SEQUENCE ASSEMBLY AND ANNOTATION OF THE BOVINE MAJOR HISTOCOMPATIBILITY COMPLEX (BOLA) CLASS IIB REGION, AND *IN SILICO* DETECTION OF SEQUENCE POLYMORPHISMS IN BOLA IIB

A Dissertation

by

CHRISTOPHER P. CHILDERS

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2006

Major Subject: Genetics

SEQUENCE ASSEMBLY AND ANNOTATION OF THE BOVINE MAJOR HISTOCOMPATIBILITY COMPLEX (BOLA) CLASS IIB REGION, AND *IN SILICO* DETECTION OF SEQUENCE POLYMORPHISMS IN BOLA IIB

A Dissertation

by

CHRISTOPHER P. CHILDERS

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Loren C. Skow
Committee Members,	James E. Womack
	Bhanu P. Chowdhary
	David L. Adelson
Chair of Genetics Faculty,	James R. Wild

December 2006

Major Subject: Genetics

ABSTRACT

Sequence Assembly and Annotation of the Bovine Major Histocompatibility Complex (BoLA) Class IIb Region, and *in silico* Detection of Sequence Polymorphisms in BoLA IIb. (December 2006)

Christopher P. Childers, B.S.; B.S., West Texas A&M University

Chair of Advisory Committee: Dr. Loren Skow

Cattle are vitally important to American agriculture industry, generating over 24.6 billion pounds of beef (by carcass weight), and 79.5 billion dollars in 2005, and over 27 billion dollars in milk sales in 2004. As of July 2006, the U.S. beef and dairy industry is comprised of 104.5 million head of cattle, 32.4 million of which were processed in 2005. The health of the animals has always been an important concern for breeders, as healthy animals grow faster and are more likely to reach market weight. Animals that exhibit natural resistance to disease do not require chemicals to stimulate normal weight gain, and are less prone to disease related wasting.

The major histocompatibility complex (MHC) is a collection of genes, many of which function in antigen processing and presentation. The bovine MHC (*BoLA)* differs from typical mammalian MHCs in that the class II region was disrupted by a chromosomal inversion into two subregions, designated *BoLA IIa* and *BoLA IIb. BoLA IIb* was transposed to a position near the centromere on bovine chromosome 23,while *BoLA IIa* retains its position in *BoLA*. Comparative sequence analysis of *BoLA IIb* with the human MHC revealed the location of the region containing the proximal inversion

breakpoint. Gene content, order and orientation of *BoLA IIb* are consistent with the single inversion hypothesis when compared to the corresponding region of the human *class II* MHC (*HLA class II*). *BoLA IIb* spans approximately 450 kb.

The genomic sequence of BoLA IIb was used to detect sequence variation through comparison to other bovine sequences, including data from the bovine genome project, and two regions in the BAC scaffold used to develop the *BoLA IIb* sequence. Analysis of the bovine genome project sequence revealed a total of 10,408 mismatching bases, 30 out of 231 polymorphic microsatellites, and 15 sequences corresponding to the validated SNP panel generated by the bovine genome sequencing project. The two overlapping regions in the *BoLA IIb* BAC scaffold were found to have 888 polymorphisms, including a total of 6 out of 42 polymorphic microsatellites indicating that each BAC derived from a different chromosome.

DEDICATION

I would like to dedicate this dissertation to my family.

ACKNOWLEDGMENTS

I would like to thank my friends and family for their constant support and encouragement. My parents Gene and Renee have always been a source of encouragement and support, especially during times when I was discouraged, and I could always count brother Gene to keep my spirits up and offer sage advise on my problems.

I would also like to thank my advisor, Dr. Loren Skow, for his patience, advice, and especially his ability to examine problems from different perspectives. I would also like to thank my committee members: Drs. David Adelson, Bhanu Chowdhary and James Womack, who provided me with insight, advice and the tools to advance my graduate education.

In addition to my committee, I would like to thank my lab mates, Deirdre Honeycutt, Leonardo Sena, and Kristi Young, who introduced me to the inner workings of the lab, and helped me in the early stages of my project. Later additions to the lab, including Krista Fritz, Gayle Linger and Jay Burzlaff also provided both technical assistance and proved to be good friends. Dr. Chowdhary's lab has always worked very closely with ours, and I would also like to thank some of the lab members there, including Glenda Goh, Candice Brinkmeyer-Langford, Terje Raudsepp, Eun-Joon Lee, Deeann Wallis, and Avni Santani.

This research was funded by USDA NRI grant 97-35205-5074 to Loren C. Skow and by funds from College of Veterinary Medicine, Texas Agricultural Experiment Station, The Institute for Bioscience Technology, and the Vice President for Research, Texas A&M University.

TABLE OF CONTENTS

Page

ABSTRACT iii

DEDICATION v

ACKNOWLEDGMENTS vi

TABLE OF CONTENTS vii

LIST OF FIGURES ix

LIST OF TABLES x

CHAPTER

I INTRODUCTION 1

History of the Major Histocompatibility Complex 1
Organization of the MHC 2
The MHC and Disease 10
Comparative Organization of the MHC 12
MHC Evolution 15
Objectives of This Study 16

II COMPARATIVE ANALYSIS OF THE BOVINE MHC CLASS IIB SEQUENCE IDENTIFIES INVERSION BREAKPOINTS AND THREE UNEXPECTED GENES 18

Overview 18
Introduction 19
Materials and Methods 20
Results 25
Discussion 33

III SEQUENCE VARIATION IN BOLA IIB 40

Overview 40
Introduction 40
Materials and Methods 42
Results 48
Discussion 54

CHAPTER Page

IV TAMU DERIVED BOLA IIB VERSUS BOVINE GENOME PROJECT BOLA IIB 57

Introduction 57
Materials and Methods 57
Results and Discussion 60

V SUMMARY AND CONCLUSION 62

LITERATURE CITED 68

APPENDIX A 83

APPENDIX B 90

APPENDIX C 93

APPENDIX D 94

VITA 101

LIST OF FIGURES

FIGURE Page

1 A diagrammatic representation of the antigen presenting MHC molecules 6

2 Characterization of the *BoLA IIb* sequence 27

3 Comparative analysis of *BoLA IIb* and *HLA class II* 32

4 Inversion breakpoint localized to a 2.4 kb region between *GCLC* and *DSB* 34

5 Overview of the BoLA IIb region.....49

LIST OF TABLES

TABLE Page

1 Comparisons of coding sequences and encoded proteins for *HLA II* and *BoLA IIb* genes 29

2 Repetitive interspersed elements present in *BoLA* IIb 31

3 Global alignments between overlapping *BoLA* IIb sequences, regions are distinguished by shading 50

4 Summary of simple sequence repeat polymorphisms between overlapping *BoLA IIb* sequences 52

5 Mapping validated SNP sequences to *BoLA IIb* 54

CHAPTER I

INTRODUCTION

History of the Major Histocompatibility Complex

The MHC was first discovered and defined through tumor graft rejection studies in mice (Little and Tyzzer 1916). If the donor mouse shared certain alleles with the recipient, the mouse would be tolerant to the transplanted tumor. If the donor and recipient mouse differed in those alleles, the tumor would be destroyed. This was the first study that found a genetic link with graft rejection. Several additional studies by Gorer confirmed and expanded the model of genetic and antigenic basis of graft rejection in mice (Gorer 1937). The presence of multiple loci that affected graft rejection increased the difficulty in studying the effects of any particular locus to the graft rejection system. During this period of time George Snell was also investigating transplant rejection in mice, simplifying the problem of studying H (histocompatibility) locus by developing a highly inbred strain of mice that was genetically identical with the exception of the H locus in question. This congenic line of mice exhibited an H locus that cosegregated with an easily determined phenotype designated *fused* (Snell 1948).

The human MHC (HLA) was first identified in 1958 (Dausett 1958), and was quickly followed by studies into leukocyte reactions in pregnant women (Payne and Rolfs 1958; van Rood *et al.* 1958). At this time, research into the HLA was restricted to studying leukocyte agglutination patterns in order to define different allotypes. In the

---

This dissertation follows the style and format of *Animal Genetics*.

following years, other techniques such as gel electrophoresis, restriction fragment length polymorphisms (RFLP), polymerase chain reaction (PCR), and DNA sequencing were developed and refined. These technologies have been used to characterize of HLA, resulting in the annotated genomic sequence in 1999 (MHC Sequencing Consortium, 1999), a greater understanding of the variation present in HLA, and the definition of haplotype structure (Dawkins *et al.* 1999; Daly *et al.* 2001; Smith *et al.* 2006). Sequencing projects in a number of other species have resulted in partial or complete MHC sequences for a variety of animals, including: mouse (Takada *et al.* 2003; Xie *et al.* 2003), rat (Hurt *et al.* 2004), dog (Debenham *et al.* 2005), cat (Yuhki *et al.* 2003), swine (Renard *et al.* 2006), gray short tailed opossum (Belov *et al.* 2006), medaka (Matsuo *et al.* 2002), shark, (Ohta *et al.* 2000; Terado *et al.* 2003), chicken, and quail (Shiina *et al.* 2004).

Organization of the MHC

The MHC contains over 260 genes, many of which have some function in the immune response. In general, the MHC of most eutherian mammals is comprised of three discrete regions (MHC Sequencing Consortium 1999; Gustafson *et al.* 2003; Hurt *et al.* 2004). These regions have been divided by functional gene content into three subregions designated class I, class II, and class III. The Class I genes are expressed on all nucleated cell types and encode glycoproteins that display processed intracellular peptides at the cell surface. The class II MHC genes present exogenously derived antigens on the cell surface to CD4+ T cells, and are expressed on B cells, macrophages, and dendritic cells. The class III MHC region is defined by its location between the class I and class II

regions, and has been characterized as the most gene dense region in the human genome (Xie *et al.* 2003). The class I and class II regions were initially defined by the presence of characteristic genes within each region, while the class III region was defined as the region between the class I and class II regions. In the regions flanking the boundaries of the class I and class II regions lies what is referred to as the extended MHC. In the human MHC the extended class I region begins with the gene encoding myelin oligodendrocyte glycoprotein (*MOG*), and extends through the histone gene family, the most distal of which is *HIST1H2AA*. The class II extended region begins with the collagen gene *COL11A2* and continues through RPL12P1 (ribosomal protein L12, pseudogene1) (Horton *et al.* 2004). In the case of the gray short tailed opossum, the class I genes are interspersed through the class II region, though the extended MHC regions in marsupials appear to be conserved with that of HLA (Belov *et al.* 2006).

The MHC also contains genes that are associated the more generalized immune response. The immune system is characterized as including two main types of response: a direct and specific response to a pathogen, known as the innate response, and a less specific but more flexible system known as the adaptive response. The MHC contains genes that function in both the innate and the acquired immune response.

The innate immune response describes the portion of the immune system that act as the initial defense against infection. Direct opposition to pathogens characterizes the innate immune system, by recognizing a few molecules that are common to many pathogens. Phagocytic cells such as macrophages and neutrophils display cell surface receptors that recognize characteristic molecules associated with pathogens, but not host cells. Once the receptors come into contact with cells exhibiting such molecules, the

receptors bind to the molecule, and the phagocyte initiates the process of engulfing and destroying the foreign cell. These types of direct responses result in very fast reaction to foreign bodies.

The complement system is one of the primary systems for recognition and removal of foreign bodies and functions to either tag pathogens for phagocytosis, or directly destroys foreign cells by perforating the cell membrane. The varied responses that the complement system can mediate show some of the flexibility of the innate immune response. The complement system consists of several components, which are involved in three different types of immune response. Three pathways have been defined to describe the complement immune response: the classical pathway, the lectin mediated pathway, and the alternative pathway. Both the classical and lectin pathways function to tag pathogens for destruction via phagocytosis, while the alternative pathway results in the lysing of pathogens directly. Four of the genes involved in the complement response are located in the class III region: *C4A*, *C4B*, *BF* and *C2* (The MHC Sequencing Consortium 1999).

One of the early responses of the immune system to a pathogen is a nonspecific reaction known as inflammation. The inflammatory response results in redness, swelling, and heat, and is also characterized by an increase in the numbers of leukocytes in the affected tissue. Tumor necrosis factor alpha (*TNFA*) is a membrane bound cytokine that functions in the inflammatory response. In addition to this, TNFA has been implicated in tumor formation (Coussens *et al.* 2002; Rao *et al.* 2006). TNFA has also been connected to predisposition to lupus erythmatosis (Jacob *et al.* 1990). Variability in the TNFA

promoter region, resulting in changes in the expression level of TNFA has also been linked to susceptibility to several diseases (Waterer and Wunderink 2003).

While the innate immune response functions as a direct response to pathogens, the adaptive immune response is characterized by a multitude of different receptors, each of which recognizes very specific molecules. The number of different receptors results in a very specific response to a broad range of pathogens. The molecules recognized by these receptors are referred to as antigens. Antigens may be derived from pathogens or the host. The adaptive immune system is designed to differentiate between antigens derived from self and non-self sources, and react accordingly. The function of the MHC molecules is to display peptide sequences on the cell surface. T cells recognize the MHC molecules, and examine the peptide bound in the antigen binding groove. If the peptide is not recognized as a part of the normal complement of the host organism, an immune response is initiated.

Class I

Class I MHC molecules are expressed on the surface of virtually all nucleated cells. Class I genes operate as functional pairs, with alpha and beta subunits forming a functional dimer (Figure 1). The alpha subunit of the class I gene contains three functional domains. Domains one and two form the antigen-binding groove and are encoded by exons two and three. Domain three contains the transmembrane domain, and is encoded by exon five, and exons six and seven encode the cytoplasmic tail. The beta subunit of the class I genes is a β2 microglobulin, which is encoded by a gene not located in the MHC. There are three subcategories of class I genes, Ia, Ib, and Ic (Hughes *et al.*

1999). The class Ia genes are also known as the classical class I genes. The classical class I genes encode for the proteins that display processed intracellular proteins to the CD8+ T cells. These genes are expressed on all nucleated cells, and display a very high level of polymorphism in the antigen binding domain. The nonclassical class I genes are also

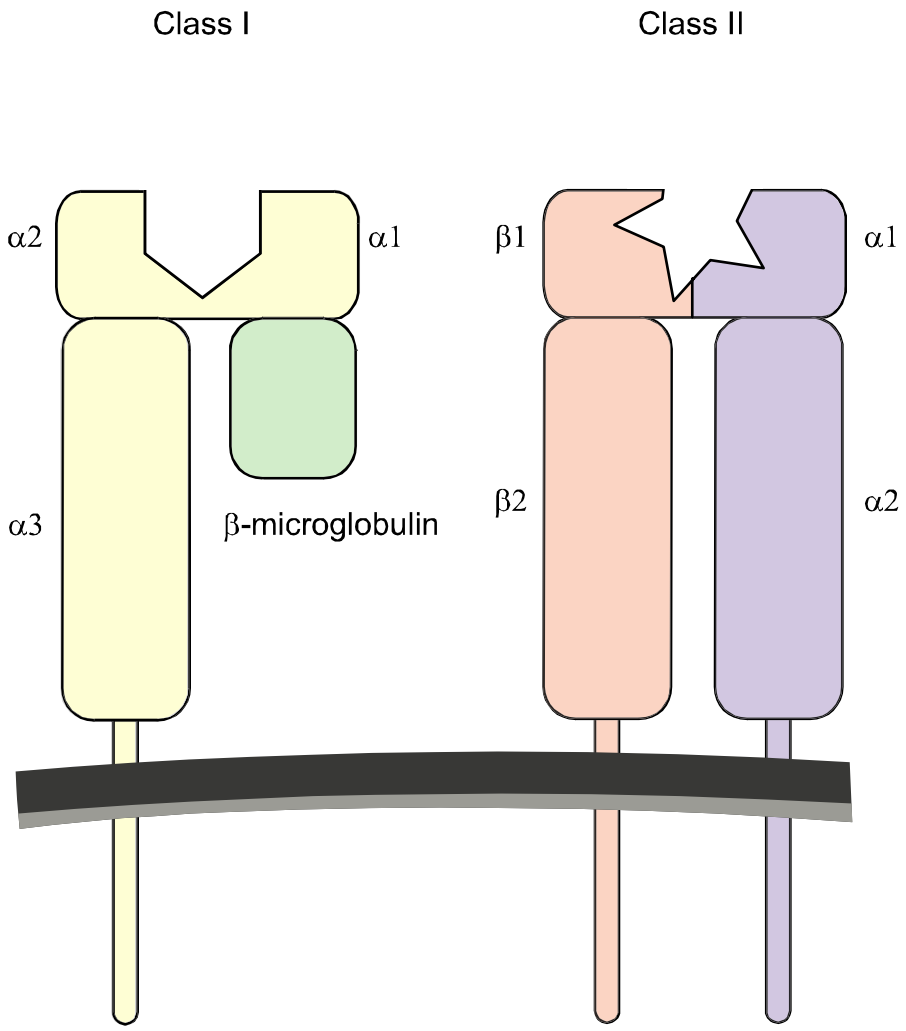


Figure 1 A diagrammatic representation of the antigen presenting MHC molecules.

referred to as class Ib and Ic genes. These genes are expressed on restricted subsets of cells, and are less polymorphic in the exons coding for antigen binding sites. The class Ic genes include *MICA*, *MICB*, and *Hfe*, and differ from the class Ib genes in that they appear to have diverged before the placental mammal radiation (Hughes *et al.* 1999). The numbers of class Ia and Ib genes varies from species to species, with 30 functional genes in mouse (Kumánovics *et al.* 2002), six functional genes in human (Hughes *et al.* 1999), and seven functional class I genes in the opossum (Belov *et al.* 2006). The class I genes are not orthologous between species (Hughes *et al.* 1989), suggesting that these genes have arisen through duplication, and subsequent duplication, deletion, and reorganization events occurred after speciation, resulting in differing sets of functional and nonfunctional class I genes in each lineage. The model developed to explain this system is named "Birth and Death" evolution (Nei and Rooney 2005).

The majority of genes present in the class I region are not MHC genes. These genes occur in conserved blocks interspersed throughout the class I region and are known as "class I framework genes". There are three main blocks of framework genes, and the regions between the blocks allow for significant reorganization events including duplication and deletion of class I genes (Amadou 1999). Framework blocks are conserved among mammalian species (Amadou 1999) including the marsupial MHC with the framework region in a single uninterrupted block (Belov *et al.* 2006).

The antigens presented by class I genes are processed into small peptides by the proteosome, a macromolecule derived from 24 subunits. Two of the subunits, proteosome subunit, beta type 8 (*PSMB8*) and proteosome subunit, beta type 9 (*PSMB9*), are located in the class II region. The processed antigens are then transferred into the endoplasmic

reticulum by the TAP complex (Karttunen *et al.* 2001), which is encoded by two genes in the class II region: transporter 1, ATP-binding cassette, sub-family B (*TAP1*) and transporter 2, ATP-binding cassette, sub-family B (*TAP2*). Following successful incorporation of antigens into the antigen binding domain, the class I proteins migrate to the cell surface for presentation to CD8+ T cells. The peptides bound to the class I antigen binding groove range from eight to ten amino acid residues in length.

Class II

The class II genes are functional as heterodimers consisting of an alpha and a beta subunit that functions to present extracellular derived antigens to CD4+ T cells (Figure 1). Unlike the class I genes, the class II gene subunits each encode for a portion of the antigen binding domain. The functional class II protein contains two transmembrane domains, whereas the class I molecule contains a single transmembrane domain, connected to the class I alpha subunit. The peptides bound to class II antigen binding grooves are generally 13 – 25 amino acid residues long, and may be longer.

The MHC class II region contains at least five groups of orthologous genes that are shared across different mammalian species: *DQ*-, *DR*, *DP*-, *DM*-, and *DO*-. In addition to these genes, the pecoran ruminants are known to contain an additional class II gene family designated *DY*- (Stone and Muggli-Cockett 1990), and the marsupial MHC contains three marsupial specific class II gene families: *DA*-, *DB*-, and *DC*- (Belov *et al.* 2006). The *DQ*-, *DR*-, and *DP*- gene families perform the primary function of presenting peptides from exogenous sources such as bacteria or parasites to CD4+ T cells. The *DP*- genes appear to be functional only in the Euarchontoglires mammals, though fragments

of these genes are still present in the genomes of other mammals (Kelley *et al.* 2005). The *DY*- genes present antigens to dendrocytes in afferent lymph, indicating that they are functional class II genes (Ballingall *et al.* 2004). The marsupial specific class II genes have been found in multiple species (Belov *et al.* 2004; Belov *et al.* 2006) and appear to differ from the eutherian class II genes. The class II genes most closely related to *DAA* and *DAB* are *DRA* and *DRB*, respectively, while DBA is most closely related to *DOA*. *DBB* does not show a clear relationship to any specific class II gene (Belov *et al.* 2006).

After the class II gene product is synthesized and aggregates as a dimer in the endoplasmic reticulum, it is bound to a third peptide known as the MHC class II invariant, covering the antigen binding site. The newly formed trimer then migrates to a vesicle designated the class II compartment, where the invariant chain is degraded to a small fragment called the class II-associated invariant-chain peptide (CLIP) that covers the antigen binding domain. The dimer synthesized from the DM genes facilitates the release of CLIP, freeing up the antigen binding site to acquire processed antigens. (Benaroch *et al.* 1995) The *DO*- genes *DOA*/*DOB* form a functional heterodimer that inhibits the activity of the *DM* heterodimer. This functions to inhibit the binding of antigens to the class II molecules (Fallas *et al.* 2004).

As in the class I region, the class II region also includes genes that do not encode receptors for antigen presentation. These include the class I antigen transport and processing genes *TAP1*, *TAP2*, *PSMB8*, and *PSMB9*, as well as the *DM*- and *DO*- genes. There are also genes in the class II region that have no obvious link to the immune response, including bromodomain containing 2 (*BRD2*), (Beck *et al.* 1992), and retinoid receptor x, beta (*RXRB*) (Fitzgibbon *et al.* 1993). These genes are found in MHCs across

the entire mammalian lineage (Horton *et al.* 2004; Debenham *et al.* 2005; Belov *et al.* 2006).

Class III

Gene content and organization of the class III region is conserved among mammalian species, and is among the most gene dense regions in the genome, containing 60 genes in human and 61 genes in mouse (Xie *et al.* 2003; Belov *et al.* 2006). The boundaries of the class III region are demarcated by the genes notch homolog 4 (*NOTCH4*) and HLA-B associated transcript 1 (*BAT1*) on the class II/III and class III/I boundaries, respectively (Xie *et al.* 2003). The class III region does not contain any antigen presenting genes, although there are genes that function in other aspects of the immune response. The complement components 2, 4, and BF (*C2*, *C4*, and *BF*, respectively) are subunits of the complement system, while tumor necrosis factor alpha (*TNFA*), lymphotoxin alpha (*LTA*), and lymphotoxin beta (*LTB*) function in the inflammatory response. In addition to these genes are several genes that have no apparent connection to the immune response. Some examples are *BAT1* (HLA-B associated transcript 1), *APOM* (apolipoprotein M), and *VARS1* (valyl-tRNA synthetase 1) (Horton *et al.* 2004). The class III region is also interesting in that it contains no pseudogenes (Xie *et al.* 2003; Horton *et al.* 2004).

The MHC and Disease

Though the MHC was first discovered through experimentation into graft rejection, the biological function of the MHC is in the immune response. The human

MHC best characterized, and many diseases have been associated with MHC genes, including: autoimmune diseases, type 1 and 2 diabetes, psoriasis, and Behcet disease, arthritis, and narcolepsy (Bodmer 1987; Hirohata and Kikuchi 2003). The non antigen presenting genes have also been associated with various diseases. Mutation in the extended class II collagen gene *COL11A2* can result in deafness (Mcguirt *et al.* 1999), while the extended class I gene *MOG* (myelin oligodendrocyte glycoprotein) is associated with multiple sclerosis (Iglesias *et al.* 2001). Variant forms of any of the TAP genes, *TAP1* (Furukawa *et al.* 1999), *TAP2* (de la Salle *et al.* 1994), or *TAPBP* (Yabe *et al.* 2002) may result in a deficiency in the expression of class I molecules on the cell surface, known as "bare lymphocyte syndrome". *BRD2* has been associated with juvenile myoclonic epilepsy (Pal *et al.* 2003). *PSMB8* and *PSMB9* have been associated with diabetes mellitus (Deng *et al.* 1995). While the class III region was initially defined as the region between the class I and II regions, there are diseases associated with various class III genes. Nonsense mutations in cytochrome P450, family 21, subfamily A, polypeptide 2 (*CYP21A2*) result in steroid 21-hydroxylase deficiency, causing congenital adrenal hyperplasia (Globerman *et al.* 1988). *NFKBIL1* (nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor-like 1) has been associated with rheumatoid arthritis (Okamoto *et al.* 2003), and tenascin XB (*TNXB*) has been linked to Ehlers-Danlos syndrome (Burch *et al.* 1997). Variant forms of LTA are associated with increased susceptibility to myocardial infarction (Ozaki *et al.* 2002). The complement genes C2, C4, and BF are also present in the class II region. While mutations in these genes would reduce or eliminate the complement response, the genes have been associated with other disorders. Variants in C4A have been associated with lupus erythematosus (Huang *et al.*

1995). Mutations in C2 and BF have been associated with age related blindness caused by macular degeneration (Gold *et al.* 2006).

The economic significance of the cattle industry has provided a strong incentive for research into disease and disease resistance in cattle, with some focus on the MHC. In cattle, *DRB3* has been associated with resistance to foot and mouth disease virus (Haghparast *et al.* 2000). The progression of bovine leukemia virus infection to bovine leukemia or lymphosarcoma is dependant on alleles containing variations in exon 2 of *BoLA-DRB3* (Xu *et al.* 1993). More often, disease resistance or susceptibility has not been attributed to a specific gene or allele. For example, bovine mastitis is one of the most significant diseases in dairy cattle, causing inflammation of the udder, with symptoms ranging from loss of milk to death, and an estimated cost of over $2 billion dollars to the dairy industry in the U.S. (Harmon 1994). Mastitis susceptibility was linked to haplotypes that include the class II *DQ*- genes (Mallard *et al.* 1998; Park *et al.* 2004). Often association studies are unable to further refine a linked region to a specific gene or mutation due to the high levels of polymorphism and multigene families that are present in the MHC.

Comparative Organization of the MHC

The MHC has been observed in all vertebrates ranging from cartilaginous fishes through birds and mammals. Recent research discovered a genomic region corresponding to the MHC in Amphioxus (Abi-Rached *et al.* 2002) which appears to be a “proto-MHC” region. Genes of the MHC are syntenic in mammals, birds, reptiles and cartilaginous

fishes, but not in osteichthyes (Kuroda *et al.* 2002), indicating that the non-syntenic organization is a derived rather than ancestral state.

The human MHC (HLA) is the most intensively studied MHC, and is considered to be the standard mammalian organization, consisting of a single cluster each for the class I, II and III regions. HLA consists of more than 260 genes, of which 128 are expressed, spanning 3.6 Mb on chromosome 6p21.3 (The MHC Sequencing Consortium 1999; Beck and Trowsdale 2000). There are functional pairs of *DR*, *DR*, and *DP* genes in the class II region. HLA class II also has numerous pseudogenes for each of the class II MHC genes.

The cow MHC differs from other characterized MHCs in that the class II region was split by a large chromosomal inversion (Andersson *et al.* 1988; Band *et al.* 1998). The portion of the class II region adjacent to the class III region was designated class IIa, and the isolated region was called class IIb. The class IIa region contains the *DRA*/*DRB* and *DQA*/*DQB* class II gene pairs. The class IIb region of BoLA contains the *DSB* gene followed by the *DYA*/*DYB* gene pair, and extending through the remainder of the classical and extended class II regions.

The mouse MHC (H2) resides on chromosome 17, and is similar in general organization to the MHC of human, with the exception of a partial duplication of the class I region that resides on the distal end of the class II region. Regions of H2 have been sequenced and characterized, and the release of the initial draft sequence of the mouse genome has increased the amount of genomic MHC sequence available (Watterson *et al.* 2002; Xie *et al.* 2003; Takada *et al.* 2003; Yuhki *et al.* 2003). The gene nomenclature for H2 and the rat MHC (RT1) differs significantly from that of other

MHCs. The H2 class II region includes a functional pair each of E-, A-, O-, and M-genes (Kumánovics *et al.* 2003). RT1 is the designation for the rat MHC, which resides on chromosome 20. RT1 does include a duplication of class I sequence proximal to the class II class III boundary, though the gene content of the duplicated regions indicates that the duplication occurred after the speciation event that would give rise to the mouse and rat (Hurt *et al.* 2004). RT1 also contains a second Db gene designated RT1-Db2 that is designated as putatively functional (Hurt *et al.* 2004).

The feline MHC (FLA) resides on chromosome B2. FLA lacks functional DQ-genes; however three copies of functional DR- genes are present. A single functional pair of DO- and DM- genes is also present (Yuhki *et al.* 2003). The class I region of FLA contains an inversion, relocating the framework genes proximal to the class I extended region (Beck *et al.* 2005).

The canine MHC (DLA) is located on chromosome 12, and is organized in a manner similar to humans, with a single class I, class II, and class III region (Debenham *et al.* 2005). DLA class II contains one pair each of functional DR-, DQ-, DO-, and DM genes.

The *Monodelphis domestica* (gray short tailed opossum) genome sequencing project represents the generation of genomic sequence for the first non-eutherian mammal, and was carried out by the Broad Institute at M.I.T. (http://www.broad.mit.edu/mammals/opossum/). The sequence spanning the MHC was found to be on a single scaffold in the second assembly, designated MonDom2 (Belov *et al.* 2006). There are three families of functional class II gene families unique to the marsupials, designated DA-, DB, and DC- (Belov *et al.* 2004; Belov *et al.* 2006). The opossum MHC exhibits an

organization very distinctive from eutherian mammals, sharing some organizational features of non-mammalian MHCs. The opossum MHC is comprised of a distinct class III region, a class I/II region that contains both class I and II MHC genes, and a separate region containing the class I associated framework genes.

The Chicken MHC is the most compact MHC described to date, with two regions, the B and RfpY loci separated by the chicken nucleolar organizing region (NOR) (Kaufman *et al.* 1999b). The B locus contains 21 expressed genes, including homologues to 12 of the genes found in HLA (Kaufman *et al.* 1999a), and spans only 92 Kb of DNA in contrast to HLA with 3.2 Mb. This minimal essential MHC is expanded to almost 200 kb in the related quail (Coturnix) (Shiina *et al.* 2004) so it is not clear whether the avian MHC is similar to the ancestral vertebrate MHC or has undergone contraction during evolution.

MHC Evolution

The struggle between pathogens that survive by infecting and compromising an individual and the immune system to thwart potential attackers provides a constant pressure for the immune system to be able to change to recognize new pathogens, and new forms of existing pathogens. Any particular allele of an antigen presenting gene is able to bind to a limited range of antigens (Chen and Parham 1989), and the more diversity in antigen presenting genes results in a greater range of antigens that may be presented. An individual that is able to recognize and respond to a pathogen has an advantage over a susceptible individual upon exposure to the pathogen.

One of the processes of MHC gene evolution is by gene duplication followed by divergence. This phenomenon is used as the primary explanation for the observation that different MHC genes that share many characteristics yet have acquired different functions, as well as the documented copy number polymorphism of certain MHC genes between species (Andersson and Rask, 1988; Yuhki *et al.* 2003). An example of this system in action is the class IIb DY- gene pair, DYA and DYB. These genes show highest similarity to the DQA1 and DQB1 genes in human, and share the same relative position and orientation as human DQA1 and DQB1. However it was not until recently that there was any evidence that these genes were functional. Ballingall found that these genes were expressed in a subset of dendritic cells, a very different pattern of expression than the DQ- genes (Ballingall *et al.* 2004). Another example is the DP- gene pair, DPA and DPB. These genes appear to be functional in humans and some primates, although in other species only pseudogenes or gene fragments remain. The DP- genes appear to have a different expression pattern, and a divergent function.

Objectives of This Study

Cattle are an economically important species, generating over $90 billion in revenue annually in the U.S. alone. Developing a foundation of knowledge about the bovine genome would help to promote further studies into factors relating to bovine health. To this end, focusing on the bovine MHC is a natural goal toward further understanding the immune response in cattle. The primary goal of this research project is to create a sequence based map of BoLA IIb. This includes construction of a sequence contig derived from a minimum tiling path of BoLA IIb, and annotation of the relevant

features in the sequence, resulting in a comprehensive map of BoLA IIb. Additionally, the BoLA IIb sequence will be screened for polymorphisms, both from overlapping BAC sequence from the initial project, and the corresponding sequence from the bovine genome project. The polymorphisms determined *in silico* will facilitate rapid and specifically targeted marker generation. The second goal of this research is to mine publicly available sequence to define polymorphisms that exist in the BoLA IIb region. The addition of a sequence map for BoLA IIb will greatly enhance the ability to continue research into the genetics of BoLA IIb.

CHAPTER II

COMPARATIVE ANALYSIS OF THE BOVINE MHC CLASS IIB SEQUENCE IDENTIFIES INVERSION BREAKPOINTS AND THREE UNEXPECTED GENES*

Overview

The bovine major histocompatibility complex (MHC) or *BoLA* is organized differently from typical mammalian MHCs in that a large portion of the *class II* region, called *class IIb*, has been transposed to a position near the centromere on bovine chromosome 23. Gene mapping indicated that the rearrangement resulted from a single inversion but the boundaries and gene content of the inverted segment have not been fully determined. Here we report the genomic sequence of *BoLA IIb*. Comparative sequence analysis with the human MHC revealed that the proximal inversion breakpoint occurred approximately 2.5 kb 3' of the *glutamate-cysteine ligase catalytic subunit* (*GCLC*) locus and the distal breakpoint occurred about 2 kb 5' from a divergent *class II DR beta*-like sequence designated *DSB*. Gene content, order and orientation of *BoLA IIb* are consistent with the single inversion hypothesis when compared to the corresponding region of the human *class II* MHC (*HLA class II*). Differences with *HLA* include the presence of a

---

* Reprinted with permission from "Comparative analysis of the bovine MHC class IIb sequence identifies inversion breakpoints and three unexpected genes." by Childers CP, Newkirk HL, Honeycutt DA, Ramlachan N, Muzney DM, Sodergren E, Gibbs RA, Weinstock GM, Womack JE, and Skow LC., 2006, Animal Genetics, 37, 121-9.

Data Deposition Footnote: Sequence deposited to GenBank under accession AY957499.

single histone *H2B* gene located between the *proteasome 9* (*PSMB9*) and *DMB* loci and a duplicated *TAP2* with a variant splice site. *BoLA IIb* spans approximately 450 kb DNA, with 20 apparently intact genes and no obvious pseudogenes. The region contains 227 simple sequence repeats and approximately 167 kb of retroviral-related repetitive DNA. Nineteen of the 20 genes identified *in silico* are supported by bovine EST data indicating that the functional gene content of *BoLA IIb* has not been diminished because it has been transposed from the remainder of *BoLA* genes.

Introduction

The MHC of most mammalian taxa consist of tightly linked genes and gene families organized much as *HLA* (Hurt *et al.* 2004; Gustafson *et al.* 2003; Wagner 2003; Yuhki *et al.* 2003; The MHC Sequencing Consortium 1999); however, two different arrangements are found in the MHC of artiodactyls. The swine MHC (*SLA*) on chromosome 7 is disrupted by the centromere such that the *class II* region is on SSC7p and the *class I* and *III* regions are on SSC7q (Chardon *et al.* 1999). All of the genes in *SLA* are very tightly linked due to recombination suppression in the vicinity of the centromere. Organization of the bovine MHC, designated *BoLA*, differs from most other mammalian species in that *class II* loci are found in two regions designated *IIa* and *IIb*. Linkage (Andersson *et al.* 1988) and cytogenetic analysis (McShane *et al.* 2001) located the *class IIb* region near the centromere at BTA23q12 and the *IIa* region near the *class I* and *III* regions at BTA23q23. Detailed mapping of BTA23 by radiation hybrid analysis (Band *et al.* 1998; Itoh *et al.* 2005) revealed that the ancestral MHC was likely disrupted by a large inversion that produced the *BoLA IIa* and *IIb* regions. Comparative studies in

other species demonstrate that that the inversion is likely a common feature of the MHC of all pecoran ruminants (Skow *et al.* 1996a). Two divergent *class II* loci, *DYA* and *DYB* (previously *DIB*) map to *BoLA IIb* (Andersson *et al.* 1988; Skow *et al.* 1996b; Ballingall *et al.* 2004) but a complete and detailed analysis of *BoLA IIb* has not been reported. Here we present the annotated genomic sequence of *BoLA IIb* and identify the breakpoints of the ancestral inversion that produced *BoLA IIb*. The sequence has been uploaded to Genbank under accession number AY957499.

Materials and Methods

Bacterial artificial chromosome (BAC) clones containing BoLA sequences were isolated from a composite TAMU bovine BAC library. The TAMU library is comprised of 60,000 clones, yielding a 3.6X coverage. The library was constructed from White Blood Cell DNA of an Angus bull designated Y6 (Cai *et al.* 1995) and augmented by DNA from Longhorn (ID-14), Brahman (ID-40) and Angus (ID- 7138) cows (C. Gill personal communication). The clones were created by digesting the DNA with BamHI and HindIII, followed by insertion of the fragments into the pBeloBAC11 vector (7.4 kbp). The library was initially screened using PCR primers designed from conserved regions of coding sequences of 14 genes in the MHC of human (*HLA*) and mouse (*H2*). Each BAC clone was end-sequenced and fingerprinted (Gustafson *et al.* 2003) to orient overlaps and verify assembly of the BAC array. BAC end sequencing reactions were modified from the standard ABI reaction for BAC scale templates by using the following parameters: 2 ul Big Dye, 2 ul Half Big Dye (Genetix USA), 0.5 ul 10X MasterAmp (Epicentre), 1 ul 10 mM primer, and 1 ug template DNA in a volume of 10 ul. The

thermal profile for the sequencing reaction was 96 C for 2 min, followed by 99 cycles of 96C for 30 sec, 50 C for 4 min, and a final hold of 65 C for 15 min. BAC fingerprinting was performed as described by Gustafson (2003), with modifications as described below. One ug of DNA for each BAC was digested to completion using BamHI. The restriction fragment patterns were then determined by running them out on 0.65% TBE agarose gels containing ethidium bromide. The fingerprint gels were run at 50 volts for 24 hours, at 4 C. After electrophoresis, digital images were taken for each gel using an Alpha Innotech ChemiImager system. Band sizes were determined using Image 3.0 (Sulston *et al.* 1988) and contig overlap was confirmed using FPC V4.7.9 (Soderlund *et al.* 1997). Fingerprint analysis was then confirmed by repeating the procedure with the restriction enzyme HindIII. The Internet contig explorer, iCE, (http://www.bcgsc.ca/ice/) was also used to verify the tiling path using the cow BAC fingerprinting data.

PCR primers were designed as needed from BAC end sequences for library screening to identify bridging BAC clones for completion of the BoLA IIb array. A single cloning gap between *DMA* and *BRD2* genes was spanned with BAC 065P20 isolated from the bovine CHORI-240 library (http://bacpac.chori.org/bovine240.htm).

DNA from each BAC was prepared using the Qiagen midiprep protocol (Qiagen, Valencia, CA). Using a sterile loop, LB agar plates containing 12.5 ug/ml Chloramphenicol were streaked with *E. coli* from glycerol stocks for each BAC clone of interest, and grown overnight at 37 C.. A single Colony from each clone was picked and used to inoculate 5ml starter cultures consisting of 2YT, containing 12.5 ug/ml Chloramphenicol, and allowed to grow for eight to ten hours at 37 C, with agitation. A 100 ml culture of 100 ml 2YT containing 12.5 ug/ml Chloramphenicol was inoculated

with 1.0 ml of the starter culture. This culture is incubated overnight at 37 C. with agitation. The cultures were split into two 50 ml tubes, and centrifuged at 645 x g (2,000 RPM on a Beckman JA-12 rotor). Following centrifugation, the media was poured off, and excess media was removed with a pipette. The cell pellets were gently resuspended in 10 ml cold Qiagen buffer P1 containing 100 ug/ml Rnase A. Following resuspension, 10 ml Qiagen buffer P2 was added to each tube, mixed thoroughly by inverting the tube four to six times and incubated at room temperature for five minutes. The mixture was then neutralized by the addition of 10 ml cold Qiagen buffer P3 and mixed thoroughly by inverting the tube four to six times. The tubes were then incubated on ice for 15 minutes, then centrifuged for 30 minutes at 10,000 x g (8,000 RPM on Beckman JA12 rotor). The supernatant was then removed into labeled tubes, then centrifuged a second time for 30 minutes at 10,000 x g (8,000 RPM on Beckman JA12 rotor), followed by removal of the supernatant into labeled tubes. During the second centrifugation, a Qiagen tip-100 was equilibrated for each clone by applying 4 ml of Qiagen Buffer QBT and allow column to empty via gravity flow. After equilibration, the samples from each clone were applied to the respective Qiagen tip-100 and allowed to empty via gravity flow. Each Qiagen tip-100 was then washed with 2 x 10 ml Qiagen buffer QC, then allowed to drain. DNA was eluted with 5 ml Qiagen buffer QF, heated to 65 C. Eluting was carried out in five aliquots of 1 ml to minimize cooling of the elution buffer. DNA for each sample was precipitated through by adding 3.5 ml room temperature isopropanol to the eluted DNA, followed by mixing and immediate centrifugation at 10,000 x g for 30 minutes at 4 C (8,000 rpm on a Beckman JA-12 rotor). The supernatant was then decanted, taking care not to dislodge the DNA pellets. Following precipitation, The DNA was resuspended in

1.5 ml 70% ethanol, then transferred to 1.5 ml microcentrifuge tubes and centrifuged at full speed for 5-10 minutes. The supernatant for each sample was carefully decanted following centrifugation, allowed to air dry for five to ten minutes and resuspended in 40 ul elution buffer (10 uM Tris-Cl pH 8.5). DNA was quantified on a 0.7% TBE agarose gel.

Shotgun libraries were produced from each BAC using the M13 adaptor method (Andersson *et al.* 1994) and sequenced to 3X average depth to yield approximately 5.9 Mb of raw sequence. Quality assessment and sequence assembly was performed using PHRED (Ewing and Green 1998; Ewing *et al.* 1998) and PHRAP (P. Green, unpublished, see http://www.phrap.org). Sequence gaps were closed by primer walking or by sequencing PCR products that spanned small gaps. Additional sequence from the bovine genome sequencing project (http://www.ncbi.nlm.nih.gov/Traces/) facilitated gap closure and full assembly. Primers for sequencing BAC templates and PCR products were designed from unique sequence at the terminal ends of each assembly contig (Appendix A).

Internal BAC sequencing reactions were modified from the standard ABI reaction for BAC scale templates as follows: 2 ul Big Dye, 2 ul Half Big Dye (Genetix USA), 0.5 ul 10X MasterAmp (Epicentre), 1 ul 10 mM primer, and 1 ug template DNA in a volume of 10 ul. The thermal profile for the sequencing reaction was 96 C for 2 min, followed by 99 cycles of 96C for 30 sec, 50 C for 4 min, and a final hold of 65 C for 15 min.

Sequencing by primer walking was continued until new sequence bridged to another assembly or failed to provide unique primers for continued walking. AssemblyLign (Accelerys) software was used for the final assembly of each BAC.

Concordance of predicted and observed HindIII restriction fragment patterns was used to validate each BAC assembly, which were then combined using AssemblyLign to produce a full sequence contig of *BoLA IIb*.

BLAST (Altschul *et al.* 1990) alignments of the repeat masked, assembled *BoLA IIb* sequence against NCBI EST and non-redundant nucleotide databases were used to identify expressed sequences and other highly conserved regions likely to contain functional genes. All sequences with >85% identity to known coding regions in *HLA* were aligned to published mRNA and amino acid sequences using Pairwise BLAST, then passed through GENSCAN (Burge and Karlin 1997) and GenomeScan (Yeh *et al.* 2001) (http://genes.mit.edu) to predict the exon structures for each candidate gene. To confirm identity, we aligned predicted translations to published protein sequences of human, mouse, rat, and when possible, cow and sheep using T-Coffee (Notredame *et al.* 2000) (http://www.ch.embnet.org/software/TCoffee.html).

The assembled sequence was analyzed by RepeatMasker (http://www.repeatmasker.org/) using the "–cow" option to identify bovine interspersed repetitive elements (LINES and SINES). Sputnik (http://espressosoftware.com/) was used to identify simple sequence repeats (SSRs) with di- to penta nucleotide motifs/SSR. The repeat masked sequence was then aligned to itself using pairwise BLAST, and the output screened for additional bovine repeat sequences that were not detected with RepeatMasker or Sputnik.

Anchored pair-wise alignment to orthologous sequences was performed to identify loss of sequence similarity as an indicator of the inversion breakpoint. The 9.2 kb non-coding sequence between the non-MHC *GCLC* gene and the *IIb DSB* gene was

repeat masked and aligned using pairwise BLAST to human sequence flanking the 3` end of *GCLC* and to sequences flanking each of the *DRB* genes in the *HLA* assembly (GenBank accession no. NT_007592). Multipipmaker (Schwartz *et al.* 2003) (http://pipmaker.bx.psu.edu) was used to generate percent identity plots for *BoLA IIb* against human, mouse, rat, and cat sequences representing the *class II* MHC and the human, mouse and rat sequences containing *GCLC*. The feline MHC *class II* consensus sequence was assembled from published sequences (GenBank accession nos. AY152826, AY152827, AY152828, AY152829, AY152830, AY152831, AY152833, AY152834, and AY152836).

Results

A total of eleven BAC clones were isolated from the TAMU and CHORI-240 bovine BAC libraries to form a contig spanning *BoLA IIb*. Four BAC clones from the TAMU library (27C10R5, 3C8R7, 14-47C8R7, and 7138-04C6R6) constituted a minimum tiling path of *BoLA IIb* and were used as initial templates for shotgun cloning and sequencing. Initial assembly identified a cloning gap of approximately 12 Kb based on comparison to *HLA* sequence between BACs 14-47C8R7 and 7138-04C6R6. This gap was spanned by a single clone (065 P20) from the CHORI-240 library and provided a template for direct sequencing of the gap between clones 14-47C8R7 and 7138-04C6R6. The final assembly of *BoLA IIb* contig contains 475,011 bp including 37,236 bp centromeric to the *IIb* boundary (Figure 2). The sequence is complete except for three presumably small gaps, located between *PSMB9* and *H2B* at base 317,766, within intron S of *COL11A2* at base 412,768, and within intron F of *HKE4* at base 439,443. The size of

each gap was difficult to estimate from the corresponding human sequence, but each is probably less than 500 bp in size. Analysis of the assembled sequence (Figure 2) identified twenty genes and includes elements of the breakpoint that disrupted the ancestral *class II* region. The breakpoint was delineated by a single BAC between two genes, *GCLC*, a non-MHC gene, and *DSB*, a divergent MHC *class II* gene separated by 9.7 kb. The *DSB* gene is followed by the divergent *class II DYA* and *DYB* loci. The remainder of the assembly contains most of genes typically found in the MHC *class II* region including the *class II DOB* gene, proteosome genes (*PSMB2* and *PSMB8*), transporter genes (*TAP1* and *TAP2*), the CLIP releasing *class II* gene pair *DMB* and *DMA*, the mitogen-activated nuclear kinase gene *BRD2* and the class II gene pair *DNA* and *DOA*. A short fragment with 78% identity to exon 3 of *HLA class II DPB* was identified approximately 9.4 kb from *DOA*, a region that contains variable numbers of pseudogenes and gene fragments of *DPA* and *DPB* in other non-primate mammals (Kelley *et al.* 2005). Analysis of this region in *BoLA IIb* failed to identify any additional sequences similar to *DPB* or *DPA*. The sequence for *BoLA IIb* extends into the *extended class II* region, and includes the *DPB* fragment and apparently intact genes for *COL11A2*, *RXRB*, *HKE4*, and *HKE6*. The assembled sequence of *BoLA IIb* presented in this paper ends in intron B of *RING1*.

In addition to the expected MHC genes and the previously uncharacterized *DSB* gene, the assembled sequence of *BoLA IIb* contains two unexpected genes, a duplicated transporter 2 gene, designated *TAP2.1* tandemly arrayed with *TAP2* and a histone 2B (*H2B*) gene located between *PSMB9* and *DMB*. *TAP2.1* is contained within a duplication of 10 Kb immediately adjacent and head to tail to *TAP2*. The predicted coding sequence

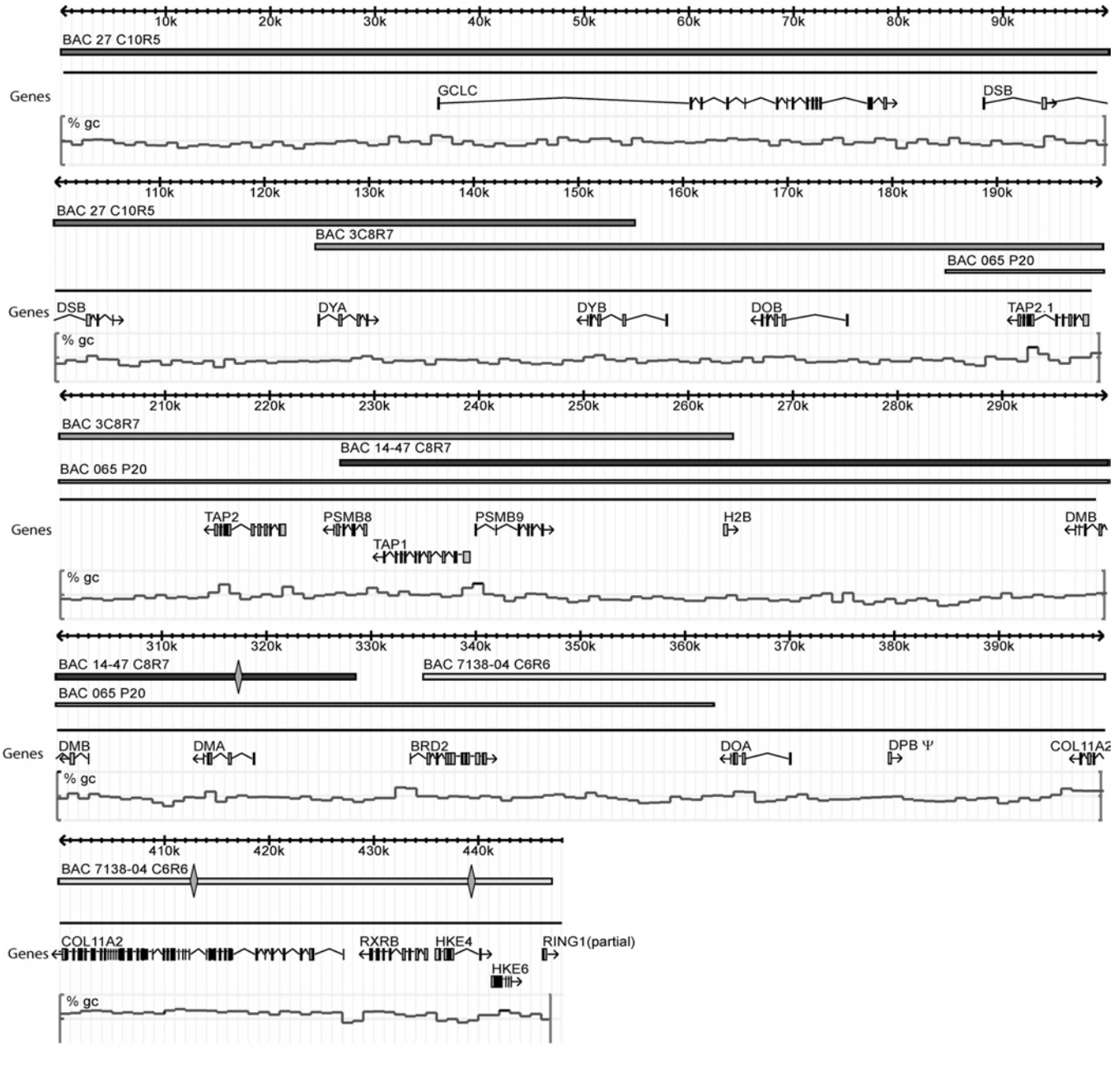


Figure 2 Characterization of the *BoLA IIb* sequence. The *class IIb* region spans roughly 450 Kb, and includes twenty potential genes. (From top to bottom) Tiling path: The shaded bars represent BACs 27C10R5, 3C8R7, 14-47C8R7, and C 7138-04C6R6 that were shotgun sequenced and assembled. Diamonds indicate the position of sequence gaps. The open bar, representing BAC 065 P20 filled the gap in the tiling path. Di-, Tri-, Tetra-, Penta-: simple sequence repeats derived by SPUTNIK and separated by motif length. Genes: predicted exons are represented by filled boxes and introns by bent lines. Arrows indicate the direction of transcription. Percent G+C: the percentage of sequence comprised of G + C using a window size of 1 Kb.

of *TAP2.1* is 98% identical to the corresponding sequence in *TAP2* and 84% identical to human *TAP2* at the amino acid level, but *TAP2.1* contains a variant splice site at the 5' end of intron I. Absence of any corresponding EST in the bovine EST database suggests that *TAP2.1* is not transcribed. The *H2B* gene displays 96% identity to the human sequence (NP_003516.1) at the amino acid level, and histone genes have never previously been characterized in *BoLA IIb*. The *H2B* sequence is present at the same position in overlapping BACs from the TAMU and CHORI-240 libraries, indicating that the location of *H2B* was not an artifact of the assembly process. Analysis of the adjacent flanking sequences of the *H2B* gene in *BoLA IIb* revealed a sharp boundary of non-*class II* and repetitive sequences consistent with an origin by retrotransposition.

Excluding *TAP2.1* and the *DPB* fragment, the transcriptional state of each of the genes in *BoLA IIb* was supported by at least one entry in the bovine EST database. Other than divergent *DSB*, *DYA* and *DYB*, each of the intact genes aligned to its orthologue in the *HLA class II* region at E values < e^{-100} at the nucleotide level and < e^{-62} at the amino acid level (Table 1). The absence of functional *class II DP* genes in *BoLA IIb* is consistent with the restriction of functional members of these class II genes to Euarchontoglires genomes and the presence of *DP* pseudogenes and gene fragments in the genomes of other placental mammals (Kelley *et al.* 2005).

The order and relative orientation of genes in *BoLA IIb* are conserved with *HLA* consistent with the inversion hypothesis for the origin of *IIb*. However, *BoLA IIb* is significantly smaller than the homologous region in *HLA*. The size difference between *BoLA IIb* and the homologous region in *HLA* is almost entirely accommodated in four segments rich in pseudogenes in *HLA* but completely or nearly absent in *BoLA IIb*

(Figure 3). The comparative lack of pseudogenes in *BoLA IIb* when compared to *HLA* may be the result of physical removal from the proximity of other members of the *class II*

Table 1 Comparisons of coding sequences and encoded proteins for *HLA II* and *BoLA IIb* genes. Columns: HSA = Human, BTA = cow, length of CDS (bp) or translation (aa), %Id = Percent sequence identity, %Sim = Percent functional similarity of amino acid residues, E-Value = Expectation value, and HSA Accession No. = Genbank accessions used in alignments.

Gene		HSA	BTA	%Id	%Sim	E-Value	HSA Accession No.
GCLC	CDS	1914	1818	0.86		0	NM_001498.2
	Protein	637	605	0.88	0.9	0	NP_001489.1
DSB	CDS	801	783	0.77		E-154	NM_002124.1
	Protein	266	260	0.66	0.75	3.00E-89	NP_002115.1
DYA	CDS	768	762	0.78		E-129	NM_002122.2
	Protein	255	253	0.66	0.8	2.00E-93	NP_002113.2
DYB	CDS	786	780	0.78		E-130	NM_002123.2
	Protein	261	259	0.71	0.8	1.00E-96	NP_002114.2
DOB	CDS	822	816	0.81		0	NM_002120.2
	Protein	273	271	0.73	0.81	E-113	NP_002111.1
TAP2.1	CDS	2112	2136	0.8		0	NM_000544.2
	Protein	703	711	0.75	0.84	0	NP_000535.2
TAP2	CDS	2112	2136	0.78		0	NM_000544.2
	Protein	703	711	0.75	0.85	0	NP_000535.2
PSMB8	CDS	819	831	0.88		0	NM_004159.3
	Protein	272	276	0.9	0.96	E-121	NP_004150.1
TAP1	CDS	2427	2250	0.81		0	NM_000593.4
	Protein	808	749	0.75	0.82	0	NP_000584.2
PSMB9	CDS	660	660	0.87		0	NM_002800.3
	Protein	219	219	0.91	0.95	E-113	NP_002791.1
H2B	CDS	381	381	0.88		E-130	NM_003525.2
	Protein	126	126	0.96	0.96	2.00E-62	NP_003516.1
DMB	CDS	792	789	0.8		E-174	NM_002118.3
	Protein	263	262	0.74	0.8	E-113	NP_002109.1
DMA	CDS	786	783	0.82		0	NM_006120.2
	Protein	261	260	0.76	0.84	E-116	NP_006111.2
BRD2	CDS	2406	2412	0.91		0	NM_005104.2
	Protein	801	803	0.97	0.97	0	NP_005095.1
DOA	CDS	753	753	0.81		E-178	NM_002119.2
	Protein	250	250	0.74	0.81	E-105	NP_002110.1
COL11A2	CDS	5211	5211	0.88		0	NM_080680.1
	Protein	1736	1736	0.94	0.95	0	NP_542411.1
RXRB	CDS	1602	1599	0.92		0	NM_021976.3
	Protein	533	532	0.98	0.98	0	NP_068811.1
HKE4	CDS	1290	1290	0.89		0	NM_006979.1
	Protein	429	429	0.91	0.93	0	NP_008910.1
HKE6	CDS	786	780	0.88		0	NM_014234.3
	Protein	261	259	0.87	0.92	E-123	NP_055049.1

gene family and consequently lack of unequal recombination necessary to produce the expansion/contraction that is typical of MHC gene families. Also, transposition to a region near the centromere may have further reduced recombination rate in *BoLA IIb*.

Repetitive sequences, primarily LINES and SINES, comprise approximately 40% of *BoLA IIb* and the region is richly populated with SSRs (Table 2). Sputnik analysis identified 229 SSRs representing 48 different motifs. Among dinucleotide motifs, (AC:GT) was the most common, comprising 58% of all dinucleotide repeats, and 21.4% of all SSRs present in *BoLA IIb*. These SSRs will likely provide a rich source of genetic markers for probing *BoLA IIb* haplotype structure and seeking health-associated genotypes.

Pairwise alignment and Multipip analysis of the non-coding region between *GCLC* and *DSB* localized the centromeric breakpoint of the ancestral inversion that generated *BoLA IIb* to a 2.3 Kb segment about equidistant between *GCLC* and *DSB* (Figure 4.). Anchored, pairwise alignment of the *GCLC-DSB* intergenic region of *BoLA IIb* to sequences 3' of human *GCLC* (GenBank accession no. NT_007592) revealed sequence similarity extending 3.7 Kb 3` of *GCLC*. Similarly, alignment of the sequence spanning the breakpoint to sequence flanking *HLA-DRB1* (GenBank accession no. NT_007592) revealed sequence similarity extending 3.2 Kb upstream of *DSB*. The remaining 2.3 Kb of intergenic sequence showed no similarity to sequences flanking either of the human genes, and most likely contains the ancestral breakpoint. About 66% of this sequence consists of degenerate LINE and SINE elements consistent with an inversion that predates pecoran ruminant divergence over 35 million years ago (Nijman *et*

Table 2 Repetitive interspersed elements present in *BoLA IIb*. Repeats identified by RepeatMasker and arranged by element class and simple sequence repeats (SSRs) identified by SPUTNIK, arranged by motif length.

Interspersed Repeats (RepeatMasker)	**Number of elements**	**Length (bp)**	**Percentage (%)**
SINEs	301	50049	11.20
MIRs	52	6901	1.54
LINEs	206	96852	21.67
LINE1	108	46192	10.33
BovB/Art2	71	44608	9.98
LINE2	22	5323	1.19
L3/CR1	5	729	0.16
LTR elements	24	11645	2.61
MaLRs	12	2468	0.55
ERVL	3	518	0.12
ERV_classI	3	3120	0.70
ERV_classII	0	0	0.00
DNA elements	37	8123	1.82
MER1_type	22	5608	1.25
MER2_type	8	1589	0.36
Unclassified	0	0	0.00
Total interspersed repeats	**568**	**166669**	**37.29**
Small RNA	3	182	0.04
Satellites	0	0	0.00
Simple repeats	68	3908	0.87
Low complexity	82	4180	0.94
Total repeats	**721**	**174764**	**39.15**
SSRs (SPUTNIK)			
Dinucleotide	82	1679	0.38
Trinucleotide	59	902	0.20
Tetranucleotide	37	535	0.12
Pentanucleotide	49	869	0.19
Total	**227**	**3985**	**0.89**

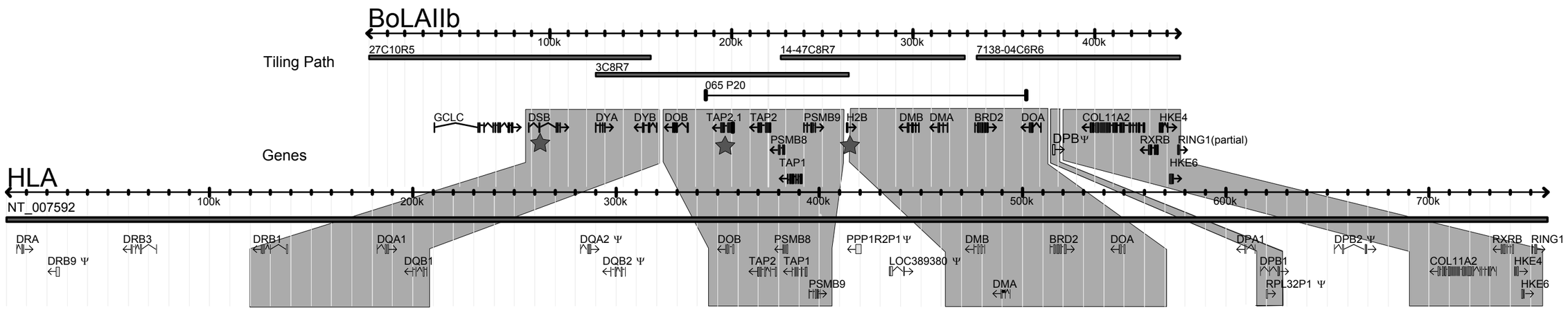


Figure 3 Comparative analysis of *BoLA IIb* and *HLA class II*. Clusters of genes are arranged in a relative orientation, and are drawn to the same scale. (From top to bottom) Tiling path: source of sequence used for annotation. RepeatMasker was used to screen for repetitive sequence using species specific parameters. Genes: drawn to scale, rectangles mark the presence of exons, bent lines denote introns, and arrows indicate direction of transcription. Stars identify genes present in *BoLA IIb* that are absent in *HLA class II*. Shaded boxes indicate regions where the gene content, organization and orientation are shared between species.

al. 2002). The predominant feature of the breakpoint region is a degraded L1 LINE element into which a Bov B element was inserted in opposite orientation. Multipip alignment across the intergenic sequence (Figure 4b) detected greater similarity between cattle and human or cat sequence than cattle to rat or mouse sequence.

Discussion

Detailed examination of the major histocompatibility complexes (*MHC*) of warm-blooded vertebrates reveals conserved synteny that belies a very dynamic region of the genome that differs in organization, size and gene number among and within species (reviewed by Kelley *et al.* 2005).

The ruminant MHC appears to be unique in that the complex was disrupted by an inversion to produce two large, syntenic clusters separated by about 20Mb of non-MHC DNA. The genomic sequence of *BoLA IIb* presented here confirms the inversion hypothesis. When inverted, the arrangement and organization of genes in *BoLA IIb* are almost perfectly conserved with genes in the homologous *HLA class II* region (Figure 3) and the similarity extends to the level of transcriptional orientation. Only the *DSB* gene immediately proximal to the inversion breakpoint is in opposite orientation compared to the presumed human orthologue. The sequence comprising *BoLA IIb* is smaller than the orthologous region in *HLA class II,* largely due to the paucity of pseudogenes in the bovine sequence. The comparative alignment of *BoLA IIb* to *HLA class II* identifies four major blocks of conserved organization separated by three segments present in *HLA* but not detected in *BoLA IIb*. Each of the three *HLA* segments contain pseudogenes absent in

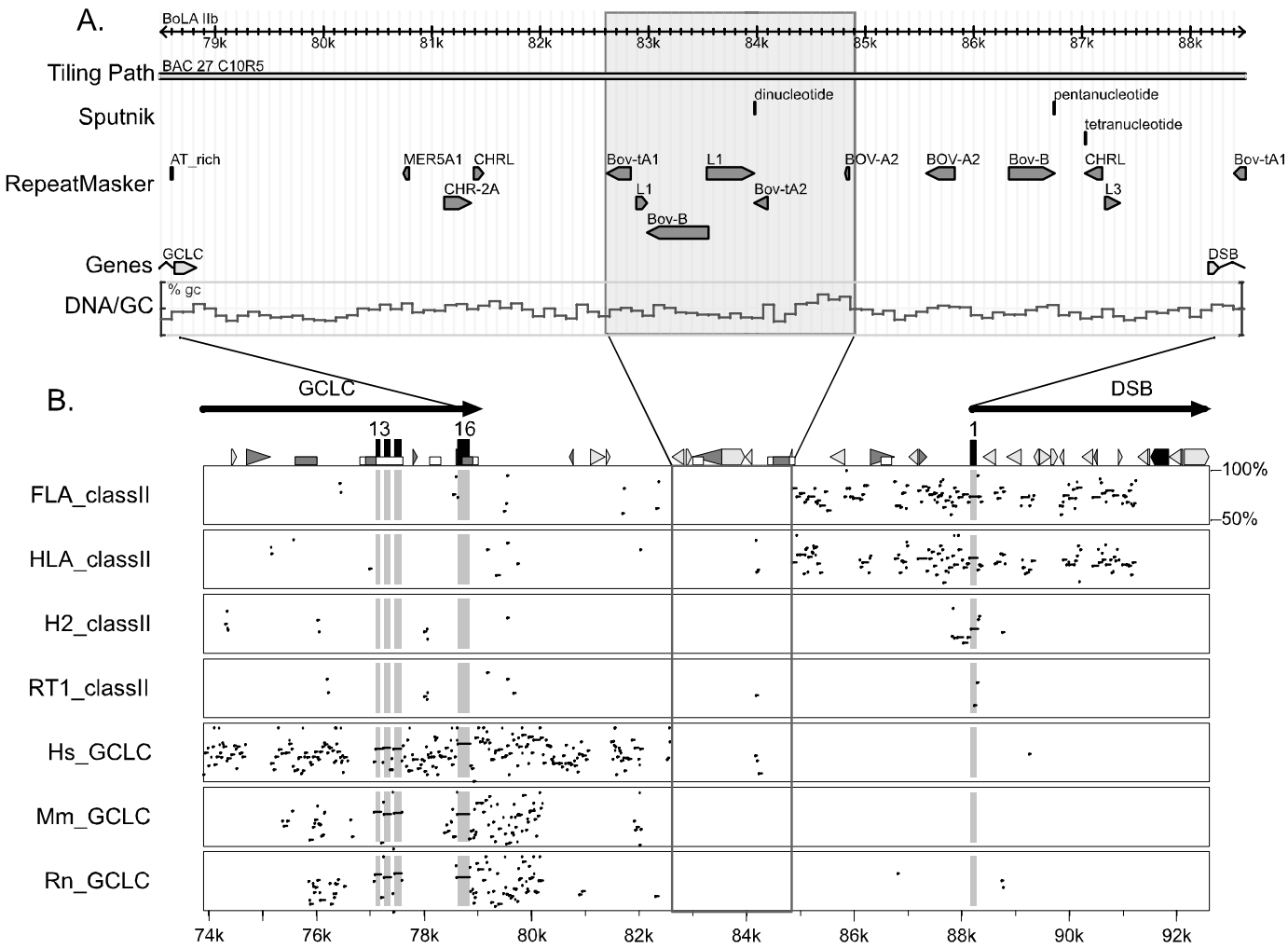


Figure 4 A. Inversion breakpoint localized to a 2.4 Kb region between *GCLC* and *DSB*. The shaded box spans a region that shows no similarity to sequences flanking *GCLC* or the *DRB* genes in other species. B. Multipip alignment spanning the inversion breakpoint proximal to *BoLA IIb*. The boxed area delineates the predicted breakpoint region. Percent identity plots are scored from 50 to 100%, and are represented by black dots. Shaded columns indicate regions corresponding to exons.

the corresponding sequence of *BoLA IIb* except for the single *DPB* fragment between *DOA* and *COL11A2*. The three *HLA* segments absent from *BoLA IIb* account for the larger size of this region of *HLA class II* compared to *BoLA IIb* and probably represent recent expansions in the primate lineage rather that a loss of sequences in the ruminant genome.

The MHC class III region has been found to be the most gene dense region in the human genome with 60 genes spanning a 700 kb region, averaging one gene every 11.6 kb (Xie *et al.* 2003). BoLA IIb was found to contain 20 genes across 447 kb, averaging one gene every 22.3 kb. HLA spans 3.6 Mb, and contains 128 expressed genes, with an average of one gene every 28.1 kb (The MHC Sequencing Consortium 1999). The latest assembly of the human genome (NCBI build 36) is estimated to contain 22,000 genes, with a total length of roughly 3,000 Mb, averaging one gene every 136.3 kb (http://www.ensembl.org/Homo_sapiens/).

Identification of a divergent member of the *DRB* gene family, *DSB*, just proximal to the inversion breakpoint was unexpected and establishes the centromeric boundary of the *BoLA IIb* region. Furthermore, the presence of a *DRB* – like gene increases to three the number of expressed but divergent *class II* genes in *BoLA IIb*. *DRB* genes typically contain six exons encoding (in order) a leader polypeptide (exon 1), two extracellular domains (exons 2 and 3), a transmembrane domain (exon 4) and small exons 5 and 6 that encode a short cytoplasmic tail. The sequence for *DSB* predicts a 783 bp mRNA derived from five exons. The first four predicted exons of *DSB* are highly conserved with *HLA-DRB1* (Table 1), while exon 5 is divergent from all previously characterized *DRB* sequences and exon 6 appears to be missing. Transcription of *DSB* is supported by a

single bovine EST sequence (AW 480953) with agreement between the coding sequence predicted from the genomic contig and the EST sequence. The divergent cytoplasmic tail predicted for *DSB* may indicate evolution of a novel function and/or tissue specificity as has been proposed for the adjacent *DYA* and *DYB* genes.

DYA and *DYB* evolved from *class II DQ*-like sequences and comprise a *BoLA class II* functional gene pair apparently unique to the pecoran ruminants and expressed predominantly in dendrocytes in afferent lymph (Stone *et al.* 1990; van der Poel *et al.* 1990; van Eijk *et al.* 1992; Mann *et al.* 1993; Wright *et al.* 1994; van Eijk *et al.* 1995; Skow *et al.* 1996a; Skow *et al.* 1996b; Ballingall *et al.* 2001; Ballingall *et al.* 2004). *DSB* may also have acquired a specialized function upon divergence. If *DSB* is functional, the identity and location of its *DRA*-like partner remains to be determined.

Comparison of chromosomal gene order and sequence appear to support paralogous relationships between the *DSB-DYA-DYB* trio in *BoLA IIb* with the *DRB1*, *DQA1* and *DQB1* segment in *HLA* (Figure 3). Multiple *DQ* and *DR* genes exist in the *BoLA IIa* region (Andersson and Rask 1988) indicating that expansion of the bovine *class II* region preceded the inversion that produced *BoLA IIb*. If the proposed origin of the *DSB-DYA-DYB* segment is correct a cluster of *DQ* and *DR* genes would mark the centromeric boundary of the *IIa* region. The origin and evolutionary history of *BoLA IIb* will be clarified further with the completion of the *BoLA IIa* sequence.

TAP2.1 and *H2B* within *BoLA IIb,* to our knowledge, are unique to the MHC of cattle and closely related species. The presence of *TAP2.1* was confirmed by PCR in cattle and bison but could not be demonstrated in deer, sheep, or goats (data not shown), indicating that the *TAP2.1* duplication may be a recent event in the bovid lineage.

Predicted translations of *TAP2.1* sequence downstream from the corrupted splice site at the 5' end of intron I using possible splice sites in intron I or a fusion sequence of exon 9, intron I and exon 10 resulted in multiple premature termination codons. If expressed, *TAP2.1* would encode a slightly truncated protein with a highly divergent carboxyl terminal end. No support for transcription of *TAP2.1* yet exists in the bovine EST database but more detailed studies in cattle and other bovids are needed to fully evaluate the functional state of this duplicated locus.

The presence of the single *H2B* gene in *BoLA IIb* is unique to ruminant genomes as no member of any histone gene family has been reported within the classical MHC of other mammals. Cattle histone genes are not precisely mapped but a large cluster of histone genes is found in the extended *HLA class I* region. *In silico* analysis indicates that the *H2B* gene in *BoLA IIb* is intact and likely expressed. The conservation of the *BoLA IIb H2B* with other *H2B* sequences will make it difficult to specifically assess the functional status of this locus in the presence of products from other, nearly identical histone genes.

Localization of one of the *BoLA IIb* inversion breakpoints to a 2.4 kb sequence between *GCLC* and *DSB* (Figure 4) represents one of the most precisely defined evolutionary breakpoints yet discovered in mammals. The preponderance of repetitive elements within this sequence may implicate these elements in the chromosomal break that produced the inversion. Similar repetitive elements, especially Alu- and Alu-like sequences, have been causally implicated in chromosomal breakage events within the primate lineage (Rudiger *et al.* 1995; Hill *et al.* 2000; Pletcher *et al.* 2000; Kehrer-Sawatzki *et al.* 2002). Alternatively, the *BoLA IIb* inversion may have arisen as a

consequence of instability associated with *class II* gene expansion. Similar rearrangements consequential to gene expansion have been reported for human chromosome 18 (Dennehey *et al.* 2004). Such a mechanism might also explain the reverse orientation of *DSB* compared to all other orthologous loci in *BoLA IIb*. Because expansion of ancestral *DRB* genes preceded the chromosomal inversion, we might expect *BoLA IIa* to contain one or more *DRB* genes immediately proximal to the centromeric boundary of *BoLA IIa*. However, expansions and contractions of *BoLA IIa* subsequent to the inversion may obscure features of the ancestral *class II* region.

Elucidation of the *BoLA IIb* sequence offers insight into the significance of MHC organization for coordinated gene function. Separation of genes in *BoLA IIb* from proximity to the remainder of *BoLA* has had no apparent disruption of function in the bovine immune response. Genes encoding bovine class I antigen proteases and transporters are only found in *BoLA IIb* and apparently function normally in processing and loading of antigen onto class I receptors. Similarly, the *DMA* and *DMB* loci in *BoLA IIb* encode proteins necessary to expose the antigen binding sites of newly synthesized *class II* heterodimers encoded by genes in *BoLA IIa*. Presumably the regulatory sequences required to drive coordinated expression of these functionally related genes have been retained in regions flanking the genes of *BoLA IIb*.

Conversely, analysis of *BoLA IIb* suggests that removal of genes from proximity to the core MHC genes may alter the evolutionary processes that shape conventional MHC. The paucity of recognizable pseudogenes that distinguishes *BoLA IIb* from the homologous region in *HLA*, which contains five pseudogenes: (GenBank accession no. NT_007592.13) could be due to suppression of the “birth and death” (Takahashi *et al.*

2000) process that characterizes *class I* and *class II* gene families. Divergence of *DSB*, *DYA* and *DYB* may have been facilitated by elimination of the homogenization process associated with gene conversion of closely linked genes, a situation that may be enhanced by reduced recombination in *BoLA IIb* due to its proximity to the centromere.

CHAPTER III

SEQUENCE VARIATION IN BOLA IIB

Overview

The genomic sequence of the BoLA IIb region as determined in Chapter II provided a basis for detecting sequence variation through comparison to other bovine sequences, including genomic sequence from the bovine genome project, revealing a total of 10,408 mismatching bases, and 15 sequences corresponding to the validated SNP panel generated by the bovine genome sequencing project. Additionally, two distinct blocks of sequence corresponding to overlapping regions in the BAC scaffold were defined, and found to have 888 polymorphisms, indicating that each BAC derived from a different chromosome.

BoLA IIb contains 227 simple sequence repeats and approximately 167 kb of retroviral-related repetitive DNA. Nineteen of the 20 genes identified *in silico* are supported by bovine EST data indicating that the functional gene content of *BoLA IIb* has not been diminished because it has been transposed from the remainder of *BoLA* genes.

Introduction

The MHC has long been recognized as a region of the genome that exhibits extremely high levels of polymorphism. HLA is the best characterized MHC, and there have been 1,603 alleles found for HLA class I genes, and 832 alleles for HLA class II genes, with some MHC genes consisting of hundreds of alleles such as HLA-DRB1, HLA-DP1, or HLA-A, HLA-B, and HLA-C (Robinson *et al.* 2003). MHC genes are

known to exhibit copy number differences between species (Yuhki *et al.* 2003; Debenham *et al.* 2005) as well as copy number differences within a species (Andersson and Rask 1988). One example is the DP- gene family. DP- genes are present as both functional genes and pseudogenes in human, exist as pseudogenes in mouse, and are not present in cattle. Similarly, the DY- genes in cattle appear to have originated from DQ- genes.

The fact that many MHC genes function in some aspect of immune response has highlighted the importance of understanding the region, including determining the organization and function of the region, as well as levels of diversity within and between individuals. The recent release of the cow genome project has provided a large amount of sequence data that may be used to screen for sequence variation. One of the secondary objectives of the cow genome project is to create a SNP dataset from different cattle breeds, in order to survey the diversity currently present among cattle breeds. This dataset includes a number of validated SNP sequences, and may provide an additional set of sequence variants for the BoLA IIb region.

The goal of this study is to estimate the level of sequence diversity present in the BoLA IIb region through the utilization of publicly available sequence data. Through this study, the discovery of motif length variant microsatellites may provide the basis for polymorphic markers in a very short time frame. Other types of sequence variations could provide the basis for developing SNP markers very tightly linked to loci of interest. Sequences that have been found to be polymorphic through computational alignments could then be screened against other animals in order to determine allele frequencies and determine the haplotype structure of BoLA IIb.

Materials and Methods

Sequencing and assembly of BoLA IIb

Bacterial artificial chromosome (BAC) clones containing BoLA sequences were isolated from a composite TAMU bovine BAC library (Cai *et al.* 1995). The TAMU library is comprised of 60,000 clones, yielding a 3.6X coverage. The library was constructed from White Blood Cell DNA of an Angus bull designated Y6 (Cai *et al.* 1995) and augmented by DNA from Longhorn (ID-14), Brahman (ID-40) and Angus (ID- 7138) cows (C. Gill personal communication). The genomic DNA was digested with the restriction enzymes BamHI and HindIII. The resulting fragments were inserted into the pBeloBAC11 vector (7.4 kbp).

The coding conserved coding regions of 14 genes in HLA and H2 were used to design primers for PCR screening the library. Each positive clone was end sequenced and fingerprinted to orient overlaps and verify the assembly of the BAC array. BAC end sequencing reactions were modified from the standard ABI reaction for BAC scale templates by using the following parameters: 2 ul Big Dye, 2 ul Half Big Dye (Genetix USA), 0.5 ul 10X MasterAmp (Epicentre), 1 ul 10 mM primer, and 1 ug template DNA in a volume of 10 ul. The thermal profile for the sequencing reaction was 96 C for 2 min, followed by 99 cycles of 96C for 30 sec, 50 C for 4 min, and a final hold of 65 C for 15 minutes. BAC fingerprinting was performed as described by Gustafson (2003), with modifications as described below. One ug of DNA for each BAC was digested to completion using BamHI. The restriction fragment patterns were then determined by running them out on 0.65% TBE agarose gels containing Ethidium Bromide. The

fingerprint gels were run at 50 volts for 24 hours, at 4 C. After electrophoresis, digital images were taken for each gel using an Alpha Innotech ChemiImager system. Band sizes were determined using Image 3.0 (Sulston *et al.* 1988) and contig overlap was confirmed using FPC V4.7.9 (Soderlund *et al.* 1997). Fingerprint analysis was then confirmed by repeating the procedure with the restriction enzyme HindIII. The Internet contig explorer, iCE, (http://www.bcgsc.ca/ice/) was also used to verify the tiling path using the cow BAC fingerprinting data.

PCR primers were designed as needed from BAC end sequences for library screening to identify bridging BAC clones for completion of the BoLA IIb array. A single cloning gap between *DMA* and *BRD2* genes was spanned with BAC 065 P20 isolated from the bovine CHORI-240 library (http://bacpac.chori.org/bovine240.htm).

DNA from each BAC was prepared using the Qiagen midiprep protocol (Qiagen, Valencia, CA). Shotgun libraries were produced from each BAC using the M13 adapter method (Andersson *et al.* 1994) and sequenced to 3X average depth to yield approximately 5.9 Mb of raw sequence. Quality assessment and sequence assembly was performed using PHRED (Ewing and Green 1998; Ewing *et al.* 1998) and PHRAP (P. Green, unpublished, see http://www.phrap.org). Sequence gaps were closed by primer walking or by sequencing PCR products that spanned small gaps. Additional sequence from the bovine genome sequencing project (http://www.ncbi.nlm.nih.gov/Traces/) facilitated gap closure and full assembly. Primers for sequencing BAC templates and PCR products were designed from unique sequence at the terminal ends of each assembly contig (Appendix A).

Internal BAC sequencing reactions were modified from the standard ABI reaction for BAC scale templates as follows: 2 ul Big Dye, 2 ul Half Big Dye (Genetix USA), 0.5 ul 10X MasterAmp (Epicentre), 1 ul 10 mM primer, and 1 ug template DNA in a volume of 10 ul. The thermal profile for the sequencing reaction was 96 C for 2 min, followed by 99 cycles of 96C for 30 sec, 50 C for 4 min, and a final hold of 65 C for 15 min.

Sequencing by primer walking was continued until new sequence bridged to another assembly or failed to provide unique primers for continued walking. AssemblyLign (Accelerys) software was used for the final assembly of each BAC. Concordance of predicted and observed HindIII restriction fragment patterns was used to validate each BAC assembly, which were then combined using AssemblyLign to produce a full sequence contig of *BoLA IIb*.

Sequence annotation of BoLA IIb

BLAST (Altschul *et al.* 1990) alignments of the repeat masked, assembled *BoLA IIb* sequence against NCBI EST and non-redundant nucleotide databases were used to identify expressed sequences and other highly conserved regions likely to contain functional genes. All sequences with >85% identity to known coding regions in *HLA* were aligned to published mRNA and amino acid sequences using Pairwise BLAST, then passed through GENSCAN (Burge and Karlin 1997) and GenomeScan (Yeh *et al.* 2001) (http://genes.mit.edu) to predict the exon structures for each candidate gene. To confirm identity, we aligned predicted translations to published protein sequences of human, mouse, rat, and when possible, cow and sheep using T-Coffee (Notredame *et al.* 2000) (http://www.ch.embnet.org/software/TCoffee.html). The assembled sequence was

screened for repetitive elements with RepeatMasker (http://www.repeatmasker.org/) using the "–species cow" option to identify bovine interspersed repetitive elements (LINES and SINES). Sputnik (http://espressosoftware.com/) was used to identify simple sequence repeats (SSRs) with di- to penta nucleotide motifs/SSR. All of the annotation data generated for the sequencing project was stored in the General Feature Format (gff) (Reese *et al.* 2000), and loaded into a Gbrowse database (Stein *et al.* 2002).

Two regions of overlapping coverage were characterized during the assembly phase, including a 31 kb overlap between BAC 27C10R7 and BAC3 C8R7 and a 37kb overlap between 3C8R7 and 14-47C8R7. The sequence from each BAC template was used for both of the overlapping regions in order to characterize polymorphisms between the BAC clones.

Other sources of BoLA IIb sequence

The cow genome project version 2 is based on a 6X coverage whole genome shotgun sequence which is available on the Baylor College of Medicine, Human Genome Sequencing Center web site, and the version 2 assembly was last modified in 2005 (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/fasta/Btau20050310-freeze/). Ensembl BLAST (http://www.ensembl.org/Bos_taurus/blastview/) was used to determine the number and location of sequences corresponding to BoLA IIb. Once contigs and scaffolds were determined to lie in the BoLA IIb region, they were downloaded, and formatted for use in alignment and polymorphism studies.

A component of the genome sequencing project was the construction of a SNP panel (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/snp). This set of sequences was

constructed by shallow depth sequencing of several different breeds of cattle, in order to generate data about the level of diversity between the cattle breeds. The sequences were then compared to the genome sequence to determine the presence of high quality polymorphisms. As of the July 24, 2006 release, more than two million SNP sequences were generated from *in silico* analysis of the genome assembly, and 114,958 SNP sequences were found through a SNP discovery project that included six different breeds of cattle. All of the data is freely available through the Baylor College of Medicine FTP site (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/snp). Each of the sequences contains 250 bases flanking the polymorphic base, resulting in a sequence of 501 bases in length. An archive of the SNP sequences was downloaded from the Baylor College of Medicine FTP site, reformatted into FASTA format, and made into a BLAST database. This database was then used to screen the BoLA IIb sequence derived in Chapter II in order to determine the presence of SNP sequences that fall in the BoLA IIb region. The BoLA IIb sequence had repetitive elements masked out via RepeatMasker using the cow species library. The SNP database was screened for BoLA IIb sequence using BLAST with a minimum word size of 200 (–W 200), in order to limit the results to nearly perfect matches.

Polymorphism detection

All of the sequences were manually aligned using the program Sequencher. Large gaps between contigs in the scaffolds required the manual gap insertion and positioning of the sequences into a new contig. This new contig was used to determine the positions, allowing for more precise alignments between the sets of sequences. The regions of

alignment between the sequences were saved into separate fasta files. Each pair of sequences was then globally aligned using Clustalw (http://www.ebi.ac.uk/clustalw/). A short PERL script was used to strip out the lines of the alignment with perfect identity, dramatically reducing the amount of data to process (source and sample input/output in appendix B). Scoring the numbers of mismatching bases was performed manually for each alignment.

A Multipipmaker alignment was constructed, using the BoLA IIb contig and the Ensembl scaffolds. The previously determined BoLA IIb sequence was used as the base for the alignment, including exon positions and RepeatMasker output data. All sequences were aligned using the "chaining" option (appendix D).

Each of the regions was then screened for simple sequence repeats (SSR) using SPUTNIK (As described in Chapter II). The data generated from this process were manually compared between the regions for each sequence to determine whether variant microsatellites exist. The output for each sequence was compared to its counterpart, and differences in motif, motif length, and overall composition were determined.

The data generated from the bovine SNP detection project includes over two million SNP sequences derived from the genome sequence and nearly 115,000 SNP sequences derived from interbreed whole genome shotgun sequencing. The data was reformatted from a comma delimited file to a file of fasta formatted sequences using a short PERL script (source and sample input/output in appendix C). The sequences were then made into a BLAST database using the program formatdb with the following options (-p F –i bovine-snps2-nr.fa –n mappedSNPs –t mappedSNPs). The query sequence was masked using RepeatMasker with the following options selected: " –

species cow, -pa 2 –xsmall". The option "xsmall" masks the repeat sequence using lowercase letters instead of N or X. BLASTN was then used to align the BoLA IIb sequence to the SNP database using options to limit the hits to long, nearly identical matches (-U –W 200 –F f).

Results

Sequence similarity between assemblies

The release of a cow genome sequence assembly opens up new possibilities for determining polymorphism rates and haplotype definition in cattle. A previously determined sequence of BoLA IIb was used to screen the bovine genome project sequence assembly V2.0 (http://www.ensembl.org/Bos_taurus/), revealing four scaffolds that span the length of BoLA IIb: Chr23.10, Chr23.11, Chr23.12, and Chr23.13. The version 2.0 assembly is in general agreement with the TAMU BoLA IIb sequence except that the positions of scaffolds 23.11 and Chr23.12 appear to be inverted in comparison to the BoLA IIb sequence (Figure 5). The Multipipmaker alignment revels that the bulk of the BoLA IIb sequence is in agreement with the version 2.0 assembly. There are numerous regions of highly polymorphic sequence and the gaps in the scaffolds produce some large gaps in the alignment. The Multipipmaker alignment verified the manually constructed contig (Appendix C).

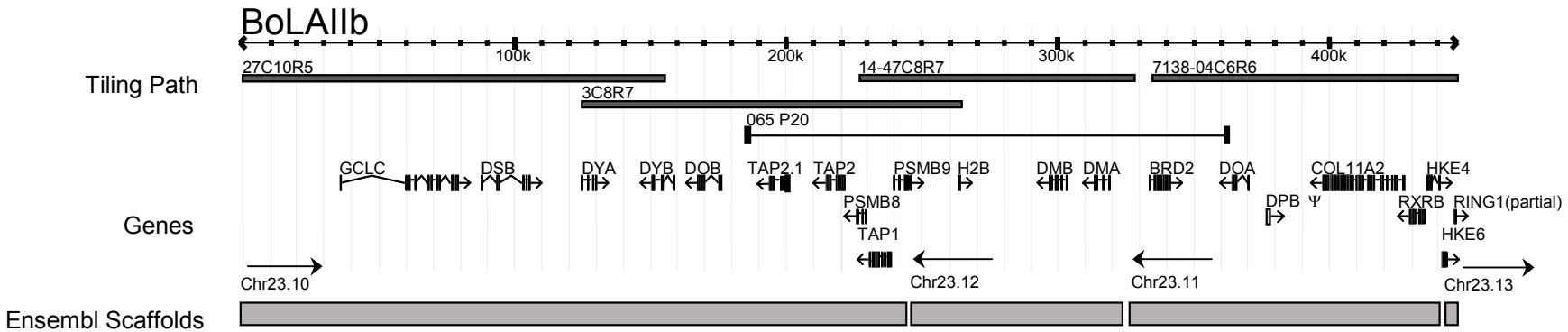


Figure 5. Overview of the BoLA IIb region. This map includes data from the Bovine Genome sequencing project. Tiling path: source of sequence used for annotation. Genes: drawn to scale, rectangles mark the presence of exons, bent lines denote introns, and arrows indicate direction of transcription. Ensembl Scaffolds: Scaffolds corresponding to BoLA IIb from the Ensembl cow assembly, Build 2.0.

A global alignment of each overlapping region was performed using Clustal. These sequences are approximately 432,500 bases in total length, and reveal that the overall alignments range from 97.7% to 99.7%, and include large regions of perfect identity (Table 3). The differences between the sequences may be due to many factors, and may be due in part to breed specific or animal specific variants. Some of the differences may be due to sequencing or assembly errors. These types of errors should be most prevalent near the ends of the contigs. The sequences at the ends of a contig typically exhibit a lower coverage depth, which may increase the odds of a miscalled sequence being displayed. As the whole genome sequence coverage is improved and subsequent assemblies are released, the number of differences should decrease with each new assembly.

Table 3. Global alignments between overlapping BoLA IIb sequences, regions are distinguished by shading. Length indicates length of sequence corresponding to the overlap, Mismatches indicates total number of mismatching bases between the two sequences, % Different, % Identity refer to the percentage difference and identity between the two sequences, respectively.

Region	Length	Mismatches	% Different	% Identity
BoLA IIb	246336	3515	1.427	98.573
Chr23.10	229532	3515	1.531	98.469
BoLA IIb	112877	2555	2.264	97.736
Chr23.11	112364	2555	2.274	97.726
BoLA IIb	71798	433	0.603	99.397
Chr23.12	73641	433	0.588	99.412
BoLA IIb	1511	5	0.331	99.669
Chr23.13	1512	5	0.331	99.669
27R5C10	31358	456	1.454	98.546
3C8R7	31415	456	1.452	98.548
3C8R7	37436	432	1.154	98.846
14-47C8R7	37434	432	1.154	98.846

SPUTNIK similarity analysis

The release of the cow genome assembly has allowed for rapid *in silico* detection of microsatellite sequences. The comparison between BoLA IIb and the Ensembl scaffolds revealed a total of 331 microsatellites spread across all four of the ovine genome project scaffolds. The majority of the repeats are shared and monomorphic between the BoLA IIb contig and the Ensembl scaffolds (Table 4). Out of a total of 55 microsatellites not shared between the two sets of sequences, 44 are only present in the BoLA IIb sequence and 11 are present only in the Ensembl sequence. The bulk of the discrepancies in SSR composition between the sequences is likely due to the limitations of SPUTNIK to recognize a repeat below a certain length. SPUTNIK has a minimum threshold of twelve bases for recognizing tandemly repeated sequence. Specifically, a minimum of six tandem motifs are required to recognize dinucleotide repeats, while four or more tandem motifs are required to recognize a trinucleotide repeat. Additionally, the Ensembl scaffolds are comprised of contigs separated by runs of Ns to denote the contig boundaries. Any data relating to the SSRs in these regions is currently unavailable, which could account for SSRs that are absent in the Ensembl sequence when compared to the corresponding sequence in the original BoLA IIb contig.

It is also possible to identify polymorphisms by analyzing overlapping sequence derived from different BAC clones present in the TAMU BoLA IIb assembly. Analysis of the two overlapping regions present in the TAMU BoLA IIb sequence revealed the presence of 42 total microsatellites, with 21 repeats in each region (Table 4). The overlap

spanning BACs 27C10R5 and 3C8R7 contains 20 shared SSR, 16 of which are monomorphic, and a single pentanucleotide repeat that is present in 3C8R7 and is absent

Table 4. Summary of simple sequence repeat polymorphisms between overlapping BoLA IIb sequences. Region name indicates the region where the overlap occurs. Length refers to the length of the overlap. Motif, Shared Mono-, Shared Poly-, Total # Shared, # Not Shared, Total indicates the motif length, number of monomorphic repeats shared between the sequences, number of polymorphic repeats shared between the sequences, total number of repeats shared between the sequences, number of sequences not shared between the sequences and total number of SSR sequences in the overlapping region, respectively.

Region Name	Length	Motif	Shared Mono-	Shared Poly-	Total # Shared	# Not Shared	Total
Chr32.10	**246336**	**Total**	**62**	**13**	**75**	**43**	**118**
		Di-	11	8	19	22	***41***
		Tri-	20	3	23	10	***33***
		Tetra-	15	0	15	4	***19***
		Penta-	16	2	18	7	***25***
Chr23.11	**112877**	**Total**	**40**	**11**	**51**	**6**	**57**
		Di-	13	10	23	3	***26***
		Tri-	8	0	8	1	***9***
		Tetra-	9	0	9	1	***10***
		Penta-	10	1	11	1	***12***
Chr23.12	**73641**	**Total**	**43**	**6**	**49**	**6**	**55**
		Di-	12	6	18	1	***19***
		Tri-	13	0	13	0	***13***
		Tetra-	9	0	9	1	***10***
		Penta-	9	0	9	4	***13***
Chr23.13	**1512**	**Total**	**1**	**0**	**1**	**0**	**1**
		Di-	1	0	1	0	***1***
		Tri-	0	0	0	0	***0***
		Tetra-	0	0	0	0	***0***
		Penta-	0	0	0	0	***0***
27C10R5 to 3C8R7	**31415**		**16**	**4**	**20**	**1**	**21**
		Di-	6	2	8	0	***8***
		Tri-	6	1	7	0	***7***
		Tetra-	2	0	2	0	***2***
		Penta-	2	1	3	1	***4***
3C8R7 to 14-47C8R7	**37436**		**16**	**2**	**18**	**3**	**21**
		Di-	5	1	6	1	***7***
		Tri-	6	0	6	0	***6***
		Tetra-	3	0	3	2	***5***
		Penta-	2	1	3	0	***3***

in 27C10R5. The overlap between BAC 3C8R7 and 14-47C8R7 contained 18 repeats that were shared between the two BACs, 16 of which were monomorphic, one dinucleotide

repeat present in BAC 14-47C8R7. Interestingly, both BAC sequences contained differing tetranucleotide repeats at the same position. In BAC 3C8R7, the repeat is made out of four repetitions of the motif “AAAT”, while in BAC 14-47C8R7 the repeat is comprised of three repetitions of the motif “ACAT”. BAC 14-47C8R7 derived from a longhorn cow, while both BAC 27C10R5 and BAC 3C8R7 originated from an Angus bull. The presence of polymorphic sequences indicates that while BACs 27C10R5 and 3C8R7 originated a single animal, one BAC was derived from one copy of chromosome 23, while the other originated with the other chromosome.

Identification of SNP sequences in Bola IIb

BLAST was used to screen the SNP dataset for sequences matching to BoLA IIb, resulting in thirteen full length matches. The distance between these sequences as mapped to the cow assembly is consistent with the corresponding distances in BoLA IIb. The search resulted in SNP sequences for three of the four scaffolds. The SNP sequences present in Chr23.10 are consistent in relative position and orientation, while the SNP sequences matching to Chr23.12 are consistent in relative distance; however the order is reversed in comparison to the BoLA IIb sequence (Table 5). The thirteen SNP sequences are dispersed throughout the BoLA IIb sequence. Nine of the thirteen sequences correspond to sequence from scaffold Chr23.10, three sequences are located in Chr23.11, and one sequence lies within Chr23.12. Six of the sequences are in intergenic regions, and six of the SNPs are in intronic sequence. Most interestingly, one of these SNP positions is present one base outside exon one of DSB, though neither of the observed variants disrupts the splice site. There is a single SNP that lies within exon 10 of BRD2

resulting in a glutamic acid to lysine change. The SNP sequences are spaced an average of 33.5 kbp apart, with a minimum of fourteen and a maximum of 93,038 bases separating SNPs. Five of the SNPs alter restriction enzyme recognition sites, including the polymorphic site in BRD2 exon 10.

Table 5. Mapping validated SNP sequences to BoLA IIb. SNP refers to the name of the SNP sequence, Map pos, Distance indicate the map position on the chromosome and distance between the adjacent sequences. BoLA Start, BoLA End, BoLA Distance refer to the start and end positions in the TAMU BoLA IIb assembly, and the distance between adjacent positions. Scaffold indicates the scaffold name in cow assembly build 2 the SNP sequence is present in.

SNP	Map pos	Distance	BoLA Start	BoLA End	BoLA Distance	Scaffold
BTA-57080	5358566		61982	62482		Chr23.10
		196			196	
BTA-57081	5358762		62178	62678		Chr23.10
		86612			88540	
BTA-57083	5445374		150717	151218		Chr23.10
		53			53	
BTA-57082	5445427		150770	151271		Chr23.10
		81692			79447	
BTA-57088	5527119		230218	230718		Chr23.10
		343			343	
BTA-57087	5527462		230561	231061		Chr23.10
		104			104	
BTA-57086	5527566		230665	231165		Chr23.10
		18			18	
BTA-57085	5527584		230683	231183		Chr23.10
		80			80	
BTA-57084	5527664		230763	231263		Chr23.10
		160052			83096	
BTA-57098	5687716		313859	314359		Chr23.12
		-208			208	
BTA-57097	5687508		314067	314567		Chr23.12
		-230			230	
BTA-57096	5687278		314297	314797		Chr23.12
		-4230			4302	
BTA-57093	5683048		318599	319099		Chr23.12
		-3			3	
BTA-57094	5683045		318602	319102		Chr23.12
		-376			377	
BTA-57095	5682669		318978	319479		Chr23.12

Discussion

The ability to screen for microsatellite sequences *in silico* enables marker development in sequences that are more likely to contain detectable variants with a minimum of sequence data. The more sequences that are available for comparison, the more likely it is to discover variations *in silico.* This allows for an initial enrichment for potential markers in sites known to be polymorphic. A similar approach may be used in the creation of SNP based markers. By determining what SNP sequences are present in the region of interest, PCR primers designed with both variants of the SNP at the terminal end of one of the primers would allow for testing for a specific allele depending on which of the primers was used. By utilizing preexisting sequence data, the development time for informational markers could be reduced from weeks to days.

The importance of the MHC in the immune response makes it an important region for developing informative markers. Developing an evenly distributed marker panel is greatly simplified with genomic sequence, making marker selection a matter of screening the sequence in regions of interest, and repeating until sufficient marker density has been obtained.

Determining haplotype structure is useful in terms of defining a minimal set of markers required for characterizing an individuals complement of alleles in a given haplotype block. One of the factors in determining haplotype structure is an overabundance of markers in the region of interest, in order to define the blocks as clearly as possible, and marker selection is facilitated by the availability of sequence. After marker variants are linked to alleles of interest, such as resistance to various diseases, determining whether an animal holds multiple resistance alleles becomes a matter of

screening for previously defined alleles. Once animals are found to contain a set of alleles of interest, the development of a mating system to maximize the benefits of this information will be simplified with the understanding of the haplotype structure.

Developing robust maps based on genomic sequence information provides the framework for specifically targeting regions of interest for further studies. This dramatically increases the speed of marker development, while allowing for data mining techniques to screen for polymorphisms *in silico*. Selecting for primers already known to contain informative polymorphisms allows for the concentration of resources in directions likely to yield results immediately.

CHAPTER IV

TAMU DERIVED BOLA IIB VERSUS BOVINE GENOME PROJECT BOLA IIB

Introduction

The bovine genome project has released an initial draft sequence based upon a 6x whole genome sequence assembly (http://www.ensembl.org/Bos_taurus/). This assembly, designated V2.0 contains four scaffolds that spans the section of cow chromosome 23 previously determined to contain the BoLA class IIb region. According to the assembly data, these four scaffolds have been positioned and oriented based on marker data. When this assembly was compared to the BoLA IIb sequence as determined in Chapter II, inconsistencies in organization were discovered (Figure 5), though the sequence organization is consistent within each scaffold (Appendix D). The difference between the two assemblies poses a problem: which of the sequences is correct, or if both are incorrect or correct, why is that the case?

Materials and Methods

Source of sequences

Bacterial artificial chromosome (BAC) clones containing BoLA sequences were isolated from a composite TAMU bovine BAC library. The tiling path was determined, and cloned were sequenced, assembled and annotated as described in Chapter II. The bovine genome project has made the second version of the genome assembly publicly available through several sources. BLAST alignments made using the BoLA IIb contig

through the Ensembl genome browser (http://www.ensembl.org/Bos_taurus/) facilitated the detection of scaffolds spanning the BoLA IIb region.

Defining differences between the assemblies

The two sequence assemblies were compared using four methods: local alignments via Ensembl and pairwise BLAST, global alignments via ClustalW and multipipmaker, as described in Chapter III. The sequences were then aligned using dotter (Sonnhammer and Durbin 1995) to produce an exhaustive sequence alignment using the default parameters.

Determining the validity of each sequence assembly

The BAC clones were designated as belonging to BoLA through probe hybridization, then a minimum tiling path was selected by utilizing the program FPC (Soderlund *et al.* 1997) to analyze the restriction fingerprint maps to determine the nature and extent of BAC clone overlap (Newkirk, unpublished). The BAC clones used to make the BoLA IIb tiling path were checked against the internet contig explorer (iCE), revealing that two of the four clones, 27R5C10 and 3R7C8 were found to be in contig 4615. The other two clones were not part of the iCE data set. BAC 065 P20 was selected for inclusion in the project to span a gap in the original BAC scaffold. Probes corresponding to the genes flanking the gap were constructed and used to screen the CHORI 240 library. After selection of the BAC clone 065 P20 to span the gap, further confirmation was performed via end sequencing, restriction fingerprinting, and screening the iCE database. End sequencing and restriction fingerprinting confirmed the overlap

between 065 P20 and the BoLA IIb scaffold, and the iCE database revealed that 065 P20 is also present in contig 4615.

Once the BACs were sequenced, the individual sequences were assembled both as individual projects, and as a single project. Each of the individually derived partial BAC sequences was compared to the other BAC sequences to determine the extent of overlap, and to verify the multiple BAC assembly.

The BCM sequenced a number of BACs at low coverage, to include in the bovine genome project. These BAC skims included the BAC from the CHORI library that was selected to fill the tiling path gap in the BoLA IIb project. The BAC skim was composed of nineteen unordered contigs, and was deposited on GenBank under accession AC156977.

All analysis performed using data from the bovine genome project was from build 2, a whole genome assembly constructed from 6X coverage whole genome shotgun sequence. All of the data is available through the Baylor College of Medicine (http://www.hgsc.bcm.tmc.edu/projects/bovine/). Clones used for sequencing were selected from multiple libraries with insert sizes ranging from 2-4 kb and 4-6 kb. All clones used in the project were sequenced bi-directionally, and each sequence was logically connected to the sequence derived from the other end sequencing reaction and the estimated insert size for that clone. This paired end information was used in the assembly phase to connect contigs that were connected by subclones, but not by contiguous sequence. These groups of non-contiguous sequence sequences were designated as scaffolds. The contigs are positioned and oriented based on paired end data.

Gaps between contigs were padded with varying amounts of the letter "N", based on estimated gap size (http://www.hgsc.bcm.tmc.edu/projects/bovine/).

Following the whole genome shotgun phase, the bovine genome project sequenced a tiling path of BACs clones to a low coverage. These "BAC skims" were then used as a scaffold to assemble the sequences generated from the shotgun sequencing phase. Each of the BAC clones selected for the tiling path were restriction mapped, and are present on the iCE browser.

Results and Discussion

The release of the initial draft genome sequence has provided a second assembly for the BoLA IIb region. When taken individually, the scaffold sequences corroborate the sequence generated from the TAMU BAC library, as noted in Chapter III, and appendix D. When the scaffolds are arranged according to the sequence map generated by the BCM, the organization of the two sequences differs (Figure 5). The internal consistency of each scaffold when compared to the TAMU BoLA IIb sequence helps to confirm both assemblies. The lack of agreement between the chromosomal order and orientation of the bovine genome project scaffolds and the TAMU BoLA IIb sequence may be due to positioning errors in the assembly phase. The differences in organization do include two of the four scaffolds that span the BoLA IIb region, and roughly 190 kb (Figure 5).

The next version of the assembly will be built from a combination of whole genome shotgun and low coverage BAC sequence, and should provide an improved overall assembly.

Once the new assembly is released, the region containing BoLA IIb may be significantly different. The addition of BAC sequence to the whole genome sequence should assist in the assembly and organization of the bovine genome sequence. Since each of the BACs sequenced was selected as part of a tiling path, its relative position in the genome is already known, causing the sequence to be clustered both by belonging to a particular BAC, and that BAC belonging to a particular region of the genome. The sequences corresponding to each of the BAC regions can also be verified by comparing the restriction map of the assembled sequence to the observed restriction map of the BAC for that region.

CHAPTER V

SUMMARY AND CONCLUSION

The MHC has been studied in a wide variety of vertebrates, and among the mammals, it exhibits a high degree of conservation in most species. The MHC of ruminants has a markedly different organization when compared to other species, in that the class II region has been disrupted, giving rise to two subregions, each of which contains functional class II genes. The class IIa regions is located adjacent to the class III region, and contains the DQ- and DR- class II gene families. The class IIb region has biological significance in that it contains ruminant specific class II genes, as well as the antigen processing genes required for the preparation and loading of antigens onto antigen presenting peptides.

The primary goal of this project was to characterize BoLA IIb at the DNA sequence level. Having the full genomic sequence of BoLA IIb provides a solid foundation for further studies, including feature annotation, marker identification and polymorphism discovery. Each of these secondary goals is greatly simplified once the genomic sequence has been defined, and each topic has provided some very interesting results.

After the sequence assembly phase of the project was completed, the BoLA IIb sequence was analyzed to characterize a wide variety of features. Through a variety of methods, the expressed genes in BoLA IIb were defined, and the coding sequences were annotated. The gene characterization phase included the use of several computational tools, including prediction tools such as GENSCAN and GenomeScan, and alignment to

known coding sequences with BLAST and Spidey. In addition to the genes that had previously been mapped to BoLA IIb, the genomic sequence revealed three genes that had not previously been detected, and are not present in the class II regions of any other characterized MHC. One of these genes is a class II gene designated DSB, which is most similar to DRB. The structure of DSB is similar to other DRB genes, with the exception of a divergent exon 5 and the absence of exon 6. This gene had not been previously characterized, and the name DSB was recommended by the BoLA Nomenclature Committee, when it was determined to be sufficiently divergent from other class II genes to warrant a new designation. DSB is the most centromeric of the BoLA genes, and therefore marks the boundary of the BoLA IIb region proximal to the inversion breakpoint.

A second unexpected result from the BoLA IIb sequence was the discovery of a 10 kb section of BoLA IIb containing the antigen processing gene TAP2 has been duplicated in BoLA IIb, giving rise to the gene TAP2.1. TAP2 and TAP2.1 have 98% sequence identity between the sequences of the two genes, although TAP2.1 does contain a SNP at the junction of intron I, which would result in a variant transcript that has not been observed in the EST database. The bovine genome project sequence contains a gap in the region of the TAP2.1 duplication, and a second sequence with high similarity to TAP2 in an undefined chromosomal scaffold, indicating that the duplication is present in both Angus and Hereford breeds. PCR amplification of TAP2.1 was successful in bison and cattle, but not in deer, sheep or goats indicating that this duplication likely occurred before speciation events giving rise to cattle and bison, and after the separation of the bovine ancestor from the other ruminants.

The third discovery that resulted from the sequence analysis was a single histone gene, H2B. Histone genes are intronless, and typically occur in large clusters. While no histone had been previously reported in any of the classical MHC regions, histone clusters have been observed in the extended class I region. The coding region of H2b appears intact, and the high copy number of histone genes coupled with the high level of conservation between the genes makes it difficult to determine the functional status of H2B.

Another interesting feature that was uncovered during the annotation phase of the project was the location of the inversion breakpoint that gave rise to the BoLA IIb region. The boundary of BoLA IIb is demarcated by the presence of DSB, the most centromeric BoLA gene. Roughly 9 kb upstream of DSB is the non-MHC gene GCLC, bracketing the region containing the inversion breakpoint.

Further analysis to refine the breakpoint region was performed, using multipipmaker to align multiple sequences to the BoLA sequence. These sequences included class II genomic MHC sequence in human, cat, mouse and rat, and GCLC genomic sequence in human, mouse and rat. The resulting alignment refined the breakpoint region to a 3.2 kb region that lacks observable similarity to the corresponding regions in other genomes. This window of sequence is composed of nearly 66% repetitive sequence including a recognition motif for *translin,* a protein that binds to specific sequence motifs that are located at DNA ends (Aoki *et al.* 1995). T*ranslin* binding sites have been found directly proximal to breakpoint sites, and have also been associated with recombination hotspots, and chromosomal translocations in lymphoid neoplasms (Aoki *et al.* 1995; Kasai *et al.* 1997; Hosaka *et al.* 2000). To date, very few non-primate

chromosomal breakpoints have been characterized at the DNA sequence level, and as more are discovered in various species, the breakpoint region in BoLA IIb may be further refined. The inability to define the breakpoint to a region smaller than 3.2 kb may indicate that further alteration of the genome sequence occurred proximal to the breakpoint site since the inversion occurred.

It is interesting to note that the three ruminant specific class II genes, DSB, DYA, and DYB are all located within 50 kb of the inversion breakpoint. The genes most closely related to DSB, DYA, and DYB in HLA are DRB1, DQA1 and DQB1, respectively.

The fact that BoLA IIa contains the DR- and DQ- genes in BoLA indicates that the region was duplicated before the inversion occurred, and at least part of the duplicated region was included in the rearrangement. The three genes in BoLA IIb could accrue mutations, as the three genes remaining in the BoLA IIa region remained fully functional. Over time, the genes were altered enough that DSB became structurally divergent from DRB, and the DY- genes became functionally distinct from the DQ- genes. This appears to be a ruminant specific example of Nei's birth and death model for gene evolution.

The annotation phase of the project was initially focused on gene annotation, then was expanded to include the defining of position, composition, and if applicable orientation of repetitive elements in BoLA IIb. This analysis included both simple sequence repeats of motifs ranging from two to five bases, as well as larger, more complex repeats including SINEs and LINEs and LTRs. Nearly 40% of BoLA IIb is comprised of repetitive sequences. Microsatellite sequences comprise 0.89% of BoLA IIb, most of which is present as dinucleotide repeats. The identification and annotation of

microsatellite sequences is of significance in that it allows for the rapid selection of markers that have a greater chance of being polymorphic. With an increase of available genomic sequence, the development of markers can be done *in silico*, dramatically reducing the time it takes to define markers, select primers, and test for specificity.

The release of build 2 of the bovine genome project allowed for confirmation of the BoLA IIb assembly, and provides additional resources to extensively screen for polymorphisms across the BoLA IIb region. The polymorphism discovery phase of the project was divided into two sections: SNP detection, and motif length polymorphisms among the microsatellites. When the TAMU BoLA IIb sequence was globally aligned to the scaffolds released by the bovine genome project, the two sets of sequences had between 97.7% and 99.7% identity. Two regions of overlapping coverage in the TAMU BoLA IIb sequencing project spanned approximately 31kb and 37kb, and analysis of both regions revealed that the overlapping BAC sequence was more than 98% the same between the two BACs in both regions. Analysis of microsatellites revealed a total of 231 SSR sequences between the two assemblies, 30 of which were polymorphic, and 55 of which were only present in one of the two assemblies. The differences in SSR composition between the two assemblies may be due an artifact of the draft sequence released by the bovine genome project. Gaps between scaffolds and within each scaffold are delineated by long runs of the letter N, and the determination of any data within the gaps is dependant on future assemblies. Another possible explanation for the presence of a microsatellite in only one sequence is that the repeat was reduced to the point that it was not recognizable in the other sequence. When the two overlapping regions in the

TAMU BoLA IIb sequence were screened, 42 SSR were found, six of which were polymorphic, and four of the microsatellites were found in only one of the sequences.

While comparisons between the two assemblies were used to discover polymorphisms, comparison was also used to confirm the assembly of both projects. One of the outcomes of the global alignments between the two assemblies was the confirmation that the assemblies were in agreement within each scaffold. The whole chromosome organization is such that the region spanning scaffolds Chr23.11 and Chr23.12 are inverted when compared to the TAMU BoLA IIb sequence. Previous mapping data and DNA fingerprint data for the BoLA IIb BACs support the TAMU assembly.

By characterizing the BoLA IIb region, this project has provided a solid foundation for future studies of the bovine MHC, including the defining of the haplotype structure of BoLA IIb. The process of discovering markers to be used in such studies is greatly simplified with the availability of genomic sequence that has been annotated for features of interest. SSR sequences may be selected by relative position to ensure even and thorough coverage, or by motif, to maximize the probability of finding motif length polymorphisms. The availability of SNP sequences from the SNP portion of the bovine genome project give us the ability to quickly screen the validated SNP sequence for matches to the BoLA IIb region. Additionally, *in silico* definition of polymorphisms can quickly provide regions of interest to focus on. The availability of whole genome sequence in cow could also facilitate further studies in other related species, and could be used in evolutionary studies to refine the period during which the inversion that gave rise to the IIb region occurred.

REFERENCES

Abi-Rached L., Gilles A., Shiina T., Pontarotti P., Inoko H. (2002) Evidence of en bloc duplication in vertebrate genomes. *Nature Genetics*. **31** (1), 100-5.

Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Amadou C. (1999) Evolution of the Mhc class I region: the framework hypothesis. *Immunogenetics*. **49** (4), 362-7.

Andersson B., Povinelli C.M., Wentland M.A., Shen Y., Muzny D.M. and Gibbs R.A. (1994) Adaptor-based uracil DNA glycosylase cloning simplifies shotgun library construction for large-scale sequencing. *Analytical Biochemistry* **218**, 300-308.

Andersson L. and Rask L. (1988) Characterization of the MHC class II region in cattle. The number of DQ genes varies between haplotypes. *Immunogenetics* **27**, 110-120.

Andersson L., Lunden A., Sigurdardottir S., Davies C.J. and Rask L. (1988) Linkage relationships in the bovine MHC region. High recombination frequency between class II subregions. *Immunogenetics* **27**, 273-280.

Aoki K., Suzuki K., Sugano T., Tasaka T., Nakahara K. and Kuge O. (1995) A novel gene, *Translin*, encodes a recombination hotspot binding protein associated with chromosomal translocations. *Nature Genetics*. **10** (2), 167-74.

Ballingall K.T., MacHugh N., Taracha E., Mertens B. and McKeever D. (2001) Transcription of the unique ruminant class II major histocompatibility complex-

DYA and DIB genes in dendritic cells. *European Journal of Immunoogy*. **31**, 82-86.

Ballingall K.T., Ellis S.A., MacHugh N.D., Archibald S.D. and McKeever D.J. (2004) The DY genes of the cattle MHC: expression and comparative analysis of an unusual class II MHC gene pair. *Immunogenetics* **55**, 748-755.

Band M., Larson J.H., Womack J.E. and Lewin H.A. (1998) A radiation hybrid map of BTA23: identification of a chromosomal rearrangement leading to separation of the cattle MHC class II subregions. *Genomics* **53**, 269-275.

Beck S., Hanson I., Kelly A., Pappin D.J., and Trowsdale J. (1992) A homologue of the *Drosophila* female sterile homeotic (fsh) gene in the class II region of the human MHC. *DNA Sequence*. **2** (4), 203-10.

Beck S. and Trowsdale J. (2000) The human major histocompatability complex: lessons from the DNA sequence. *Annual Review of Genomics and Human Genetics*. **1**, 117-37.

Beck T., Menninger J., Murphy W., Nash W., O'Brien S., and Yuhki N. (2005) The feline major histocompatibility complex is rearranged by an inversion with a breakpoint in the distal class I region, *Immunogenetics* **56**, 702 – 709.

Belov K., Deakin J.E., Papenfuss A.T., Baker M.L., Melman S.D., Siddle H. V., *et al*. (2006) Reconstructing an ancestral mammalian immune supercomplex from a marsupial major histocompatibility complex. *PLoS Biology* **4**(3), e46

Belov K., Lam M.K., and Colgan D.J. (2004) Marsupial MHC class II ß genes are not orthologous to the eutherian ß gene families. *Journal of Heredity* **95**, 338-345

Benaroch P., Yilla M., Raposo G., Ito K., Miwa K., Geuze H.J., *et al*. (1995) How MHC class II molecules reach the endocytic pathway. *EMBO Journal*. **14** (1), 37-49.

Bodmer W.F. (1987) The HLA system: structure and function. *Journal of Clinical Pathology*. **40** (9), 948-58.

Burch G.H., Gong Y., Liu W., Dettman R.W., Curry C.J., Smith L., *et al*. (1997) Tenascin-X deficiency is associated with Ehlers-Danlos syndrome. *Nature Genetics*. **17**, 104-108.

Burge C. and Karlin S. (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**, 78-94.

Cai L., Taylor J.F., Wing R.A., Gallagher D.S., Woo S.S. and Davis S.K. (1995) Construction and characterization of a bovine bacterial artificial chromosome library. *Genomics* **29**, 413-425.

Chardon P., Rogel-Gaillard C., Peelman L., Yerle M., Velten F.W., Renard C., *et al*. (1999) Physical organization of the pig major histocompatibility complex class II region. *Immunogenetics* **50**, 344-348.

Chen B.P., and Parham P. (1989) Direct binding of influenza peptides to class I HLA molecules. *Nature* **337**, 743 – 745.

Coussens P.M., Colvin C.J., Wiersma K., Abouzied A., and Sipkovsky S. (2002) Gene expression profiling of peripheral blood mononuclear cells from cattle infected with *Mycobacterium paratuberculosis*. *Infection and Immunity*. **70**(10): 5494-5502.

Daly M.J., Rioux J.D., Schaffner S.F., Hudson T.J., and Lander E.S. (2001) High-resolution haplotype structure in the human genome. *Nature Genetics*. **29** (2), 229-32.

Dausset J. (1958) Iso-leuco-antianticorps *Acta Haematologica*. **20**(1-4), 156-66.

Dawkins R., Leelayuwat C., Gaudieri S., Tay G., Hui J., Cattley S., *et al*. (1999) Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. *Immunological Reviews*. **167**, 275-304.

Debenham S.L., Hart E.A., Ashurst J.L., Howe K.L., Quail M.A., Ollier W.E., *et al*. (2005) Genomic sequence of the class II region of the canine MHC: comparison with the MHC of other mammalian species. *Genomics*. **85**(1), 48-59.

de la Salle H., Hanau D., Fricker D., Urlacher A., Kelly A., Salamero J., *et al*. (1994) Homozygous human TAP peptide transporter mutation in HLA class I deficiency. *Science* **265**, 237-241.

Deng G.Y., Muir A., Maclaren N.K., and She J.X. (1995) Association of LMP2 and LMP7 genes within the major histocompatibility complex with insulin-dependent diabetes mellitus: population and family studies. *American Journal of Human Genetics*. **56** (2), 528-34.

Dennehey B.K., Gutches D.G., McConkey E.H. and Krauter K.S. (2004) Inversion, duplication, and changes in gene context are associated with human chromosome 18 evolution. *Genomics* **83,** 493-501.

Ewing B. and Green P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**, 186-194.

Ewing B., Hillier L., Wendl M.C. and Green P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8**, 175-185.

Fallas J.L., Tobin H.M., Lou O., Guo D., Sant'Angelo D.B., and Denzin L.K. (2004) Ectopic expression of HLA-DO in mouse dendritic cells diminishes MHC class II antigen presentation. *Journal of Immunology*. **173** (3),1549-60.

Fitzgibbon J., Gillett G.T., Woodward K.J., Boyle J.M., Wolfe J., and Povey S. (1993) Mapping of RXRB to human chromosome 6p21.3. *Annals of Human Genetics*. **57** (Pt 3), 203-9.

Furukawa H., Murata S., Yabe T., Shimbara N., Keicho N., Kashiwase K., *et al.* (1999) Splice acceptor site mutation of the transporter associated with antigen processing-1 gene in human bare lymphocyte syndrome. *Journal of Clinical Investigation*. **103** (5), 755-8.

Globerman H., Amor M., Parker K.L., New M.I., and White P.C. (1988) Nonsense mutation causing steroid 21-hydroxylase deficiency. *Journal of Clinical Investigation*. **82** (1), 139-44.

Gold B., Merriam J.E., Zernant J., Hancox L.S., Taiber A.J., Gehrs K., *et al.* (2006) Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. *Nature Genetics*. **38** (4), 458-62.

Gorer P.A. (1937) The genetic and antigenic basis for tumor transplantation. *Journal of Pathology & Bacteriology*. **44**: 691-697.

Gustafson A.L., Tallmadge R.L., Ramlachan N., Miller D., Bird H., Antczak D.F.*, et al.* (2003) An ordered BAC contig map of the equine major histocompatibility complex. *Cytogenetic and Genome Research* **102**, 189-195.

Haghparast A., Wauben M.H., Grosfeld-Stulemeyer M.C., van Kooten P., and Hensen E.J. (2000) Selection of T-cell epitopes from foot-and-mouth disease virus reflects the binding affinity to different cattle MHC class II molecules. *Immunogenetics*. **51** (8-9), 733-42.

Harmon R.J. (1994) Physiology of mastitis and factors affecting somatic cell counts. *Journal of Dairy Science*. **77** (7), 2103-12.

Hill A.S., Foot N.J., Chaplin T.L. and Young B.D. (2000) The most frequent constitutional translocation in humans, the t(11;22)(q23;q11) is due to a highly specific alu-mediated recombination. *Human Molecular Genetics* **9**, 1525-1532.

Hirohata S., and Kikuchi H. (2003) Behcet's disease. *Arthritis Research & Therapy*. **5** (3), 139-46.

Horton R., Wilming L., Rand V., Lovering R.C., Bruford E.A., Khodiyar V.K., *et al.* (2004) Gene map of the extended human MHC. *Nature Reviews Genetics*. **5** (12), 889-99.

Hosaka T., Kanoe H., Nakayama T., Murakami H., Yamamoto H., Nakamata T., *et al.* (2000) Translin binds to the sequences adjacent to the breakpoints of the TLS and CHOP genes in liposarcomas with translocation t(12;6). *Oncogene*. **19** (50), 5821-5.

Huang D.F., Siminovitch K.A., Liu X.Y., Olee T., Olsen N.J., Berry C., *et al.* (1995) Population and family studies of three disease-related polymorphic genes in systemic lupus erythematosus. *Journal of Clinical Investigation*. **95** (4), 1766-72.

Hughes A.L., and Nei M. (1989) Evolution of the major histocompatibility complex: independent origin of nonclassical class I genes in different groups of mammals. *Molecular Biology and Evolution*. **6** (6), 559-79.

Hughes A.L., Yeager M., Ten Elshof A.E., and Chorney M.J. (1999) A new taxonomy of mammalian MHC class I molecules. *Immunology Today*. **20** (1), 22-6.

Hurt P., Walter L., Sudbrak R., Klages S., Muller I., Shiina T.*, et al.* (2004) The genomic sequence and comparative analysis of the rat major histocompatibility complex. *Genome Research* **14**, 631-639.

Iglesias A., Bauer J., Litzenburger T., Schubart A., and Linington C. (2001) T- and B-cell responses to myelin oligodendrocyte glycoprotein in experimental autoimmune encephalomyelitis and multiple sclerosis. *Glia* **36** (2), 220-34.

Itoh T., Watanabe T., Ihara N., Mariani P., Beattie C.W., Sugimoto Y., *et al.* (2005) A comprehensive radiation hybrid map of the bovine genome comprising 5593 loci. *Genomics*. **85**, 413-424.

Jacob C.O., Fronek Z., Lewis G.D., Koo M., Hansen J.A., and McDevitt H.O. (1990) Heritable major histocompatibility complex class II-associated differences in production of tumor necrosis factor alpha: relevance to genetic predisposition to systemic lupus erythematosus. *Proceedings of the National Academy of Science USA*. **87**(3):1233-7.

Karttunen J., Lehner P., Gupta S., Hewitt E., and Cresswell P. (2001) Distinct functions and cooperative interaction of the subunits of the transporter associated with antigen processing (TAP). *Proceedings of the National Academy of Science USA*. **98** (13), 7431-6.

Kasai M., Matsuzaki T., Katayanagi K., Omori A., Maziarz R.T., Strominger J.L. *et al.* (1997) The translin ring specifically recognizes DNA ends at recombination hot spots in the human genome. *Journal of Biological Chemistry*. **272** (17), 11402-7.

Kehrer-Sawatzki H., Schreiner B., Tanzer S., Platzer M., Muller S. and Hameister H. (2002) Molecular characterization of the pericentric inversion that causes differences between chimpanzee chromosome 19 and human chromosome 17. *American Journal of Human Genetics* **71**, 375-388.

Kelley J., Walter L. and Trowsdale J. (2005) Comparative genomics of major histocompatibility complexes. *Immunogenetics* **56**, 683-695.

Kaufman J., Milne S., Gobel T.W., Walker B.A., Jacob J.P., Auffray C. *et al.* (1999a) The chicken B locus is a minimal essential major histocompatibility complex. *Nature*. **401** (6756), 923-5.

Kaufman J., Jacob J., Shaw I., Walker B., Milne S., Beck S. *et al.* (1999b) Gene organisation determines evolution of function in the chicken MHC. *Immunological Reviews*. **167**, 101-17.

Kumánovics A., Madan A., Qin S., Rowen L., Hood L., and Fischer Lindahl K. (2002) QUOD ERAT FACIENDUM: sequence analysis of the H2-D and H2-Q regions of 129/SvJ mice. *Immunogenetics*. **54** (7), 479-89.

Kumánovics A., Takada T., and Lindahl K. (2003) Genomic organization of the mammalian MHC. *Annual Review of Immunology* **21**, 629-657

Kuroda N., Figueroa F., O'hUigin C., and Klein J. (2002) Evidence that the separation of Mhc class II from class I loci in the zebrafish, *Danio rerio*, occurred by translocation, *Immunogenetics*, **54**, 418 – 430.

Little C.C., and Tyzzer E.E. (1916) Further experimental studies on the inheritance of susceptibility to a transplantable carcinoma (JA) of the Japanese waltzing mouse. *Journal of Medical Research* **33,** 393-427.

Mallard B.A., Dekkers J.C., Ireland M.J., Leslie K.E., Sharif S., Vankampen C.L., *et al.* (1998) Alteration in immune responsiveness during the peripartum period and its ramification on dairy cow and calf health. *Journal of Dairy Science*. **81** (2), 585-95.

Mann A.J., Abraham L.J., Cameron P.U., Robinson W., Giphart M.J. and Dawkins R.L. (1993) The caprine MHC contains DYA genes. *Immunogenetics* **37**, 292-295.

Matsuo M.Y., Asakawa S., Shimizu N., Kimura H., and Nonaka M. (2002) Nucleotide sequence of the MHC class I genomic region of a teleost, the medaka (*Oryzias latipes*). *Immunogenetics*. **53** (10-11), 930-40.

McGuirt W.T., Prasad S.D., Griffith A.J., Kunst H.P., Green G.E., Shpargel K.B., *et al.* (1999) Mutations in COL11A2 cause non-syndromic hearing loss (DFNA13). *Nature Genetics* **23**(4), 413-9.

McShane R.D., Gallagher D.S., Jr., Newkirk H., Taylor J.F., Burzlaff J.D., Davis S.K., *et al.* (2001) Physical localization and order of genes in the class I region of the bovine MHC. *Animal Genetics* **32**, 235-239.

The MHC Sequencing Consortium. (1999) Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401**, 921-923.

Nei M. and Rooney A.P. (2005) Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics*. **39**, 121-52.

Nijman I.J., van Tessel P. and Lenstra J.A. (2002) SINE retrotransposition during the evolution of the Pecoran ruminants. *Journal of Molecular Evolution* **54**, 9-16.

Notredame C., Higgins D.G., and Heringa J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**, 205-217.

Ohta Y., Okamura K., McKinney E.C., Bartl S., Hashimoto K., and Flajnik M.F. (2000) Primitive synteny of vertebrate major histocompatibility complex class I and class II genes. *Proceedings of the National Academy of Science USA*. **97**(9), 4712-7.

Okamoto K., Makino S., Yoshikawa Y., Takaki A., Nagatsuka Y., Ota M., *et al.* (2003) Identification of I kappa BL as the second major histocompatibility complex-linked susceptibility locus for rheumatoid arthritis. *American Journal of Human Genetics*. **72**(2), 303-12.

Ozaki K., Ohnishi Y., Iida A., Sekine A., Yamada R., Tsunoda T., *et al.* (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*. **32** (4), 650-4.

Pal D.K., Evgrafov O.V., Tabares P., Zhang F., Durner M., and Greenberg D.A. (2003) BRD2 (RING3) is a probable major susceptibility gene for common juvenile myoclonic epilepsy. *American Journal of Human Genetics*. **73** (2), 261-70.

Park Y.H., Joo Y.S., Park J.Y., Moon J.S., Kim S.H., Kwon N.H., *et al.* (2004) Characterization of lymphocyte subpopulations and major histocompatibility complex haplotypes of mastitis-resistant and susceptible cows. *Journal of Veterinary Science*. **5**(1), 29-39.

Payne R. and Rolfs M.R. (1958) Fetomaternal leukocyte incompatibility. *Journal of Clinical Investigation*. **37**(12), 1756-63.

Pletcher M.T., Roe B.A., Chen F., Do T., Do A., Malaj E., *et al.* (2000) Chromosome evolution: the junction of mammalian chromosomes in the formation of mouse chromosome 10. *Genome Research* **10**, 1463-1467.

Rao V., Gao F., Chen B., Jacobs W.R., and Glickman M.S. (2006) Trans-cyclopropanation of mycolic acids on trehalose dimycolate suppresses *Mycobacterium tuberculosis* -induced inflammation and virulence. *Journal of Clinical Investigation*. **116**(6):1660-7.

Reese M.G., Hartzell G., Harris N.L., Ohler U., Abril J.F., and Lewis S.E. (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Research* **10**, 483-501.

Renard C., Hart E., Sehra H., Beasley H., Coggill P., Howe K., *et al.* (2006) The genomic sequence and analysis of the swine major histocompatibility complex. *Genomics*. **88** (1), 96-110.

Robinson J., Waller M.J., Parham P., de Groot N., Bontrop R., Kennedy L.J., *et al.* (2003) IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Research* **31**, 311-314.

Rudiger N.S., Gregersen N. and Kielland-Brandt M.C. (1995) One short well conserved region of Alu-sequences is involved in human gene rearrangements and has homology with prokaryotic chi. *Nucleic Acids Research* **23**, 256-260.

Schwartz S., Elnitski L., Li M., Weirauch M., Riemer C., Smit A., *et al.* (2003) MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Research* **31**, 3518-3524.

Shiina T., Shimizu S., Hosomichi K., Kohara S., Watanabe S., Hanzawa K., *et al.* (2004) Comparative genomic analysis of two avian (quail and chicken) MHC regions. *Journal of Immunology*. **172**(11), 6751-63.

Skow L.C. and Nall C.A. (1996) A second polymorphism in exon 2 of the BoLA-DYA gene. *Animal Genetics* **27**, 216-217.

Skow L.C., Snaples S.N., Davis S.K., Taylor J.F., Huang B. and Gallagher D.H. (1996) Localization of bovine lymphocyte antigen (BoLA) DYA and class I loci to different regions of chromosome 23. *Mammalian Genome* **7**, 388-389.

Smith W.P., Vu Q., Li S.S., Hansen J.A., Zhao L.P., Geraghty D.E. (2006) Toward understanding MHC disease associations: partial resequencing of 46 distinct HLA haplotypes. *Genomics*. **87**(5), 561-71.

Snell G.D. (1948) Methods for the study of histocompatibility genes. *Journal of Genetics*. **49**, 87-108.

Soderlund C., Longden I., and Mott R. (1997) FPC: a system for building contigs from restriction fingerprinted clones. *Computer Applications in the Biosciences* **291**, 523-535.

Sonnhammer E.L., and Durbin R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*. **167**(1-2), GC1-10.

Stein L.D., Mungall C., Shu S., Caudy M., Mangone M., Day A., *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Research*. **12**(10), 1599-610.

Stone R.T. and Muggli-Cockett N.E. (1990) Partial nucleotide sequence of a novel bovine major histocompatibility complex class II beta-chain gene, BoLA-DIB. *Animal Genetics* **21**, 353-360.

Sulston J., Mallett F., Staden R., Durbin R., Horsnell T., and Coulson A. (1988) Software for genome mapping by fingerprinting techniques. *Computer Applications in the Biosciences* **4**, 125-132.

Takada T., Kumanovics A., Amadou C., Yoshino M., Jones E.P., Athanasiou M., *et al.* (2003) Species-specific class I gene expansions formed the telomeric 1 mb of the mouse major histocompatibility complex. *Genome Research*. **13**(4), 589-600.

Takahashi K., Rooney A.P. and Nei M. (2000) Origins and divergence times of mammalian class II MHC gene clusters. *Journal of Heredity* **91**, 198-204.

Terado T., Okamura K., Ohta Y., Shin D.H., Smith S.L., Hashimoto K., *et al.* (2003) Molecular cloning of C4 gene and identification of the class III complement region in the shark MHC. *Journal of Immunology*. **171**(5), 2461-6.

van der Poel J.J., Groenen M.A., Dijkhof R.J., Ruyter D. and Giphart M.J. (1990) The nucleotide sequence of the bovine MHC class II alpha genes: DRA, DOA, and DYA. *Immunogenetics* **31**, 29-36.

van Eijk M.J., Russ I., Beever J.E. and Lewin H.A. (1992) Polymorphism in exon 2 of the bovine lymphocyte antigen (BoLA)-DYA gene. *Animal Genetics* **23**, 476.

van Eijk M.J., Beever J.E., Da Y., Stewart J.A., Nicholaides G.E., Green C.A. *et al.* (1995) Genetic mapping of BoLA-A, CYP21, DRB3, DYA, and PRL on BTA23. *Mammalian Genome* **6**, 151-152.

Van Rood J.J., Eernisse J.G., Van Leeuwen A. (1958) Leucocyte antibodies in sera from pregnant women. *Nature* **181**(4625), 1735-6.

Wagner J.L. (2003) Molecular organization of the canine major histocompatibility complex. *Journal of Heredity* **94**, 23-26.

Waterer G.W., and Wunderink R.G. (2003) Science review: Genetic variability in the systemic inflammatory response. *Critical Care*. **7**(4):308-14.

Waterston R.H., Lindblad-Toh K., Birney E., Rogers J., Abril J.F., Agarwal P., *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562.

Wright H., Ballingall K.T. and Redmond J. (1994) The DY sub-region of the sheep MHC contains an A/B gene pair. *Immunogenetics* **40**, 230-234.

Xie T., Rowen L., Aguado B., Ahearn M.E., Madan A., Qin S., *et al.* (2003) Analysis of the gene-dense major histocompatibility complex class III region and its comparison to mouse. *Genome Research*. **13**(12), 2621-36.

Xu A., van Eijk M.J., Park C. and Lewin H.A. (1993) Polymorphism in BoLA-DRB3 exon 2 correlates with resistance to persistent lymphocytosis caused by bovine leukemia virus. *Journal of Immunol*. **151** (12), 6977-85.

Yabe T., Kawamura S., Sato M., Kashiwase K., Tanaka H., Ishikawa Y., *et al.* (2002) A subject with a novel type I bare lymphocyte syndrome has tapasin deficiency due to deletion of 4 exons by Alu-mediated recombination. *Blood*. **100** (4), 1496-8.

Yeh R.F., Lim L.P. and Burge C.B. (2001) Computational inference of homologous gene structures in the human genome. *Genome Research* **11**, 803-816.

Yuhki N., Beck T., Stephens R.M., Nishigaki Y., Newmann K. and O'Brien S.J. (2003) Comparative genome organization of human, murine, and feline MHC class II region. *Genome Research* **13**, 1169-1179.

APPENDIX A
LIST OF PRIMERS USED IN THE SEQUENCING AND ASSEMBLY OF BOLA IIB

Template: BAC clone used for template, Name: name of primer, Sequence: sequence of primer.

Template	Name	Sequence
14-47C8R7	PCRVUAT_15_5`	CCTGAAAATAAGCCTGGAGAGTGG
14-47C8R7	PCRVUAT_5_5`	AGCCATCCTGTTGACTTGTGAAGG
14-47C8R7	T21_4_5-3`	CCATCCTGTTGACTTGTGAAGGGC
14-47C8R7	T21_4_5-5`	CACCAACTCTCACCTTCCTTTACTG
14-47C8R7	VUAT 23 3'	TGATGAGCAGCAAAGTGAAGAAGG
14-47C8R7	VUAT10 5`	AACCACAGTTTTCGGGAATGGAAC
14-47C8R7	VUAT10-3`_(6164604)	AATCTCTGGTTGGGGACCTAACATC
14-47C8R7	VUAT10-5`	AACCACAGTTTTCGGGAATGGAAC
14-47C8R7	VUAT10-5`	GATGACCAGAGGTATTGCTGAAGC
14-47C8R7	VUAT10-5`_(6164605)	TGATGACCAGAGGTATTGCTGAAGC
14-47C8R7	VUAT13-3`_(6032123)	ACACTGATTTGTATGAACCTGGTGG
14-47C8R7	VUAT13-3`_(6164606)	GGAGCTAAGAGAATAAAAGACGAAGA
14-47C8R7	VUAT13-5`	GAAGTGCCTGAAAATAAGCCTGGAGAGTGG
14-47C8R7	VUAT13-5`_(6032124)	GCCTGAAAATAAGCCTGGAGAGTGG
14-47C8R7	VUAT13-5`_(6164607)	CAAGAGTACTGGATGATGAACTCACT
14-47C8R7	VUAT133`	AAGCGTCCTTGGATTTGG
14-47C8R7	VUAT135`	TGGGATTTCCAGGCAAGAG
14-47C8R7	VUAT14-3`_(6032125)	GGAGTGAAGATGTTTGTAGAGTGGC
14-47C8R7	VUAT14-3`_(6164608)	TAGGAATGGAGGACAGACTCAGGAG
14-47C8R7	VUAT15-5`	TTGGGGCTGTTATGACTGCTGTTG
14-47C8R7	VUAT16-3`	GCATAGCACAGAGTGACTAATG
14-47C8R7	VUAT203`	TTTCCACTCAAGTCGGTG
14-47C8R7	VUAT205`	CCCCAAACGGATGACATAC
14-47C8R7	VUAT21_3`	CTCAGTTTTGATGTGGTATTGTTG
14-47C8R7	VUAT21_5`	ATGTGACTGTGGTGGTGGTGGGGG
14-47C8R7	VUAT213`	AACAAGGTCCTGCTGTGTGG
14-47C8R7	VUAT215`	TAGTGAGGCTGGGCATCATC
14-47C8R7	VUAT23-3	CGTGTATGTGGTATCTTAGTTTGC
14-47C8R7	VUAT23-5`_(6164609)	ACCACCCTTATGGCAGAAAACGAAG
14-47C8R7	VUAT26-5`	TCCAGTTAGGTTTCCTCATCAGAC
14-47C8R7	VUAT26-5`	TCTCCAGTTAGGTTTCCTCATC
14-47C8R7	VUAT26IN.A	ATTTTCCACAGGCACAGGCTAACG
14-47C8R7	VUAT26IN.B	TCTGATGAGGAAACCTAACTGGAG
14-47C8R7	VUAT27-3`_(6032126)	TTATTTGAAAGACACCCAAGCAAGC
14-47C8R7	VUAT273`	TGGCACAGTGGTAAAGAATC
14-47C8R7	VUAT275`	TCAAGGGACACTTCAAGACTG
14-47C8R7	VUAT28-3`_(6164610)	CTATTCACGAAAAGACCCTGATGCTG
14-47C8R7	VUAT30_3`	GGAAAACACCCTTGTCCTCAAATC
14-47C8R7	VUAT30-5`	ATGATGCTGATAGATTTGCTTGACGCAGGG
14-47C8R7	VUAT303`	TGTCTGACTCTTTGCGAGC
14-47C8R7	VUAT305`	AGATTTGCTTGACGCAGG
14-47C8R7	VUAT4_3`	ACAGGGTGGGGTTCAGGGAGGTGG
14-47C8R7	VUAT4_5`	GAGTTTATTAGGAAATGTTGATGG

Template	Name	Sequence
14-47C8R7	VUAT43`	AGAATCCACCTTGCGATG
14-47C8R7	VUAT45`	AGTGCTGTGCCAATCTCTGC
14-47C8R7	VUAT5-5`	AAAGCACCCCAGCCAAGGAGGTAAGTTGTG
14-47C8R7	VUAT5-5`_(6032127)	CCATCCTGTTGACTTGTGAAGGGC
14-47C8R7	VUAT5-5`_(6164611)	CCATCCTGTTGACTTGTGAAGGGC
14-47C8R7	VUAT8 5`	AGCCAGGGAGAGAGTTCTTACCAG
14-47C8R7	VUAT8-3`_(6032128)	AGGGATTGAACCTCCATTCTCAG
14-47C8R7	VUAT8-5`_(6032129)	AAGCCAGGGAGAGAGTTCTTACCAG
14-47C8R7	VUATSP6F	CTGAATCTTGTAGAAAGGGAGGAGC
14-47C8R7	VUATSP6R	TATGAAGTGTGAGGAGGGGTTGGCGAAC
27C10R5	1718VUAB17-PCR	GAAGTGGGGGCAGAGATTCG
27C10R5	1718VUAB18-PCR	GGAGTGGTGTGTGGCTTACGAC
27C10R5	B16_18X12_5-3`	CAAAGCCCAGAAGAGAAGTGAGCAC
27C10R5	B16_18X12_5-5`	AGGATTCACTGATTTAGCCACCCTG
27C10R5	DRBEX2-INA	TGCGTGCCGTTGGAGAAGTGACAC
27C10R5	DRBEX2-INB	TGTCACTTCTCCAACGGCACGCAG
27C10R5	DRBEX3-INB	ACAGGATACACAGTCACGGTAGGC
27C10R5	LS3-3`	ACTGGCATTTCTCTGGGAGGGAATC
27C10R5	LS33`	ATTTCAACCCCTGGTCAG
27C10R5	PCR_B17X18F	GCATCCCCCACTTCTCCTTTC
27C10R5	PCR_B17X18F	AAGACCCAGTTTTTCCCACAGC
27C10R5	PCR_BIG17	CAGGGCAAGTCCGAGTGTCAC
27C10R5	PCR_BIG18	GGAGTGGTGTGTGGCTTACGAC
27C10R5	PCR-VUAB17-5`	GAGGAATGTAAAGGACTGTGAGAC
27C10R5	PCR-VUAB18-3`	TCAAAGAGCCAACCATCAAG
27C10R5	VUAB10-5`	AAGAAGACACCAGAACGCTACCAGAAGGTC
27C10R5	VUAB12-RAT	CCAAACAGCAGTGGGCAAAATGTAAG
27C10R5	VUAB13-3`	GATGCTTTATTGCGACTGTGCTGCCAAG
27C10R5	VUAB13-3`_(6164598)	ACTTCTCAGGAGTCTGTGGTTTCAG
27C10R5	VUAB13-5`	AGTGAGACCAATGGACTGGGTTTAG
27C10R5	VUAB15_3`	GTGGATGCTGGTGAAGACTTTGGG
27C10R5	VUAB15_5`	GCTGCCCCACCAACTACCATTAGC
27C10R5	VUAB153`	GTCCTTGGGATTTTCCAG
27C10R5	VUAB155`	CCCACCAACTACCATTAGC
27C10R5	VUAB16-3`	AGGGGCAGAGAACAAAGAGGACCG
27C10R5	VUAB16-3`	AAAGAAATGTGGACGGATGGAGGC
27C10R5	VUAB16-3`_(6032118)	AAAGAAATGTGGACGGATGGAGGC
27C10R5	VUAB16-5`_(6164599)	ACTTTCTGTCATCAAGGGTGGT
27C10R5	VUAB17_3`	TTAGTCCAGGTTTTTCAACATAG
27C10R5	VUAB17_5`	GACTACTTTTTGCTTCTTGGTATG
27C10R5	VUAB17-3`	GCGTCATCTTCACTGTTGCTTGCC
27C10R5	VUAB17-3`C	TGGAAATGTGAATCTATGCGCGGCTC
27C10R5	VUAB17-3`D	GCGTCATCTTCACTGTTGCTTGC
27C10R5	VUAB17-3`F	TCATCTTCACTGTTGCTTGCCTC
27C10R5	VUAB17-5`_(6032119)	ATTGTGACTAATGGTTTACCGTCG
27C10R5	VUAB17-5`_(6164600)	TGAGTCAACATTTTCTCCAGTTTGC
27C10R5	VUAB17-5`NEW	GTAAGTGTGCTGGCGTCTGTGT
27C10R5	VUAB17+18F	AGGAGTGGTGTGTGGCTTACGAC
27C10R5	VUAB17+18R	TGGCTGAAGTGGGGGCAGAGATTCG
27C10R5	VUAB173`A	GCGTCATCTTCACTGTTGCTTGC
27C10R5	VUAB175`	TCCCACACCCCTCTGACCTTTC
27C10R5	VUAB17PLUS3`	CGTCATCTTCACTGTTGCTTGCCTC
27C10R5	VUAB17PLUS5`	TGAGTCAACATTTTCTCCAGTTTGC

Template	Name	Sequence
27C10R5	VUAB18_3`	ATCCTCAAAGAGCCAACCATCAAGC
27C10R5	VUAB18_5`	ACCGTGACACTTGGACGAGAACCTTC
27C10R5	VUAB18_INTERNAL	GAGGAGTGGTGTGTGGCTTACGAC
27C10R5	VUAB18-3`	TGAGCACTCGCAACAGAATGACGC
27C10R5	VUAB18-3`_(6032120)	TGAGCACTCGCAACAGAATGACGC
27C10R5	VUAB18-3`_(6164601)	ATGTCTCATCCTCAAAGAGCCAACC
27C10R5	VUAB18-5`B	CGTGAAGCGGTGGTATGGAAAC
27C10R5	VUAB18-5`E	ACCGTGACACTTGGACGAGAAC
27C10R5	VUAB183`	AAGACCCACTTTTTCCCACAGC
27C10R5	VUAB185`A	ACCGTGACACTTGGACGAGAAC
27C10R5	VUAB185`B	GCGATTTTGTCAGTATGCGGTG
27C10R5	VUAB19_3`	TCCATAAGGGCATTACAGACCAAG
27C10R5	VUAB19_5`	AGAGTTCTGTTTTCCTCCTATCC
27C10R5	VUAB19-3`_(6032121)	CTCCCTCCTACGGTAAGTCCATAAG
27C10R5	VUAB19-3`E	AAGATGGTATGGGGAGTAGACCG
27C10R5	VUAB19-3`F	CCTCCTACGGTAAGTCCATAAGGG
27C10R5	VUAB19-3`H	GCCACGAGGAAACCAGAAACAG
27C10R5	VUAB19-5`A	GGCAAGGAATAAGGAGAGGAGGAC
27C10R5	VUAB193`	AAGGGCATTACAGACCAAG
27C10R5	VUAB195`	GTTCTGTTTTCCTCCTATCCTC
27C10R5	VUAB20-3`_(6032122)	CTACCCTTCTTCGGCAAAACTCACC
27C10R5	VUAB20-5`_(6164602)	CATGGCTGCAGTCAATTAGTTT
27C10R5	VUAB21_5`	GCACTTGTTTTACTGAAAGCGTTAC
27C10R5	VUAB21-3`A	CCACATCAATGAGAAACTCGCAC
27C10R5	VUAB21-5`	GCACTTGTTTTACTGAAAGCGTTAC
27C10R5	VUAB21-5`	GCTTTGTCATTTCAGTGATTCAGGGACC
27C10R5	VUAB21-5`B	GCAAATGGAAAAAGACCTGGC
27C10R5	VUAB21-5`D	CACCCCACTCATTCTTCTGTTGTG
27C10R5	VUAB21-5`G	GCAAATGGAAAAAGACCTGGC
27C10R5	VUAB213`	ACACAGCGACAGAGAGTAGC
27C10R5	VUAB215`	TGACCACCAGGAAGGAATG
27C10R5	VUAB22-3`_(6164603)	CTGTCCTTTCCCCATTGTATAGTC
27C10R5	VUAB22-3`C	AAGACTGAAGGCAAAAGAAGAGGG
27C10R5	VUAB223`A	AAGACTGAAGGCAAAAGAAGAGGG
27C10R5	VUAB23_3`	CAGCAATGAAGACCTGGAATAGCC
27C10R5	VUAB23-3`G	GCAGCAATGAAGACCTGGAATAGC
27C10R5	VUAB23-3`H	ACCCCTGCCATTCATTGAGG
27C10R5	VUAB233`A	ACCCCTGCCATTCATTGAGG
27C10R5	VUAB233`B	GCAGCAATGAAGACCTGGAATAGC
27C10R5	VUAB7-3`	TCAAGTGTTATTCCCTTTCACGGTTACG
27C10R5/3C8R7	DIBGAP1	TGTGGCAGGGGAGTCTTCATTACAGC
27C10R5/3C8R7	DIBGAP2	CCTGTCAGACTGTAGTAGAAGCCAGAAAAG
27C10R5/3C8R7	DIBGAP3	TACTGGTATTGACAGGAAAGTGTG
27C10R5/3C8R7	DIBGAP3F_(6105413)	TCGGAGGCTGAGATAAAGAGAGATG
27C10R5/3C8R7	DIBGAP4INTERNAL_(6105414)	CTCTTCAATCCAATCCATCAAAAGC
3C8R7	2131-3`	CCAAGGCTATGGTTTTTTCCAGTG
3C8R7	LS1-3`	GCCTCGGATGGTAGTTTTTCATTTGG
3C8R7	LS1-5`	GGAAAAGTAGAAGCAATCCAAAGGTGCG
3C8R7	LS1-5`_(6032115)	GGTATTTTCTCCTTACGCATCTGTGC
3C8R7	LS13`	TCCGACTCTTTGTAACCCC
3C8R7	LS15`	GGCTCATTTTGGAGGTAGAG
3C8R7	LS2-3`	GCTATGTGTGGTGACAAAGAGTAACTGG
3C8R7	LS2-3`	GAGGAAGGTAGAAGGTGATGAACTATCTGC

Template	Name	Sequence
3C8R7	LS2-3`_(6032116)	AAGGTGGGAAGAATGGAGAGGGTAG
3C8R7	LS2-5`	TGTCTTTTTGTCAGTTCCCACCAGG
3C8R7	LS2-5`_(6032117)	CTGAATAATCAAACCTCGTAAAGC
3C8R7	LS23`	CGTTGGAGAACACTCTTGAG
3C8R7	LS25`	TCCACTCTGATGATACTCACG
3C8R7	PCRLS23	ATGCTCCTGAACTGTGGCGTTG
3C8R7	PCRLS25	GCTGTTTGTTCTGTTTGTTCTGCC
3C8R7	PCRLS25	CTCAGGTCTTGGACAAACCGTG
3C8R7	PCRVUAX_15_3`	CCACCCACTCCAGTATTCTTGC
3C8R7	PCRVUAX_21_5`	CCTTCCCTGTCTTTTTGTCAGTTC
3C8R7	PCRVUAX143	GCTATGGTAAAATGGATACTGCCG
3C8R7	PCRVUAX145	ATTGAATGTGTGATGCCCGC
3C8R7	PCRVUAX163	CCCTGTCCTTTATCATCTCCTGG
3C8R7	PCRVUAX193	TGAATCATCAGGAAAGCCCCTC
3C8R7	PCRVUAX195	AACTGCCAGAGTCCACCCAAAC
3C8R7	TAP2-GENF	TTTCCCAGTCACGACGTTGTTTCAGGTGTACAGCAAAGTGAA
3C8R7	TAP2-GENR	CAAGGTCCCATTGTATAGCACA
3C8R7	TAP2EX2AMBF	GGCGGCGAATGGAGTAAATGGATG
3C8R7	TAP2EX2AMBR	CCAGGAGAACAGAGTCTTGATGTG
3C8R7	TAP2EX9	GGTGAGGGACAACATCACCTAC
3C8R7	TAP2EXON9	GAGGGACAACATCACCTACGGC
3C8R7	TAP2GEN-LONG-F	TTTCAGGTGTACAGCAAAGTGAATCAGT
3C8R7	VUAX10-5	AGTTCTCGCCTAAACCCCATCCAC
3C8R7	VUAX10-5`_(6164612)	AGTAACTTCTTTCCCCTTCGTCTCC
3C8R7	VUAX11-3`_(6032130)	CCAGTTCCCTCTTTCATCTTACAGTC
3C8R7	VUAX11-5`_(6164613)	GGCGACCATGAGTTTGTTTTCTATATCT
3C8R7	VUAX12 3`	AGGCTATGGTTTTTTCCAGTGGTC
3C8R7	VUAX12 3`	TAGAAATCCCTATGTTCCTCTGG
3C8R7	VUAX12 5`	TTCTCTTCCTGCCTGACTTCAAC
3C8R7	VUAX12-3`_(6032131)	TAGAAATCCCTATGTTCCTCTGGC
3C8R7	VUAX14_5`	CAAGCGTGAAGTGACAGCACATTC
3C8R7	VUAX14-3`	TTTGAAACAGACGCCACC
3C8R7	VUAX14-5`	CAGCCAGGAGTTTGAAGATTGTGCC
3C8R7	VUAX14-5`	CTCAGGGTCCTTATGGTATTGATGG
3C8R7	VUAX143`	ATAGGGTGAGGGTGTAGTGG
3C8R7	VUAX145`	AAGCGTGAAGTGACAGCAC
3C8R7	VUAX15 5`	TTTGTTTCCAGTTAGCCTATTGTC
3C8R7	VUAX15-3	GAAAGGAGTACGTCAAGGCTGTATA
3C8R7	VUAX163`	TTGGTTGGTAATGTCTCTGC
3C8R7	VUAX165`	GAAGCCACCACAACATTG
3C8R7	VUAX18 3`	TTGGAAGAAAGGGAGGTCCTGGTG
3C8R7	VUAX18 5`	CTTATCCATCGTGGGCATTGTATG
3C8R7	VUAX18-3`_(6032132)	TTGGAAGAAAGGGAGGTCCTGGTG
3C8R7	VUAX18-5`_(6032133)	CTTATCCATCGTGGGCATTGTATG
3C8R7	VUAX19-3`_(6032134)	TCATTACAAGGACTACCCAGAATC
3C8R7	VUAX193`	CCAAAGAAGACATACAGATGGC
3C8R7	VUAX195`	GTGGTTTCTCTGCGTTTTC
3C8R7	VUAX20-3`_(6164614)	TATGACGAAGAAGAACGGAAACGCC
3C8R7	VUAX21 5`	GTGATAAGTGAATAGTGTTAGTCCAC
65P20	BACGAP-A5`	ACTGTGGTTGCTCTGATTGTGTGAC
65P20	BACGAP-B3`	TGTATTTATGGGGTGGGATGTCGC
65P20	BACGAP-B5`	AGACCTTGGCGGCAATACTGAAAG
65P20	BACGAP-C3`	TCAAAGTCCTCAAAGCAGGGTTGG

Template	Name	Sequence
65P20	BACGAP-C5`	AAGTGGACAGACAGGTTAGGTGAC
65P20	BACGAP-D3`	GGAGGTGATGGGGTTCATTTACTG
65P20	GAP-CONTIG-BIG-F	TATGGGAACAGGGTGAGTGTGGTG
65P20	GAP-CONTIG-BIG-R	AAGAAATGCTGCCACCTTGAGACC
65P20	GAP-F2	CTCCCAAATCTACCTCTACCAATC
65P20	GAP-R2	TACCCTGAAGGAAACTCTCAATGG
65P20	GAP4R	CGAAGGCTCTGTGTAGATACGCAG
65P20	GAP5F	GAGGGACAATGATAAGTCAAGGGG
65P20	GAP5R	TGCCATTCGTAGTTCCTCCATTGG
65P20	GAPF3	GAGCCCCCCAACCACAAACATTAC
65P20	GAPF4	TAGTCCACGGGGTCACAAACAGTC
65P20	GAPF6	GGCAAAATCTGGGCAAACATCAAC
65P20	GAPR3	TACTTGTTGTGGGGAGTCACG
65P20	GAPSTUFF-F	CTCTGAGGTGGGAAAGATTGAGTG
65P20	GAPSTUFF-R	AATCTCCTTCTCGGACTCC
65P20	GNLSTUFF2R	GGAACTGTGGTTGCTCTGATTGTG
65P20	GNLSTUFF3F	AAGTCCTCAAAGCAGGGTTGGGTC
65P20	GNLSTUFF3R	AAGTGGACAGACAGGTTAGGTGACTGCC
65P20	INIT-GAP-F	TGCTTCCCCTACCTGGTTCAATCCC
65P20	INIT-GAP-R	ACCTCTTATCGCCTTTTCTACACCCAC
65P20	SCAFFOLD30228GAP1F	GCAAACCCTCACAACTCCAGGTC
65P20	SCAFFOLD30228GAP1R	TGTTCCCCTTGGCTTGGTGC
65P20	SCAFFOLD30228GAP2R	CAGGGCTGCTACGAAGGAAAG
65P20	SCAFFOLD30228GAP2R	GGCAACTAAACAGAGACCACGG
7138-04	7138-04_12_3`	AAACCTTCGTGTAGACTTCCGTTG
7138-04	7138-04_12_5`	AACCTGTAGAACGGAGTAACCTCG
7138-04	7138-04-14_5`	CATCTTATCTCACCCCTACATCTAC
7138-04	CONTIG 16 3'	TCCTCACTCAACCTACTCGCCTTC
7138-04	CONTIG 16 5`	ATCGGGGTCACTCGGGGAAGAGGC
7138-04	CONTIG10-3`	GACCATAAAGAAGGCTGAGCACTG
7138-04	CONTIG10-5`	TGCTATGCTATGCTAAGTCGCTTC
7138-04	CONTIG13 5'	TACAGGCATACCTCAGAGATACTG
7138-04	CONTIG13-5	CACCTTCTCTCATCACACATCCAC
7138-04	CONTIG14-3`	CCCACCACTCCTTCAACTATTATC
7138-04	CONTIG15-3`	TCCTTCTTCTCTCGTTCCCTACTC
7138-04	CONTIG15-5`	ATCTTTCCTTCCCTCATCCACAGG
7138-04	CONTIG17 3`	ATCCTAAACTCCATAATCCCTCCC
7138-04	CONTIG17 5`	ATCTTTCCTTCCCTCATCCACAGG
7138-04	CONTIG18 3'	ACTTCCTTGGGTAATCCACCATTC
7138-04	LS4-3`	AGGAGATACAAGAAGGAGCCAGCCCCG
7138-04	RXRBEX6-FOR	AAAGACAAAGACGGGGATGGGGAG
7138-04	RXRBEX6-REV	TGAGAGAAGGAGGCGATGAGCAGC
7138-04	U7138-04-22_3`	ATCTTTCCTTCCCTCATCCACAGG
7138-04	U7138-04-22_5`	TCCTTCTTCTCTCGTTCCCTACTC
7138-04	U7138-04-23_3`	TTCCTTGGGTAATCCACCATTCTG
7138-04	VUAS 11 3`	CCCTTGTCTTGGTGTCTCTGACTC
7138-04	VUAS 11 5`EXTENDED	AAGGGGTTCAGAGGTCAGGTGAGC
7138-04	VUAS 14 3`	GTTTTTGCTACTTTAGGTGATGGC
7138-04	VUAS 16 3`	TTTTCTCTGCCTACACGCACCACC
7138-04	VUAS 16 5`	CCACTCTGTTGAGCAAGCCTTTTG
7138-04	VUAS 17 3`	GGAGAAGGGACACAGAGTGAGTAT
7138-04	VUAS 18 3`	GAGACTCTGTCCTTGTTAGAAAAC
7138-04	VUAS 18 5`	GCAGCATATGGGTTCACGTA

Template	Name	Sequence
7138-04	VUAS 3 5`	GGCACCTCCTAACATTTTATCTCC
7138-04	VUAS 4 3`	GCTACCGAAGCCTGTGTGCCCCAG
7138-04	VUAS 4 5`	GCCTTTAGGAGAGTTCTTTAGTCC
7138-04	VUAS 5 3`	CCACAAAAAAAGATTCCTTGACAG
7138-04	VUAS 5 5`	TTCTCAAGTCACCCCTGGCTAATG
7138-04	VUAS 9 3`	TGGCGGATGGGGGAAGGGTGTCAGG
7138-04	VUAS10-14-5`	GCCACAGAAAAGCACAGAAACTTCAGTG
7138-04	VUAS11_3`	TGAGTGACCAGTGCCCCGATTCTG
7138-04	VUAS11_5`	GCTGACCTTCGTGCCTTCTTTTCC
7138-04	VUAS11-3`	GACCTTCGTGCCTTCTTTTCCAGC
7138-04	VUAS11-3`	GTCTCTGACTCTTCTTTCTCCCCCAG
7138-04	VUAS11-3`_(6105416)	GTCTTGGTGTCTCTGACTCTTCTTTC
7138-04	VUAS11-5`5`_(6105415)	GAAGGGGTTCAGAGGTCAGGTGAGC
7138-04	VUAS11-5NEW	TAAACGGGCAGGTGGTGAGGAGCAG
7138-04	VUAS115VER3`	ACTGGAGTCTGTTTGCCTCTTTCC
7138-04	VUAS11PLUS-5	TAGGATGCTTTGTCTCTGTGGGGGAG
7138-04	VUAS11PLUS-5	ACCCCTTGTCTTGGTGTCTCTGACTC
7138-04	VUAS12-15-3`	AAGGGCTGACTACAGAGGATGAGCG
7138-04	VUAS12-15-5`	AATGTCTGGCTGTGACTGAGTGTCC
7138-04	VUAS12-3NONREP	AGCCTCCTCTCTGTCCTGTAACCCC
7138-04	VUAS12-5	CCAAAATGTCTGGCTGTGACTGAGTGTCC
7138-04	VUAS12-5`_(6105424)	ATTCGGGGTTCTGAAGGGGTCACAAG
7138-04	VUAS12-5`_4	TAGACACTGACAGGTTGAGAGGGGCTACAC
7138-04	VUAS12-5`NEW	TGATGTCCTCTCGCCCCGTATTTC
7138-04	VUAS13 5`	ATTACAAGTATCCCACCAGTGTTG
7138-04	VUAS13-3`A	AGAGACTCCTCTGCGGCTTTC
7138-04	VUAS13-3`B	TCCCAGGGGTTCAGCGACA
7138-04	VUAS13-5`	CAGAAGTGCTGGTGCTGCCAAAGC
7138-04	VUAS14_3`	GTGGTTTTTGCTACTTTAGGTGATGGC
7138-04	VUAS14-3`_(6105426)	TTGCCTGCTCAGTGCGACTCCTACC
7138-04	VUAS14-3`_3	ATCAGGTAGGAGGGTAGAGAGTAG
7138-04	VUAS14-3`EXT	GAAGAGGGGATTTCTGGTGATGGC
7138-04	VUAS14-5`_(6105425)	ACTCTGTATCATCGCAAGGGTAACC
7138-04	VUAS15-3`	ACCAAGAAGGGCTGACTACAGAGGATGAG
7138-04	VUAS15-5`	GCTGGTTGTATTAGTGAAACTTGCGG
7138-04	VUAS16_5`	CTACAGGCATACTTTGTTTTATTG
7138-04	VUAS16-3`	TTCTCTGCCTACACGCACCACCTCACCCAC
7138-04	VUAS16-3`_(6105418)	GTCTCTGTCTTACTTACCCTTCGTC
7138-04	VUAS16-5`	TCAAGCCCTTTACCCTCTCTAACC
7138-04	VUAS16-5`	GTTGCCTCAAGCCCTTTACCCTCTCTAACC
7138-04	VUAS16-5`_(6105417)	ACCACTCTGTTGAGCAAGCCTTTTG
7138-04	VUAS16PLUS3`	TTTTCTCTGCCTACACGCACCACCTCAC
7138-04	VUAS16PLUS5`	TCTGTTGAGCAAGCCTTTTGATACC
7138-04	VUAS17_3`	AGGTCAGGGCAGGAGGTTGGC
7138-04	VUAS17-3`_(6105422)	AGAAGGGACACAGAGTGAGTATGGG
7138-04	VUAS17-3`A	ACGGTCTGCGTCTGTCTGTGTTGG
7138-04	VUAS17-3`B	CATTTTTCAGGGTGATACTGGTGC
7138-04	VUAS17-5`_(6105421)	AGAGTTTGTGAGGGTCAGGGCAAGGC
7138-04	VUAS17_6-3`	AGAAGGGACACAGAGTGAGTATGGGG
7138-04	VUAS17_6-5`	CCACACAGAGAAAGGAGATACAAGAAGG
7138-04	VUAS18-3`	GAGACTCTGTCCTTGTTAGAAAAC
7138-04	VUAS18-3`EXT	CATCAGTCCCTCCAATGAACACCC
7138-04	VUAS18-5`	TATTCCCCCAGCCTCCCTCTGCTTG

Template	Name	Sequence
7138-04	VUAS18-5`_(6105423)	TATTCCCCCAGCCTCCCTCTGCTTG
7138-04	VUAS18-8`	ATTCCCCCAGCCTCCCTCTGCTTG
7138-04	VUAS18-PLUS-5`	TATTCCCCCAGCCTCCCTCTGCTTG
7138-04	VUAS19-3`	GAAGCCTGATGGTTCCTTACTGAAAG
7138-04	VUAS4-5`EXT	CCCAACCGAGACAGATACTCATTC
7138-04	VUAS5-3`	TAAGCCCCCACTCTGGTAAGGTTC
7138-04	VUAS5-5`_2	CAACTAAGACCTGGCAGAGC
7138-04	VUAS6 3`	ACACTCTGGATCTTCGACCCT
7138-04	VUAS6_3`	ATCCAGCCACCAGCCCTGTCCACC
7138-04	VUAS6-17-3`	ACGGTCTGCGTCTGTCTGTGTTGGG
7138-04	VUAS6-17-5`	AAAGGAGATACAAGAAGGAGCCAGCCCC
7138-04	VUAS6-3`_(6105420)	AAAGGAGATACAAGAAGGAGCCAGC
7138-04	VUAS6-5`_(6105419)	AAGGGCGAGGTTGACTTACATCCTGC
7138-04	VUAS9-3`_(6105427)	GCTCATACAGGCTCAGTTCCCACG
7138-04	VUAS9PLUSRE3`	AAGTGGGAAAGAGGCTGTGAGAAGG
7138-04	VUAST7F	GCACTGCTGATGGTCTTGTGAATAG
7138-04	VUAST7R	AAAACTCCTTCTCTTCCTACCTGCC

APPENDIX B

PERL SCRIPT USED TO OUTPUT ONLY SECTIONS OF CLUSTAL ALIGNMENT CONTAINING MISMATCHES

```
#!/usr/bin/perl -w
use strict;


#Sample input:
#
#vuab_overlapping_with_vuax      AGCCCCTATGGAGGTGAAGACATCGTGGGTGAGTGTACAGTTGAGGGGTG 400
#vuax_overlapping_with_vuab      AGCCCCTATGGAGGTGAAGACATCGTGGGTGAGTGTACAGTTGAGGGGTG 400
#                                ************************************************
#
#vuab_overlapping_with_vuax      TGGGTTTACAATTGTGAAGAATTTTTAAAATTTTTATTTTATATTGGAGT 450
#vuax_overlapping_with_vuab      TGGGTTTACAATTGTGAAGAATTTTTTAAATTTTTATTTTATATTGGAGT 450
#                                ************************** ***********************
#Sample output:
#
#vuab_overlapping_with_vuax      TGGGTTTACAATTGTGAAGAATTTTTAAAATTTTTATTTTATATTGGAGT 450
#vuax_overlapping_with_vuab      TGGGTTTACAATTGTGAAGAATTTTTTAAATTTTTATTTTATATTGGAGT 450
#                                ************************** ***********************
$/ = "\n\n";

while (<>)
{
    my @record = split (/\n/,$_);
    my $star_count = $record[-1] =~ tr/\*/\*/;
#    print "$star_count\n";
    if ($star_count != 50)
    {
        print "$_";
    }
}
################################################################################
```

Sample Input/Output:

Sample Input:

```
CLUSTAL W (1.83) multiple sequence alignment


vuab_overlapping_with_vuax      TACTCTTTGATGATTCCTCCTATGTATGATATCCTTCTGCTCCCTTTCGA 50
vuax_overlapping_with_vuab      TACTCTTTGATGATTCCTCCTATGTATGATATCCTTCTGCTCCCTTTCGA 50
                                **************************************************

vuab_overlapping_with_vuax      ATCTATGCCCTTAGATAATTCAAAACAAGGGAGTTTTCCAGGCTTCTTAC 100
vuax_overlapping_with_vuab      ATCTATGCCCTTAGATAATTCAAAACAAGGGAGTTTTCCAGGCTTCTTAC 100
                                **************************************************

vuab_overlapping_with_vuax      AGGTCTCTTCCTAAAAATGCTTCAGCTGGCAACTGAGGTGTCAGCTCAGG 150
vuax_overlapping_with_vuab      AGGTCTCTTCCTAAAAATGCTTCAGCTGGCAACTGAGGTGTCAGCTCAGG 150
                                **************************************************

vuab_overlapping_with_vuax      GAATTTCTCTGATGGGCTGAGACACGATAGAGCAGGGTGAGGTGTGGACT 200
vuax_overlapping_with_vuab      GAATTTCTCTGATGGGCTGAGACACGATAGAGCAGGGTGAGGCGTGGACT 200
                                ****************************************** *******

vuab_overlapping_with_vuax      GCTCCAACATGATTTCTTCCAGCAGTTCTCTTTAGACCACCTTCCGGGAG 250
vuax_overlapping_with_vuab      GCTCCAACATGATTTCTTCCAGCAGTTCTCTTTAGACCACCTTCCGGGAG 250
                                **************************************************

vuab_overlapping_with_vuax      AGGCACCACTTGGAACAGCCACTCCTGAGGATACCCTTGGAGGAGGAGGA 300
vuax_overlapping_with_vuab      AGGCACCACTTGGAACAGCCACTCCTGAGGATACCCTTGGAGGAGGAGGA 300
                                **************************************************

vuab_overlapping_with_vuax      GGATGAAGAAAGCTCTGATTCTGAGGGCTCTCACTCTGGCCACCATGATG 350
vuax_overlapping_with_vuab      GGATGAAGAAAGCTCTGATTCTGAGGGCTCTCACTCTGGCCACCATGATG 350
                                **************************************************

vuab_overlapping_with_vuax      AGCCCCTATGGAGGTGAAGACATCGTGGGTGAGTGTACAGTTGAGGGGTG 400
vuax_overlapping_with_vuab      AGCCCCTATGGAGGTGAAGACATCGTGGGTGAGTGTACAGTTGAGGGGTG 400
                                **************************************************

vuab_overlapping_with_vuax      TGGGTTTACAATTGTGAAGAATTTTTAAAATTTTTATTTTATATTGGAGT 450
vuax_overlapping_with_vuab      TGGGTTTACAATTGTGAAGAATTTTTTAAATTTTTATTTTATATTGGAGT 450
                                ************************** ***********************

vuab_overlapping_with_vuax      ATGGTGGCTCAGATGGTAAAGAATCTGCCTGCAATGCAGGAGACCCAGGT 500
vuax_overlapping_with_vuab      ATGGTGGCTCAGATGGTAAAGAATCTGCCTGCAATGCAGGAGACCCAGGT 500
                                **************************************************

vuab_overlapping_with_vuax      TTGATCCCT-GGATTGGGAAGATCCCCTGGAGAAGGGAATGGCTACCCAC 549
vuax_overlapping_with_vuab      TTGATCCCTTGGATTGGGAAGATCCCCTGGAGAAGGGAATGGCTACCCAC 550
                                ********* ****************************************
```

Sample Output:

```
CLUSTAL W (1.83) multiple sequence alignment

vuab_overlapping_with_vuax      GAATTTCTCTGATGGGCTGAGACACGATAGAGCAGGGTGAGGTGTGGACT 200
```

```
vuax_overlapping_with_vuab      GAATTTCTCTGATGGGCTGAGACACGATAGAGCAGGGTGAGGCGTGGACT 200
                                **************************************** *******

vuab_overlapping_with_vuax      TGGGTTTACAATTGTGAAGAATTTTTAAAATTTTTATTTTATATTGGAGT 450
vuax_overlapping_with_vuab      TGGGTTTACAATTGTGAAGAATTTTTTAAATTTTTATTTTATATTGGAGT 450
                                ************************** ************************

vuab_overlapping_with_vuax      TTGATCCCT-GGATTGGGAAGATCCCCTGGAGAAGGGAATGGCTACCCAC 549
vuax_overlapping_with_vuab      TTGATCCCTTGGATTGGGAAGATCCCCTGGAGAAGGGAATGGCTACCCAC 550
                                ********* ****************************************

vuab_overlapping_with_vuax      TTTATTATAAGTATAACTTC-AAATTTGATTGTTTGCTTCAGAAATTTGG 1048
vuax_overlapping_with_vuab      TTTATTATAAGTATAACTTCCAAATTTGATTGTTTGCTTCAGAAATTTGG 1050
                                ******************* ****************************

vuab_overlapping_with_vuax      TATCCTGCGTCCCTGTTCTTATCTTTCTGCTTTGGCGTGGCCACGCACCA 2248
vuax_overlapping_with_vuab      TATCCTGCGTCCCTGTTCTTATCTTTCTGCTTTGGCGTGGCCACACACCA 2250
                                ******************************************* *****

vuab_overlapping_with_vuax      GCTGACCACGTCGGCACTTACGGCACAAATGTCTACCAGACGTACGGCGC 2298
vuax_overlapping_with_vuab      GCTGACCACGTGGGCACTTACGGCACAAATGTCTACCAGACGTACGGCGC 2300
                                *********** ************************************

vuab_overlapping_with_vuax      TTCCTCCCCCGGGTAGCTAGTCTTCCCCCTCCCCGACAC-TCTAAATTGC 2597
vuax_overlapping_with_vuab      TTCCTCCCCCGGGTAGCTAGTCTTCCCCCTCCCCGACACCTTTAAGAGGC 2600
                                ************************************ * ***   **

vuab_overlapping_with_vuax      TCCCCCTTTCTATTTCATTTCCTGGCAAATACCCAGTCCTCAGCTACAGG 2647
vuax_overlapping_with_vuab      TCCACCTTTCTATTTCATTTCCTGGCAAATACCCAGTCCTTGGCTACAGA 2650
                                *** ***********************************  *******

vuab_overlapping_with_vuax      TTTAATCTTGAAATATCCCTCCCCTGAGTTCCAAGAACCCACTCCTTGCA 2697
vuax_overlapping_with_vuab      TTTAATGTTGAAATATCCCTCCCCTGAGTTCCAAGAACCCACTCCTTGCA 2700
                                ****** ****************************************

vuab_overlapping_with_vuax      ACTCATAAAACAACTCAGAGTTTGGAATTTGTTTTTTTTTAACATTTTTT 2897
vuax_overlapping_with_vuab      ACTCATAAAAGAACTCAGAGTTTGGAATTTGTTTTTTTT-AACATTTTTT 2899
                                ********** ************************** **********
```

APPENDIX C

PERL SCRIPT USED TO REFORMAT BOVINE SNP DATA

FROM COMMA DELIMITED TO FASTA FORMAT

```
#!/usr/bin/perl -w
use strict;

my $error_list = "These records were not formatted correctly: \n\n";

while(<>)
{
    chomp $_;
    my @line = split (/,/, $_);
    if (defined $line[0] && defined $line[1] && defined $line[2] &&
defined $line[3] && defined $line[4] && defined $line[5] && defined
$line[7] && defined $line[8])
    {
        print ">$line[0]_$line[1]-$line[2]-$line[3]-$line[4]-
$line[5]\n$line[7]$line[8]\n\n";
    }
}
print $error_list;
#########################################################################
```

Sample input:

```
BTA-
15465,VSWHP1D0766A.scf,Limousin,chr1,3194,A/G,63,ACACACACACACACACACTT
CAAGTTGCACAGTGTTCTGCAACACAAGAGGAGGCTGTGTCTTGGGAAGGAAGAGACCAGAGAGGGATCCC
TGGCAGAGGTCTCGCCATAGGACTTCCAAAGGCATGGAGCACCAGCTGCTG
ACATTCAGTTGTTAAGTTCAGCTGATGTTTGCTGTCAAGCCGCCGGGGTCTTATAGTGTGAATGCCCCCTC
TGTCACATCAAAGTTCTCACACCTGAATGTCTTGG,TGAGTGTCTTGGGGAAACAGCAGTGGGCAGGAGAA
GCTTTTCTTCCCTGTTCTTACCATTCCTTTCCTGAGGAGAAGTGCTTAATAGCTT
AATAGCTCAGTCGTGCCCAACTCTTTGTGACCCCGTGGACTGTAGCCTTCCCAGCTCCCATGTCCATGGGA
TTCTCTAGGCAAGAATACTGGAGTGGGTTGCCATTCCCTTCTCCAGGGGATCTTCCACACTCAGGGATCGA
ACCTGTGTCTTCTGTGAT,1,aatgtcttgg,tgagtgtctt,63 59 58 63
59 60 60 60 63 60,58 63 63 58 60 63 60 59 60 60,866_1,1
```

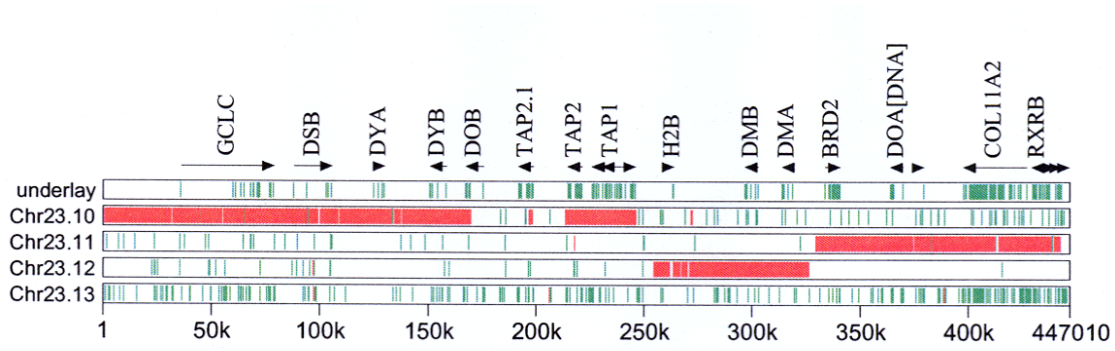
Sample output:

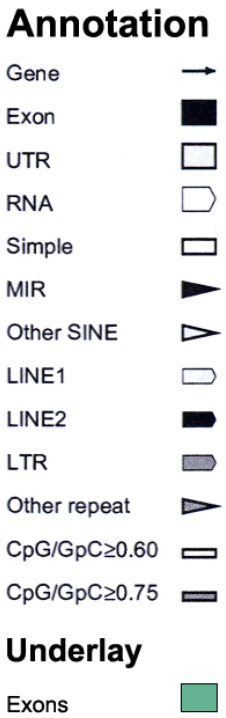
```
>BTA-15465_VSWHP1D0766A.scf-Limousin-chr1-3194-A/G
ACACACACACACACACACACTTCAAGTTGCACAGTGTTCTGCAACACAAGAGGAGGCTGTGTCTTGGGAAG
GAAGAGACCAGAGAGGGATCCCTGGCAGAGGTCTCGCCATAGGACTTCCAAAGGCATGGAGCACCAGCTGC
TGACATTCAGTTGTTAAGTTCAGCTGATGTTTGCTGTCAAGCCGCCGGGGTCTTA
TAGTGTGAATGCCCCCTCTGTCACATCAAAGTTCTCACACCTGAATGTCTTGGTGAGTGTCTTGGGGAAAC
AGCAGTGGGCAGGAGAAGCTTTTCTTCCCTGTTCTTACCATTCCTTTCCTGAGGAGAAGTGCTTAATAGCT
TAATAGCTCAGTCGTGCCCAACTCTTTGTGACCCCGTGGACTGTAGCCTTCCCAG
CTCCCATGTCCATGGGATTCTCTAGGCAAGAATACTGGAGTGGGTTGCCATTCCCTTCTCCAGGGGATCTT
CCACACTCAGGGATCGAACCTGTGTCTTCTGTGAT
```

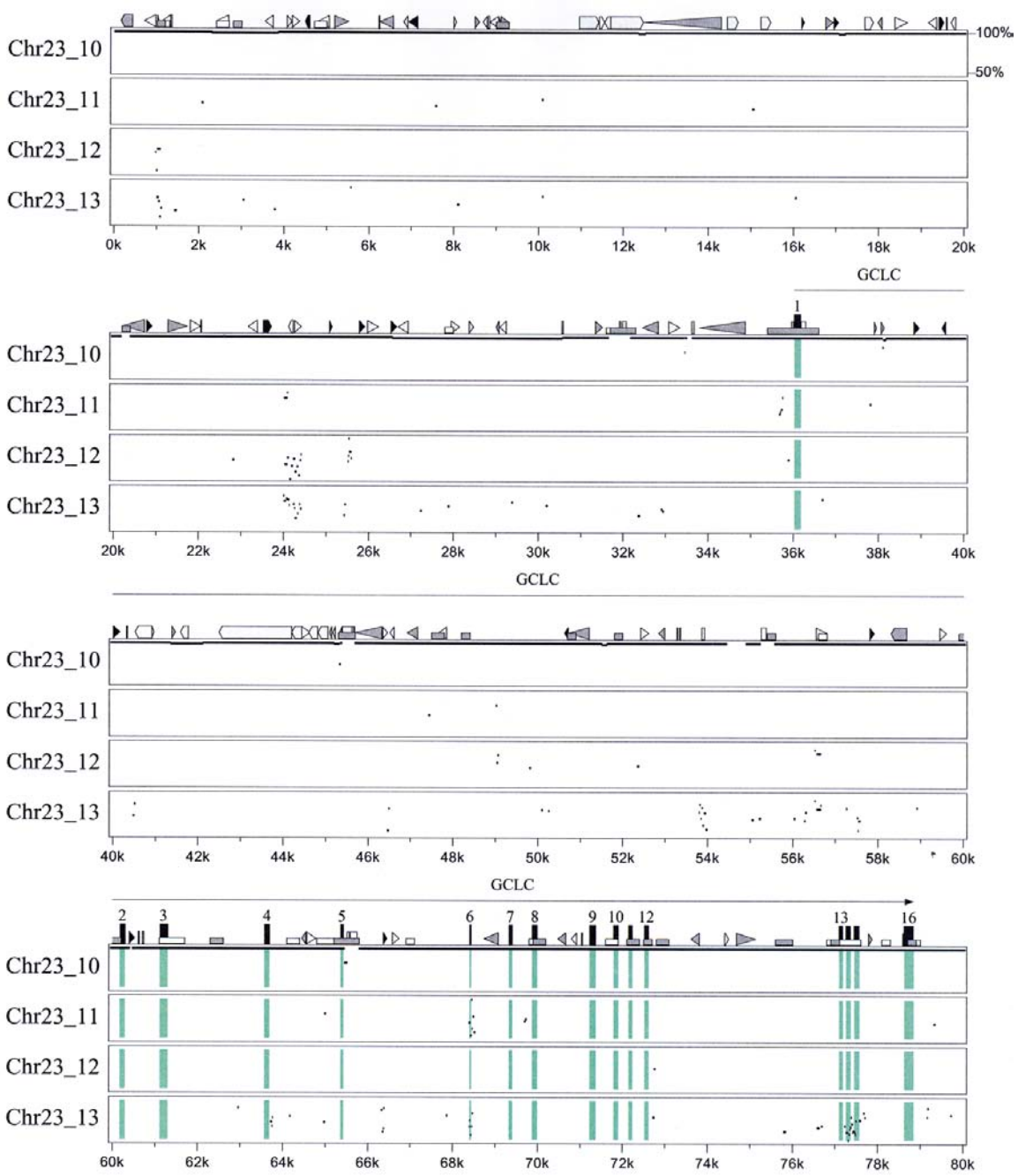
APPENDIX D

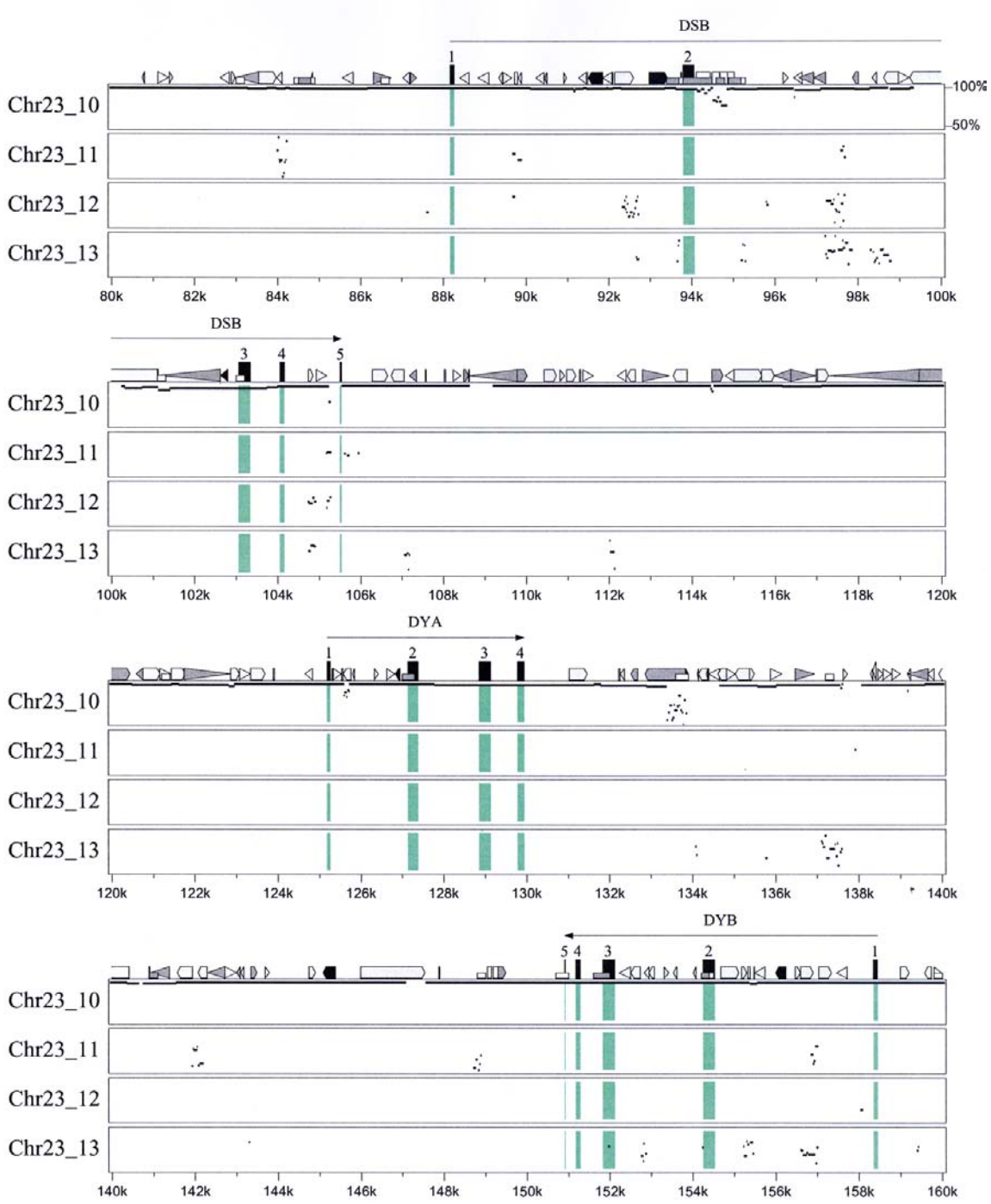
MULTIPIPMAKER PLOT OF TAMU BOLA IIB AGAINST BOVINE GENOME PROJECT SCAFFOLDS SPANNING BOLA IIB

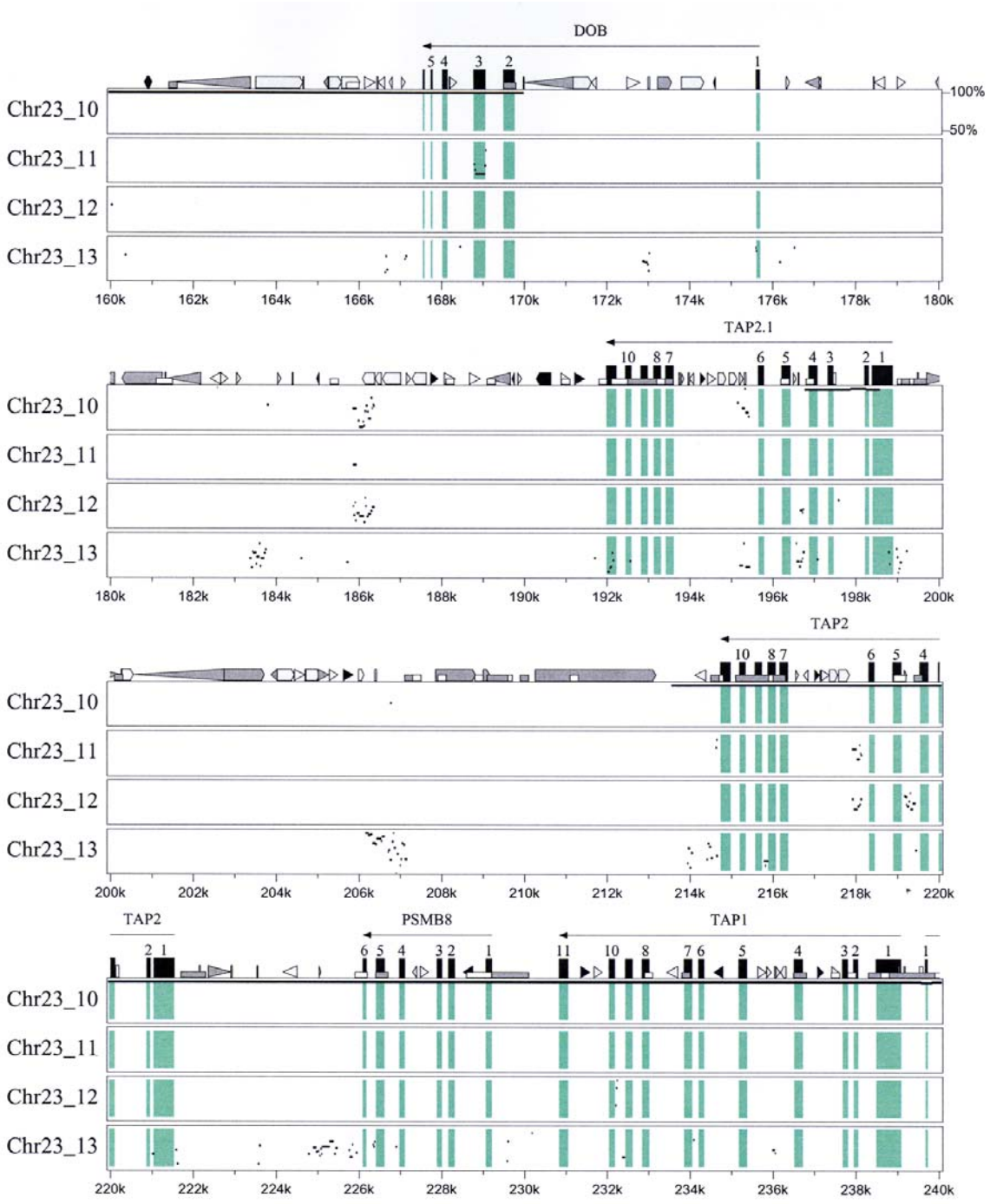
Percent identity plots are scored from 50 to 100%, and are represented by black dots. Green columns indicate regions corresponding to exons.

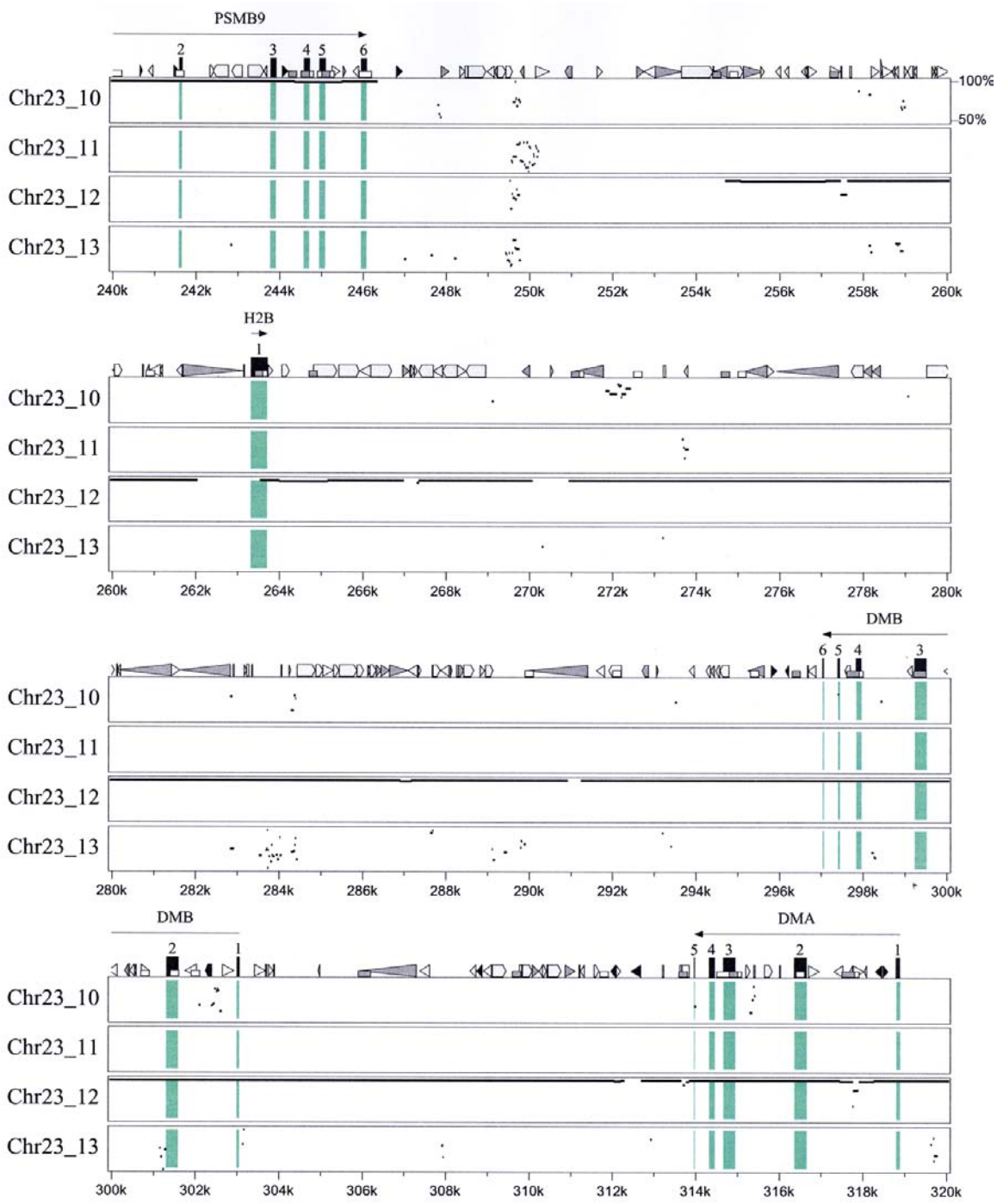


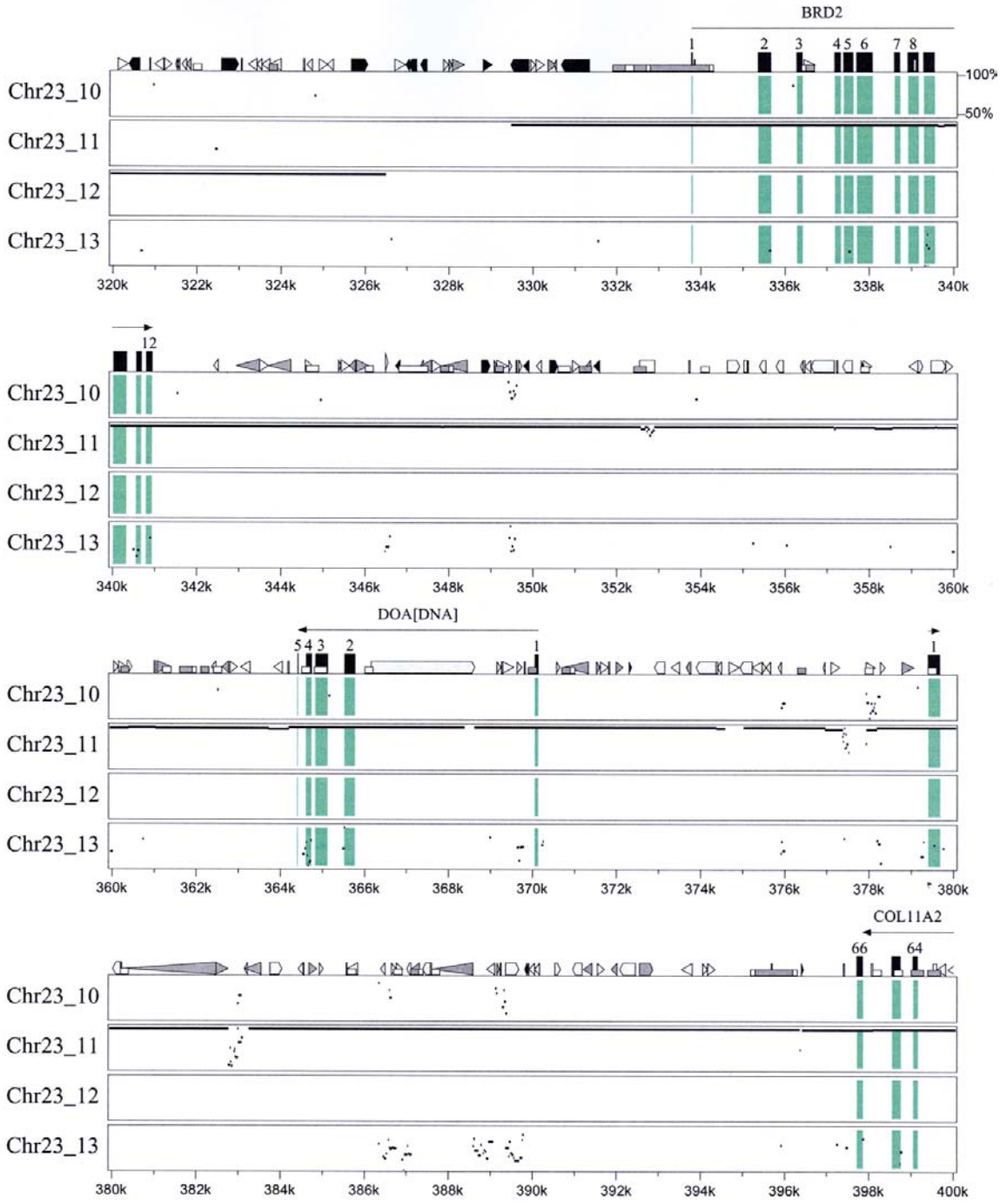


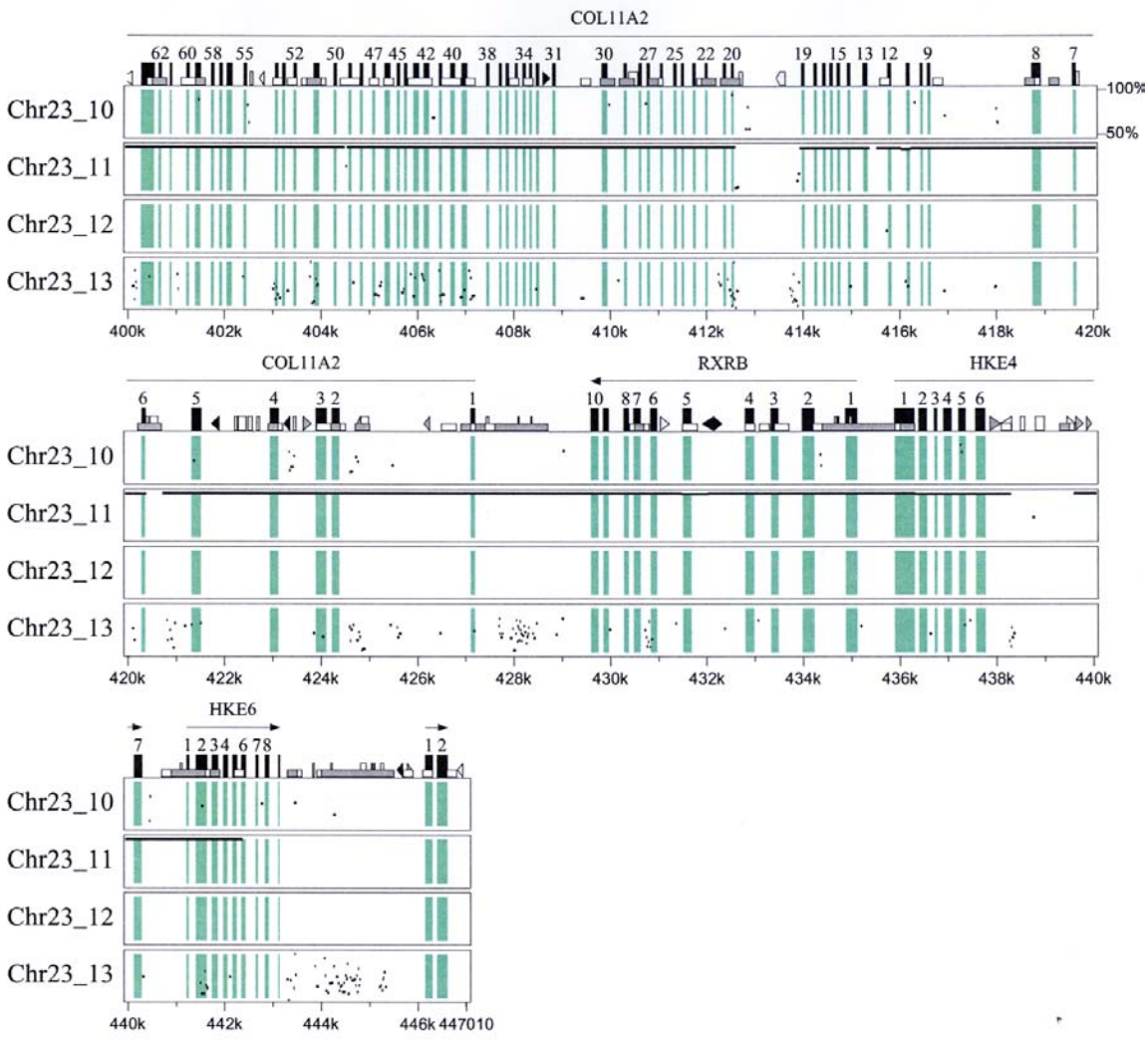

Chr23_10
Chr23_11
Chr23_12
Chr23_13
100%
50%
0k 2k 4k 6k 8k 10k 12k 14k 16k 18k 20k
GCLC
1
20k 22k 24k 26k 28k 30k 32k 34k 36k 38k 40k
GCLC
40k 42k 44k 46k 48k 50k 52k 54k 56k 58k 60k
GCLC
2 3 4 5 6 7 8 9 10 12 13 16
60k 62k 64k 66k 68k 70k 72k 74k 76k 78k 80k


DSB
1
2
100%
50%
Chr23_10
Chr23_11
Chr23_12
Chr23_13
80k
82k
84k
86k
88k
90k
92k
94k
96k
98k
100k
DSB
3
4
5
Chr23_10
Chr23_11
Chr23_12
Chr23_13
100k
102k
104k
106k
108k
110k
112k
114k
116k
118k
120k
DYA
1
2
3
4
Chr23_10
Chr23_11
Chr23_12
Chr23_13
120k
122k
124k
126k
128k
130k
132k
134k
136k
138k
140k
DYB
5
4
3
2
1
Chr23_10
Chr23_11
Chr23_12
Chr23_13
140k
142k
144k
146k
148k
150k
152k
154k
156k
158k
160k


DOB
5 4 3 2
1
100%
50%
Chr23_10
Chr23_11
Chr23_12
Chr23_13
160k 162k 164k 166k 168k 170k 172k 174k 176k 178k 180k
TAP2.1
10 8 7
6 5 4 3
2 1
Chr23_10
Chr23_11
Chr23_12
Chr23_13
180k 182k 184k 186k 188k 190k 192k 194k 196k 198k 200k
TAP2
10 8 7
6 5 4
Chr23_10
Chr23_11
Chr23_12
Chr23_13
200k 202k 204k 206k 208k 210k 212k 214k 216k 218k 220k
TAP2
PSMB8
TAP1
2 1
6 5 4 3 2 1
11 10 8 7 6 5 4 3 2 1 1
Chr23_10
Chr23_11
Chr23_12
Chr23_13
220k 222k 224k 226k 228k 230k 232k 234k 236k 238k 240k


PSMB9
2
3
4
5
6
100%
50%
Chr23_10
Chr23_11
Chr23_12
Chr23_13
240k
242k
244k
246k
248k
250k
252k
254k
256k
258k
260k
H2B
1
Chr23_10
Chr23_11
Chr23_12
Chr23_13
260k
262k
264k
266k
268k
270k
272k
274k
276k
278k
280k
DMB
6
5
4
3
Chr23_10
Chr23_11
Chr23_12
Chr23_13
280k
282k
284k
286k
288k
290k
292k
294k
296k
298k
300k
DMB
DMA
2
1
5
4
3
2
1
Chr23_10
Chr23_11
Chr23_12
Chr23_13
300k
302k
304k
306k
308k
310k
312k
314k
316k
318k
320k


BRD2
1 2 3 4 5 6 7 8
100%
50%
Chr23_10
Chr23_11
Chr23_12
Chr23_13
320k 322k 324k 326k 328k 330k 332k 334k 336k 338k 340k
12
Chr23_10
Chr23_11
Chr23_12
Chr23_13
340k 342k 344k 346k 348k 350k 352k 354k 356k 358k 360k
DOA[DNA]
5 4 3 2 1 1
Chr23_10
Chr23_11
Chr23_12
Chr23_13
360k 362k 364k 366k 368k 370k 372k 374k 376k 378k 380k
COL11A2
66 64
Chr23_10
Chr23_11
Chr23_12
Chr23_13
380k 382k 384k 386k 388k 390k 392k 394k 396k 398k 400k


COL11A2
62 60 58 55 52 50 47 45 42 40 38 34 31 30 27 25 22 20 19 15 13 12 9 8 7
100%
50%
Chr23_10
Chr23_11
Chr23_12
Chr23_13
400k 402k 404k 406k 408k 410k 412k 414k 416k 418k 420k
COL11A2
RXRB
HKE4
6 5 4 3 2 1 10 8 7 6 5 4 3 2 1 1 2 3 4 5 6
Chr23_10
Chr23_11
Chr23_12
Chr23_13
420k 422k 424k 426k 428k 430k 432k 434k 436k 438k 440k
HKE6
7 1 2 3 4 6 7 8 1 2
Chr23_10
Chr23_11
Chr23_12
Chr23_13
440k 442k 444k 446k 447010

VITA

Name: Christopher P. Childers

Address: Department of Veterinary Integrative Biosciences,
Texas A&M University, College Station, TX 77843-4458, USA

Telephone: 979-845-3283

FAX: 979-845-9972

Email: cchilders@cvm.tamu.edu

Education

Degree	Year	Institution
Ph.D. (Genetics)	2006	Texas A&M University
B.S. (Computer Science)	1999	West Texas A&M University
B.S. (Biology)	1998	West Texas A&M University

Professional Experience

West Texas A&M University
Teaching Assistant: Programming Fundamentals, Department of Computer Science,
Fall 1998

Texas A&M University
Teaching Assistant: Genetics, Department of Genetics/Biochemistry,
Fall 2000 – Spring 2001

Texas A&M University
Teaching Assistant: Biomedical Genetics, Department of Veterinary Integrative Biosciences, Fall 2001 – 2006

Texas A&M University
Graduate Research Assistant, Department of Veterinary Integrative Biosciences,
Fall 2000 – 2006

Publications

C. P. Childers, H. L. Newkirk, D. A. Honeycutt, N. Ramlachan, D. M. Muzney, E. Sodergren, R. A. Gibbs, G. M. Weinstock, J. E. Womack, L. C. Skow. Comparative analysis of the bovine *MHC class IIb* sequence identifies inversion breakpoints and three unexpected genes. 2006 Animal Genetics 37(2):121-129.