases protein is not sufficiently active to generate valid results despite optimisation. Murine RAG proteins (215) were a kind gift of Dr K. Hiom.

##### b) High Motility Group (HMG) Proteins

The ubiquitous nuclear chromatin-associated HMG1/2 proteins make an important contribution to RAG/RSS interactions both *in vitro* and *in vivo* (215, 237-240) and are thought to act as a DNA-bending accessory factor (239, 241). In particular cleavage of 23-spacer RSSs (23-RSSs) is stimulated more than tenfold in the presence of HMG1/2. It is believed that HMG1/2 may stabilise the RAG-1/2/DNA interaction (239).

##### c) Divalent Cations

The progress of RAG/RSS interactions is highly dependent on the nature of the divalent ion available within the reaction. The most physiologically relevant cation is  $Mg^{2+}$  (242), which allows nicking of single double-stranded RSS substrates in the presence of murine RAG-1/-2. However, hairpin formation, the second stage of processing in  $Mg^{2+}$  is not possible without the introduction of a second, alternative length spacer RSS substrate. As discussed previously (section 1.8.4) the alternative spacer lengths represent single and double turns in the DNA helix and the conformation of the RAG protein complexes that associate at each spacer type are thought to allow interaction only with RAG proteins of the alternative conformation, therefore enforcing the 12/23 rule.

Hairpin formation within a single RSS substrate *is* possible in the presence of  $Mn^{2+}$  when 12/23 reactions are uncoupled. This is thought to be the result of an alteration in the geometry of the RAG-1/2 active site in the presence of this ion. Although the presence of this ion may not best simulate the *in vivo* environment during V(D)J



recombination the use of  $Mn^{2+}$  during the generation of hairpins in the presence of a single RSS is commonly used throughout the literature (72, 242).

#### d) Oligonucleotide Substrates

Oligonucleotides were gel purified prior to radiolabelling, annealing and use within the assay in order to remove incomplete oligonucleotides (n-1, n-2, etc) that may have formed during synthesis and may produce spurious banding during visualisation of assay products.

#### 5.4.3 Oligonucleotide Design

To provide the most realistic test of components of the llama recombination system in the *in vitro* assay, oligonucleotides representative of the recombination signal sequences found within the llama  $V_H$  and  $V_{HH}$  repertoire were designed. Our analysis of germline gene segment expression within a Unilever cDNA library in section 3.18 indicates that sequences derived from gene segment  $V_{HH3}$  or similar make up a larger proportion of cDNA sequences than any of the other isolated variable gene segments in this chapter. Single substrates derived from the RSSs associated with germline variable gene segments  $V_{HH3}$  (representing heavy chain gene segments) and  $V_{H1}$  (representing classical gene segments) were therefore synthesised. Although these sequences are broadly similar, specific differences are present, particularly within the coding flank and spacer sequence (table 5.1).

$V_{III3}$  and  $V_{II1}$  contain all the sequence elements that have been shown to be crucial to recombination in the better-characterised human and murine recombination systems (section 3.10). In addition their coding flank sequences upstream of the heptamer are approximately representative of the coding flanks isolated from the six V region germline clones reported in Chapter 3. It is not certain that these recombination signal sequences are utilised *in vivo* but the close similarity of the germline sequences to expressed cDNAs (section 3.18) suggests that recombination signal sequences similar if not identical to  $V_{H1}$  and  $V_{HH3}$  are involved in successful recombination events. Although no significant differences are apparent between RSS signals derived from the isolated germline clones (figure 5.4) the use of sequences derived from both llama  $V_H$  ( $V_{H1}$ ) and  $V_{HH}$  ( $V_{HH3}$ ) provide the opportunity to observe any subtle differences in RSS processing between the two antibody types that may occur *in vitro*. A murine derived 23-RSS control was also used in these studies, the sequence of which is shown in table 5.2.

**Table 5.1 Summary of Recombination Signal Sequences derived from V<sub>H</sub>, D and J Gene Segment Data.** Colours indicate runs of nucleotides that might either promote (Blue) or inhibit (Red) nucleotide deletion (64, 100)(section 5.14). The trinucleotide adjacent to the heptamer sequence is shown in **bold**. The functional significance of each nucleotide within the RSS is described in section 1.8.3.

Sequence	Coding End	Heptamer	Spacer	Nonamer
J <sub>H</sub> 1	TCGAGATACCTGT <b>AGC</b>	CACACTG	GTCTGCCTGCCTGGCCCCAGTG	CACAAAACCC
J <sub>H</sub> 2	TGCGTCC <b>AAAGCATTG</b>	CACAGGG	ACATAGTCCC GGCTCTCCCCCAG	ACATAAACCC
J <sub>H</sub> 3	CAGTAGTCATA <b>CTCAT</b>	CACAGCG	CCACGGGCCCCGTTAGGTGCTGT	GCAAAAACCC
J <sub>H</sub> 4	TATTC <b>AAACTGGGGGT</b>	CACATTG	TGACAACTGTGTCAAGACCCCAG	GCAAATGCT
J <sub>H</sub> 5	CAGGAACCC <b>AAAGTCAG</b>	CACAACG	CCACGGGCCCCGTTAGGTGCTGT	GCAAAAACCC
V <sub>HH</sub> 1	ATTACTGTGCAGC <b>CAGA</b>	CACAGTG	AGGGGAAGTCATTGTGAGCCCAG	ACAAAAACA
V <sub>HH</sub> 2	ATTACTGTGCAGC <b>CAGA</b>	CACAGTG	AGGGGAAGTCATTGTGAGCCCAG	ACAAAAACA
V <sub>HH</sub> 3	ATTACTGTAATGC <b>CAGA</b>	CACAGTG	AGGGGAAGTCATTGTGAGCCCAG	ACAAAAACCC
V <sub>HH</sub> 4	ATTACTGTGC <b>AAAAGA</b>	CACAGTG	AGGGGAAGTCATTGTGAGCCCAG	ACAAAAACCC
V <sub>HH</sub> 5	TGTAT <b>TACTGTGCGCA</b>	CACTGTG	AGGGGAAGTCATTGTGAGCCCAG	AAAAAAACCC
V <sub>H</sub> 1	ATTACTGTGCAGC <b>CAGA</b>	CACAGTG	AGGGGAAGTCGGTGTGAGCCCAG	ACACAATCC
D <sub>H</sub> 1 5'	CTGT <b>GGCTCCAGTTAG</b>	CACTGCG	GTTCCCAGCTCA	GCCAAAACCC
D <sub>H</sub> 1 3'	CAGTGCTAACTGG <b>GAGC</b>	CACAGTG	ACTGACAACCTCT	ACAAAAACT

**Figure 5.4 Comparison of the sequence of the llama recombination substrates.** Differences shown in bold.

	Coding flank	Heptamer	23-Spacer	Nonamer	Signal
V <sub>HH</sub> 3 (V <sub>HH</sub> )	ATTACTGT <b>TAATGC</b> CAGA	CACAGTG	AGGGGAAGTC <b>CAT</b> GTGAGCCCAG	ACAAAAACCC	<b>TGCTC</b>
V <sub>H</sub> 1 (V <sub>H</sub> )	ATTACTGT <b>GCAGC</b> CAGA	CACAGTG	AGGGAAAGTC <b>CGG</b> TGTGAGCCCAG	ACACAATCC	<b>TGCAGG</b>

**Table 5.2 Sequence of recombination substrates used in this study.** Sequences given in this table refer only to the 'top' strand of the recombination substrate. Reverse complements of these sequences were also used to generate double stranded substrates

Substrate	Coding Ends	Heptamer	Spacer	Nonamer	Signal
23 control	GATCTGGCCTGTCTTA	CACAGTG	GTAGTACTCCACTGTCTGGCTGT	ACAAAAACCC	TGCAG
12 control	GATCTGGCCTGTCTTA	CACAGTG	CTACAGACTGGA	ACAAAAACCC	TGCAG
V <sub>HH</sub> 3	ATTACTGTAATGC <b>CAGA</b>	CACAGTG	AGGGGAAGTCATTGTGAGCCCAG	ACAAAAACCC	TTGCTC
V <sub>HH</sub> 3F	GATCTGGCCTGTCT <b>TA</b>	CACAGTG	AGGGGAAGTCATTGTGAGCCCAG	ACAAAAACCC	TTGCTC
V <sub>H</sub> 1	ATTACTGTGCAGC <b>CAGA</b>	CACAGTG	AGGGAAAGTCGGTGTGAGCCCAG	ACACAATCC	TGCAGG
V <sub>H</sub> 1F	GATCTGGCCTGTCT <b>TA</b>	CACAGTG	AGGGAAAGTCGGTGTGAGCCCAG	ACACAATCC	TGCAGG

## **5.5 Analysis of Llama Recombination Signal Sequences**

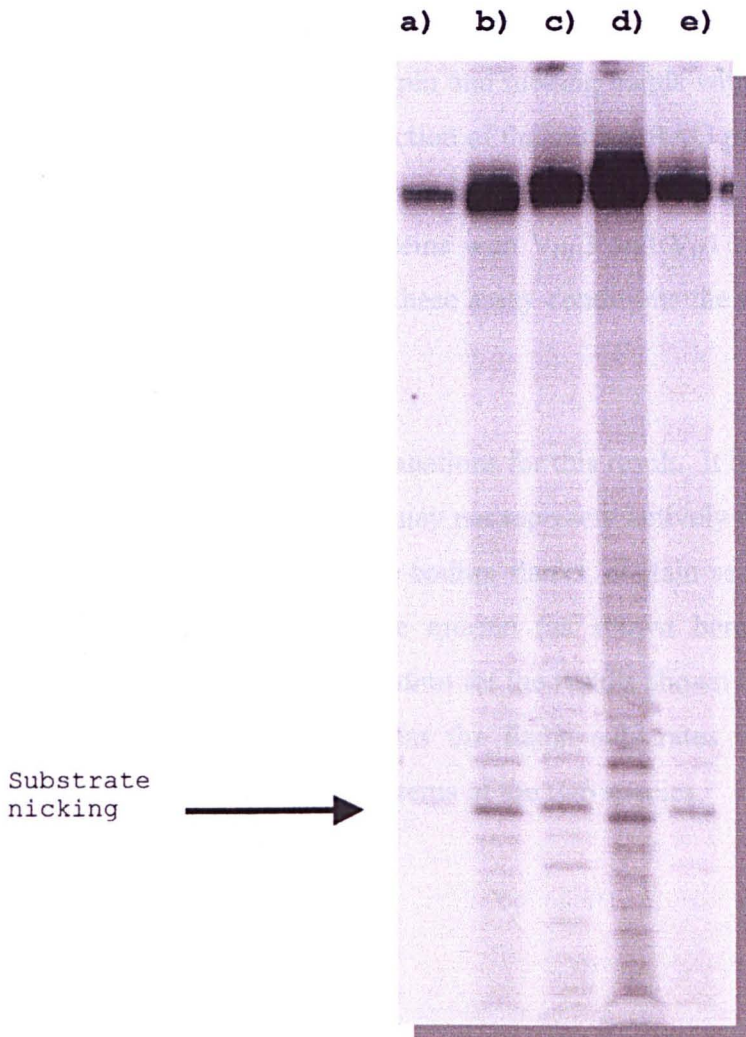
Specific sequence requirements for *in vitro* substrate cleavage have been identified (52, 83) which, while almost certainly relevant to *in vivo* recombination, are of known importance to *in vitro* assays. The first three bases of the heptamer are essential for recombination during cleavage with purified RAG proteins. By contrast substrates that lack a recognisable nonamer are able to undergo a reduced level of cleavage. The composition of the coding sequence flanking recombination signal sequences can have considerable effects on the efficiency of the cleavage reactions (236).

Initial experiments (sections 5.6-5.9) sought to examine the ability of murine RAG-1/2 proteins to cleave llama RSSs.

For this purpose murine RAG proteins were initially incubated with murine oligonucleotide controls (in this case 12-RSSs) and two llama substrates V<sub>HH3</sub> and V<sub>H1</sub>, to test the ability of the protein to induce the first step of recombination, nicking. Substrates were incubated in the presence of both magnesium and manganese ions. The results of these experiments are shown in sections 5.10-15.12

## **5.6 Initial Murine RAG Processing of Llama Recombination Signal Sequences**

This first gel (figure 5.5) examines only the initial recognition and processing event of RAG-RSS interaction, nick formation in the presence of Mg<sup>2+</sup> ions. Recombination substrates were incubated in the presence of recombinant murine RAG-1/-2 proteins. The llama RSSs are clearly nicked by murine proteins, suggesting a high degree of similarity between the recombination systems of both species. This gel therefore provides direct evidence of the conservation of RAG-RSS interaction between species at a functional level.

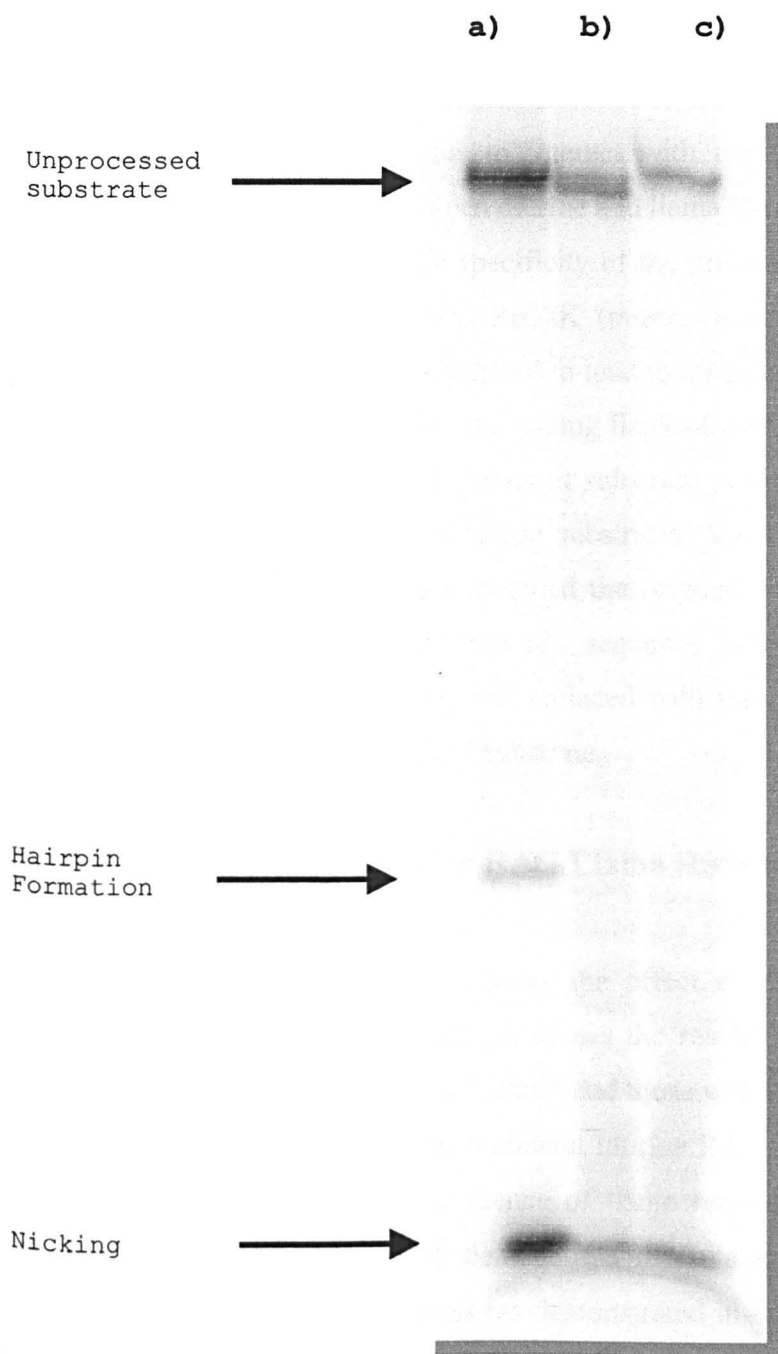


**Figure 5.5** Nicking of llama RSS DNA substrates by murine RAG-1/-2 in the presence of  $Mg^{2+}$ . Lane a) 23-RSS nicking positive control Lane b)  $V_{HH}3$ , Lane c)  $V_{HH}3F$ , Lane d)  $V_{H1}$ , Lane e)  $V_{H1}F$ . Variations in the nicked product size and additional bands may represent slight differences in RAG protein specificity at the heptamer (12% Polyacrylamide Gel)

## **5.7 Interaction of Murine RAG Proteins with Llama RSSs in the Presence of Manganese**

The second gel (figure 5.6) shows the results of processing of llama recombination signal sequences derived from the regions downstream of the variable (V) gene segments in the presence of manganese and recombinant murine RAG-1 and RAG-2 proteins. The presence of both hairpin and nicking bands within the positive control lane testifies to the successful interaction of the murine RAG proteins with the control RSS substrate. While bands corresponding to nicked products are present after incubation of the murine RAG proteins with  $V_{HH3}$  and  $V_{H1}$  no hairpin formation is present. This demonstrates that in these assay conditions the murine proteins cannot process llama RSSs fully.

There are a number of possible explanations for this result. It is important to re-iterate the possibility that these sequences may not represent actively recombining RSSs, and that the sequences, in particular the coding flanks, contain sequence differences that prevent recombination both in the murine (as shown here) and llama systems. However, the most probable explanation for the results shown here is that the inability of murine RAG proteins to process the llama substrates is the result of subtle differences in the recombination systems of the two species.



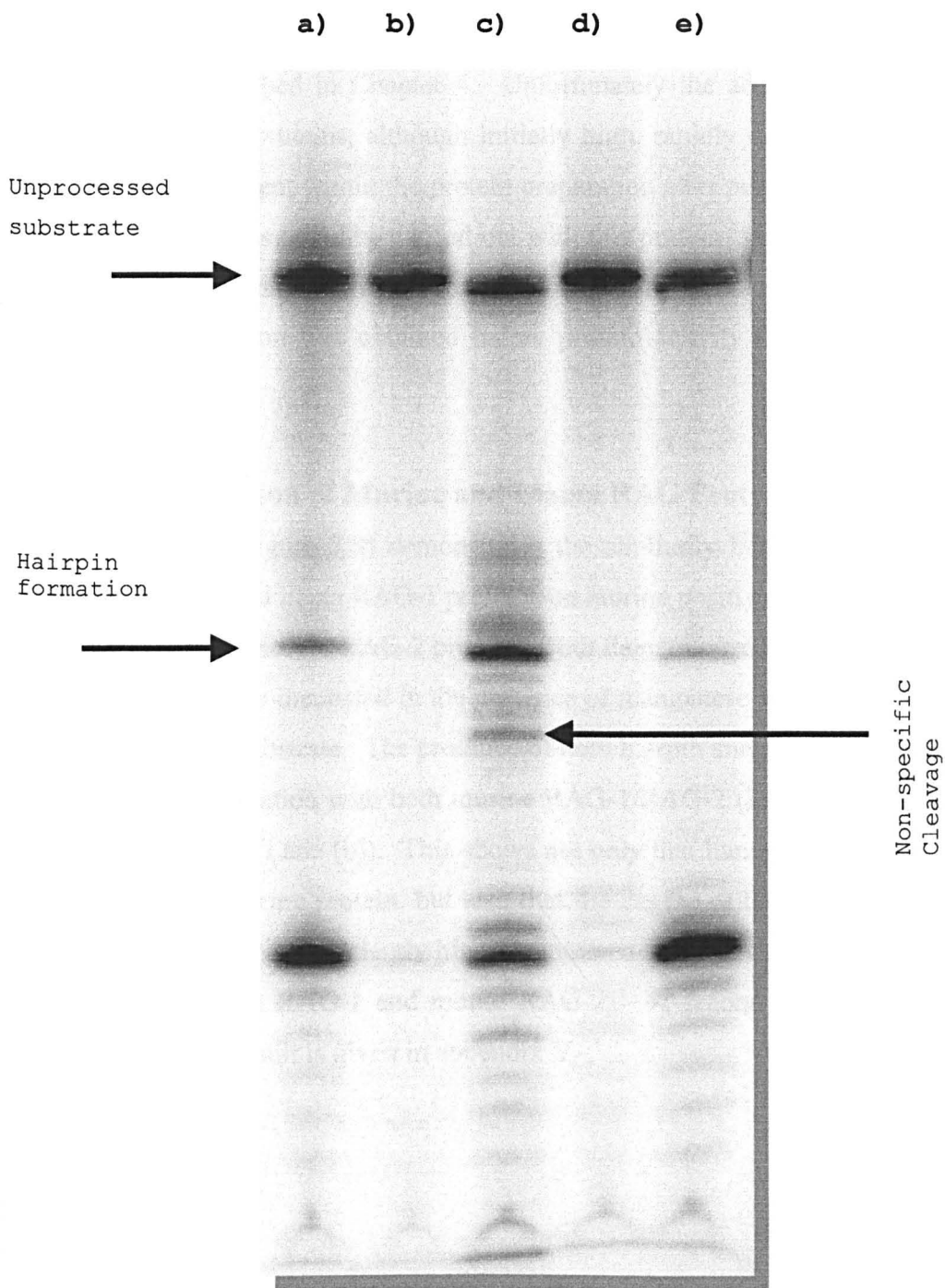
**Figure 5.6 Ability of murine RAG-1/-2 to cleave llama recombination substrates in the presence of  $Mn^{2+}$ , in lane a) murine control 23-RSS, in lane b)  $V_{HH}3$  RSS and in lane c)  $V_{H1}$  RSS (12% Polyacrylamide Gel)**

## **5.8 Substitution of Coding Flank**

Given that murine RAG proteins were unable to fully process the llama classical and heavy chain only derived recombination substrates, attempts were made to restore the ability of the murine proteins to interact with the llama sequences. Chapter 4 describes the differences between murine and llama RAG-1 amino acid sequences that may lead to differences in the specificity of the proteins. In section 4.10 amino acid differences such as V626I and R628K (murine numbering) in the vicinity of the H609L mouse mutation are predicted to lead to a similar coding flank sensitivity. To test whether differences within the coding flank of the llama recombination substrates were responsible for the differences in substrate processing observed in figures 5.5 and 5.6 further llama recombination substrates, V<sub>HH</sub>3F and V<sub>H</sub>1F (table 5.2) were synthesised. These substrates included the original heptamer, spacer, nonamer and signal ends as derived from primary sequence data but differed in that the 16 nucleotide llama coding flank was replaced with the murine coding flank sequence from the murine positive control substrate.

## **5.9 The Effect on Murine RAG/Llama RSSs of Coding Flank Substitution**

The third gel (figure 5.7) examines the effect of the coding flank substitutions described in section 5.8. This gel shows the results of the incubation of both the original llama recombination substrates and those with substituted coding flanks in the presence of manganese and recombinant murine RAG-1/-2 proteins. The restoration of hairpin formation in the presence of llama recombination substrates containing murine coding flanks clearly demonstrate that the inability of murine proteins to generate hairpins in llama RSSs (as demonstrated in figure 5.5 and lane (b) of figure 5.7) is a result of coding flank differences.



**Figure 5.7 Restoration of hairpinning through the substitution of coding flanks in the presence of  $Mn^{2+}$**  a) Murine 23-RSS Positive Control, b)  $V_{HH3}$  original flank, c)  $V_{HH3F}$  (substituted flank), d)  $V_{H1}$  original flank, and e)  $V_{H1F}$  (substituted flank). Nicking in this reaction is variable (12% polyacrylamide gel)

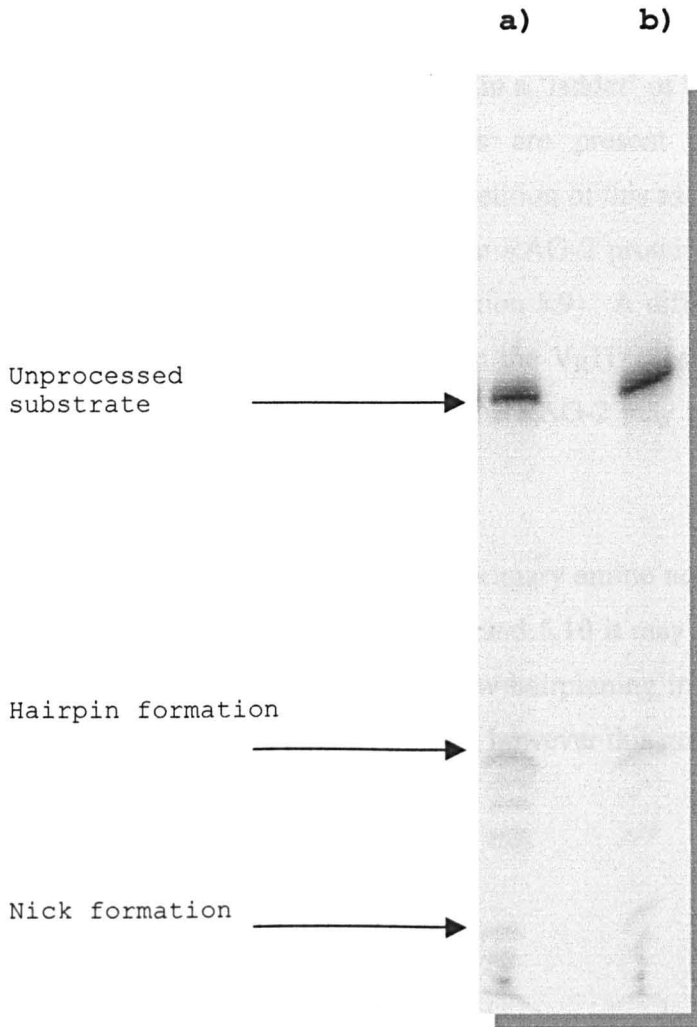


### **5.10 Llama RAG-1/Mouse RAG-2**

After examination of the effects of murine RAG proteins on llama recombination signals further studies were conducted using the llama RAG-1/mouse RAG-2 protein preparation described in Chapter 4. Unfortunately the activity of the llama RAG-1/mouse RAG-2 proteins, although initially high, rapidly decreased, perhaps due to contaminants present within the protein preparation after purification. Therefore only preliminary studies could be carried out with this protein. As the results presented in subsequent sections illustrate, information regarding the activity of the llama/mouse protein combination was obtained before protein activity dropped below a useable level.

### **5.11 Comparison of Murine and Llama RAG Protein Activity**

The fourth gel (figure 5.8) demonstrates the similarity between the activity of the murine RAG-1 and llama RAG-1 proteins on murine positive control RSSs when both are coupled with murine RAG-2 protein. Both llama/mouse and mouse/mouse protein combinations were incubated in the presence of manganese and a 23bp-spacer murine positive control substrate. The presence of both hairpin and nicked products is clearly shown after incubation with both murine RAG-1/RAG-2 (Lane (a)) and llama RAG-1/murine RAG-2 (Lane (b)). This shows not only that llama RAG-1 acts in a similar manner to the murine protein, but also that the degree of homology between murine and llama RAG-1 is sufficiently high for interaction between the proteins of different species (i.e. llama RAG-1 and mouse RAG-2). A second gel of lower resolution, confirming this result is given in appendix V.

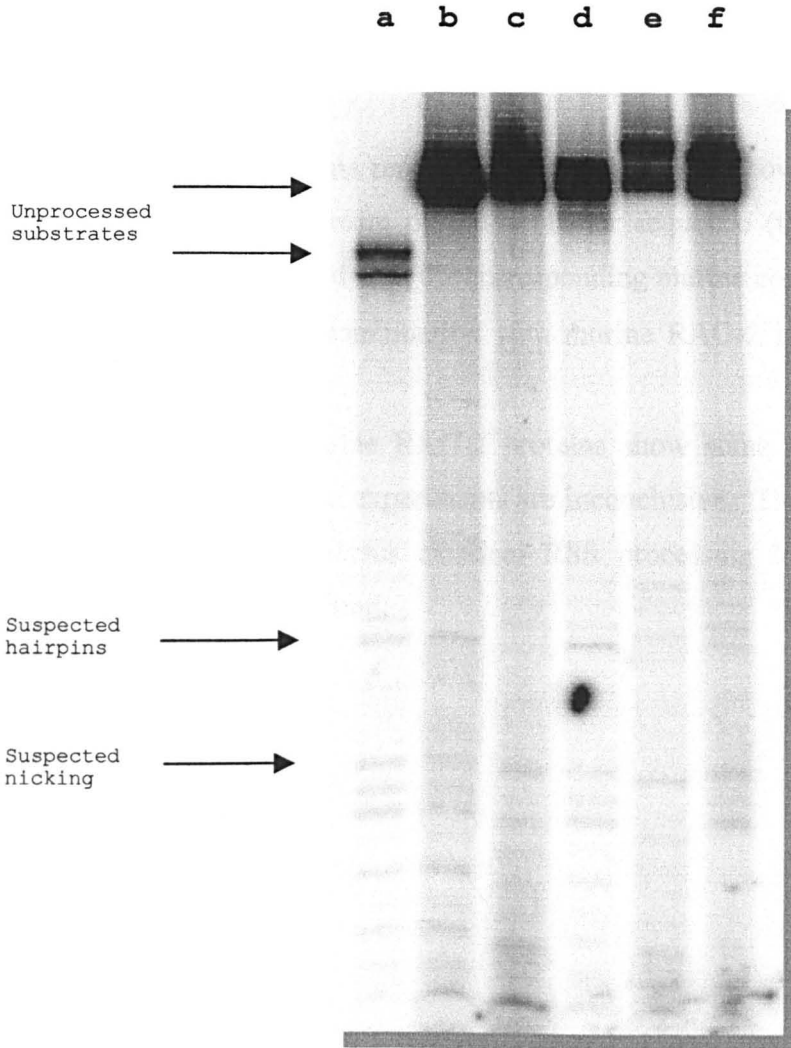


**Figure 5.8 Processing of control 12-RSS with murine RAG-1/-2 (lane a) and llama RAG-1/murine RAG-2 (lane b).** Substrates were processed in the presence of  $Mn^{2+}$  (12% Polyacrylimide Gel). A second lower resolution gel is provided in appendix V.

## 5.12 Processing of Llama Recombination Substrates with Llama RAG-1/Murine RAG-2.

Figure 5.9 shows the effects of llama RAG-1 protein in combination with murine RAG-2. The proteins show some specific activity but also a degree of non-specific nicking and possible hairpin formation resulting in a 'ladder' of bands within the gel. Probable specific hairpin and nicking bands are present although the rapid deterioration of protein activity did not allow repetition of this assay. The gel appears to show similar activity of llama RAG-1/murine RAG-2 proteins to that previously shown for murine proteins alone (figure 5.7, section 5.9). A difference with possible significance is the absence of hairpinning within the  $V_{H1F}$  substrate. This suggests that the llama protein in combination with murine RAG-2 may be unable to process the RSS even after coding flank substitution.

Given the differences within the llama RAG-1 primary amino acid sequence (section 4.10) and the observations made in sections 5.9 and 5.10 it may have been predicted that the llama/murine RAG proteins would allow hairpinning in the presence of the original llama coding flanks (i.e.  $V_{HH3}$  and  $V_{H1}$ ), however this gel suggests that this is not the case.



**Figure 5.9 Processing of llama recombination substrates with llama RAG-1/murine RAG-2.** Lane a and b show murine control RSSs (12 and 23 spacer respectively), lane c and d show  $V_{HH3}$  and  $V_{HH3F}$  while lanes e and f show  $V_{H1}$  and  $V_{H1F}$  respectively (15% Polyacrylamide Gel).

### 5.13 Recombination Assays – Summary of Findings

- Core ‘murine’ RAG proteins are able to interact with and generate single stranded nicks within oligonucleotide substrates representative of both heavy chain antibody and classical antibody RSSs *in vitro*.
- Core ‘murine’ RAG proteins are unable to reproduce the second stage of V(D)J recombination within llama heavy chain and classical recombination substrates *in vitro*.
- Hairpin formation within llama recombination substrates is, however, possible if the region immediately upstream of the heptamer sequence (the coding flank) within the substrate is replaced with the corresponding murine control sequence.
- Truncated llama RAG-1 in combination with murine RAG-2 is able to process murine RSS successfully.
- Combined llama RAG-1/murine RAG-2 proteins show some activity at llama RSSs although results of these experiments are inconclusive. There may be some difference in heavy chain versus classical RSS processing but further, more thorough investigation is required.

## 5.14 Discussion

In this chapter the process of V(D)J recombination has been explored through the development of an *in vitro* recombination system that allows detailed analysis of the initial stages of recombination. Deriving recombination signal sequences from data obtained in previous chapters the high level of conservation between recombination events in different species is demonstrated. Murine RAG proteins are found to both bind to and interact with llama RSSs *in vitro* through the production of single stranded nicks in llama-derived recombination substrates. While this process does not confirm the preservation of absolute RAG-dependent sequence specificity between species it is clear that the llama recombination process is driven by RAG-dependent recombination signal sequence interaction, as has been previously shown in the mouse (215). Furthermore these results illustrate that there are no apparent fundamental differences between the recombination events taking place in the generation of heavy chain and classical llama antibodies, at least in the case of the representative recombination signal sequences used in this study. The assay used in this chapter provides strong evidence for the utilisation of variable gene segments such as those associated with the  $V_{HH3}$  and  $V_{H1}$  recombination substrates during *in vivo* antibody generation (section 1.4 point (2)). The results described in this chapter also suggest that the llama immune system has subtly refined the recombination mechanisms characterised in other species such as the mouse and human (section 1.4 point (5)). The finding that llama RAG proteins can direct the first stages of V(D)J recombination make the use of completely novel recombination mechanisms by the llama unlikely (section 1.4 point (4)).

Assuming that the RSSs obtained through genomic library screening are representative of actively recombining sites within the llama immunoglobulin heavy chain locus this chapter describes specific differences in the second stage of recombination, that of hairpin formation, between murine and llama RSS. The finding that sequence differences within the coding flank of the RSS are responsible for this effect echos a previous finding of coding flank sensitivity reported in a mutant murine form of RAG-1 (92). This confirms the hypothesis formed in section 4.10 where it is suggested that a number of specific amino acid differences between llama and murine RAG-1 sequences, in the vicinity of a murine mutation known to lead to

coding flank sensitivity may have a similar effect on the action of the llama RAG-1 protein. If this assumption is to be accepted, one or more of three possible evolutionary events may have taken place during the divergence of the recombination systems of the two species. The llama RSS or RAG gene sequence (or, indeed both) may have evolved to interact fully and enable completion of recombination. It is conceivable that the differences between the initial stages of recombination reported here may in some way allow, or favour the generation of heavy chain antibodies. One possibility stems from the work of Ann Feeney and co-workers (66, 100) who have shown that processing of recombination interfaces such as that present within the extended CDR3 of the llama heavy chain antibody proceeds in a manner highly dependent on the nature of the coding flank nucleotide composition (an indication of the possible effects within the llama RSSs isolated in this thesis is given in table 5.1). Specific N deletions and P additions have been shown to occur at particular coding flank sequences. It is possible, therefore, that the differences in llama and murine coding flanks represent an adaptation of the llama recombination process in order to favour processing events such as nucleotide addition that in turn favour the formation of the extended CDR3. If this is the case the significance of such differences is unclear.

## 5.15 Future Work

Future investigation of llama V(D)J recombination would initially involve the use of llama RAG-1/murine RAG-2 proteins of a higher specific activity. It is unclear what degree of the non-specificity exhibited by this protein preparation may be the result of the combination of proteins of different species. A preparation of murine RAG proteins carried out in parallel with the llama RAG purification showed lower than normal activity and a degree of non-specificity suggesting that further preparation of llama/murine proteins may be beneficial.

The full recreation of the llama V(D)J recombination process will require the isolation of the full llama RAG-2 protein sequence perhaps through rescreening of the llama genomic library described in section 2.3 using a probe derived from the sequence given in section 4.6. Once this sequence is available the simultaneous expression of truncated llama RAG-1 and -2 proteins can be undertaken. This will allow more complete investigation of the coding flank sensitivities described in sections 4.10 and 5.9.

A more comprehensive understanding of the events taking place during llama V(D)J recombination would require the investigation of other processes governed by terminal deoxynucleotidyl transferase and double strand repair proteins. For this reason the isolation of the llama equivalents of these genes would be a priority. The isolation of the TdT gene sequence would have the added value of allowing examination of the upstream regulatory elements that may upregulate N nucleotide addition (section 1.8.6 and discussed in section 7.5).



## Chapter 6

### Llama Immunoglobulin Constant Region Genes.

#### 6.1 Abstract

The immunoglobulin constant region genes encode the invariant portion of the antibody. The three constant immunoglobulin domains are responsible for interaction with cells of the immune system via Fc receptors and for activation of innate immune defences through the complement pathway. Camelid heavy chain antibodies are characterised by the absence of the first constant heavy domain ( $C_{H1}$ ), reducing antibody length and preventing interactions between both heavy/light chains and variable/constant domains. The analysis of an unrearranged llama genomic DNA library has enabled isolation of a number of immunoglobulin constant genes, and more specifically, demonstrated that different llama hinge types are the products of separate constant region genes. The exon/intron arrangement of a llama classical *IgG* gene has been determined and, in addition, *C<sub>H1</sub>*-like sequence is shown to be present upstream of the hinge exon in two llama IgG isotypes representing long and short hinge heavy chain antibodies. The acceptor splice sites adjacent to both hinge and *C<sub>H1</sub>* exons are found to adhere to established consensi whereas the donor splice site flanking the *C<sub>H1</sub>* exon is mutated in the heavy chain but not classical *C<sub>H1</sub>* genes. These results predict that splicing of the *C<sub>H1</sub>* domain sequence to the hinge exon is prohibited as a result of this mutation, leading to the production of antibodies lacking the  $C_{H1}$  domain.

#### 6.2 Introduction

cDNA sequences derived from llama IgG constant regions have been isolated previously (193, 243). Although these sequences demonstrate the structural similarities between llama Fc regions and those of other species they do not provide an explanation for the absence of the  $C_{H1}$ . It has been speculated that this domain is either absent from the germline, or spliced out during RNA processing (175). Through the screening of the genomic library using cDNA sequence corresponding to a typically conserved region of the  $C_{H2}$  domain an insight into the mechanism of  $C_{H1}$  domain exclusion has been gained. The isolation of llama constant region genes also

allows the confirmation of an isotype-based gene structure similar to the constant genes of other species. In addition comparison of germline sequences encoding heavy chain antibody and classical heavy chains constant domains is possible.

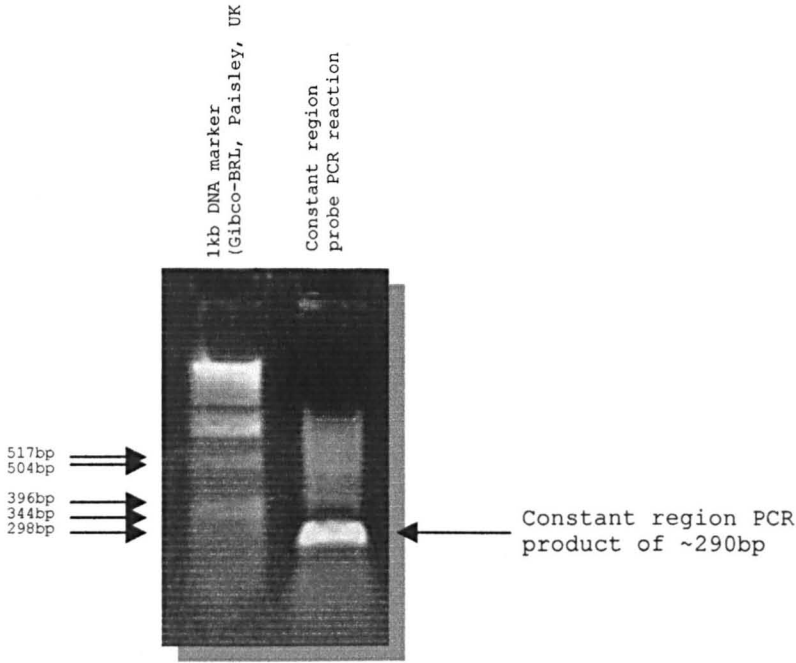
### **6.3 Screening Strategy**

Constant region clones were isolated through the screening of a genomic library (section 2.4) with a specific constant region probe designed from conserved sequences within both heavy chain and classical cDNA sequences. The generation of this probe is described in section 2.4.4 and the PCR product amplified during probe generation is shown in figure 6.1.

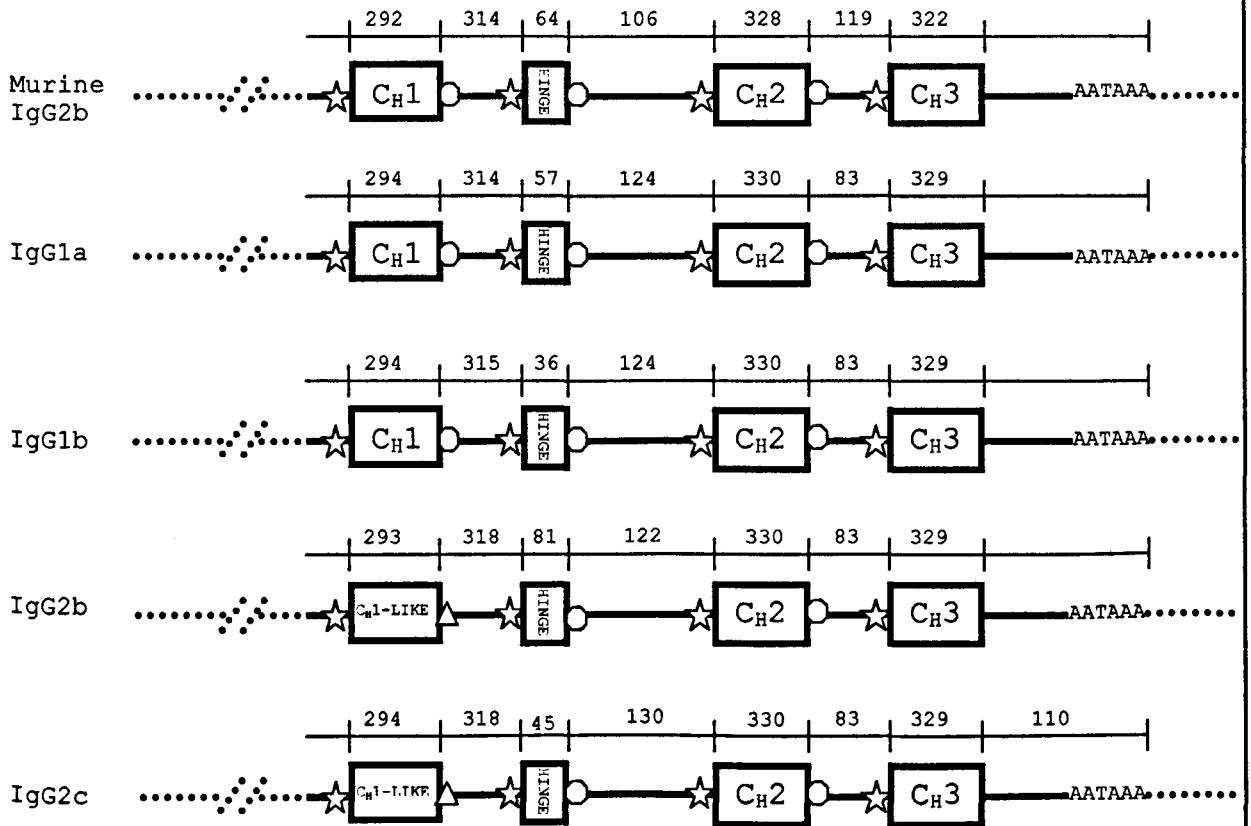
### **6.4 Sequence and General Organisation of Llama Gamma Immunoglobulin Constant Region Genes.**

The four clones isolated during library screening with the constant region probe were sequenced and the main features of each gene was determined by comparison to other mammalian IgG genes and llama cDNA sequences, and by splice site prediction (figures 6.2. and 6.3). Llama hinge sequences associated with both heavy chain and classical llama antibodies have been published previously (175), and C<sub>H</sub>1, C<sub>H</sub>2 and C<sub>H</sub>3 sequence data was also available from two separate sources of llama cDNA sequence (175, 243). This sequence data was compared with the germline sequence of the isolated clones to determine the exon/intron arrangement of each gene locus (figure 6.2). A full alignment of the nucleotide sequences of the four isolated heavy chain genes is given below in figure 6.3

**Figure 6.1 PCR to generate a constant region probe specific to the C<sub>2</sub> region of the llama constant gene.** The PCR product of the desired size was cut from the gel and radiolabelled for use as a hybridisation probe.



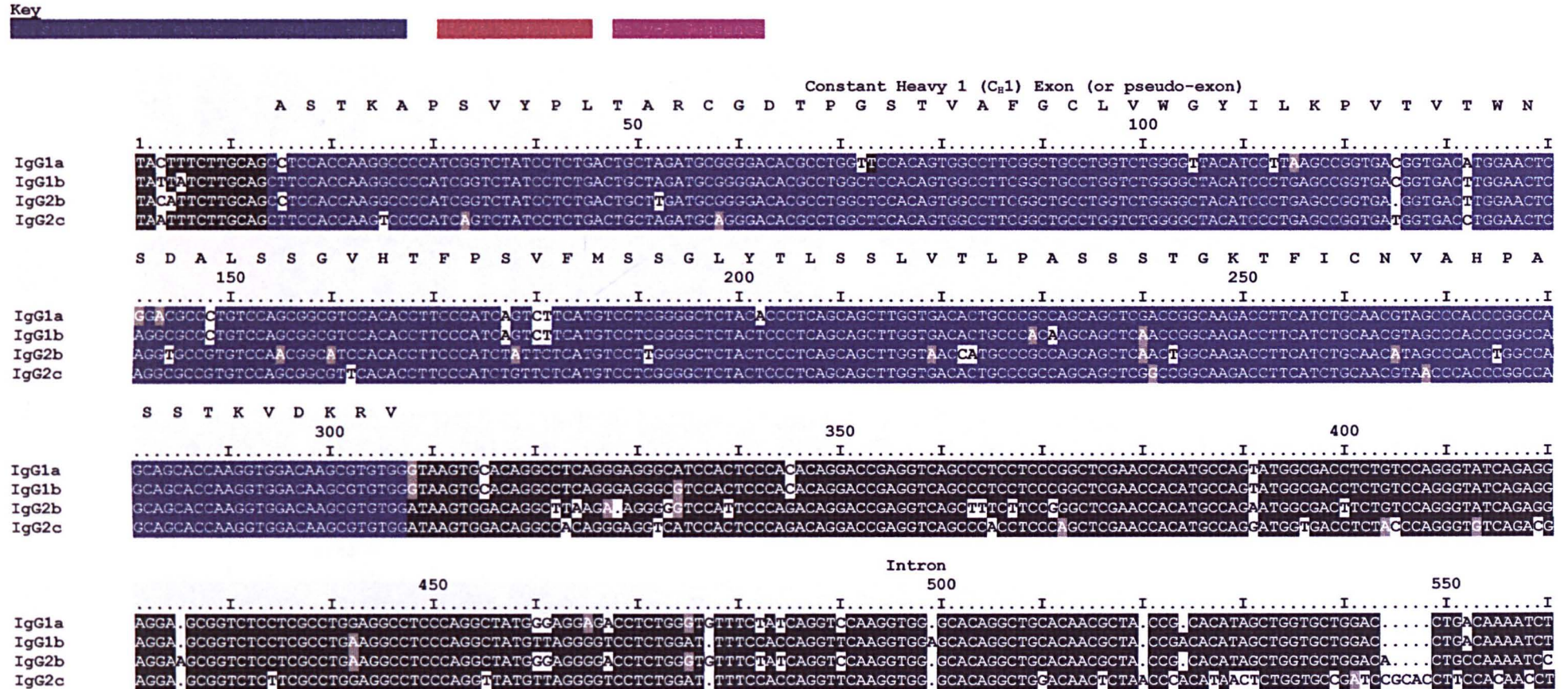
**Figure 6.2 Basic layout of Llama Constant Genes including splice sites and exon/intron lengths.** The layout of the murine IgG2b constant locus is also shown for comparison



**Key**

- Donor splice site consensus GT
- △ Mutated donor splice site consensus AT
- ☆ Acceptor splice site AG

**Figure 6.3 Sequence of Four Llama constant region clones aligned using ClustalW method.** Isotypes IgG1a and IgG1b (Genbank accession nos AF132603 and AF305955) are classical while IgG2b and IgG2c (Genbank accession nos AF132604 and AF132605) are heavy chain only by virtue of a G/A substitution at nucleotide 308 (by this alignment). The IgG1a inferred amino acid sequence is also shown.



Hinge

L K T P Q P Q S Q P E C R

600 650 700

IgG1a  
 IgG1b  
 IgG2b  
 IgG2c

Intron

C P K C P 750 800

IgG1a  
 IgG1b  
 IgG2b  
 IgG2c

Constant heavy 2 (C<sub>H</sub>2) Exon

P E L L G G P S V F I F P P K P K D V L S I S G R P E V T C V V V D V G Q E D P E V S F

850 900 950

IgG1a  
 IgG1b  
 IgG2b  
 IgG2c

N W Y I D G A E V R T A N T R P K E E Q F N S T Y R V V S V L P I Q H Q D W L T G K E F K C K

1000 1050 1100

IgG1a  
 IgG1b  
 IgG2b  
 IgG2c



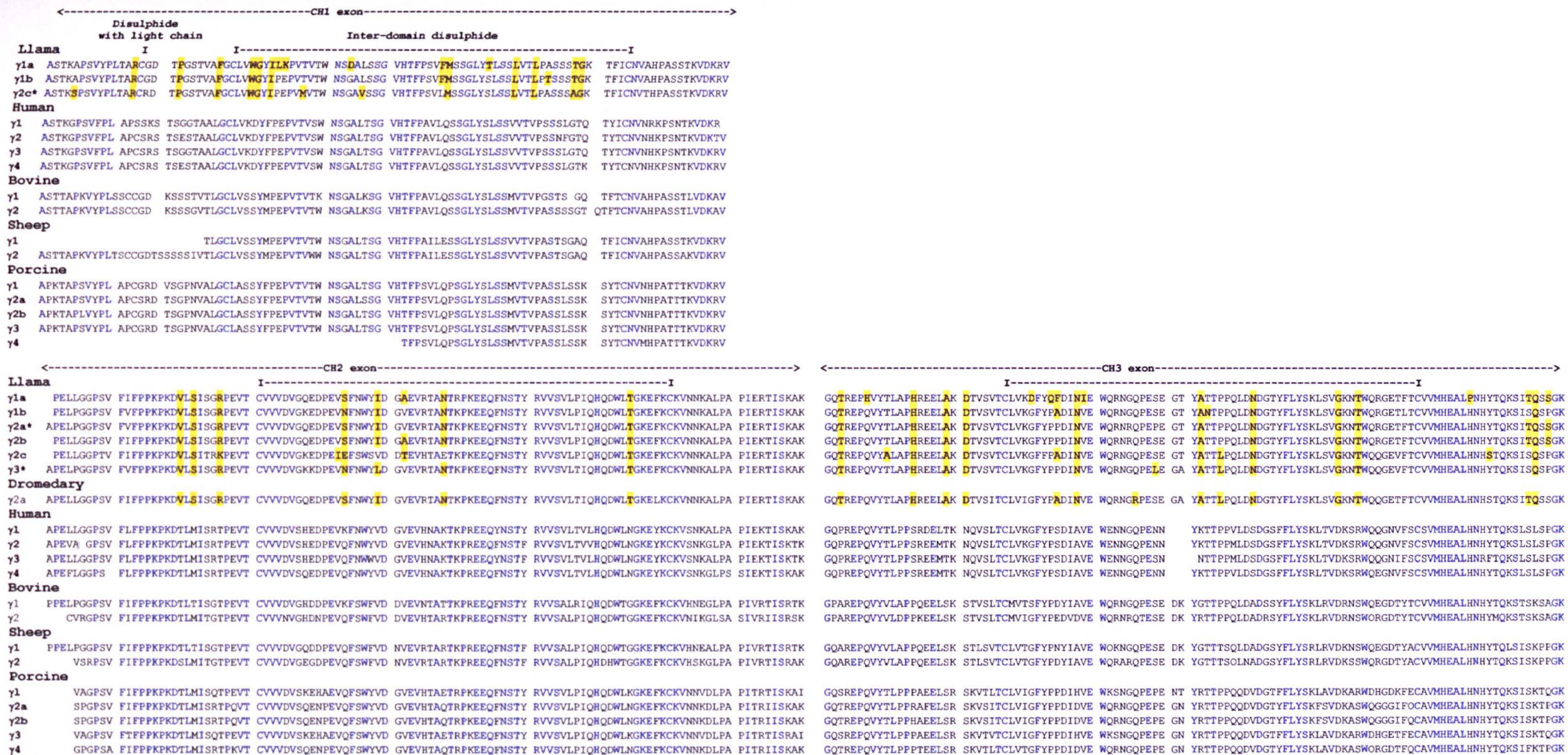
Comparisons indicate that two clones isolated in this chapter encode the constant region sequence of llama immunoglobulin isotypes IgG1a and IgG1b respectively while the remaining clones encode isotypes IgG2b and IgG2c (175, 193). IgG2c is an unpublished isotype that differs from other llama isotypes through its ability to bind protein A, but not protein G (193). In all cases the exon layout and length is comparable to that found in other constant region genes sequenced (for example, those found in human and mouse), with  $C_{H1}$  (or pseudo- $C_{H1}$ ), hinge,  $C_{H2}$  and  $C_{H3}$  domains being encoded by separate exons. While a  $C_{H1}$  pseudo-exon with full coding potential is present within the llama IgG2c germline sequence, a stop codon is located within the corresponding region of the IgG2b gene (due to an A/T substitution at position 51). Given that neither isotype includes an expressed  $C_{H1}$  domain this mutation is probably the result of the recent loss of selective pressure due to  $C_{H1}$  exon exclusion. A comparison of the amino acid sequence derived from each constant region gene is given in figure 6.4.





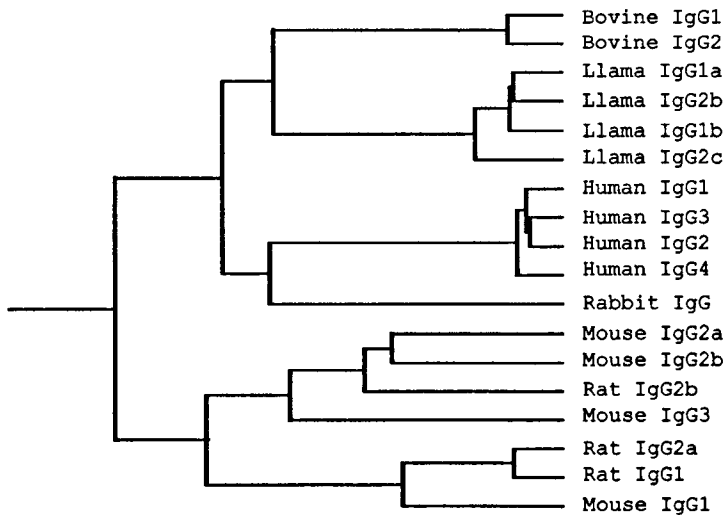
## 6.5 Comparison of Llama Constant Genes with other Mammalian IgG Genes

By comparing the sequence of the constant region genes isolated in this chapter with the constant region sequences of other species it is possible to identify sequence differences that may result in different effector functions within the llama immunoglobulin Fc region. It is also possible to identify Fc regions from other species that share greatest homology to sequences presented here. These may represent close evolutionary relatives of the llama genes. The coding (exon) nucleotide sequences of the germline llama constant genes reported here were compared with the sequence of other functional mammalian genomic constant genes. Genes examined include human IgG1-4, mouse IgG1, 2a, 2b and 3, rat IgG1, 2a and 2b, rabbit IgG and bovine IgG1 and 2 (Accession Numbers Z17370, J00230, D78345, K01316, J00453, J00470, J00461, J00451, M28670, M28669, M28671, L29172, X16701 and X16702 respectively) (figure 6.5). In the case of the IgG2b and IgG2c clones the *C<sub>H1</sub>*-like sequences were not compared as these are assumed to be non-coding by virtue of a downstream splice donor mutation (section 6.6). This relationship between all these constant region genes is described in the phylogenetic tree given in figure 6.6. This analysis provides some clues as to the evolution of the heavy chain antibody isotypes. The llama constant genes show greatest similarity to bovine and human genes. Interestingly, the two heavy chain isotypes are less similar to one another than they are to the llama classical constant isotypes.



**Figure 6.5 Comparisons of llama immunoglobulin constant genes with those of other species.** Species compared are human, cow, sheep, pig and dromedary. Accession numbers are given in the text. Constant region domains are labelled as are conserved cysteines required for disulphide bond formation. Residues highlighted in yellow are unique to llama sequences at this position. Residues in blue are conserved in all species shown. It is clear from this alignment that discrete regions within each domain are strongly conserved. The alignment is intended only to give an indication of regions of conservation along the protein. Actual sequences are given for illustrative purposes only.

\* Llama IgG2c C<sub>H1</sub> sequence is not expressed. Llama γ2a and γ3 C<sub>H2</sub> and C<sub>H3</sub> sequences are from (193)



**Figure 6.6** Rooted phylogenetic tree comparing the evolutionary relationship between the nucleotide sequence of the constant region domains of a number of species and isotypes (excluding the hinge exon, C<sub>H</sub>1/hinge intron and hinge/C<sub>H</sub>2 intron). Genbank accession numbers from top X16701, X16702, AF13603, AF305955, AF132604, AF132605, Z17370, D78345, J00230, K01316, L29172, J00470, J00461, M28671, X00915, M28669, M28670 and J00453)

## 6.6 Characteristics of the C<sub>H</sub>1 exons

The C<sub>H</sub>1 domain of classical immunoglobulin gamma lies adjacent to the variable domain within the heavy chain polypeptide and interacts directly with the constant domain of the light chain (C<sub>L</sub>). The C<sub>H</sub>1 domain exons of each of the llama immunoglobulin gamma isotypes isolated in this chapter are located at positions typical of mammalian constant region genes, roughly 320bp upstream of the hinge exon. The two coding C<sub>H</sub>1 exons of the IgG1 isotypes each encode 98 amino acids, contain only five residue differences (figure 6.4) and are 96% identical to one other at nucleotide level. The five differences include substitutions at positions 164, 186 and 194a (Kabat numbering) which have also been reported within llama IgG1a and IgG1b cDNA sequences. However, the germline *L. glama* sequences contain additional differences at positions 154 and 155 not reported from cDNA sequences (193). This apparent disparity may be the result of errors introduced during generation of cDNA sequences by these workers. In both sequences the crucial cysteine residues at positions 125 and 208 are conserved, enabling the characteristic and structurally crucial formation of an internal disulphide bond. Comparison of the two expressed C<sub>H</sub>1 exons with the pseudo-C<sub>H</sub>1 exons indicates considerable identity and therefore suggests that loss of this domain was a relatively recent evolutionary step. Greatest similarity lies between the two pseudo-exons and the IgG1b C<sub>H</sub>1 exon, where nucleotide identity is 95-96%, perhaps suggesting a closer evolutionary relationship between heavy chain isotypes and IgG1b. Given the shorter hinge of the IgG1b isotype this closer evolutionary relationship suggests that the initial heavy chain antibody had only a short hinge (as remains the case with the IgG2c isotype) before evolving by further duplication and mutation into a longer hinge such as that found associated with the hinge of IgG2b.

### 6.6.1 Splice Sites

The 3' flanking region of the C<sub>H</sub>1 exon of all human and murine *IgG* constant genes is characterised by the presence of a donor splice site immediately adjacent to the coding sequence. Each donor splice site contains a dinucleotide 'GT' consensus immediately 3' of each constant gene exon. This splice site consensus is found associated with each exon of the classical llama constant genes (IgG1a and IgG1b) reported here. However, this consensus is absent from the 3' end of the C<sub>H</sub>1 exons of both llama

heavy chain constant genes (IgG2b and IgG2c) in which the 'GT' dinucleotide is replaced with AT (nucleotides 308-9 in figure 6.3). Mutation of this dinucleotide at other mammalian exon/intron boundaries has been shown to inactivate 3' cleavage and exon joining during *in vitro* splicing (244). It is probable, therefore, that splicing of the  $C_{H1}$  exon to the hinge exon is inhibited *in vivo* (section 6.11).

## 6.7 Hinge Sequences

The IgG hinge typically acts as a flexible spacer between Fc and Fab antibody regions and provides cysteine residues that generate inter-heavy chain disulphide bonds. Although the hinge is not directly responsible for effector functions such as complement activation and Fc receptor interaction (245-247), differences in hinge length influence the segmental motion of antigen binding (248) and this in turn may affect the ability of the antibody to interact with multivalent antigens (249). The hinge exons of the four germline clones vary considerably, both in length and composition. The hinges also share little similarity with those found in other species. The IgG1a isotype contains a hinge of 19 amino acids while that of the IgG1b isotype is only 12 residues in length. 35 amino acids make up the hinge of the IgG2b isotype, and 12 residues that of IgG2c. The extended hinge of IgG2b heavy chain antibodies is thought to compensate for the absence of the  $C_{H1}$  domain within the antibody structure, increasing the distance between the variable domain and the Fc portion of the antibody so that antigen interaction and effector functions can take place unimpeded. The repeated Pro-Xaa motif (Xaa being either Lys, Gln or Glu) found within the hinge of the IgG2b isotype is also found in the Ton B protein of *E. coli* where it acts as a rigid rod-like spacer (3, 250). If the llama IgG2b hinge assumes a similar structure *in vivo* it may act as a spatial replacement for the  $C_{H1}$  domain, effectively holding the  $V_{HH}$  away from the Fc portion of the antibody, providing space for antigen interaction. The hinge exons of both classical isotypes contain three cysteine residues, while heavy chain hinges both contain two cysteines. In both cases these residues are likely to be involved in heavy chain dimer formation.

## 6.8 Fc Fragments

The Fc portion of immunoglobulin consists of  $C_{H2}$  and  $C_{H3}$  domains, each encoded by separate exons within the germline. The  $C_{H2}$  and  $C_{H3}$  domain exons reported here,

like those of the  $C_H1$  domain show a high level of similarity between isotypes (the four  $C_H2$  exons show 88-98% identity, while the  $C_H3$  regions are 93-96% identical at nucleotide level). Within the  $C_H2$  exon the highest level of amino acid identity is between the IgG1a and IgG2b isotypes where there is only one residue difference (Arg/Lys at position 364). The most dissimilar  $C_H2$  domain belongs to IgG2c where 10 of 18 residues between position 286 and 307 within the core  $C_H2$  exon deviate from the llama consensus. A full comparison of the inferred amino acid sequences with those of other species is given in figure 6.5

Binding of antibody Fc regions to Fc receptors allows interaction between antibodies and the cells of the immune system such as macrophages, neutrophils and B-cells. Binding of antibodies to Fc receptors may lead to a diverse range of effects (251). Three types of Fc receptor interact with classical immunoglobulin gamma, Fc $\gamma$ RI, Fc $\gamma$ RII and Fc $\gamma$ RIII. The binding site for Fc receptors within the Fc portion of human IgG typically involves Leu248, Ser252 and surrounding residues (252). Llama IgG1a and IgG2b contain both these residues, while the remaining isotypes contain one of the two key residues (IgG2c contains Leu248 and IgG1b, Ser252). Other residues possibly involved in this interaction are Gly335 (within  $C_H2$ ) and Lys383 (within  $C_H3$ ), both of which are conserved in all characterised llama isotypes. The motif Leu247-Leu-Gly-Gly-Pro251 is crucial for human and murine Fc $\gamma$ RI receptor interaction (252) and therefore likely to play a similar role in other mammalian species such as the llama. This motif is conserved in all the reported llama germline constant sequences other than that encoding the IgG1b isotype. In the case of the IgG1b isotype the significance of the Leu235Pro substitution within this motif is unknown.

The Fc portion of the antibody also plays a role in the initiation of the classical complement pathway through the binding of the Fc region to the C1q component of complement (253). Complement activation ultimately leads to formation of a membrane attack complex capable of lysing invading pathogens. C1q binding to llama IgG Fc regions may be possible through the conservation of Glu337, Lys339, Lys341 and Pro350 residues known to be important to human C1q binding (254, 255). Complement activation and Fc $\gamma$ RI interactions are also dependent to a lesser degree

on the C-terminal end of the C<sub>H</sub>2 which is again highly conserved between the llama isotypes reported in this chapter and other species.

Binding of llama immunoglobulin gamma to *Staphylococcus*-derived protein A and G, was a crucial step in the original isolation of heavy chain antibodies (1). However the IgG2c isotype described in this chapter does not bind protein G (193). Key residues in human and murine IgG binding to protein A and G include 265-267 and 327-330 within the C<sub>H</sub>2 and 464-467 in the C<sub>H</sub>3 domain (252). Of these Met265, His464, His466 and His467 are thought particularly crucial to protein G binding. However, the ability of three of the isotypes reported here to bind protein G despite Ser265 and Tyr467 suggests that these residues are not always crucial. The inability of IgG2c to bind protein G is therefore presumed to be the result of Thr267 and Ser467. The position of each of these key residues and motifs is shown in figure 6.4

## 6.9 Evolution of Constant Domains

Immunoglobulin constant domains have typically been shown to demonstrate less overall sequence conservation than their variable counterparts (159). Although the elements that constitute the basic IgG fold (a domain consisting of seven  $\beta$ -strands connected by bends and helices, stabilised by disulphide bonds) are preserved the lengths of the loops connecting  $\beta$ -pleated segments generally vary more than the corresponding loops within variable domains (159). This difference in level of conservation is best understood if placed in the context of domain function. The variable domains must form an antigen-binding interface, and the high level of conservation within framework regions is the result of the constraints of building a cleft or surface suited to such interactions. Variability is therefore largely confined to the CDRs. In the case of constant domains the wide range of effector functions for which they may be responsible leads to a more diverse range of sequences.

Between the various constant domains characterised in different species it seems that evolutionary change within the first (in this case C<sub>H</sub>1) and last (C<sub>H</sub>3) domains generally occurs at the most linear rate while greater and non-constant rates of change are observed for the more internal (C<sub>H</sub>2) domains. It is argued that the high level of C<sub>H</sub>1 conservation is a result of interaction with BiP and the light chain, although it is



unclear why such conservation is usually also found in the last ( $C_{H3}$ ) domain (159). The llama constant domains reported here follow this general pattern of conservation although the  $C_{H3}$  domains are generally less conserved, and less similar to those of other mammalian species than those of the  $C_{H2}$  domain.

The alignments and phylogenetic trees given in figures 6.5 and 6.6 give an idea of the level of conservation between the various domains and also indicate residues unique to camelid antibodies at various positions.

### **6.10 Constant Region Gene Isolation – Summary of Findings**

- Different llama heavy chain and classical IgG are encoded by separate constant region genes or isotypes
- The absence of the  $C_{H1}$  domain from the heavy chain antibody is presumed to be most likely the result of a single nucleotide splice donor mutation
- Llama IgG is likely to interact with both Fc receptors and complement components

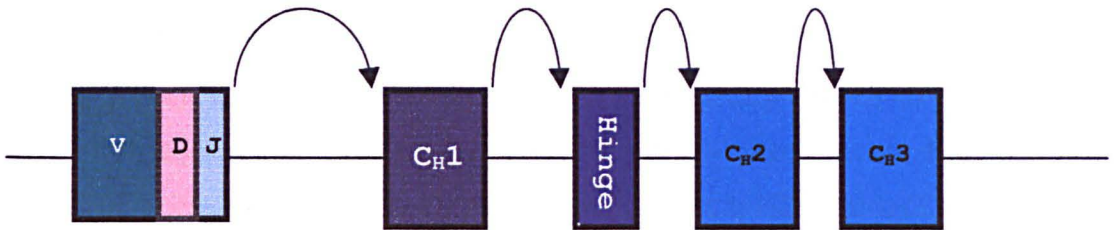
## 6.11 Discussion

The data presented within this chapter shows that separate, discrete constant region gene loci encode the various expressed llama antibody isotypes. Further, the germline sequence data reported here strongly suggests that splicing of the  $C_{H1}$  exon to the hinge exon is inhibited during IgG2b and IgG2c RNA processing. This relates directly to section 1.4 point (3) and shows clearly that the smaller size of the heavy chain antibody is the result of a post-transcriptional deletion of the  $C_{H1}$  domain sequence. During splicing of heavy chain antibody RNA transcripts it is likely that donor splice sites adjacent to the rearranged  $V(D)J$  subunits show preference for acceptor splice sites flanking the hinge exon above those flanking the non-spliced  $C_{H1}$ -like exon (figure 6.7). This would explain the absence of the  $C_{H1}$  domain in the expressed protein. It is conceivable that the choice of acceptor splice site chosen by the  $V(D)J$ -associated donor site may represent a level of control of heavy chain antibody generation (section 1.4 point (6)) in a manner similar to alternative splicing. If the acceptor site adjacent to the  $C_{H1}$ -like exon of the IgG2b or IgG2c is favoured above the hinge acceptor then the level of functional heavy chain antibodies produced may be reduced.

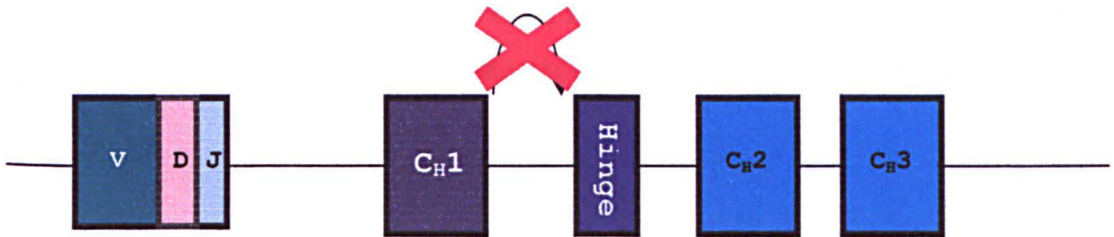
Splice site modification and the use of alternative splice sites is common within immunoglobulin genes. The primary immunoglobulin isotype expressed by developing B-cells, IgM, has both secretory and membrane forms encoded by separate exons, the expression of which is determined by alternative use of splice sites (256). Alternative splicing is presumed to be responsible for the generation of membrane and secretory forms of other immunoglobulin isotypes. Another example of the role of splice site recognition came to light during the discovery of a hinge deletion variant of porcine IgA (163). Rather than the splice donor mutation discovered in this thesis, a splice acceptor mutation (AA rather than AG) was found to result in a two amino acid hinge variant. This porcine study provides a precedent for the exclusion of an immunoglobulin constant gene exon due to a splice site mutation reported in this chapter.

**Figure 6.7 Proposed mechanism of splice site knockout in germline IgG2b and IgG2c genes.** The A/G substitution prevents splicing between C<sub>H</sub>1 and hinge exons. Splicing between recombined V(D)J and C<sub>H</sub>1 is then prevented by the presence of the hinge acceptor site. This acceptor site is preferred by the V(D)J donor site which consequently interacts with it.

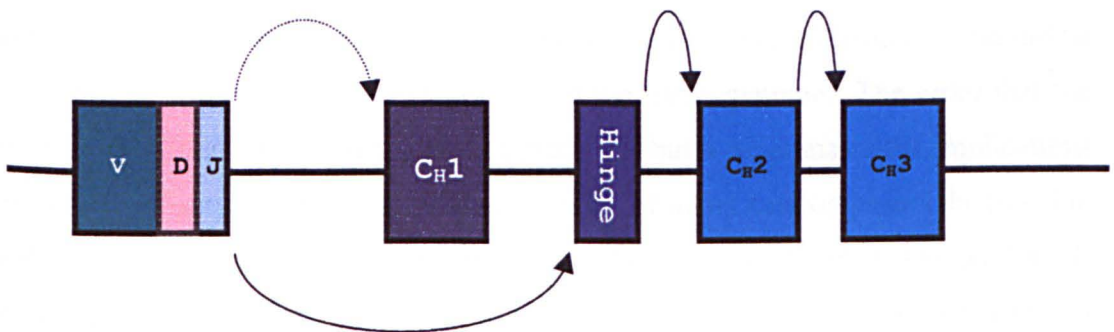
a) Conventional Splicing (IgG1a and IgG1b). Four introns are spliced out.



b) Conventional splicing is blocked by mutated splice donor.



c) Heavy Chain Only Splicing (IgG2b and IgG2c). The splice donor site 3' of the recombined V(D)J unit must have a stronger ability to associate with the acceptor site 5' of the hinge than that 5' of the pseudo-C<sub>H</sub>1 exon.



By comparison to known sequence motifs it has been possible to demonstrate the presence of a number of protein motifs within the constant domains isolated in this chapter (section 6.8). The presence of these motifs suggests that the llama Fc regions are competent binders of Fc $\gamma$  receptors and complement C1q. Residues believed responsible for the inability of llama IgG2c to bind protein G are also described (section 6.8).

Comparisons between the four llama constant region genes isolated in this thesis and those of other species (section 6.9) have demonstrated a high level of sequence conservation between isotypes and allowed speculation as to the evolutionary development of the heavy chain antibody isotypes. The data presented in this chapter suggests that an initial splice mutation (G to A) led to heavy chain antibody generation from a precursor isotype probably most closely related to present day IgG1a. Subsequent duplication events are then likely to have occurred in order to generate the three heavy chain antibody isotypes (IgG2a, IgG2b and IgG2c) observed today. The presence of longer hinge exons within the llama IgG2b and dromedary IgG2a C<sub>H1</sub> pseudo-exons suggests that these isotypes may represent evolutionary improvements on a precursor heavy chain isotype that may have resembled the IgG2c isotype.

## **6.12 Future Work**

Ultimately it would be useful to isolate the constant region gene encoding the IgG2a isotype. Workers in Belgium (257) have isolated the germline constant gene encoding the dromedary IgG2a isotype although no other germline camel isotypes have yet been reported. As is the case with the variable genes reported in chapter 3 it would be useful to link the various IgG isotypes within the llama genome. The order that the genes occur within a llama heavy chain immunoglobulin locus may have implications for the role of isotype switching on the generation of llama antibodies and the possible problems encountered during llama B-cell development described in section 7.6. It would also be interesting to characterise the germline gene structure of other camelid antibody isotypes such as IgM for comparison to those described in this thesis. The full understanding of the mechanisms involved in llama antibody generation must await a fuller understanding of the process of isotype switching in general. The

specificity of this system is poorly understood and components of the switching machinery have only recently begun to be reported (258-260). It is, however, clear that the production of heavy chain antibodies, and indeed the type of heavy chain antibodies, depends on this process.

Although the mutated splice acceptor described in this chapter should result in the absence of the C<sub>H1</sub> domain from the final heavy chain antibody gene product (244) it would be interesting to investigate further the process of splicing within heavy chain and classical llama IgG transcripts. Such experiments would allow investigation of the possible splice acceptor choice made by the V(D)J-associated donor splice sites (section 6.6.1 and 6.11). The splicing events leading to the generation of fully processed IgG mRNAs can be examined through the process of *in vitro* splicing. In such a system plasmid constructs containing exons and introns derived from the different constant genes would be constructed to transcribe synthetic radiolabelled RNAs representative of those generated *in vivo*. Such synthetic RNAs could then be incubated with mammalian nuclear extracts containing all proteins necessary for splicing. The results of *in vitro* splicing can then be visualised by electrophoresis and autoradiography.

The results described in this chapter have been published in a peer reviewed journal (Woolven, B.P., L.G. Frenken, P. van der Logt and P.J. Nicholls. 1999. The structure of the llama heavy chain constant genes reveals a mechanism for heavy-chain antibody formation. *Immunogenetics* 50:98.).

## Chapter 7 General Discussion.

### 7.1 Overview

In the preceding chapters data is presented pertaining to the mechanisms of antibody generation in the llama. In this chapter the significance of these results is discussed in the more global context of the developing llama immune system as a whole. A number of models are described that may be relevant to llama antibody generation *in vivo*.

### 7.2 A Diverse Range of Heavy Chain Variable Gene Segments Provide the Llama with Unknown Evolutionary Benefits

The discovery of five heavy chain gene segments in the llama (chapter 3) and larger numbers in the dromedary (190) suggest that considerable evolutionary pressure has led to multiple duplications of a primordial variable gene segment. These duplications have led to the development of a subset of variable gene segments that are able to generate a diverse range of heavy chain antibodies. It is probable that the initial evolutionary event that led to the generation of heavy chain antibodies was the single splice site mutation described in chapter 6. The original heavy chain antibodies would, therefore, have lacked the C<sub>H</sub>1 domain but must, at first, have utilised the existing variable gene segment repertoire. At least one of the classical variable gene segments present in the genome at the time of this mutation must have been sufficiently 'V<sub>HH</sub>-like' to encode a variable domain which could not only bind antigen independent of a V<sub>L</sub> domain but also confer some selective advantage on the animal in which the mutation had occurred. Are such heavy chain gene segment precursors present within the genomes of other species?

The full complement of variable gene segments has been reported and genetically linked only in the human (31). Within the human repertoire only one expressed gene segment, V<sub>H</sub>3-49 (Genbank accession no AB019438), contains any of the key residues characteristic of heavy chain antibody (Phe37). It is unlikely that this residue alone would allow a similar gene segment in the llama to generate heavy chain antibodies with sufficient heavy chain immunoglobulin characteristics to be of benefit to the animal. In a further attempt to identify sequence representative of a possible heavy chain variable gene segment precursor the complete human heavy chain locus

has been analysed (data not shown) with 'tBlastx' analysis software in all six reading frames in an attempt to find similarity to the variable gene segments reported in chapter 3. No significant matches have been found.

### **7.3 Llama Immunoglobulin Heavy Chain Variable Gene Segments are Encoded in the Germline and May Be Components of a Single Llama Immunoglobulin Locus.**

The isolation of specific variable gene segments encoding both heavy chain and classical immunoglobulin variable domains provides an answer to perhaps the most rudimentary of questions that are posed by the discovery of unique heavy chain antibody types in the llama. The work described in chapter 3 is the first to show that llama heavy chain antibodies are the result of the recombination of specific, germline, variable gene segments. The unique polypeptides encoded by these gene segments contain a number of amino acids that make interaction of the heavy chain variable domain with the light chain improbable. Given that five of the eight variable gene segments isolated in this study are of the heavy chain type, it is likely that such gene segments make up a considerable proportion of the llama variable gene segment repertoire. These findings are consistent with those of other groups (3) which have also demonstrated the presence of specific heavy chain antibody-encoding gene segments within the dromedary (206). Through the isolation of larger numbers of variable gene segments this group (190) predict a ratio of approximately 1:1.2  $V_{HH}$  to  $V_H$  encoding gene segments within the dromedary genome.

It is tempting to assume that the classical and heavy chain gene segments isolated in this thesis form part of a single llama immunoglobulin heavy chain locus. A single locus is responsible for the generation of the entire heavy chain polypeptide repertoire in all studied mammalian species although the precise mechanisms used to generate diversity differ to some degree (section 1.13). Despite this non-functional variable gene segments have been isolated outside the conventional human heavy chain locus (261, 262) and so the possibility that the gene segments reported here may be distributed throughout the llama genome cannot be ruled out. If this were the case it may allow differential control of gene segments at different chromosomal locations. The generation and mapping of clones from a larger cosmid, or yeast artificial

chromosome (YAC) genomic library (as suggested in section 3.22) would enable confirmation of the relative locations of different gene segments. This strategy has proved successful in the full characterisation and linking of the equivalent human locus (263).

If functional variable gene segments are not confined to a single locus, both D-J and constant region genes would also need to be present within each locus in order to generate full length antibodies. It is possible that heavy chain variable gene segments and constant genes that exclude the C<sub>H1</sub> domain (section 6.4) may have developed together and in isolation from other immunoglobulin genes. This would have required an initial large-scale duplication of a major portion of the original heavy chain immunoglobulin locus. Given the close evolutionary relationship between the llama and other mammals established both by conventional means (169) and through the analysis of the *rag* genes reported in this thesis (section 4.7) such a major genetic event seems unlikely.

#### **7.4 Controlling Heavy Chain Immunoglobulin Generation**

The levels of various immunoglobulin classes within llama sera fluctuate considerably. Heavy chain antibodies typically make up anywhere between 5-20% of the total serum IgG. Serum levels of heavy chain immunoglobulins are considerably higher in the closely related dromedary where up to 70% of IgG is of the heavy chain type. How are these fluctuations and differences in serum antibody levels brought about?

The level of heavy chain antibody within the llama serum depends on: a) the quantity of heavy chain antibody produced by each B-cell, b) the number of B-cells currently generating heavy chain antibody and c) the rate of clearance and degradation of the antibody. Any model that predicts the ways in which serum levels of heavy chain immunoglobulins are controlled must therefore address two questions. Firstly given that a B-cell is committed to production of a heavy chain antibody, can the levels of antibody produced by that individual B-cell production be raised or lowered? Secondly can the antibody type (i.e. heavy chain or classical) produced by the B-cell be influenced in any way?



To produce more or less antibody a B-cell may regulate the transcription of the rearranged immunoglobulin gene. Such regulation may be stimulated through processes such as exposure to antigen and/or T-cell help. If more copies of the rearranged gene are transcribed, all other things being equal, more antibody will be produced. The different transcription factor binding motifs described in section 3.9 may be central to this level of antibody generation control. If the immune system requires more antibodies of a particular family, it may stimulate the B-cell population to upregulate expression of, or activate, a particular transcription factor. This factor can then bind the promoter region of the rearranged gene (if the gene promoter has the required motif), thereby increasing the level of transcription.

Other, more poorly characterised levels of control may also regulate antibody production by individual B-cells. For example, mRNA degradation rates will affect the number of transcribed copies of the immunoglobulin that are translated while changes in chromatin accessibility may either promote or inhibit transcription.

To understand how the B-cell may control the type of antibody it produces it is important to remember that heavy chain antibodies are unique, not only in their structural characteristics, but also in the nature of the genetic elements encoding them. Classical antibodies utilise a broad range of variable gene segments that can then be expressed as part of any number of immunoglobulin classes or isotypes. The situation during heavy chain generation is more complex. Heavy chain antibodies can also be expressed as a number of immunoglobulin isotypes (such as those described in chapter 6). Heavy chain antibodies cannot, however, utilise a complete range of variable gene segments. Rearrangement using a specific subset of variable gene segments (such as  $V_{HH1}$ - $V_{HH5}$  isolated in Chapter 3) is required to generate a functional heavy chain antibody. If the type of antibody produced by a llama B-cell is to be regulated the choice of variable gene segment as well as the choice of isotype must be taken into consideration.

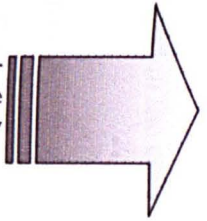
The choice of isotype generated by the B-cell is determined by the mechanism of isotype switching (264). Specific signals (such as exposure to an interleukin) have been shown to stimulate the switching of B-cell antibody production from one isotype

to another. Isotype switching may therefore represent an important level of control in the production of both classical and heavy chain antibodies. For example, a particular stimulus may result in a switch from the production of a long hinge heavy chain antibody to a short hinge heavy chain isotype. What is less clear is the result of a switch from a classical to a heavy chain isotype or vice versa, as the variable domain associated with each isotype must remain the same. This is discussed further in section 7.6

Mechanisms by which particular variable gene segments are selected for rearrangement are not well understood. What is certain is that variable gene segment usage in systems such as the human or mouse, is biased at different developmental stages towards particular variable gene segments (265). Were such biases to act on gene segments in the llama, levels of heavy chain antibody production would vary according to whether the preferred gene segment was of the heavy chain or classical type. If such biases do take place in the llama B-cell how might they come about?

One potential source of variable gene segment bias is the process of 'sterile' transcription. Conventional, post-rearrangement transcription is not the only form of transcription that has been detected within the heavy chain immunoglobulin locus ((266) and section 1.7.4). Sterile transcription is thought to 'open up' the local chromatin structure associated with variable gene segments prior to V(D)J recombination. This may allow preferential access of the recombination apparatus (including the RAG proteins) to particular families of immunoglobulin genes. It is assumed that sterile transcription is regulated in the same manner as conventional transcription, through the binding of specific transcription factors to particular promoter elements. If this is the case specific signals received by the B-cell during development (and before V(D)J recombination) may upregulate expression of, or activate, a particular transcription factor. If a motif for this transcription factor is present within the promoter of a particular gene segment, sterile transcription of this gene segment may bias recombination in favour of the use of this gene. At least two families of transcription factor motifs are described in section 3.9. Sterile transcription may control the level at which each of these families takes part in V(D)J recombination. The mechanisms by which llama antibody generation may be controlled are summarised in figure 7.1.

Signals are sent to the B-cell in order to stimulate heavy chain antibody production



### Llama B-cell

Once a signal is received the B-cell may respond in three ways

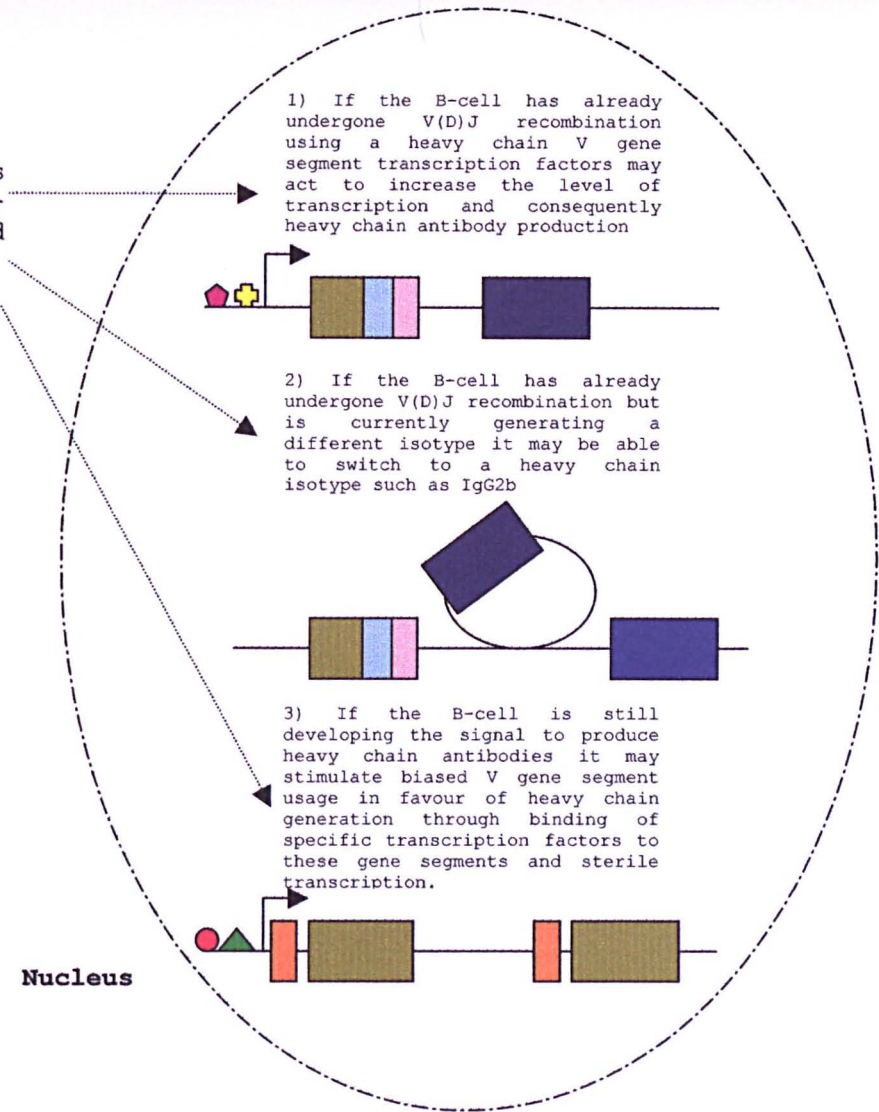


Figure 7.1 A hypothetical model describing the possible ways in which a B-cell may be able to generate heavy chain antibodies after receiving a signal to do so. Any or all of these mechanisms could act within a particular B-cell

## **7.5 Generation of an Extended CDR3 – Antigen Receptor Selection or Antibody Generation?**

The absence of both the  $C_{H1}$  domain and the light chain have major effects on the structure and functional capacities of the llama heavy chain antibody. However it is the unusually extended CDR3 region of the  $V_{HH}$  domain that may have the greatest significance during  $V_{HH}$ /antigen interaction. The protuberance of the resultant H3 loop from the  $V_{HH}$  surface has been demonstrated through crystallographic studies (172, 173) and may allow the antibody to interact with clefts within the antigen. If the extended CDR3 is crucial to the action of the  $V_{HH}$  domain, how might it be generated?

The extended CDR3 found within camelid heavy chain antibody cDNA sequences may result simply from the selection of heavy chain immunoglobulin molecules best able to interact with antigen during affinity maturation ((10) and section 1.4 point (1)). If, on encounter with a foreign molecule, those heavy chain antibodies best able to interact with the antigen are those containing an extended H3 loop, then these B-cells will be stimulated to proliferate. Proliferation of B-cells will therefore lead to a misrepresentative number of B-cells producing extended CDR3 heavy chain antibodies within the serum. The nature of CDR sequences isolated through the generation of cDNAs does not therefore directly correlate with events taking place during V(D)J recombination. The importance of this affinity maturation with respect to the nature of isolated cDNA sequences within the mouse has been discussed by other groups (267). The study of cDNAs derived from bone marrow (where B-cells develop) rather than peripheral blood lymphocytes (as used in the generation of the Unilever database) could provide an interesting comparison to the cDNA sequences used in the course of this thesis.

Even if B-cells generating antibodies with extended H3 loops are favoured during affinity maturation it is still reasonable to speculate as to the origin of the extended CDR3. The action of affinity maturation only reduces the likelihood that the llama recombination system produces a longer CDR3 more frequently than the recombination system of other species. Instead, the unique structure of the heavy chain antibody is such that antibodies with long CDR3s are better able to bind antigen and therefore positively selected. In the mouse or human, antibodies with long

CDR3s would not be able to interact with antigen successfully (because of light chain association) and so would never be positively selected. The question of how such extended CDR3s (even if generated in only small numbers) are produced must still be considered, even if the mechanisms are not unique to the llama.

Mechanisms that may lead to the generation of the extended CDR3 include differences in D-J and V-DJ recombination. These could involve the utilisation of multiple diversity gene segments or DIR segments as described in section 1.8.6. An increased level of TdT activity within the B-cell may lead to more frequent addition of random nucleotides at recombination junctions (section 1.8.5). Additionally, subtle differences in the resolution of coding end hairpin may lead to the inclusion of greater numbers of palindromic sequences at these junctions. RSS-like sequences downstream or upstream of established recombination signals could compete for RAG interaction, leading to recombination that allows the inclusion of sequence normally relegated to the signal joint. Alternatively, such elongated CDR3s may not result from recombination and associated processes. Instead V, D and J gene segments containing more sequence able to contribute to the CDR3 could be present within the germline (although no evidence to support this possibility was found during this project). Also, an insertion/deletion model similar to that proposed by Nguyen and co-workers (190) could act to add nucleotides to the CDR3.

Examining first mechanisms that may apply to the generation of extended CDR3s in all mammalian species, the possible role of an increased level of TdT activity during llama B-cell development cannot be ruled out. The link between TdT expression and variable gene segment transcription has long been known through the identification of a common transcription factor binding site (binding members of the Ikaros transcription factor family (268)) located upstream of TdT and variable gene segment promoters. The same transcription factor binding site is present upstream of all the isolated variable gene segments isolated in chapter 3. The requirement for immunoglobulins containing longer H3 loops could conceivably be met by the simultaneous upregulation of transcription at both TdT (in order to increase levels of N nucleotide addition) and variable gene segments (during sterile transcription). In this model an Ikaros-related transcription factor may promote both nucleotide addition and accessibility of the heavy chain locus to recombination. To complicate matters

further the expression of TdT in the mouse has recently been linked to biased variable gene segment utilisation (269)

Llama-specific B-cell mechanisms that may increase CDR3 length include the differences in coding flank sensitivity of the RAG proteins reported in section 5.14. Nucleotide addition by TdT is dependent on the composition of the region immediately upstream of the RSS heptamer element (66). All llama and dromedary coding flanks characterised thus far differ from those murine coding flanks (chapters 3 and 6, (206) and (270)) utilised in previously reported recombination experiments. It is possible that such variable gene segment coding flanks (typically with nucleotides AGA adjacent to the heptamer) allow greater TdT activity than the conventional coding flanks found in other species. However, there is no evidence that particular coding flanks are favoured by heavy chain gene segments over classical gene segments.

Another possibility is that the llama may possess 'modified' RAG proteins that interact with coding flanks in a novel manner thereby promoting either N or P nucleotide addition (Chapter 5).

## 7.6 A Model for Llama Heavy Chain Antibody Secretion

In Chapters 1 and 6 of this thesis the mechanisms understood to lead to the production and secretion of antibodies within both developing and mature B-cells are discussed. The ability of the camelid heavy chain antibody to leave the endoplasmic reticulum and travel to the B-cell membrane in the absence of an accompanying light chain is thought to result from the absence of a  $C_{H1}$  domain (figure 1.11). An explanation for the absence of the  $C_{H1}$  domain in the translated llama immunoglobulin gene product is given in Chapter 6 where a single nucleotide substitution is described, resulting in the splicing out of the  $C_{H1}$  exon during post-transcriptional processing. What remains unclear, however, is the manner in which variable domains encoded by heavy chain gene segments are trafficked through the B-cell during production of alternative isotypes.

If the presence of a single llama heavy chain immunoglobulin locus, similar in layout to the human locus, is assumed (section 7.1), heavy chain gene segments can presumably be linked to other constant genes such as those encoding IgA and IgM isotypes. All of these isotypes conventionally contain  $C_{H1}$  domains and so will presumably interact with BiP (section 1.10). However, the unique nature of the variable domain encoded by a heavy chain variable gene segment means that interaction with a light chain should not be possible. If light chains are unable to replace the BiP/ $C_{H1}$  domain interaction then these isotypes will be unable to leave the endoplasmic reticulum. In most cases such immunoglobulins will presumably be degraded by the B-cell.

This model does, however predict a potential problem during llama B-cell development. As the human B-cell matures IgM becomes the first immunoglobulin to be expressed after successful rearrangement of the V, D and J gene segments. Every successful V(D)J rearrangement is therefore expressed initially as a membrane-bound IgM antibody. Once the B-cell reaches maturity isotype switching may take place so that the initial IgM form of the antibody is replaced with another class, such as IgG2. If the llama B-cell operates in the same manner, how does secretion of heavy chain antibody ever take place?

There may be three ways in which this problem is avoided. Firstly BiP interaction with the IgM C<sub>H1</sub> domain alone (as is presumed to occur when a V<sub>HH</sub> is present during successful rearrangement), rather than simultaneous interaction with both C<sub>H1</sub> and variable domains may be insufficient to hold the polypeptide within the ER. If this is the case heavy chain IgM polypeptides should be able to reach the B-cell surface. Secondly the llama C<sub>μ</sub> gene (that which encodes the IgM isotype) may contain splice mutations similar to those afflicting the C<sub>γ2b</sub> and C<sub>γ2c</sub> genes and described in Chapter 6. In this case the heavy chain IgM polypeptide would lack the C<sub>H1</sub> domain and therefore escape interaction with BiP. However, in evolutionary terms the near-simultaneous exclusion of two constant gene C<sub>H1</sub> domains seems unlikely. The final possible explanation relies on the presence of a separate heavy chain antibody-encoding locus (section 7.1). Such a locus may lack a C<sub>μ</sub> gene altogether so that the first isotype expressed would be a heavy chain antibody. How this final model could account for the complex cell signalling pathways that are controlled by membrane bound IgM during B-cell development is not clear.

## 7.7 Summary

In this chapter a number of models have been described that may explain a number of facets of llama antibody generation. These include:

- Speculation as to the layout of the llama heavy chain immunoglobulin locus or loci.
- A description of the possible levels at which heavy chain llama antibodies are generated.
- A discussion of the possible mechanisms of generation of the extended CDR3 in heavy chain antibodies.
- Models of heavy chain antibody secretion. Ways in which the heavy chain antibodies may be secreted by the B-cell are discussed.



## 7.8 Final Conclusions

In this thesis a number of questions that arise from the original discovery of the llama heavy chain antibody are addressed (section 1.4). The data presented here describes a llama immune system broadly similar to our own and utilising mechanisms such as V(D)J recombination that are well characterised in the human and mouse. The principal genetic differences that lead to the generation of heavy chain antibodies are described (Chapters 3 and 6) and the first *in vitro* study to investigate non-murine V(D)J recombination finds llama and murine recombination to be highly similar (Chapters 4 and 5).

Although the presence of distinct heavy chain antibody gene segments and a mutated splice site explain the origins of the major structural differences between classical and heavy chain immunoglobulins the finer details of how llama antibody generation is controlled remain unknown. Chapter 7 of this thesis describes the ways in which heavy chain antibodies generation may be regulated within the llama immune system, how heavy chain antibodies might be secreted by the llama B-cell and how the extended CDR3 may be formed. This is achieved principally through analogy to the current understanding of these processes in other species. It is clear that until fundamental immunological processes such as affinity maturation, coding flank processing and isotype switching are better understood, be it in mouse, human or llama questions regarding the generation of llama heavy chain antibodies will remain.

## Chapter 8 References

1. Hamers-Casterman, C., T. Atarhouch, S. Muyldermans, G. Robinson, C. Hamers, E. B. Songa, N. Bendahman, and R. Hamers. 1993. Naturally occurring antibodies devoid of light chains. *Nature* 363:446.
2. Sheriff, S., and K. L. Constantine. 1996. Redefining the minimal antigen-binding fragment [news; comment]. *Nature Structural Biology* 3:733.
3. Muyldermans, S., and M. Lauwereys. 1999. Unique single-domain antigen binding fragments derived from naturally occurring camel heavy-chain antibodies. *J Mol Recognit* 12:131.
4. Casterman, C., and R. Hamers. 1993. Immunoglobulins devoid of light chains US1993000106944, US.
5. Davies, J., and L. Riechmann. 1994. 'Camelising' human antibody fragments: NMR studies on V<sub>H</sub> domains. *Febs Letters* 339:285.
6. Davies, J., and L. Riechmann. 1996. Single antibody domains as small recognition units: design and in vitro antigen selection of camelized, human V<sub>H</sub> domains with improved protein stability. *Protein Engineering* 9:531.
7. Riechmann, L. 1996. Rearrangement of the former VL interface in the solution structure of a camelised, single antibody V<sub>H</sub> domain. *Journal Of Molecular Biology* 259:957.
8. Lauwereys, M., M. Arbabi Ghahroudi, A. Desmyter, J. Kinne, W. Hölzer, E. De Genst, L. Wyns, and S. Muyldermans. 1998. Potent enzyme inhibitors derived from dromedary heavy-chain antibodies. *Embo Journal* 17:3512.
9. Wernery, U., and O. Kaaden. 1995. *Infectious Diseases of Camelids*. Blackwell Science.
10. Tarlinton, D. M., and K. G. C. Smith. 2000. Dissecting affinity maturation: a model explaining selection of antibody-forming cells and memory B cells in the germinal centre. *Immunology Today* 21:436.
11. Marquart, M., and D. J. 1982. The three-dimensional structure of antibodies. *Immunology Today* 3:160.
12. Ward, E. S., D. Gussow, A. D. Griffiths, P. T. Jones, and G. Winter. 1989. Binding activities of a repertoire of single immunoglobulin variable domains secreted from *Escherichia coli* [see comments]. *Nature* 341:544.

13. Wu, T. T., and E. A. Kabat. 1970. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *Journal Of Experimental Medicine* 132:211.
14. Chothia, C., J. Novotný, R. Brucoleri, and M. Karplus. 1985. Domain association in immunoglobulin molecules. The packing of variable domains. *Journal Of Molecular Biology* 186:651.
15. Davies, D. R., E. A. Padlan, and S. Sheriff. 1990. Antibody-antigen complexes. *Annual Review Of Biochemistry* 59:439.
16. Wilson, I. A., R. L. Stanfield, J. M. Rini, J. H. Arevalo, U. Schulze-Gahmen, D. H. Fremont, and E. A. Stura. 1991. Structural aspects of antibodies and antibody-antigen complexes. *Ciba Foundation Symposium* 159:13.
17. Rock, E. P., P. R. Sibbald, M. M. Davis, and Y. H. Chien. 1994. CDR3 length in antigen-specific immune receptors. *Journal Of Experimental Medicine* 179:323.
18. Greenwood, J., M. Clark, and H. Waldmann. 1993. Structural motifs involved in human IgG antibody effector functions. *European Journal Of Immunology* 23:1098.
19. Sondermann, P., R. Huber, V. Oosthuizen, and J. Uwe. 2000. The 3.2-A crystal structure of the human IgG1 Fc fragment-Fcγ<sub>3</sub> complex. *Nature* 406:267.
20. Parslow, T. G., D. L. Blair, W. J. Murphy, and D. K. Granner. 1984. Structure of the 5' ends of immunoglobulin genes: a novel conserved sequence. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 81:2650.
21. Mizushima-Sugano, J., and R. G. Roeder. 1986. Cell-type-specific transcription of an immunoglobulin kappa light chain gene in vitro. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 83:8511.
22. Annweiler, A., M. Müller-Immerglück, and T. Wirth. 1992. Oct2 transactivation from a remote enhancer position requires a B-cell-restricted activity. *Molecular And Cellular Biology* 12:3107.

23. Clerc, R. G., L. M. Corcoran, J. H. LeBowitz, D. Baltimore, and P. A. Sharp. 1988. The B-cell-specific Oct-2 protein contains POU box- and homeo box-type domains. *Genes And Development* 2:1570.
24. Landolfi, N. F., X. M. Yin, J. D. Capra, and P. W. Tucker. 1988. A conserved heptamer upstream of the IgH promoter region octamer can be the site of a coordinate protein-DNA interaction. *Nucleic Acids Research* 16:5503.
25. Kemler, I., E. Schreiber, M. M. Müller, P. Matthias, and W. Schaffner. 1989. Octamer transcription factors bind to two different sequence motifs of the immunoglobulin heavy chain promoter. *Embo Journal* 8:2001.
26. Eaton, S., and K. Calame. 1987. Multiple DNA sequence elements are necessary for the function of an immunoglobulin heavy chain promoter. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 84:7634.
27. Webb, C. F., C. Das, S. Eaton, K. Calame, and P. W. Tucker. 1991. Novel protein-DNA interactions associated with increased immunoglobulin transcription in response to antigen plus interleukin-5. *Molecular And Cellular Biology* 11:5197.
28. Buchanan, K. L., S. I. Hodgetts, J. Byrnes, and C. F. Webb. 1995. Differential transcription efficiency of two Ig V<sub>H</sub> promoters *in vitro*. *Journal Of Immunology* 155:4270.
29. Buchanan, K. L., E. A. Smith, S. Dou, L. M. Corcoran, and C. F. Webb. 1997. Family-specific differences in transcription efficiency of Ig heavy chain promoters. *Journal Of Immunology* 159:1247.
30. Kabat, E. A., T. T. Wu, H. M. Perry, K. S. Gottesman, and C. Foeller. 1991. *Sequences of Proteins of Immunological Interest*. NIH Publication No. 91-3242, US Department of Health and Human Services, PHS, NIH, Bethesda, MD.
31. Matsuda, F., K. Ishii, P. Bourvagnet, K. Ki, H. Hayashida, T. Miyata, and T. Honjo. 1998. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus [see comments]. *Journal Of Experimental Medicine* 188:2151.
32. Schroeder, H. W. J., M. A. Walter, M. H. Hofker, A. Ebens, K. Willems van Dijk, L. C. Liao, D. W. Cox, E. C. Milner, and R. M. Perlmutter. 1988. Physical linkage of a human immunoglobulin heavy chain variable region

- gene segment to diversity and joining region elements. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 85:8196.
33. Siebenlist, U., J. V. Ravetch, S. Korsmeyer, T. Waldmann, and P. Leder. 1981. Human immunoglobulin D segments encoded in tandem multigenic families. *Nature* 294:631.
  34. Corbett, S. J., I. M. Tomlinson, E. L. L. Sonnhammer, D. Buck, and G. Winter. 1997. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, minor D segments or D-D recombination. *Journal Of Molecular Biology* 270:587.
  35. Okamura, K., H. Ishiguro, Y. Ichihara, and Y. Kurosawa. 1993. Comparison of nucleotide sequences from upstream of the DQ52 gene to the S mu region of immunoglobulin heavy-chain gene loci between *Suncus murinus*, mouse and human. *Molecular Immunology* 30:461.
  36. Ichihara, Y., H. Matsuoka, and Y. Kurosawa. 1988. Organization of human immunoglobulin heavy chain diversity gene loci. *Embo Journal* 7:4141.
  37. Max, E. E. 1999. *Immunoglobulins: Molecular Genetics in Fundamental Immunology*. Lippincott-Raven.
  38. Ravetch, J. V., U. Siebenlist, S. Korsmeyer, T. Waldmann, and P. Leder. 1981. Structure of the human immunoglobulin mu locus: characterization of embryonic and rearranged J and D genes. *Cell* 27:583.
  39. Shimizu, A., N. Takahashi, Y. Yaoita, and T. Honjo. 1982. Organization of the constant-region gene family of the mouse immunoglobulin heavy chain. *Cell* 28:499.
  40. Kirsch, I. R., C. C. Morton, K. Nakahara, and P. Leder. 1982. Human immunoglobulin heavy chain genes map to a region of translocations in malignant B lymphocytes. *Science* 216:301.
  41. Tucker, P. W., K. B. Marcu, N. Newell, J. Richards, and F. R. Blattner. 1979. Sequence of the cloned gene for the constant region of murine gamma 2b immunoglobulin heavy chain. *Science* 206:1303.
  42. Tucker, P. W., J. L. Slightom, and F. R. Blattner. 1981. Mouse IgA heavy chain gene sequence: implications for evolution of immunoglobulin hinge axons. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 78:7684.

43. Calame, K., J. Rogers, P. Early, M. Davis, D. Livant, R. Wall, and L. Hood. 1980. Mouse Cmu heavy chain immunoglobulin gene segment contains three intervening sequences separating domains. *Nature* 284:452.
44. Reth, M. G., and F. W. Alt. 1984. Novel immunoglobulin heavy chains are produced from DJH gene segment rearrangements in lymphoid cells. *Nature* 312:418.
45. Yancopoulos, G. D., and F. W. Alt. 1985. Developmentally controlled and tissue-specific expression of unrearranged V<sub>H</sub> gene segments. *Cell* 40:271.
46. Dreyer, W. J., and J. C. Bennett. 1965. The molecular basis of antibody formation: a paradox. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 54:864.
47. Hozumi, N., and S. Tonegawa. 1976. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 73:3628.
48. Early, P., H. Huang, M. Davis, K. Calame, and L. Hood. 1980. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: V<sub>H</sub>, D and J<sub>H</sub>. *Cell* 19:981.
49. Tonegawa, S. 1988. Somatic generation of immune diversity. *Bioscience Reports* 8:3.
50. Gellert, M. 1992. V(D)J recombination gets a break. *Trends In Genetics* 8:408.
51. Akira, S., K. Okazaki, and H. Sakano. 1987. Two pairs of recombination signals are sufficient to cause immunoglobulin V-(D)-J joining. *Science* 238:1134.
52. Hesse, J. E., M. R. Lieber, K. Mizuuchi, and M. Gellert. 1989. V(D)J recombination: a functional definition of the joining signals. *Genes And Development* 3:1053.
53. Akamatsu, Y., N. Tsurushita, F. Nagawa, M. Matsuoka, K. Okazaki, M. Imai, and H. Sakano. 1994. Essential residues in V(D)J recombination signals. *Journal Of Immunology* 153:4520.
54. Alt, F. W., T. K. Blackwell, R. A. DePinho, M. G. Reth, and G. D. Yancopoulos. 1986. Regulation of genome rearrangement events during lymphocyte differentiation. *Immunological Reviews* 89:5.

55. Yancopoulos, G. D., S. V. Desiderio, M. Paskind, J. F. Kearney, D. Baltimore, and F. W. Alt. 1984. Preferential utilization of the most J<sub>H</sub>-proximal V<sub>H</sub> gene segments in pre-B-cell lines. *Nature* 311:727.
56. Reth, M. G., S. Jackson, and F. W. Alt. 1986. V<sub>H</sub>DJ<sub>H</sub> formation and DJ<sub>H</sub> replacement during pre-B differentiation: non-random usage of gene segments. *Embo Journal* 5:2131.
57. Oettinger, M. A., D. G. Schatz, C. Gorka, and D. Baltimore. 1990. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. *Science* 248:1517.
58. Schatz, D. G., M. A. Oettinger, and D. Baltimore. 1989. The V(D)J recombination activating gene, RAG-1. *Cell* 59:1035.
59. Agrawal, A., Q. M. Eastman, and D. G. Schatz. 1998. Transposition mediated by RAG-1 and RAG-2 and its implications for the evolution of the immune system [see comments]. *Nature* 394:744.
60. Hiom, K., M. Melek, and M. Gellert. 1998. DNA transposition by the RAG-1 and RAG-2 proteins: a possible source of oncogenic translocations [see comments]. *Cell* 94:463.
61. Bernstein, R. M., S. F. Schluter, H. Bernstein, and J. J. Marchalonis. 1996. Primordial emergence of the recombination activating gene 1 (RAG-1): sequence of the complete shark gene indicates homology to microbial integrases. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 93:9454.
62. Roth, D. B., P. B. Nakajima, J. P. Menetski, M. J. Bosma, and M. Gellert. 1992. V(D)J recombination in mouse thymocytes: double-strand breaks near T cell receptor delta rearrangement signals. *Cell* 69:41.
63. Schlissel, M., A. Constantinescu, T. Morrow, M. Baxter, and A. Peng. 1993. Double-strand signal sequence breaks in V(D)J recombination are blunt, 5'-phosphorylated, RAG-dependent, and cell cycle regulated. *Genes And Development* 7:2520.
64. Ramsden, D. A., and M. Gellert. 1995. Formation and resolution of double-strand break intermediates in V(D)J rearrangement. *Genes And Development* 9:2409.
65. McBlane, J. F., D. C. van Gent, D. A. Ramsden, C. Romeo, C. A. Cuomo, M. Gellert, and M. A. Oettinger. 1995. Cleavage at a V(D)J recombination signal

- requires only RAG-1 and RAG-2 proteins and occurs in two steps. *Cell* 83:387.
66. Nadel, B., and A. J. Feeney. 1995. Influence of coding-end sequence on coding-end processing in V(D)J recombination. *Journal Of Immunology* 155:4322.
  67. Roth, D. B., C. Zhu, and M. Gellert. 1993. Characterization of broken DNA molecules associated with V(D)J recombination. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 90:10788.
  68. van Gent, D. C., K. Mizuuchi, and M. Gellert. 1996. Similarities between initiation of V(D)J recombination and retroviral integration [see comments]. *Science* 271:1592.
  69. Roth, D. B., J. P. Menetski, P. B. Nakajima, M. J. Bosma, and M. Gellert. 1992. V(D)J recombination: broken DNA molecules with covalently sealed (hairpin) coding ends in scid mouse thymocytes. *Cell* 70:983.
  70. Zhu, C., and D. B. Roth. 1996. Mechanism of V(D)J recombination. *Cancer Surveys* 28:295.
  71. Sadofsky, M. J., J. E. Hesse, J. F. McBlane, and M. Gellert. 1993. Expression and V(D)J recombination activity of mutated RAG-1 proteins [published erratum appears in *Nucleic Acids Res* 1994 Feb 11;22(3):550]. *Nucleic Acids Research* 21:5644.
  72. Hiom, K., and M. Gellert. 1997. A stable RAG-1-RAG-2-DNA complex that is active in V(D)J cleavage. *Cell* 88:65.
  73. Akamatsu, Y., and M. A. Oettinger. 1998. Distinct roles of RAG-1 and RAG-2 in binding the V(D)J recombination signal sequences. *Molecular And Cellular Biology* 18:4670.
  74. Swanson, P. C., and S. Desiderio. 1998. V(D)J recombination signal recognition: distinct, overlapping DNA-protein contacts in complexes containing RAG-1 with and without RAG-2. *Immunity* 9:115.
  75. Difilippantonio, M. J., C. J. McMahan, Q. M. Eastman, E. Spanopoulou, and D. G. Schatz. 1996. RAG-1 mediates signal sequence recognition and recruitment of RAG-2 in V(D)J recombination. *Cell* 87:253.
  76. Nagawa, F., K. Ishiguro, A. Tsuboi, T. Yoshida, A. Ishikawa, T. Takemori, A. J. Otsuka, and H. Sakano. 1998. Footprint analysis of the RAG protein



- recombination signal sequence complex for V(D)J type recombination. *Molecular And Cellular Biology* 18:655.
77. Spanopoulou, E., F. Zaitseva, F. H. Wang, S. Santagata, D. Baltimore, and G. Panayotou. 1996. The homeodomain region of Rag-1 reveals the parallel mechanisms of bacterial and V(D)J recombination. *Cell* 87:263.
  78. Swanson, P. C., and S. Desiderio. 1999. RAG-2 promotes heptamer occupancy by RAG-1 in the assembly of a V(D)J initiation complex. *Molecular And Cellular Biology* 19:3674.
  79. Gerstein, R. M., and M. R. Lieber. 1993. Coding end sequence can markedly affect the initiation of V(D)J recombination. *Genes And Development* 7:1459.
  80. Boubnov, N. V., Z. P. Wills, and D. T. Weaver. 1995. Coding sequence composition flanking either signal element alters V(D)J recombination efficiency. *Nucleic Acids Research* 23:1060.
  81. Ezekiel, U. R., T. Sun, G. Bozek, and U. Storb. 1997. The composition of coding joints formed in V(D)J recombination is strongly affected by the nucleotide sequence of the coding ends and their relationship to the recombination signal sequences. *Molecular And Cellular Biology* 17:4191.
  82. Ezekiel, U. R., P. Engler, D. Stern, and U. Storb. 1995. Asymmetric processing of coding ends and the effect of coding end nucleotide composition on V(D)J recombination. *Immunity* 2:381.
  83. Ramsden, D. A., J. F. McBlane, D. C. van Gent, and M. Gellert. 1996. Distinct DNA sequence and structure requirements for the two steps of V(D)J recombination signal cleavage. *Embo Journal* 15:3197.
  84. Willett, C. E., J. J. Cherry, and L. A. Steiner. 1997. Characterization and expression of the recombination activating genes (*rag-1* and *rag-2*) of zebrafish. *Immunogenetics* 45:394.
  85. Roman, C. A., S. R. Cherry, and D. Baltimore. 1997. Complementation of V(D)J recombination deficiency in RAG-1(-/-) B cells reveals a requirement for novel elements in the N-terminus of RAG-1. *Immunity* 7:13.
  86. McMahan, C. J., M. J. Difilippantonio, N. Rao, E. Spanopoulou, and D. G. Schatz. 1997. A basic motif in the N-terminal region of RAG-1 enhances V(D)J recombination activity. *Molecular And Cellular Biology* 17:4544.

87. McMahan, C. J., M. J. Sadofsky, and D. G. Schatz. 1997. Definition of a large region of RAG-1 that is important for coimmunoprecipitation of RAG-2. *Journal Of Immunology* 158:2202.
88. Rodgers, K. K., Z. Bu, K. G. Fleming, D. G. Schatz, D. M. Engelman, and J. E. Coleman. 1996. A zinc-binding domain involved in the dimerization of RAG-1. *Journal Of Molecular Biology* 260:70.
89. Bellon, S. F., K. K. Rodgers, D. G. Schatz, J. E. Coleman, and T. A. Steitz. 1997. Crystal structure of the RAG-1 dimerization domain reveals multiple zinc-binding motifs including a novel zinc binuclear cluster. *Nature Structural Biology* 4:586.
90. Cuomo, C. A., and M. A. Oettinger. 1994. Analysis of regions of RAG-2 important for V(D)J recombination. *Nucleic Acids Research* 22:1810.
91. Sadofsky, M. J., J. E. Hesse, D. C. van Gent, and M. Gellert. 1995. RAG-1 mutations that affect the target specificity of V(D)j recombination: a possible direct role of RAG-1 in site recognition. *Genes And Development* 9:2193.
92. Roman, C. A., and D. Baltimore. 1996. Genetic evidence that the RAG-1 protein directly participates in V(D)J recombination through substrate recognition. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 93:2333.
93. Steen, S. B., J. O. Han, C. Mundy, M. A. Oettinger, and D. B. Roth. 1999. Roles of the dispensable portions of RAG-1 and RAG-2 in V(D)J recombination. *Molecular And Cellular Biology* 19:3010.
94. Fugmann, S. D., I. J. Villey, L. M. Ptaszek, and D. G. Schatz. 2000. Identification of two catalytic residues in RAG-1 that define a single active site within the RAG-1/RAG-2 protein complex. *Mol Cell* 5:97.
95. Kim, D. R., Y. Dai, C. L. Mundy, W. Yang, and M. A. Oettinger. 1999. Mutations of acidic residues in RAG-1 define the active site of the V(D)J recombinase. *Genes Dev* 13:3070.
96. Gilfillan, S., A. Dierich, M. Lemeur, C. Benoist, and D. Mathis. 1993. Mice lacking TdT: mature animals with an immature lymphocyte repertoire [published erratum appears in Science 1993 Dec 24;262(5142):1957]. *Science* 261:1175.
97. Komori, T., A. Okada, V. Stewart, and F. W. Alt. 1993. Lack of N regions in antigen receptor variable region genes of TdT-deficient lymphocytes

[published erratum appears in *Science* 1993 Dec 24;262(5142):1957]. *Science* 261:1171.

98. Kepler, T. B., M. Borrero, B. Rugerio, S. K. McCray, and S. H. Clarke. 1996. Interdependence of N nucleotide addition and recombination site choice in V(D)J rearrangement. *Journal Of Immunology* 157:4451.
99. Lafaille, J. J., A. DeCloux, M. Bonneville, Y. Takagaki, and S. Tonegawa. 1989. Junctional sequences of T cell receptor gamma delta genes: implications for gamma delta T cell lineages and for a novel intermediate of V-(D)-J joining. *Cell* 59:859.
100. Nadel, B., and A. J. Feeney. 1997. Nucleotide deletion and P addition in V(D)J recombination: a determinant role of the coding-end sequence. *Molecular And Cellular Biology* 17:3768.
101. Sanz, I., S. S. Wang, G. Meneses, and M. Fischbach. 1994. Molecular characterization of human Ig heavy chain DIR genes. *Journal Of Immunology* 152:3958.
102. Gellert, M. 1992. Molecular analysis of V(D)J recombination. *Annual Review Of Genetics* 26:425.
103. Tuaille, N., A. B. Miller, P. W. Tucker, and J. D. Capra. 1995. Analysis of direct and inverted DJH rearrangements in a human Ig heavy chain transgenic minilocus. *Journal Of Immunology* 154:6453.
104. Sanz, I. 1991. Multiple mechanisms participate in the generation of diversity of human H chain CDR3 regions. *Journal Of Immunology* 147:1720.
105. Yamada, M., R. Wasserman, B. A. Reichard, S. Shane, A. J. Caton, and G. Rovera. 1991. Preferential utilization of specific immunoglobulin heavy chain diversity and joining segments in adult human peripheral blood B lymphocytes. *Journal Of Experimental Medicine* 173:395.
106. Brezinschek, H. P., R. I. Brezinschek, and P. E. Lipsky. 1995. Analysis of the heavy chain repertoire of human peripheral B cells using single-cell polymerase chain reaction. *Journal Of Immunology* 155:190.
107. Moore, B. B., and K. Meek. 1995. Recombination potential of the human DIR elements. *Journal Of Immunology* 154:2175.
108. Malipiero, U. V., N. S. Levy, and P. J. Gearhart. 1987. Somatic mutation in anti-phosphorylcholine antibodies. *Immunological Reviews* 96:59.

109. Betz, A. G., C. Rada, R. Pannell, C. Milstein, and M. S. Neuberger. 1993. Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: clustering, polarity, and specific hot spots. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 90:2385.
110. Milstein, C. 1986. From antibody structure to immunological diversification of immune response. *Science* 231:1261.
111. Golding, G. B., P. J. Gearhart, and B. W. Glickman. 1987. Patterns of somatic mutations in immunoglobulin variable genes. *Genetics* 115:169.
112. Kolchanov, N. A., V. V. Solovyov, and I. B. Rogozin. 1987. Peculiarities of immunoglobulin gene structures as a basis for somatic mutation emergence. *Febs Letters* 214:87.
113. Betz, A. G., M. S. Neuberger, and C. Milstein. 1993. Discriminating intrinsic and antigen-selected mutational hotspots in immunoglobulin V genes. *Immunology Today* 14:405.
114. Jacob, J., G. Kelsoe, K. Rajewsky, and U. Weiss. 1991. Intracлонаl generation of antibody mutants in germinal centres [see comments]. *Nature* 354:389.
115. Jacob, J., and G. Kelsoe. 1992. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. II. A common clonal origin for periarteriolar lymphoid sheath-associated foci and germinal centers. *Journal Of Experimental Medicine* 176:679.
116. McKean, D., K. Huppi, M. Bell, L. Staudt, W. Gerhard, and M. Weigert. 1984. Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 81:3180.
117. Berek, C., and C. Milstein. 1988. The dynamic nature of the antibody repertoire. *Immunological Reviews* 105:5.
118. Weiss, U., R. Zobelein, and K. Rajewsky. 1992. Accumulation of somatic mutants in the B cell compartment after primary immunization with a T cell-dependent antigen. *European Journal Of Immunology* 22:511.
119. Chow, L. T., R. E. Gelinas, T. R. Broker, and R. J. Roberts. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* 12:1.

120. Berget, S. M., C. Moore, and P. A. Sharp. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 74:3171.
121. Black, D. L. 1995. Finding splice sites within a wilderness of RNA. *Rna* 1:763.
122. Hedley, M. L., and T. Maniatis. 1991. Sex-specific splicing and polyadenylation of dsx pre-mRNA requires a sequence that binds specifically to tra-2 protein in vitro. *Cell* 65:579.
123. Manley, J. L., and R. Tacke. 1996. SR proteins and splicing control. *Genes And Development* 10:1569.
124. Green, M. R. 1986. Pre-mRNA Splicing. *Annual Review of Genetics* 20:671.
125. Rio, D. C. 1993. Splicing of pre-mRNA: mechanism, regulation and role in development. *Current Opinion In Genetics And Development* 3:574.
126. Zhuang, Y., and A. M. Weiner. 1986. A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell* 46:827.
127. Séraphin, B., L. Kretzner, and M. Rosbash. 1988. A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5' cleavage site. *Embo Journal* 7:2533.
128. Siliciano, P. G., and C. Guthrie. 1988. 5' splice site selection in yeast: genetic alterations in base-pairing with U1 reveal additional requirements. *Genes And Development* 2:1258.
129. Sawa, H., and J. Abelson. 1992. Evidence for a base-pairing interaction between U6 small nuclear RNA and 5' splice site during the splicing reaction in yeast. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 89:11269.
130. Zeitlin, S., and A. Efstratiadis. 1984. *In vivo* splicing products of the rabbit beta-globin pre-mRNA. *Cell* 39:589.
131. Keller, E. B., and W. A. Noon. 1984. Intron splicing: a conserved internal signal in introns of animal pre-mRNAs. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 81:7417.
132. Reed, R., and T. Maniatis. 1985. Intron sequences involved in lariat formation during pre-mRNA splicing. *Cell* 41:95.
133. Mount, S. M. 1982. A catalogue of splice junction sequences. *Nucleic Acids Research* 10:459.

134. Umen, J. G., and C. Guthrie. 1995. Prp16p, Slu7p, and Prp8p interact with the 3' splice site in two distinct stages during the second catalytic step of pre-mRNA splicing. *Rna* 1:584.
135. Smith, C. W., E. B. Porro, J. G. Patton, and B. Nadal-Ginard. 1989. Scanning from an independently specified branch point defines the 3' splice site of mammalian introns. *Nature* 342:243.
136. Zhuang, Y., and A. M. Weiner. 1990. The conserved dinucleotide AG of the 3' splice site may be recognized twice during in vitro splicing of mammalian mRNA precursors. *Gene* 90:263.
137. Haas, I. G., and M. Wabl. 1983. Immunoglobulin heavy chain binding protein. *Nature* 306:387.
138. Bole, D. G., L. M. Hendershot, and J. F. Kearney. 1986. Posttranslational association of immunoglobulin heavy chain binding protein with nascent heavy chains in nonsecreting and secreting hybridomas. *Journal Of Cell Biology* 102:1558.
139. Hendershot, L., D. Bole, G. Köhler, and J. F. Kearney. 1987. Assembly and secretion of heavy chains that do not associate posttranslationally with immunoglobulin heavy chain-binding protein. *Journal Of Cell Biology* 104:761.
140. Dul, J. L., and Y. Argon. 1990. A single amino acid substitution in the variable region of the light chain specifically blocks immunoglobulin secretion. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 87:8135.
141. Ma, J., J. F. Kearney, and L. M. Hendershot. 1990. Association of transport-defective light chains with immunoglobulin heavy chain binding protein. *Molecular Immunology* 27:623.
142. Knittler, M. R., and I. G. Haas. 1992. Interaction of BiP with newly synthesized immunoglobulin light chain molecules: cycles of sequential binding and release. *Embo Journal* 11:1573.
143. Haas, I. G. 1991. BiP--a heat shock protein involved in immunoglobulin chain assembly. *Current Topics In Microbiology And Immunology* 167:71.
144. Blond-Elguindi, S., S. E. Cwirla, W. J. Dower, R. J. Lipshutz, S. R. Sprang, J. F. Sambrook, and M. J. Gething. 1993. Affinity panning of a library of

- peptides displayed on bacteriophages reveals the binding specificity of BiP. *Cell* 75:717.
145. Muyldermans, S., T. Atarhouch, J. Saldanha, J. A. Barbosa, and R. Hamers. 1994. Sequence and structure of V<sub>H</sub> domain from naturally occurring camel heavy chain immunoglobulins lacking light chains. *Protein Engineering* 7:1129.
  146. Uhr, J. W., J. Dancis, E. C. Franklin, M. S. Finkelstein, and E. W. Lewis. 1962. The antibody response to bacteriophage in newborn premature infants. *Journal of Clinical Investigation* 41:1508.
  147. Pernis, B., L. Forni, and A. L. Luzzati. 1977. Synthesis of multiple immunoglobulin classes by single lymphocytes. *Cold Spring Harbor Symposia On Quantitative Biology* 41 Pt 1:175.
  148. Gearhart, P. J. 1977. Non-sequential expression of multiple immunoglobulin classes by isolated B-cell clones. *Nature* 269:812.
  149. Davis, M. M., S. K. Kim, and L. E. Hood. 1980. DNA sequences mediating class switching in alpha-immunoglobulins. *Science* 209:1360.
  150. Dunnick, W., T. H. Rabbitts, and C. Milstein. 1980. An immunoglobulin deletion mutant with implications for the heavy-chain switch and RNA splicing. *Nature* 286:669.
  151. Kataoka, T., T. Kawakami, N. Takahashi, and T. Honjo. 1980. Rearrangement of immunoglobulin gamma 1-chain gene and mechanism for heavy-chain class switch. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 77:919.
  152. Kataoka, T., T. Miyata, and T. Honjo. 1981. Repetitive sequences in class-switch recombination regions of immunoglobulin heavy chain genes. *Cell* 23:357.
  153. Dunnick, W., B. E. Shell, and C. Dery. 1983. DNA sequences near the site of reciprocal recombination between a c-myc oncogene and an immunoglobulin switch region. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 80:7269.
  154. Nikaïdo, T., Y. Yamawaki-Kataoka, and T. Honjo. 1982. Nucleotide sequences of switch regions of immunoglobulin C epsilon and C gamma genes and their comparison. *Journal Of Biological Chemistry* 257:7322.

155. Kinoshita, K., J. Tashiro, S. Tomita, C. G. Lee, and T. Honjo. 1998. Target specificity of immunoglobulin class switch recombination is not determined by nucleotide sequences of S regions. *Immunity* 9:849.
156. Schäcke, H., W. E. Müller, V. Gamulin, and B. Rinkevich. 1994. The Ig superfamily includes members from the lowest invertebrates to the highest vertebrates [letter]. *Immunology Today* 15:497.
157. Bork, P., L. Holm, and C. Sander. 1994. The immunoglobulin fold. Structural classification, sequence patterns and common core. *Journal Of Molecular Biology* 242:309.
158. Adema, C. M., L. A. Hertel, R. D. Miller, and E. S. Loker. 1997. A family of fibrinogen-related proteins that precipitates parasite-derived molecules is produced by an invertebrate after infection. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 94:8691.
159. Hsu, E. 1994. The variation in immunoglobulin heavy chain constant regions in evolution. *Seminars In Immunology* 6:383.
160. Dard, P., S. Huck, J. P. Fripiat, G. Lefranc, A. Langaney, M. P. Lefranc, and A. Sanchez-Mazas. 1997. The IGHG3 gene shows a structural polymorphism characterized by different hinge lengths: sequence of a new 2-exon hinge gene. *Human Genetics* 99:138.
161. Wilson, M. R., A. Marcuz, F. van Ginkel, N. W. Miller, L. W. Clem, D. Middleton, and G. W. Warr. 1990. The immunoglobulin M heavy chain constant region gene of the channel catfish, *Ictalurus punctatus*: an unusual mRNA splice pattern produces the membrane form of the molecule. *Nucleic Acids Research* 18:5227.
162. Wilson, M. R., D. A. Ross, N. W. Miller, L. W. Clem, D. L. Middleton, and G. W. Warr. 1995. Alternate pre-mRNA processing pathways in the production of membrane IgM heavy chains in holostean fish. *Developmental And Comparative Immunology* 19:165.
163. Brown, W. R., I. Kacskovics, B. A. Amendt, N. B. Blackmore, M. Rothschild, R. Shinde, and J. E. Butler. 1995. The hinge deletion allelic variant of porcine IgA results from a mutation at the splice acceptor site in the first C alpha intron. *Journal Of Immunology* 154:3836.
164. Patel, H. M., and E. Hsu. 1997. Abbreviated junctional sequences impoverish antibody diversity in urodele amphibians. *Journal Of Immunology* 159:3391.



165. Greenberg, A. S., D. Avila, M. Hughes, A. Hughes, E. C. McKinney, and M. F. Flajnik. 1995. A new antigen receptor gene family that undergoes rearrangement and extensive somatic diversification in sharks. *Nature* 374:168.
166. Reynaud, C. A., A. Dahan, V. Anquez, and J. C. Weill. 1989. Somatic hyperconversion diversifies the single V<sub>H</sub> gene of the chicken with a high incidence in the D region. *Cell* 59:171.
167. Becker, R. S., and K. L. Knight. 1990. Somatic diversification of immunoglobulin heavy chain VDJ genes: evidence for somatic gene conversion in rabbits. *Cell* 63:987.
168. Hinds, K. R., and G. W. Litman. 1986. Major reorganization of immunoglobulin V<sub>H</sub> segmental elements during vertebrate evolution. *Nature* 320:546.
169. Young, J. Z. 1981. *The Life of Vertebrates*. Clarendon Press.
170. Fowler, M. E. 1998. *Medicine and Surgery of South American Camelids: Llama, Alpaca, Vicuna, Guanaco*. Iowa State University Press, Iowa.
171. van der Linden, R. H., L. G. Frenken, B. de Geus, M. M. Harmsen, R. C. Ruuls, W. Stok, L. de Ron, S. Wilson, P. Davis, and C. T. Verrips. 1999. Comparison of physical chemical properties of llama V<sub>HH</sub> antibody fragments and mouse monoclonal antibodies. *Biochimica Et Biophysica Acta* 1431:37.
172. Spinelli, S., L. Frenken, D. Bourgeois, L. de Ron, W. Bos, T. Verrips, C. Anguille, C. Cambillau, and M. Tegoni. 1996. The crystal structure of a llama heavy chain variable domain [letter] [see comments]. *Nature Structural Biology* 3:752.
173. Desmyter, A., T. R. Transue, M. A. Ghahroudi, M. H. Thi, F. Poortmans, R. Hamers, S. Muyldermans, and L. Wyns. 1996. Crystal structure of a camel single-domain V<sub>H</sub> antibody fragment in complex with lysozyme [see comments]. *Nature Structural Biology* 3:803.
174. Decanniere, K., A. Desmyter, M. Lauwereys, M. A. Ghahroudi, S. Muyldermans, and L. Wyns. 1999. A single-domain antibody fragment in complex with RNase A: non-canonical loop structures and nanomolar affinity using two CDR loops. *Structure* 7:361.

175. Vu, K. B., M. A. Ghahroudi, L. Wyns, and S. Muyldermans. 1997. Comparison of llama V<sub>H</sub> sequences from conventional and heavy chain antibodies. *Molecular Immunology* 34:1121.
176. Seligmann, M., E. Mihaesco, J. L. Preud'homme, F. Danon, and J. C. Brouet. 1979. Heavy chain diseases: current findings and concepts. *Immunological Reviews* 48:145.
177. Greenhalgh, P., and L. A. Steiner. 1995. Recombination activating gene 1 (*rag-1*) in zebrafish and shark. *Immunogenetics* 41:54.
178. Roux, K. H., A. S. Greenberg, L. Greene, L. Strelets, D. Avila, E. C. McKinney, and M. F. Flajnik. 1998. Structural analysis of the nurse shark (new) antigen receptor (NAR): molecular convergence of NAR and unusual mammalian immunoglobulins. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 95:11804.
179. Maniatis, T., J. Sambrook, and E. F. Fritsch. 1989. *Molecular Cloning : A Laboratory Manual*. Cold Spring Harbour Laboratory, U.S.
180. Cromie, K. D., C. P. E. Van der Logt, S. C. Williams, M. Van-der-Vaart, and H. Peters. 1998. BioRecognition Antibody Fragments. Unilever Research, Colworth, p. 43.
181. Brunak, S., J. Engelbrecht, and S. Knudsen. 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *Journal Of Molecular Biology* 220:49.
182. Quandt, K., K. Frech, H. Karas, E. Wingender, and T. Werner. 1995. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research* 23:4878.
183. Wingender, E., X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Pruss, I. Reuter, and F. Schacherer. 2000. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Research* 28:316.
184. Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22:4673.
185. Felstein, J. 1989. PHYLIP -- Phylogeny Inference Package (Version 3.2). *Cladistics* 5:164.

186. Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389.
187. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal Of Molecular Biology* 215:403.
188. Winter G, Griffiths A. D, Hawkins R. E, Hoogenboom H. R. 1994 Making antibodies by phage display technology. *Annual Reviews in Immunology*. 12:433-55.
189. Clackson T, Hoogenboom H. R, Griffiths A. D, Winter G. 1991 Making antibody fragments using phage display libraries. *Nature* 352:624-8
190. Nguyen, V. K., R. Hamers, L. Wyns, and S. Muyldermans. 2000. Camel heavy-chain antibodies: diverse germline V(H)H and specific mechanisms enlarge the antigen-binding repertoire. *Embo Journal* 19:921.
191. Harmsen, M. M., R. C. Ruuls, T. A. Niewold, L. G. J. Frenken, and B. de Geus. 2000. Camelid heavy-chain V regions consist of at least four distinct subfamilies some of which reveal novel sequence features. *In Press*.
192. Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research* 16:10881.
193. Vu, K. B. 1999. Llama Antibodies and Engineering of Single-domain Antibody Fragments. In *Laboratorium Ultrastructuur en Algemene Biologie*. Vrije Universiteit Brussel, Brussels, p. 126.
194. Willems van Dijk, K., H. W. J. Schroeder, R. M. Perlmutter, and E. C. Milner. 1989. Heterogeneity in the human Ig V<sub>H</sub> locus. *Journal Of Immunology* 142:2547.
195. Decanniere, K., S. Muyldermans, and L. Wyns. 2000. Canonical antigen-binding loop structures in immunoglobulins: more structures, more canonical classes? *Journal Of Molecular Biology* 300:83.
196. Wang, J. H., A. Nichogiannopoulou, L. Wu, L. Sun, A. H. Sharpe, M. Bigby, and K. Georgopoulos. 1996. Selective defects in the development of the fetal and adult lymphoid system in mice with an Ikaros null mutation. *Immunity* 5:537.

197. Andres, V., M. D. Chiara, and V. Mahdavi. 1994. A new bipartite DNA-binding domain: cooperative interaction between the cut repeat and homeo domain of the cut homeo proteins. *Genes And Development* 8:245.
198. Piette, J., and M. Yaniv. 1987. Two different factors bind to the alpha-domain of the polyoma virus enhancer, one of which also interacts with the SV40 and c-fos enhancers. *Embo Journal* 6:1331.
199. Hannon, R., T. Evans, G. Felsenfeld, and H. Gould. 1991. Structure and promoter activity of the gene for the erythroid transcription factor GATA-1. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 88:3004.
200. Mermod, N., T. J. Williams, and R. Tjian. 1988. Enhancer binding factors AP-4 and AP-1 act in concert to activate SV40 late transcription in vitro. *Nature* 332:557.
201. Molnar, A., and K. Georgopoulos. 1994. The Ikaros gene encodes a family of functionally diverse zinc finger DNA-binding proteins. *Molecular And Cellular Biology* 14:8292.
202. Akira, S., H. Isshiki, T. Sugita, O. Tanabe, S. Kinoshita, Y. Nishio, T. Nakajima, T. Hirano, and T. Kishimoto. 1990. A nuclear factor for IL-6 expression (NF-IL6) is a member of a C/EBP family. *Embo Journal* 9:1897.
203. Pierrou, S., M. Hellqvist, L. Samuelsson, S. Enerback, and P. Carlsson. 1994. Cloning and characterization of seven human forkhead proteins: binding site specificity and DNA bending. *Embo Journal* 13:5002.
204. Sun, X. H., and D. Baltimore. 1991. An inhibitory domain of E12 transcription factor prevents DNA binding in E12 homodimers but not in E12 heterodimers [published erratum appears in Cell 1991 Aug 9;66(3):423]. *Cell* 64:459.
205. Klempnauer, K. H., and A. E. Sippel. 1987. The highly conserved amino-terminal region of the protein encoded by the v-myb oncogene functions as a DNA-binding domain. *Embo Journal* 6:2719.
206. Nguyen, V. K., S. Muyldermans, and R. Hamers. 1998. The specific variable domain of camel heavy-chain antibodies is encoded in the germline. *Journal Of Molecular Biology* 275:413.
207. Kleinfeld, R., R. R. Hardy, D. Tarlinton, J. Dangl, L. A. Herzenberg, and M. Weigert. 1986. Recombination between an expressed immunoglobulin heavy-

- chain gene and a germline variable gene segment in a Ly 1+ B-cell lymphoma. *Nature* 322:843.
208. Komori, T., H. Sugiyama, and S. Kishimoto. 1989. A novel V<sub>H</sub>DJH to J<sub>H</sub> joining that induces H chain production in an Ig-null immature B cell line. *Journal Of Immunology* 143:1040.
209. Yélamos, J., N. Klix, B. Goyenechea, F. Lozano, Y. L. Chui, A. González Fernández, R. Pannell, M. S. Neuberger, and C. Milstein. 1995. Targeting of non-Ig sequences in place of the V segment by somatic hypermutation. *Nature* 376:225.
210. Milstein, C., M. S. Neuberger, and R. Staden. 1998. Both DNA strands of antibody genes are hypermutation targets. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 95:8791.
211. Yelamos, J., N. Klix, B. Goyenechea, F. Lozano, Y. L. Chui, A. Gonzalez Fernandez, R. Pannell, M. S. Neuberger, and C. Milstein. 1995. Targeting of non-Ig sequences in place of the V segment by somatic hypermutation. *Nature* 376:225.
212. Millstein, L., P. Eversole-Cire, J. Blanco, and J. M. Gottesfeld. 1987. Differential transcription of *Xenopus* oocyte and somatic-type 5 S genes in a *Xenopus* oocyte extract. *Journal Of Biological Chemistry* 262:17100.
213. Zaller, D. M., and L. A. Eckhardt. 1985. Deletion of a B-cell-specific enhancer affects transfected, but not endogenous, immunoglobulin heavy-chain gene expression. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 82:5088.
214. Fugmann, S. D., A. I. Lee, P. E. Shockett, I. J. Villey, and D. G. Schatz. 2000. The RAG proteins and V(D)J recombination: complexes, ends, and transposition. *Annual Reviews in Immunology* 18:495.
215. van Gent, D. C., J. F. McBlane, D. A. Ramsden, M. J. Sadofsky, J. E. Hesse, and M. Gellert. 1995. Initiation of V(D)J recombination in a cell-free system. *Cell* 81:925.
216. van Gent, D. C., J. F. McBlane, D. A. Ramsden, M. J. Sadofsky, J. E. Hesse, and M. Gellert. 1996. Initiation of V(D)J recombinations in a cell-free system by RAG-1 and RAG-2 proteins. *Current Topics In Microbiology And Immunology* 217:1.

217. Hansen, J. D., and S. L. Kaattari. 1995. The recombination activation gene 1 (RAG-1) of rainbow trout (*Oncorhynchus mykiss*): cloning, expression, and phylogenetic analysis. *Immunogenetics* 42:188.
218. Hansen, J. D., and S. L. Kaattari. 1996. The recombination activating gene 2 (RAG-2) of the rainbow trout *Oncorhynchus mykiss*. *Immunogenetics* 44:203.
219. Fuschiotti, P., N. Harindranath, R. G. Mage, W. T. McCormack, P. Dhanarajan, and K. H. Roux. 1993. Recombination activating genes-1 and -2 of the rabbit: cloning and characterization of germline and expressed genes. *Molecular Immunology* 30:1021.
220. Bernstein, R. M., S. F. Schluter, D. F. Lake, and J. J. Marchalonis. 1994. Evolutionary conservation and molecular cloning of the recombinase activating gene 1. *Biochemical And Biophysical Research Communications* 205:687.
221. Greenhalgh, P., C. E. Olesen, and L. A. Steiner. 1993. Characterization and expression of recombination activating genes (RAG-1 and RAG-2) in *Xenopus laevis*. *Journal Of Immunology* 151:3100.
222. Silver, D. P., E. Spanopoulou, R. C. Mulligan, and D. Baltimore. 1993. Dispensable sequence motifs in the RAG-1 and RAG-2 genes for plasmid V(D)J recombination. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 90:6100.
223. Sadofsky, M. J., J. E. Hesse, and M. Gellert. 1994. Definition of a core region of RAG-2 that is functional in V(D)J recombination. *Nucleic Acids Research* 22:1805.
224. Callebaut, I., and J. P. Mornon. 1998. The V(D)J recombination activating protein RAG-2 consists of a six-bladed propeller and a PHD fingerlike domain, as revealed by sequence analysis. *Cellular And Molecular Life Sciences* 54:880.
225. Rodgers, K. K., I. J. Villey, L. Ptaszek, E. Corbett, D. G. Schatz, and J. E. Coleman. 1999. A dimer of the lymphoid protein RAG-1 recognizes the recombination signal sequence and the complex stably incorporates the high mobility group protein HMG2. *Nucleic Acids Research* 27:2938.
226. Maina, C. V., P. D. Riggs, A. G. d. Grandea, B. E. Slatko, L. S. Moran, J. A. Tagliamonte, L. A. McReynolds, and C. D. Guan. 1988. An *Escherichia coli*

- vector to express and purify foreign proteins by fusion to and separation from maltose-binding protein. *Gene* 74:365.
227. Bush, G. L., A. M. Tassin, H. Friden, and D. I. Meyer. 1991. Secretion in yeast. Purification and in vitro translocation of chemical amounts of prepro-alpha-factor. *Journal Of Biological Chemistry* 266:13811.
228. Schlissel, M. S., and P. Stanhope-Baker. 1997. Accessibility and the developmental regulation of V(D)J recombination. *Seminars In Immunology* 9:161.
229. Sleckman, B. P., J. R. Gorman, and F. W. Alt. 1996. Accessibility control of antigen-receptor variable-region gene assembly: role of cis-acting elements. *Annual Review Of Immunology* 14:459.
230. D'Agostaro, G., E. Hevia, G. E. Wu, and H. Murialdo. 1985. Site-directed cleavage of immunoglobulin gene segments by lymphoid cell extracts. *Canadian Journal Of Biochemistry And Cell Biology* 63:969.
231. Hesse, J. E., M. R. Lieber, M. Gellert, and K. Mizuuchi. 1987. Extrachromosomal DNA substrates in pre-B cells undergo inversion or deletion at immunoglobulin V-(D)-J joining signals. *Cell* 49:775.
232. Kameyama, K., M. Mochizuki, K. Tanaka, and K. Sugimoto. 1993. Convenient plasmid for extrachromosomal DNA recombination in mouse cells. *Biochemical And Biophysical Research Communications* 192:1327.
233. Pan, P. Y., M. R. Lieber, and J. M. Teale. 1997. The role of recombination signal sequences in the preferential joining by deletion in DH-JH recombination and in the ordered rearrangement of the IgH locus. *International Immunology* 9:515.
234. van Gent, D. C., D. A. Ramsden, and M. Gellert. 1996. The RAG-1 and RAG-2 proteins establish the 12/23 rule in V(D)J recombination. *Cell* 85:107.
235. Hiom, K., and M. Gellert. 1998. Assembly of a 12/23 paired signal complex: a critical control point in V(D)J recombination. *Mol Cell* 1:1011.
236. Hiom, K. 2000. Personal Communication.
237. Kwon, J., A. N. Imbalzano, A. Matthews, and M. A. Oettinger. 1998. Accessibility of nucleosomal DNA to V(D)J cleavage is modulated by RSS positioning and HMG1. *Mol Cell* 2:829.

238. West, R. B., and M. R. Lieber. 1998. The RAG-HMG1 complex enforces the 12/23 rule of V(D)J recombination specifically at the double-hairpin formation step. *Molecular And Cellular Biology* 18:6408.
239. van Gent, D. C., K. Hiom, T. T. Paull, and M. Gellert. 1997. Stimulation of V(D)J cleavage by high mobility group proteins. *Embo Journal* 16:2665.
240. Mo, X., T. Bailin, S. Noggle, and M. J. Sadofsky. 2000. A highly ordered structure in V(D)J recombination cleavage complexes is facilitated by HMG1. *Nucleic Acids Research* 28:1228.
241. Sawchuk, D. J., F. Weis-Garcia, S. Malik, E. Besmer, M. Bustin, M. C. Nussenzweig, and P. Cortes. 1997. V(D)J recombination: modulation of RAG-1 and RAG-2 cleavage activity on 12/23 substrates by whole cell extract and DNA-bending proteins. *Journal Of Experimental Medicine* 185:2025.
242. Santagata, S., V. Aidinis, and E. Spanopoulou. 1998. The effect of Me<sup>2+</sup> cofactors at the initial stages of V(D)J recombination. *Journal Of Biological Chemistry* 273:16325.
243. Frenken, L. G. J. 1999. Personal Communication.
244. Aebi, M., H. Hornig, R. A. Padgett, J. Reiser, and C. Weissmann. 1986. Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. *Cell* 47:555.
245. Aase, A., I. Sandlie, L. Norderhaug, O. H. Brekke, and T. E. Michaelsen. 1993. The extended hinge region of IgG3 is not required for high phagocytic capacity mediated by Fc gamma receptors, but the heavy chains must be disulfide bonded. *European Journal Of Immunology* 23:1546.
246. Brekke, O. H., B. Bremnes, R. Sandin, A. Aase, T. E. Michaelsen, and I. Sandlie. 1993. Human IgG3 can adopt the disulfide bond pattern characteristic for IgG1 without resembling it in complement mediated cell lysis. *Molecular Immunology* 30:1419.
247. Michaelsen, T. E., O. H. Brekke, A. Aase, R. H. Sandin, B. Bremnes, and I. Sandlie. 1994. One disulfide bond in front of the second heavy chain constant region is necessary and sufficient for effector functions of human IgG3 without a genetic hinge. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 91:9243.
248. Schumaker, V. N., M. L. Phillips, and D. C. Hanson. 1991. Dynamic aspects of antibody structure. *Molecular Immunology* 28:1347.



249. Cooper, L. J., A. R. Shikhman, D. D. Glass, D. Kangisser, M. W. Cunningham, and N. S. Greenspan. 1993. Role of heavy chain constant domains in antibody-antigen interaction. Apparent specificity differences among streptococcal IgG antibodies expressing identical variable domains. *Journal Of Immunology* 150:2231.
250. Evans, J. S., B. A. Levine, I. P. Trayer, C. J. Dorman, and C. F. Higgins. 1986. Sequence-imposed structural constraints in the TonB protein of *E. coli*. *Febs Letters* 208:211.
251. Sandor, M., and R. G. Lynch. 1993. The biology and pathology of Fc receptors. *Journal Of Clinical Immunology* 13:237.
252. Burton, D. R. 1985. Immunoglobulin G: functional sites. *Molecular Immunology* 22:161.
253. Kishore, U., and K. B. M. Reid. 2000. C1q: structure, function, and receptors. *Immunopharmacology* 49:159.
254. Duncan, A. R., and G. Winter. 1988. The binding site for C1q on IgG. *Nature* 332:738.
255. Brekke, O. H., T. E. Michaelsen, A. Aase, R. H. Sandin, and I. Sandlie. 1994. Human IgG isotype-specific amino acid residues affecting complement-mediated cell lysis and phagocytosis. *European Journal Of Immunology* 24:2542.
256. Early, P., J. Rogers, M. Davis, K. Calame, M. Bond, R. Wall, and L. Hood. 1980. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* 20:313.
257. Nguyen, V. K., R. Hamers, L. Wyns, and S. Muyldermans. 1999. Loss of splice consensus signal is responsible for the removal of the entire C(H)1 domain of the functional camel IGG2A heavy-chain antibodies. *Molecular Immunology* 36:515.
258. Borggreffe, T., L. Masat, M. Wabl, B. Riwar, G. Cattoretti, and R. Jessberger. 1999. Cellular, intracellular, and developmental expression patterns of murine SWAP-70. *European Journal Of Immunology* 29:1812.
259. Borggreffe, T., M. Wabl, A. T. Akhmedov, and R. Jessberger. 1998. A B-cell-specific DNA recombination complex. *Journal Of Biological Chemistry* 273:17025.

260. Longacre, A., and U. Storb. 2000. A Novel Cytidine Deaminase affects Antibody Diversity. *Cell* 102:541
261. Nagaoka, H., K. Ozawa, F. Matsuda, H. Hayashida, R. Matsumura, M. Haino, E. K. Shin, Y. Fukita, T. Imai, R. Anand, and e. al. 1994. Recent translocation of variable and diversity segments of the human immunoglobulin heavy chain from chromosome 14 to chromosomes 15 and 16. *Genomics* 22:189.
262. Tomlinson, I. M., G. P. Cook, N. P. Carter, R. Elaswarapu, S. Smith, G. Walter, L. Buluwela, T. H. Rabbitts, and G. Winter. 1994. Human immunoglobulin V<sub>H</sub> and D segments on chromosomes 15q11.2 and 16p11.2. *Human Molecular Genetics* 3:853.
263. Cook, G. P., I. M. Tomlinson, G. Walter, H. Riethman, N. P. Carter, L. Buluwela, G. Winter, and T. H. Rabbitts. 1994. A map of the human immunoglobulin V<sub>H</sub> locus completed by analysis of the telomeric region of chromosome 14q. *Nature Genetics* 7:162.
264. Zhang, K. 2000. Immunoglobulin class switch recombination machinery: progress and challenges. *Clin Immunol* 95:1.
265. Komori, T., Y. Minami, N. Sakato, and H. Sugiyama. 1993. Biased usage of two restricted V<sub>H</sub> gene segments in V<sub>H</sub> replacement. *European Journal Of Immunology* 23:517.
266. Blackwell, T. K., M. W. Moore, G. D. Yancopoulos, H. Suh, S. Lutzker, E. Selsing, and F. W. Alt. 1986. Recombination between immunoglobulin variable region gene segments is enhanced by transcription. *Nature* 324:585.
267. Gu, H., D. Tarlinton, W. Müller, K. Rajewsky, and I. Förster. 1991. Most peripheral B cells in mice are ligand selected. *Journal Of Experimental Medicine* 173:1357.
268. Hahm, K., P. Ernst, K. Lo, G. S. Kim, C. Turck, and S. T. Smale. 1994. The lymphoid transcription factor LyF-1 is encoded by specific, alternatively spliced mRNAs derived from the Ikaros gene. *Molecular And Cellular Biology* 14:7111.
269. Tuaillon, N., and J. D. Capra. 2000. Evidence that terminal deoxynucleotidyltransferase expression plays a role in Ig heavy chain gene segment utilization. *Journal Of Immunology* 164:6387.
270. Muyltermans, S. 2000. Personal Communication.

# Appendices

## Appendix I Bacterial Strains, Media, Buffers and Solutions

(in alphabetical order)

Unless otherwise stated all chemicals and buffer components were obtained from Sigma-Aldrich, Poole, UK

### i. Bacterial Host Strains

Bacterial strains used in this thesis are detailed in table AP.1

Strain Name	Protocol	Characteristics	Genotype
XL-1 Blue (Stratagene, La Jolla, US)	RAG-1 Probe cloning (section 2.4.5 and 4.3.1)	General <i>E.coli</i> -derived cloning strain	<i>recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F<sup>+</sup>proAB lacI<sup>q</sup>ZDM15 Tn10 (Tet<sup>r</sup>)]</i>
XL-1 Blue MRA (P2) (Stratagene, La Jolla, US)	Library Bacteriophage propagation (section 2.3.6)	allows <i>spi</i> (sensitive to P2 inhibition) selection of recombinant phage	$\Delta$ ( <i>mcrA</i> )183 $\Delta$ ( <i>mrcCB</i> - <i>hsdSMR</i> - <i>mrr</i> )173 <i>endA1 supE44 thi-1 gyrA96 relA1 lac 1</i> (P2 lysogen)
TOP 10 (Invitrogen, Carlsbad, US)	Subcloning of RAG-1 bacteriophage clone and others (section 2.5.2)	General <i>E.coli</i> -derived cloning strain	F <sup>+</sup> <i>mcrA</i> $\Delta$ ( <i>mrr</i> - <i>hsdRMS</i> - <i>mcrBC</i> ) $\Phi$ 80 <i>lacZAM15</i> $\Delta$ <i>lacX74 recA1 deoR araD139</i> $\Delta$ ( <i>ara-leu</i> )7697 <i>galU galK rpsL</i> (Str <sup>s</sup> ) <i>endA1 nupG</i>
DH10Bac (Gibco-BRL, Paisley UK)	Cloning of partial llama RAG-1 gene (section 2.5.2)	Contains the bMON14272 bacmid and a helper plasmid to allow pFastBac (section APx) insert transposition	F <sup>+</sup> <i>mcrA</i> D( <i>mrr</i> - <i>hsdRMS</i> - <i>mcrBC</i> ) $\Phi$ 80 <i>dlacZDM15 DlacX74 deoR recA1 endA1 araD139 D(ara, leu)</i> 7697 <i>galU galK 1-rpsL nupG/bMON14272 /pMON7124</i>

Table AP.1 Strains of Bacteria Used in This Thesis

### ii. Denaturing solution (Plaque Lifting)

Denaturing solution was used to prepare filters for pre-hybridisation after plaque lifting and was prepared as detailed in table AP.2

Reagent	Concentration
NaCl	1.5M
NaOH	0.5M
Sterilise by autoclaving.	

Table AP.2 Components of denaturing solution

### iii. Denhardt's Reagent (100 x)

Denhardt's reagent was used as a component of prehybridisation and hybridisation buffers and made up as detailed in table AP.3

Reagent	Volume/Concentration
Ficoll 400 (Sigma-Aldrich, Poole, UK)	2g
Polyvinyl pyrrolidone (Sigma-Aldrich, Poole, UK)	2g
Bovine serum albumin (Fraction V) (Sigma-Aldrich, Poole, UK)	2g

Table AP.3 Components of Denhardt's Reagent

### iv. Hybridisation Solutions

1) Oligonucleotide prehybridisation solution was prepared as detailed in table AP.4a

Reagent	Concentration
SSC (section APx)	6 x
Denhardt's reagent (section APx)	10 x
Denatured salmon sperm DNA	50µg/ml

Table AP.4a Components of oligonucleotide prehybridisation solution

2) Prehybridisation solution for use with PCR-derived probes was prepared as described in table AP.4b

Reagent	Concentration
SSC (section APx)	6 x
Denhardt's reagent (section APx)	5 x
SDS	0.5%
Denatured salmon sperm DNA	50µg/ml

Table AP.4b Components of PCR-probe prehybridisation solution

### v. Insect Media

Insect media was made up under sterile conditions as detailed in Table AP.5

Media Component	Percentage Composition
TNM-FH (Gibco-BRL, Paisley, UK) supplemented with:	
Lipid Concentrate (100x) (Gibco-BRL, Paisley, UK)	1%
Fetal Calf Serum (Gibco-BRL, Paisley, UK)	10%

Table AP.5 Components of supplemented insect cell media used for growth of Sf9 insect cells during baculovirus propagation. Note Serum-free media was used during Sf9 transfection.

## vi. Loading Dye – Agarose

DNA loading dye was made up as a 5x stock as detailed in table AP.6

Reagent	Volume/Concentration
Tris-HCl pH7.5	10mM
EDTA (disodium, dihydrate)	50mM
Ficoll® 400	10%
Bromophenol Blue	0.25%
Xylene Cyanol FF	0.25%

Table AP.6 Composition of DNA loading dye for agarose gel loading

## vii. Loading Dye SDS PAGE

Protein loading dye was made up as a 5x stock as detailed in table AP.7

Reagent	Concentration
Tris-HCl (pH 6.8)	0.625M
SDS	10% (w/v)
Sucrose	50% (w/v)
Mercaptoethanol	10% (v/v)
Bromophenol Blue	0.25%

AP.7 Components of SDS PAGE loading dye

## viii. Neutralising solution (Plaque Lifting)

Neutralising solution was used after denaturing solution prior to filter pre-hybridisation and prepared as detailed in table AP.8

Reagent	Concentration
NaCl	1.5M
Tris Cl (pH7.2)	0.5M
Na <sub>2</sub> EDTA	0.001M
<b>Sterilise by autoclaving</b>	

Table AP.8 Components of neutralising solution

## ix. Oligonucleotide Elution Buffer

Oligonucleotide elution buffer was made up as detailed in table AP.9

Reagent	Concentration
Magnesium Acetate	10mM
Sodium Acetate pH 7.0	500mM
EDTA	1mM

Table AP.9 Composition of Oligonucleotide Elution Buffer for gel purification of recombination substrate oligonucleotides

### x. Polyacrylamide Gels

Polyacrylamide gels were made up as detailed in table AP.10. Gels were poured immediately after addition of TEMED and allowed to set for 1 hour before pre-running.

Reagent	Volume
SequaGel Buffer	3ml
SequaGel Concentrate (containing acrylamide mix)	14.4ml
SequaGel Diluent	12.6ml
10% Ammonium Persulphate Solution )	0.24ml
TEMED	12 $\mu$ l

**Table AP.10 Components of a 12% Sequencing Gel.** Note: Quantities can be varied to obtain other acrylamide percentages (for example for oligonucleotide purification).

### xi. Selective Plates

Agar (15g) was added to 2xTY media (1 litre) and the solution autoclaved and cooled to 55°C. Antibiotic was then added to the specified concentration (see table AP.11). Media was then poured into 90mm diameter petri dishes and allowed to harden.

Antibiotic	Final Concentration
Ampicillin	100 $\mu$ g/ml
Tetracyclin	12.5 $\mu$ g/ml
Kanamycin	30 $\mu$ g/ml
Gentamycin	7 $\mu$ g/ml

Table AP.11. Antibiotic Concentrations within selective plates

### xii. SM Phage Buffer

SM Buffer was made up as described in Table AP.12

Reagent	Volume
NaCl	5.8g
MgSO <sub>4</sub> · 7H <sub>2</sub> O	2g
Tris-HCl (pH 7.5)	50ml
Gelatin (2% (w/v))	5ml
dH <sub>2</sub> O	to final volume 1 litre (Autoclave)

Table AP.12 Composition of Phage SM Buffer

### xiii. SSC (20x)

2xSSC and 6xSSC were made up from a 20xSSC stock the composition of which is described in table AP.13

Reagent	Volume
NaCl	175.3g
Sodium Citrate	88.2g
dH <sub>2</sub> O	800ml
NaOH (10N)	to pH 7.0
dH <sub>2</sub> O	to final volume 1 litre

**Table AP.13 Composition of 20x SSC.** Note stock solution was diluted 1 in 10 prior to use. 6x SSC was used during hybridisation

### xiv. TBE 10x buffer

TBE buffer was made up as detailed in table AP.14

Reagent	Volume
Tris Base	107.8g
Boric Acid	~55g (to correct pH)
EDTA (disodium, dhydrate)	7.44g
dH <sub>2</sub> O	To volume 1 litre
Adjust pH to 8.3. Dilute 1 in 10 for agarose gel and running buffer concentration	

**Table AP.14 Composition of 10x stock of TBE buffer.**

### xv. TY Media and Agar (2x)

Made up as specified in table AP.15

Reagent	Volume
Bactotryptone	16g
Yeast Extract	10g
NaCl	5g
dH <sub>2</sub> O	to final volume 1 litre
Add 15g of agar for 2xTY agar plates	

**Table AP.15 Composition of 2xTY Media**

## Appendix II DNA Quantification

The following formulae were used to calculate DNA concentrations

**dsDNA**      1OD unit at 260nm = 50µg/ml

$$\mu\text{g}/\mu\text{l} = \frac{(\text{OD } 260) \times (40) \times (\text{Dilution factor})}{1000}$$

**ssDNA (i.e. oligonucleotides)**      1OD unit at 260nm = 40µg/ml

$$\mu\text{g}/\mu\text{l} = \frac{(\text{OD } 260) \times (40) \times (\text{Dilution factor})}{1000}$$



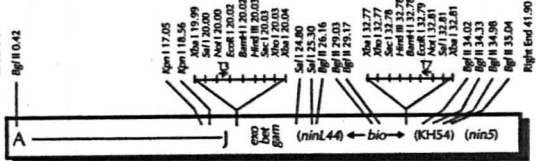


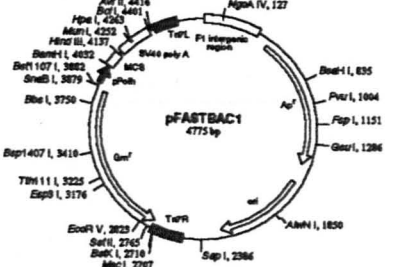
Vector	Type	Purpose	Vector Map (derived from manufacturers literature)
Lambda DASH II (Stratagene, La Jolla, US)	Bacteriophage Lamda replacement vector for cloning large fragments of genomic DNA	Genomic Library Construction	
pCRII-TOPO® (Invitrogen, Carlsbad, US)	Rapid Cloning of PCR products with 3' deoxythymidine (T) overhangs (for example those derived from Taq-based PCR)	PCR product cloning (table 2.1)	
pCR-Blunt II-TOPO (Invitrogen, Carlsbad, US)	Rapid cloning of blunt ended PCR products (such as those derived from Pfu-based PCR)	PCR product cloning (table 2.1)	
pFastBac (Gibco-BRL, Paisley UK)	Expression vector for baculovirus/insect cell transfection system	Expression of truncated llama RAG-1 protein (table 2.1 and section 2.7)	

Table AP.16 Table summarising characteristics of cloning vectors used in this thesis

Table AP.16b Origin of Bacteriophage Clones Derived from Library Screening

	Target Gene					
	Variable Gene Segment		Joining Gene Segment		Recombination-Activating Gene	
	Stratagene Library	In-house Library	Stratagene Library	In-house Library	Stratagene Library	In-house library
Round 1	72	19	3	1	1	1
Round 2	28	8	2	0	1	1
Round 3	5	3	1	0	1	0

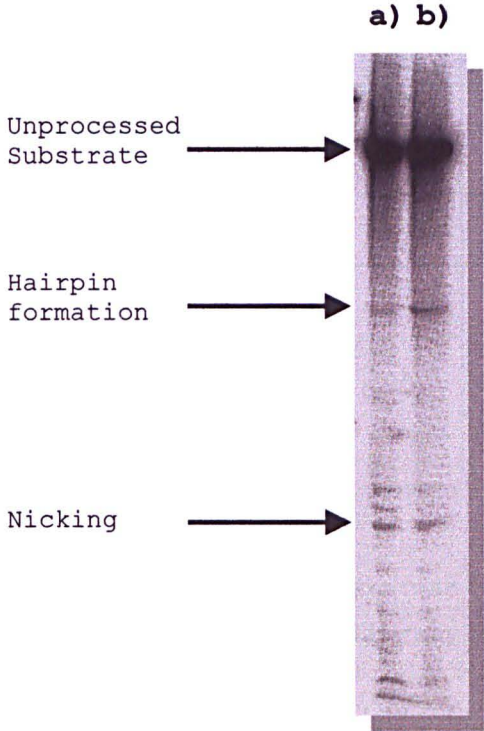
non-duplicated and sequencable

Note: Stratagene Library Titre was  $\sim 5 \times 10^6$   
 In-house constructed library was  $\sim 1.4 \times 10^6$





**Appendix V Comparison of Llama and Murine RAG Protein Activities.**



**Figure AP.2**  
A second lower resolution gel demonstrating the similar activities of murine RAG-1/-2 (lane a) and llama RAG-1/murine RAG-2 proteins (lane b) on a murine control 12-RSS in the presence of Mn<sup>2+</sup>.

## Appendix VI Origin of Bacteriophage Clones Derived from Library Screening

	Target Gene							
	Variable Gene Segment		Joining Gene Segment		Constant Region Genes		Recombination-Activating Gene	
	Stratagene Library	In-house Library	Stratagene Library	In-house Library	Stratagene Library	In-house library	Stratagene Library	In-house library
Rnd 1	72	19	3	1	4	3	1	1
Rnd 2	28	8	2	0	3	1	1	1
Rnd 3*	5 (includes V <sub>H</sub> H1, V <sub>H</sub> H2, V <sub>ψ</sub> 2 and V <sub>H</sub> 1)	3 (includes V <sub>H</sub> H3-V <sub>H</sub> H5 and V <sub>ψ</sub> 1)	1 (includes J region genes described in section 3.17)	0	3 (includes IgG1b, IgG2b and IgG2c)	1 (includes IgG1a)	1	0

\*non-duplicated and sequencable

Note:           Stratagene Library Titre was  $\sim 5 \times 10^6$   
                   In-house constructed library was  $\sim 1.4 \times 10^6$

## BRIEF COMMUNICATION

Benjamin P. Woolven · Leon G.J. Frenken  
Paul van der Logt · Peter J. Nicholls

## The structure of the llama heavy chain constant genes reveals a mechanism for heavy-chain antibody formation

Received: 15 April 1999 / Revised: 1 July 1999

**Key words** Llama · splice · CH1 · Antibody · Mutation

Heavy-chain-only antibodies make up a considerable proportion of class G immunoglobulin (IgG) within camelid sera and are characterized by the absence of the first constant heavy domain (CH1), a region encoded by a separate exon within conventional antibody loci. The loss of this domain reduces the length of the antibody and prevents interactions that would normally take place between the heavy and light chains and between variable and constant domains. The analysis of an unrearranged llama genomic DNA library has enabled us to confirm that different llama hinge types are the products of separate constant region genes. The exon/intron arrangement of a llama classical IgG gene has been determined and, in addition, we show that *CHI*-like sequence is present upstream of the hinge exon in two llama IgG isotypes representing the long- and short-hinge heavy-chain-only antibodies. The acceptor splice sites adjacent to both hinge and *CHI* exons are found to adhere to established consensi, whereas the donor splice site flanking the *CHI* exon is mutated in the heavy chain but not classical *CH* genes. We predict that splicing of the CH1 domain sequence to the hinge exon is prohibited as a result of this mutation, leading to the production of antibodies lacking the CH1 domain.

The discovery of functional heavy-chain antibodies lacking not only L chains but also the entire first constant domain in *Camelidae* (Hamers-Casterman et al. 1993) contradicted the notion that all antibodies within the vertebrate phylum have a common fundamental structure, consisting invariably of two identical H and two identical L chains. The variable domains of heavy-chain-only antibodies contain a number of consistent amino acid substitutions that cause not only structural differences, but also provide an alternative method of antibody-antigen binding (Davies and Riechmann 1996; Desmyter et al. 1996; Spinelli et al. 1996).

It is unclear how the different domains of the heavy-chain antibody are encoded in the germline. Variable domain sequences from both conventional and heavy-chain camelid antibodies have similar 3' recombination signal sequences (RSS), and both conform to the human consensus (Nguyen et al. 1998), suggesting a shared pathway of *V(D)J* recombination. A full understanding of the heavy-chain-only antibody formation mechanism must await sequence data from regions flanking *D* and *J* gene segments, but work thus far supports a model in which all products of *V(D)J* recombination are indiscriminately spliced onto the various constant region genes.

The absence of the CH1 domain is crucial to the overall structure of the heavy-chain antibody. No mechanism has been proposed to explain the manner in which a *CH* gene may be transcribed and expressed while lacking this region. cDNA sequences derived from llama heavy-chain-only clones (unpublished data) indicate that unique hinge, *CH2*, and *CH3* sequences must be present within the germline. Possible explanations for the absence of this domain may include the presence of a *CH* gene within the llama genome containing a partially or completely deleted *CHI* exon or the specific removal of the *CHI* exon during the mRNA splicing event (Vu et al. 1997).

A genomic library was constructed from llama (*Lama glama*) testicular DNA using the bacteriophage  $\lambda$ DASH II cloning vector (Stratagene Ltd., Cambridge,

L.G.J. Frenken · P. van der Logt  
Biorecognition Unit, Unilever Research, Colworth House,  
Sharnbrook, Bedfordshire, MK44 1LQ, UK

P.J. Nicholls  
Research School of Biosciences, University of Kent at  
Canterbury, Canterbury, Kent, CT2 7NJ, UK

B.P. Woolven (✉)  
Biorecognition Unit, Unilever Research, Colworth House,  
Sharnbrook, Bedfordshire, MK44 1LQ, UK  
e-mail: ben.p.woolven@unilever.com,  
Tel.: +44-1234-248086,  
Fax: +44-1234-222552

UK). This library was screened using a radioactive probe specific to the llama *CH2* sequence (unpublished data). Three clones were isolated and sequenced with a range of primers specific to *CH2*, hinge, and intronic regions. The llama classical, long-hinge heavy-chain-only and short-hinge heavy-chain-only genes are represented by these clones, referred to as B1, B2, and B3 respectively.

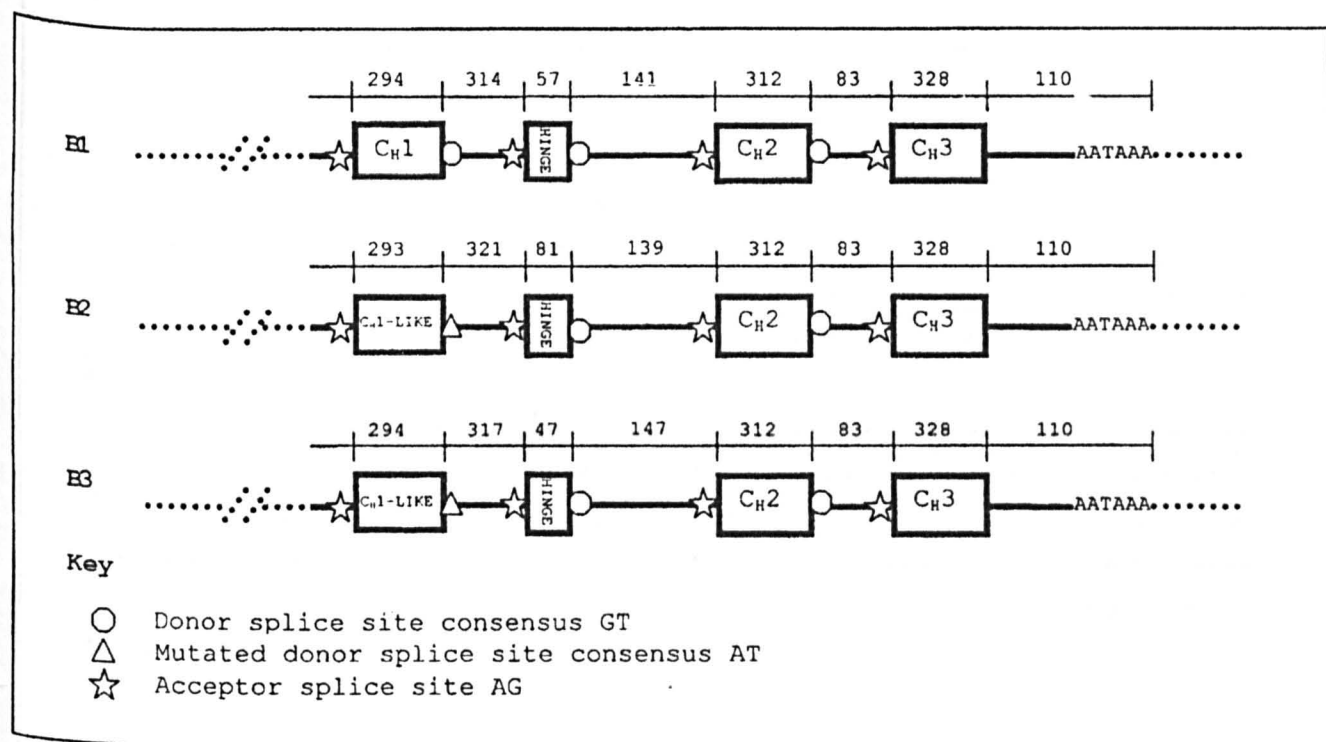
Differences in both hinge and *CH2* sequence have enabled identification of the three antibody types and confirm that each clone represents a separate gene and, consequently, isotype. The existence of separate isotypes would suggest that the heavy-chain-only antibodies arose either from duplication and subsequent mutation of a complete classical *CH* gene, or independently, through duplications of a primordial single exon gene similar to the human or mouse light chain constant exon (*CL*). This mechanism has previously been suggested to explain the evolution of other *CH* genes (Honjo and Matsuda 1995).

Clones B1-3 include not only the *CH2*, *CH3*, and hinge exons of their respective genes but also sequence representative of *CH1* and *CH1*-like exons (Fig. 1). The extent of the *CH1*-like exons and therefore position of splice site consensi was ascertained by comparison with the llama classical *IgG* cDNA sequence (unpublished data) and by analysis of the sequence by splice site prediction software (Brunak 1991). The apparent similarity

between the *CH1* domain of the classical *IgG* (B1) and that found in clones B2 and B3 suggest that a double duplication of a llama heavy chain gene may have been a relatively recent evolutionary event. Comparison of exon sequences further supports this possibility. The three *CH2* exons show 88-98% identity, while the *CH3* regions are 93-96% identical. The *CH1*-like region of B2 contains a stop codon (due to an A/T substitution at position 51) and a frame-shift mutation (position 124) preventing translation of a functional *CH1* domain (Fig. 2). By contrast B3 contains no stop codons or frame-shifts, perhaps indicating that the long-hinge heavy-chain-only antibody evolved as a result of duplication and mutation of the short-hinge heavy-chain-only variant.

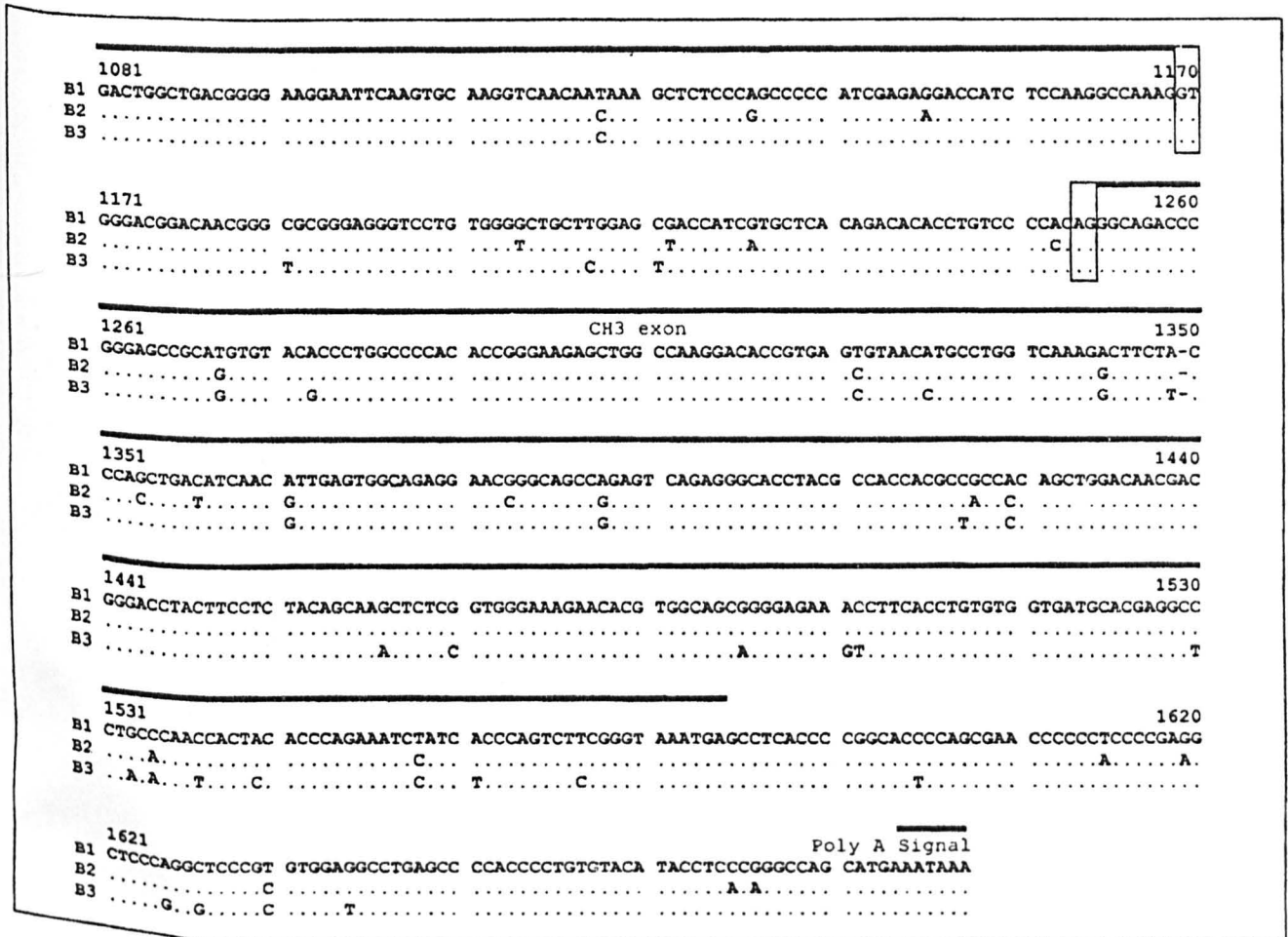
The 3' flanking region of the *CH1* exon of all human and mouse *IgG* constant genes is characterized by the presence of a donor splice site immediately adjacent to the coding sequence. The GT dinucleotide splice consensus found in all known functional *IgG* germline sequences is present in the classical llama clone sequence (B1) but not in either heavy-chain clones (B2, B3) in which the GT motif is replaced with AT. Mutation of this dinucleotide at other mammalian exon/intron boundaries has been shown to inactivate 3' cleavage and exon joining during *in vitro* splicing (Aebi et al. 1987). It is probable, therefore, that splicing of the *CH1* exon to the hinge exon is inhibited *in vivo*. We propose that donor splice sites adjacent to the rearranged *V(D)J* subunits show preference for acceptor splice sites flanking the hinge exon above those flanking the *CH1*-like exon. This would explain the absence of the *CH1* domain in the expressed protein. *In vitro*

Fig. 1 Diagram to illustrate exon/intron arrangement of clones B1-3. Scale shows length of exons and introns in base pairs. Dashed lines indicate unsequenced flanking regions









**Fig. 2** Nucleotide sequence of classical (B1, GenBank accession number AF132603), long-hinge heavy-chain-only (B2, AF132604) and short-hinge heavy-chain-only (B3, AF132605) germline llama constant region genes. *Boxes* indicate dinucleotide splice site consensus and mutations. *Bold lines* represent exons, exon-like regions, and polyadenylation site. A *dot* indicates identity to classical sequence. A *dash* indicates a gap introduced to improve alignment.

splicing experiments coupled with full characterization of the llama germline *IgG* loci should enable conformation of this mechanism.

**Acknowledgments** We would like to thank Linda Johnson at the Ashdown Llama Farm and Andrew Starnes for the generous gift of llama testicular material. This work was supported by the Biotechnology and Biological Sciences Research Council, Unilever Research Plc, and The Royal Society of Great Britain. We, the authors of the above manuscript, declare that the experiments described within are in compliance with the current laws of the country in which they were performed.

## References

Aebi M, Hornig H, Weissmann C (1987) 5' cleavage site in eukaryotic pre-mRNA splicing is determined by the overall 5' splice region, not by the conserved 5' GU. *Cell* 50:237-246

- Brunak S, Engelbrecht J, Knudsen S (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J Mol Biol* 220:49-65
- Davies J, Riechmann L (1996) Affinity improvement of single antibody VH domains: residues in all three hypervariable regions affect antigen binding. *Immunotechnology* 2:169-179
- Desmyter A, Transue TR, Ghahroudi MA, Thi MH, Poortmans F, Hamers R, Muyldermans S, Wyns L (1996) Crystal structure of a camel single-domain VH antibody fragment in complex with lysozyme. *Nat Struct Biol* 3:803-811
- Hamers-Casterman C, Atarhouch T, Muyldermans S, Robinson G, Hamers C, Songa EB, Bendahman N, Hamers R (1993) Naturally occurring antibodies devoid of light chains. *Nature* 363:446-448
- Honjo T, Matsuda F (1995) Immunoglobulin heavy chain loci of mouse and human. In: Honjo T, Alt FW (eds) *Immunoglobulin genes*. Academic Press, London, pp145-171
- Nguyen VK, Muyldermans S, Hamers R (1998) The specific variable domain of camel heavy-chain antibodies is encoded in the germline. *J Mol Biol* 275:413-418
- Spinelli S, Frenken L, Bourgeois D, de Ron L, Bos W, Verrips T, Anguille C, Cambillau C, Tegoni M (1996) The crystal structure of a llama heavy chain variable domain. *Nat Struct Biol* 3:752-757
- Vu BK, Ghahroudi MA, Wyns L, Muyldermans S (1997) Comparison of Llama  $V_H$  sequences from conventional and heavy chain antibodies. *Mol Immunol* 34:1121-31