Determination of the genetic mechanisms responsible for generating diversity in the cattle NK cell receptor repertoire

Nicholas D Sanderson
Imperial College London, Department of Medicine

September 17, 2014

Thesis submitted for PhD degree

0.1 Abstract

Cattle have expanded the *KIR* gene repertoire, a polymorphic and polygenic immunoglobulin family that encode Natural Killer cell receptors specific to MHC class I ligands. In humans, KIR are important mediators of innate immunity to viral pathogens such as HCV and HIV, and there is potential for exploiting cattle *KIR* diversity as a means of improving animal health. Cattle *KIR* expansion has occurred independently to humans, the result is a cattle *KIR* haplotype (CKH) with a completely different gene content. Successful sequencing and assembly of the CKH using whole genome techniques has failed. To interrogate cattle KIR, their function and comparative evolution, the content of a CKH must be established, then the extent of polymorphism and gene presence/absence variation can be studied.

In this project the first CKH has been sequenced and assembled using second generation sequencing of BAC clones. This has provided a reference sequence for whole genome sequence data to be aligned revealing the *KIR* content of different *Bovidae* species, including the aurochs, the ancestor to all domesticated cattle. Furthermore genome capture and enrichment was performed to determine polymorphic and polygenic *KIR* variation within 24 different cattle genomes. The sheep *KIR* haplotype (SKH) was sequenced using PacBio of BAC clones to enable comparative analysis with cattle.

The CKH has expanded through block duplications resulting in 16 discrete *KIR* loci. The haplotype is dominated by functional inhibitory receptor genes and the attenuated remains of activating *KIR*. Predicted similarity between aurochs and modern CKH suggests *KIR* blocks expanded through natural selection and not artificial selection generated through centuries of domestication. Comparative analysis of the SKH and CKH reveals that sheep have independently expanded at least five of the shared *KIR* that cattle have expanded. Cattle *KIR* are extremely polymorphic, with diversity focused within the Ig domains, regions predicted to interact with ligand.

0.2 Acknowledgements

Firstly, thanks to John Hammond for his supervision and guidance during my PhD project, it has been a great pleasure, a lot of fun and never a dull moment within the Immunogenetics group. I would like to thank the other members of the group past and present for their help and support during my project, specifically I would like to thank Mark Gibson for conducting the lab-side of the capture experiment with great skill and patience. I would also like to thank Alasdair Allan for help and providing reagents and templates within the lab.

I would like to thank Mick Watson for giving me the chance to learn new analysis techniques within Ark Genomics at the Roslin Institute. Furthermore I would also like to thank Andrew Warry and Giles Weaver for their help during the beginning of my bioinformatics education.

I would like to thank our collaborators, Paul Norman, Peter Parham and Libby Guethlein at Stanford for their help in conceiving the project and co-authoring the manuscript for the KIR haplotype paper. I would also like to thank David MacHugh and Steven Park for their generous donation of aurochs raw genome sequences.

I would also like to the thank my Imperial College supervisors Salim Khakoo and Mike Skinner for their help and advice during my project.

0.3 Declaration of originality

I declare that the work in this thesis is the result of my own work during my PhD project. All figures, tables and illustrations have been conducted by me unless stated otherwise.

0.4 Copyright Declaration

Contents

List of Figures

List of Tables

0.4.1 List of abbreviations

μl micro litres

3′ 3 prime

5′ 5 prime

BAC Bacterial artificial chromosome

Bota Bos taurus

bp base pair

CDR complementary determining regions

CKH Cattle KIR haplotype

CNV Copy number variation

CT Cytoplasmic tail domain

D Domain

FCAR Fc fragment of IgA receptor gene

gb giga basepairs (1,000,000,000 bp)

HF Holstein-Friesian

ITIM immunoreceptor tyrosine-based inhibition motif

kb kilo basepairs (1000 bp)

KIR Killer cell immunoglobulin-like receptor

KLR killer cell lectin-like receptor

KU Kuchinoshima-Ushi

LILR leukocyte immunoglobulin-like receptor

LRC Leukocyte receptor complex

mb mega basepairs (1,000,000 bp)

MHC Major histocompatability complex

MID Molecular identitifier

mya million years ago

NCR1 Natural cytotoxicity triggering receptor 1

NK Natural Killer

NKC Natural Killer Complex

OLC Overlap layout consensus

PacBio Pacific biosciences

PCR polymerase chain reaction

s seconds

SHP-1/2 Src homology region 2 domain-containing phosphatase-1/2

SKH Sheep KIR haplotype

SMRT single molecule real time sequencing

TCR T-cell receptor

TM transmembrane domain

ZMW Zero mode waveguide

1 Chapter 1. Introduction

1.1 Natural Killer cells

Natural killer (NK) cells are large granular lymphocytes of the innate immune system that express a diverse range of inhibitory and activating receptors. NK cells display cytotoxicity alongside the ability to produce cytokines [34]. They were first described in the early 1970s by Professor Rolf Kiessling and colleagues who described cells with natural cytotoxicity that recognise cells with missing self [55, 77, 78]. NK cells are now recognised as important immune effectors and regulators that are involved in the successful prevention or retardation of tumours [21] and several viral diseases including cytomegalovirus [4,137] influenza virus [87], herpes simplex virus [136], hepatitis C virus [76] and HIV-1 [93].

Upon activation, cytotoxic NK cells release the membrane disrupting protein perforin, this perforates the target cell to enable passage of cytotoxic granzyme proteases into the target cell cytoplasm and initiate apoptosis [48,64,70,138,139]. An alternative pathway has been described that shows perforin-independent NK cytotoxicity using the cell death ligands FasL and TRAIL [143].

NK cells interact with host cells using an array of receptor ligand combinations in order to recognise self molecules that also convey the infectious status of the cell. The receptors convey either activating or inhibitory signals which, once bound to expressed ligands, can initiate or retard NK cell function respectively. The receptor ligands include the major histocompatibility complex (MHC) class I molecules, which display peptides processed within the cell. Expressed peptides displayed by the MHC class I molecules are constantly surveyed by CD8+ cytotoxic T-cells (CTLs) using an antigen specific T-cell receptor (TCR). Recognition and binding of the TCR to the MHC class I peptide complex leads to cytotoxic killing of the host cell. To avoid detection by CTLs, intracellular pathogens have generated mechanisms to down-regulate the cell surface expression of the MHC class I molecules. NK cells kill host cells that do not express MHC class I on the cell surface [71] or altered MHC class I that express non-self peptide [46], Figure 1. To enable continual recognition of their rapidly evolving polymorphic and often polygenic ligands, the NK cell receptor gene complexes contain considerable diversity.

1.1.1 NK cell receptor diversity

In many species the genes encoding the NK cell receptor families are polymorphic and polygenic generating diverse genetic complexes within populations. Humans and simian primates have expanded the killer-cell immunoglobulin-like receptor

(*KIR*) genes, located within the leukocyte receptor complex (LRC), resulting in multiple different gene complexes containing both gene presence absence variation and polymorhisms. The KIR are cell surface NK cell receptors containing two or three Ig domains that recognise and specifically bind ligands, the majority of which are MHC class I molecules. KIR have been shown to interact with MHC class I molecules through cellular adhesion and functional assays using cellular and non-cellular targets [42, 142], direct interaction has been studied using x-ray crystallography and mutagenesis studies [15, 25, 47, 141]. A critical number of receptors are required to interact with their ligand in order to breach the threshold required to initiate an inhibitory or activating pathway. Therefore, the ligand requires sufficient down-regulation to prevent inhibitory receptors clustering enough to signal. Alternatively, enough peptide will need to be processed and expressed to alter the receptor-MHC interaction [46].

In humans, the heterogeneous development of NK cells between individuals is a result of allelic variation of MHC class I genes and *KIR*. The importance of variable NK cell receptor repertoires has been highlighted by resistance to certain pathogens through possessing the correct KIR-MHC class I combination. For example the *KIR3DL1*004* allele alongside the HLA-Bw4 epitope slows the progression of HIV infection to AIDS in comparison to HLA-Bw6, which is not a ligand for KIR3DL1 [3, 53]. Therefore, MHC class I diversity is largely responsible for shaping the evolution of *KIR* and maintaining their existence within populations. Pathogen selection pressures have driven MHC class I evolution and therefore indirectly affected *KIR* evolution [110].

These pathogen and MHC class I mediated selection pressures have independently driven NK cell receptor gene expansion and diversification in different species. Humans, chimpanzees [75], orangutans [57], macaques [12] and cattle [94] have expanded *KIR* gene complexes. In contrast, equine and murine genomes contain expanded killer cell lectin-like receptor A *KLRA* gene complexes, also called *Ly49*. The *KLRA* map to the NK Complex (NKC), this contains several more c-type lectin receptor genes and is located on chromosome 6 in both species [131, 146, 147]. Furthermore the prosimians have expanded the CD94/NKG2 family of genes which also map to the NKC, proving a third route for generating NK cell receptor diversity [6]. The CD94/NKG2, KIR and KLRA receptors are structurally different yet occupy the same function in NK cells, providing an excellent example of convergent evolution. In humans and mice, evolution of NK cell receptors in concert with MHC class I diversification has resulted in increased resistance to certain pathogens [4, 38, 62, 76, 80, 136]. Therefore, the generation of NK cell receptor diversity has resulted in greater fitness of

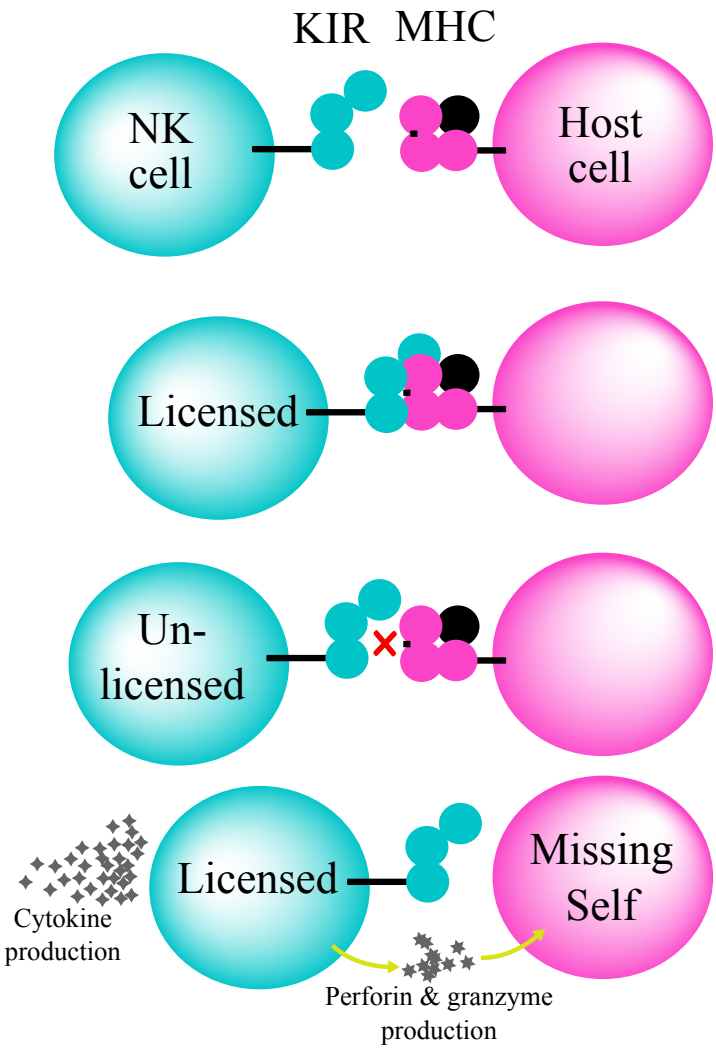


Figure 1: Diagram of NK cell education and killing of cell non expressing self MHC molecule. NK cell expresses inhibitory receptors in this instance a KIR molecule. Host cells express MHC class I molecules. To be able to kill, NK cells need to engage an inhibitory receptor to a self MHC class I molecule and become licensed. NK cells are hyporesponsive or "un-licensed" if they do not engage inhibitory receptors, a result if the inhibitory receptors do not recognise self MHC class I. NK cells kill host cells missing self MHC class I by cytokine and perforin/granzyme production.

the species. However, several species have been identified that have not expanded NK cell receptor genes that are currently known, including but not limited to seals [59], bats [149] and pigs [118]. Within an individual, NK cells express different combinations of receptors that are acquired during the maturation of the NK cells in a process termed education or licensing.

1.1.2 NK cell receptor acquisition

NK cell receptor acquisition occurs during maturation from haematopoietic progenitor cells [95, 123, 124] and results in cells expressing all the possible combinations of receptors in their genetic repertoire [95]. However, certain receptor combinations that result in functional NK cells are selected for during NK cell education and these dominate the NK population [145]. Functional NK cells can be activated and are capable of cytotoxicity or cytokine production in humans, however non-functional NK cells are incapable of becoming activated. The first receptors expressed by NK cells during maturation are the lectin-like receptor/C-type lectin heterodimer CD94/NKG2A [132]; in humans this is followed with expression of the KIR and in mice the C-type lectin KLRA [96]. In humans, individuals with a single strong KIR-MHC class I interaction develop a KIR dominant NK cell repertoire. However, if there are several strong KIR-MHC class I interactions the NK cell repertoire will be NKG2A dominant [145]. This is also true of individuals with only weak or no KIR-MHC class I binding, creating a balance of NK receptor to MHC class I that pivots around NKG2A [145]. After maturation the heterogeneous NK population then go through "education" or "licensing" to prevent potential self-reactive NK cells and become "licensed".

1.1.3 NK cell education

The process of NK cell education is still unclear and there are several proposed models that account for the responsive and hyporesponsive NK cells in peripheral blood [67]. The "arming" or "licensing" model requires an inhibitory receptor to recognise self MHC class I before becoming functional. The "disarming" model predicts that without inhibitory receptors that recognise self MHC class I, NK cells become "anergic" and therefore are no longer functional [67, 115]. The "cis-interaction" model requires an inhibitory receptor to bind to its ligand on the same target cell surface in order to prevent "licensing" from another source [23, 67]. The "rheostat" model is an amalgamation of the arming and disarming models that places the activation states onto a continuum, therefore the cellular response depends on the strength of the inhibitory receptor contact

during education [16, 67]. The education process, whichever theory is subscribed to, licenses NK cells to only activate when they have previously engaged an inhibitory receptor. This prevents auto-immune NK cells that cannot recognise self from being activated. Successful education of NK cells results in cells that can become activated, with cytotoxic NK cell subsets displaying a CD56-(dim) phenotype and cytokine producing NK subsets are CD56+(bright) upon activation [34]. The complexity of NK cell receptors and their ligands has made the process of education essential, furthermore expanded and differentiated inhibitory receptors that are specific to host ligands can generate potent NK cell responses better capable of killing virally infected cells. However, there is evidence that uneducated or "unlicensed" NK cells play an important role in fighting viral infection [107].

1.1.4 Cattle NK cells

Cattle are a species of huge economic importance, exploiting the natural immunity provided by their NK cells could improve animal health and boost productivity. Cattle NK cells have been implicated in various diseases that have important economic and zoonotic consequences including *Mycobacterium bovis* (bovine tuberculosis) [39] and bovine herpes virus 1 (BHV-1) [33]. Therefore, to understand the complex and sophisticated interactions involved in cattle NK cell functions, cattle *KIR* need to be understood. The expansion of *KIR* in cattle is a result of gene duplication from the X-lineage, which has remained monogenic in humans as *3DX1*. The reverse is true in humans that have expanded genes from the L-lineage which remains monogenic in cattle [56]. Therefore cattle are unique in that they are the only known species outside primates to have expanded *KIR*.

1.2 The killer-immunoglobulin-like receptors

KIR encode for activating and inhibitory receptors with two or three immunoglobulin-like (Ig) domains, Figure 2. The Ig domains are named from n-terminus to c-terminus D0, D1 and D2 respectively and interact with their MHC class I ligands. Signalling from activating KIR receptors occurs from within the transmembrane domain, a basic residue (usually lysine) interacts with an aspartic acid in DAP12 [49, 121] causing cross-linking of KIR and DAP12 [82]. This causes phosphorylation of DAP12 resulting in recruitment of ZAP-70 and Syk proteins implementing the signalling pathway required for cellular activation. Inhibitory KIR signalling is mediated by the immunoreceptor tyrosine-based inhibition motif (ITIM), which has the canonical residue sequence of I/VxYxxL/V [7]. During

KIR clustering the tyrosine residues within the motifs are subject to phosphorylation by Src family kinases with the resulting SHP-1 and SHP-2 recruitment initiating a pathway to inhibit NK cell function [7]. KIR clustering occurs when multiple receptors engage with mutliple ligands within the NK cell host cell synapse. Therefore KIR signalling is not binary; it involves many receptors to generate a response.

1.2.1 *KIR* nomenclature

The signaling ability of the KIR receptor is denoted in the name assigned to the gene encoding it along with the number of domains it contains. KIR nomenclature is defined by the components of the gene, making the naming descriptive and discerning. Naming uses the number of domains (2D or 3D), the signalling potential of the receptor (inhibitory is L for long tail, activating is S for short tail) and the chronological discovery order of the gene (1-5). Therefore the first three domain inhibitory *KIR* discovered in humans is called *3DL1* and the second two domain KIR gene discovered is called *2DS2* etc. Not all *KIR* haplotypes contain the same complement of genes; the LRC in which they are contained is gene dense, polymorphic and variable.

1.2.2 LRC gene structure and diversity

The LRC in humans contains the *KIR* complex that is flanked by the leukocyte immunoglobulin-like receptor genes (*LILR*) at the centromeric end and the Fc fragment of IgA receptor gene (*FCAR*) at the telomeric end, Figure 3. There are two forms of human *KIR* haplotypes, a gene consistent "A" haplotype and a gene variable "B" haplotype, with human *KIR* haplotype "A" containing fewer genes compared to the "B" haplotype. Haplotype "A" is hypothesised as having a more inhibitory role with a single activating receptor. This haplotype has greater potential for generating educated or licensed NK cells that are more potent at fighting infections. The "B" haplotype has a more activating role with as many as five activating *KIR* [110]. This haplotype is believed to have a greater role during pregnancy; NK cells reshape the uterine arteries for trophoblast implantation with a population expressing higher prevalence of activating receptor phenotypes [99]. Both these haplotypes are found in all human populations and it is believed that a balance of the two haplotypes within a population is required to effectively fight infection and mediate placental development simultaneously [1, 109]. The multitude of *KIR* loci is indicative of gene duplication and expansion which can be explained by genetic mechanisms including homologous and non-homologous

recombination.

1.2.3 Genetic recombination and generation of diversity

Whilst CTL TCRs undergo somatic VDJ recombination to recognise MHC class I with peptide, NK cells are only equipped with receptors that are germline encoded. Therefore, during simian primate evolution it has been an advantage to duplicate and diversify these inherited NK receptors several times in order to broaden the spectrum of NK receptor recognition ability. This results in a more comprehensive range of NK cell receptors that recognise MHC class I and therefore contribute to the immune response. This suggests that as the major ligand for KIR, MHC class I has driven the expansion of the KIR genes [57].

To generate diversity in NK receptor gene families, various genetic mechanisms have been employed. An example in humans is the generation of the fusion gene *KIR2DL5A/3DPA* is thought to have occurred due to misalignment of *KIR* genes on homologous chromosomes during synapsis of meiosis. This resulted in unequal crossing over producing a haplotype with a novel fusion gene and the duplication of *KIR2DL4* and *KIR3DL1* [92]. This non-reciprocal recombination mechanism has been proposed for generating diversity on the *KIR* locus by creating new fusion genes and also duplicating genes that can subsequently diverge through point mutations [92].

The *KIR* cluster contains several highly homologous genes that share large quantities of nucleotide sequence. It is believed that these properties of the *KIR* cluster enhance the likelihood of misalignment during synapsis and help drive the rapid expansion of *KIR* in humans [92]. Further human KIR haplotype diversity is believed to have occurred due to homologous recombination. In humans this reciprocal recombination mechanism is thought to switch alleles between haplotypes, one example is *3DL1/S1* that share a locus. Haplotypes containing this locus always have either *3DL1* or *3DS1*, and therefore these alleles define the haplotype. The other genes on the *KIR 3DL1/S1* haplotypes are not fixed however and it is believed that homologous recombination of parent chromosomes during synapsis has caused these alleles to be transferred between haplotypes creating new haplotypes with varying allelic content [102]. The mechanisms generating a dynamic *KIR* haplotype sequence enable continued recognition of the equally dynamic MHC class I molecules that are under constant pathogen mediated selection pressures.

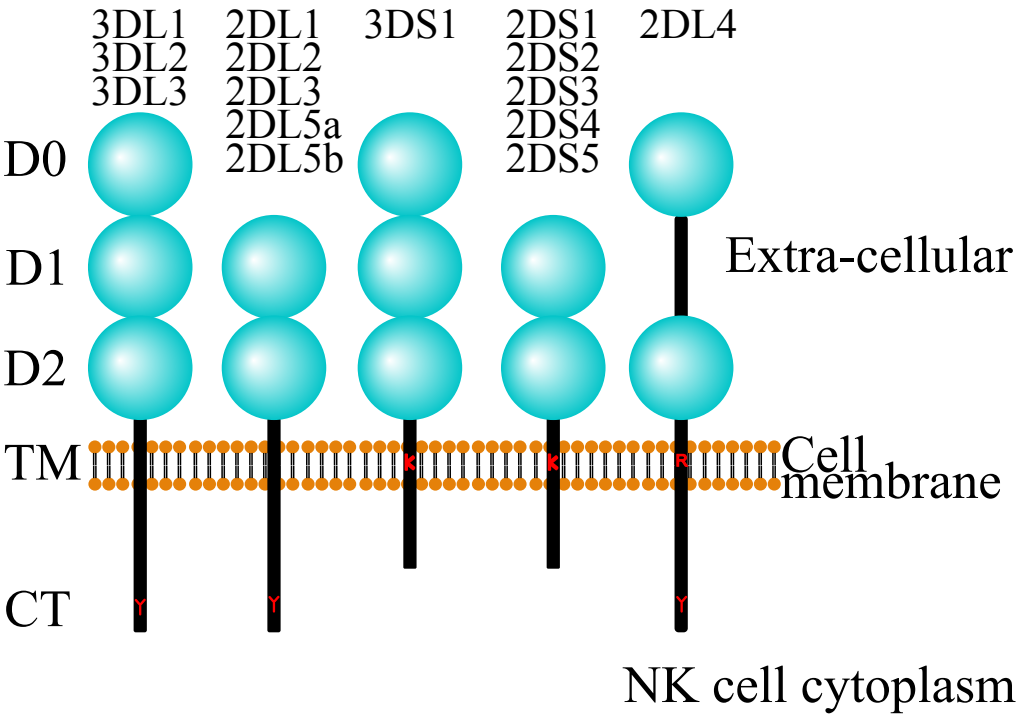


Figure 2: Diagram of different human KIR molecule forms. The names of the different human *KIR* are shown over the forms. Light blue circles represent Ig domains, all 3 Ig KIR have the D0-D1-D2 form and all 2 Ig KIR have the D1-D2 form, expet for 2DL4 that has the D0-D2 form. The cell membrane phospholipid bi-layer is shown and the KIR transmembrane domain (TM) passes through this. Activating KIR contain a basic lysine (K) residue within the TM. Inhibitory receptors contain a tyrosine (Y) residue within the cytoplasmic tail domain (CT). The 2DL4 receptor contains a basic arginine (R) residue within the TM and has a long CT containing a tyrosine residue.

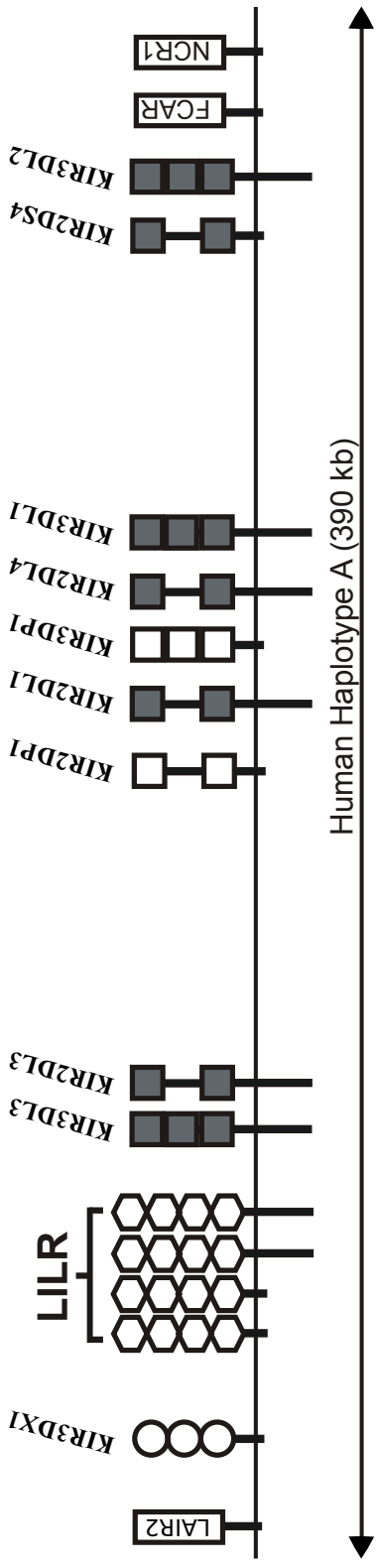


Figure 3: Diagram of human LRC gene structure. Each *KIR* is shown in its expressed molecular form. Squares represent Ig domains, grey are functionally intact and clear is a pseudogene. Inhibitory and activating genes are shown with short and long tails.

1.3 Major Histocompatability Complex

The major ligands for the KIR and KLRA receptors are the polymorphic MHC class I molecules. These glycoproteins express endogenously processed peptides to the cell surface where the peptides can be surveyed by host immune receptors. The TCR is the product of somatic genetic recombination to generate a specific immunoglobulin receptor capable of only detecting non-self peptides expressed on self MHC molecules. T-cells have undergone positive and negative selection to firstly only weakly bind self-MHC molecules and secondly only strongly bind MHC molecules expressing non-self peptides. The major role of MHC class I is to display the internal health of the cell, pathogen peptides are detected by CTLs that kill the infected cell. To escape this process, intracellular pathogens have developed mechanisms of down-regulating the MHC class I molecules to prevent detection by T-cells. The role of NK cells is to detect this "missing-self" and kill cells that are suppressing MHC class I expression.

The class I molecules consists of a heavy chain containing three extracellular α domains, with the $\alpha 3$ proximal to the membrane surface and $\alpha 1$ and $\alpha 2$ forming a cleft that contains the oligopeptide projected away from the cell. The MHC class I molecule associates with the conserved β_2-microglobulin protein, which combined, forms a stable heterodimer on the cell surface. The TCR interacts with regions of the $\alpha 1$ and $\alpha 2$ domains surrounding the cleft to establish a connection with self MHC, but contains hypervariable complementary determining regions (CDRs) that specifically recognises certain foreign peptide sequences. Crystal structures show the D1 and D2 domains make contact with MHC class I in two domain KIR [15, 19, 47]. The D0 in three domain KIR extends towards the β_2-microglobulin protein contacting a less polymorphic region of the molecule in what has been described as "innate sensing" the MHC class I molecule [141], Figure 4.

There are three classical MHC class I genes in humans, MHC-A, MHC-B and MHC-C, which maintain a regular genomic structure in all human genomes but display an unrivalled level polymorphism. Each locus has multiple families of alleles which encode differing detectable MHC class I serotypes such as Bw4, A3/11, C1 and C2. Each serotype encodes a shared epitope structure despite further polymorphisms between the alleles within each serotype, furthermore these epitopes are not necessarily specific to a single locus. The Bw4 epitope is found on certain MHC-A and MHC-B molecules and the C1 epitope is found on certain MHC-C and two MHC-B allotypes [97]. There are three non-classical MHC class I genes, MHC-E, MHC-F and MHC-G that are located within the same genomic

region as the classical class I genes. MHC-E and MHC-G loci encode few allotypes and are the ligands specifically for NK cell receptors CD94/NKG2D and *KIR2DL4* respectively. MHC-E expresses the leader peptide sequence of other classical class I MHC molecules and MHC-G is a secreted form that does not get expressed on the cell surface.

The human KIR are specific to different MHC class I epitopes [111]. The type II KIR, namely 3DL1/S1 and 3DL2, recognise the Bw4 and A3/11 epitopes respectively. Whereas the type III KIRs largely recognises the C1/2 epitopes, 2DL2 recognising C1 and 2DL1, 2DS1, 2DL3 recognising C2 [97]. *KIR2DS4* is the only activating gene found on the human A haplotype and recognises the A3/11 epitopes and some C1 and C2 epitopes. The MHC-A and B molecules do not all contain KIR specific epitopes, therefore some MHC-A and B allotypes are unrecognisable by the KIR. However, the MHC-C molecules all contain either C1 or C2 epitopes that are recognisable by KIR receptors. During human evolution MHC-C has co-evolved with and been influenced by NK cell receptors to fulfil an NK cell dependent role [104], whilst the evolution of MHC-A and B genes has largely been influenced by the TCR. This highlights the importance of NK cell receptors which have heavily affected the evolution of a classical class I gene.

1.3.1 Decoy MHC class I proteins

The *herpesvirdae* family is a group of viruses containing large DNA genomes. These viruses have the capacity to encode MHC class I analogue proteins that mimic the shape of the NK cell receptor ligands on the surface of the infected cell. The virus suppresses expression of MHC class I to prevent detection by the the specific T-cell response and expresses MHC class I decoy proteins to prevent detection of "missing-self". Therefore, the virus is subverting the inhibitory signals generated by the receptor to evade NK cell killing.

One of the first examples of this viral subversion was identified within the mouse model between murine cytomegalovirus (MCMV) resistant C57BL/6 and the susceptible BALBc and 129/J mice strains [4, 125], Figure 5. MCMV expresses m157, an MHC class I analogue protein that interacts with the lectin-like inhibitory receptor Ly49I, which is expressed by the MCMV sensitive 129/J mouse strain amongst others. The MCMV resistant C57BL/6 mice strain expresses the activating Ly49H receptor that recognises m157 and causes NK cell activation, resulting in killing of the virally infected cell via cytokine production and cytotoxicity. It is believed that m157 has evolved within MCMV in order to subvert the NK cell response and evade detection by the immune system. To counter this it is believed that Ly49H has evolved to detect this decoy protein

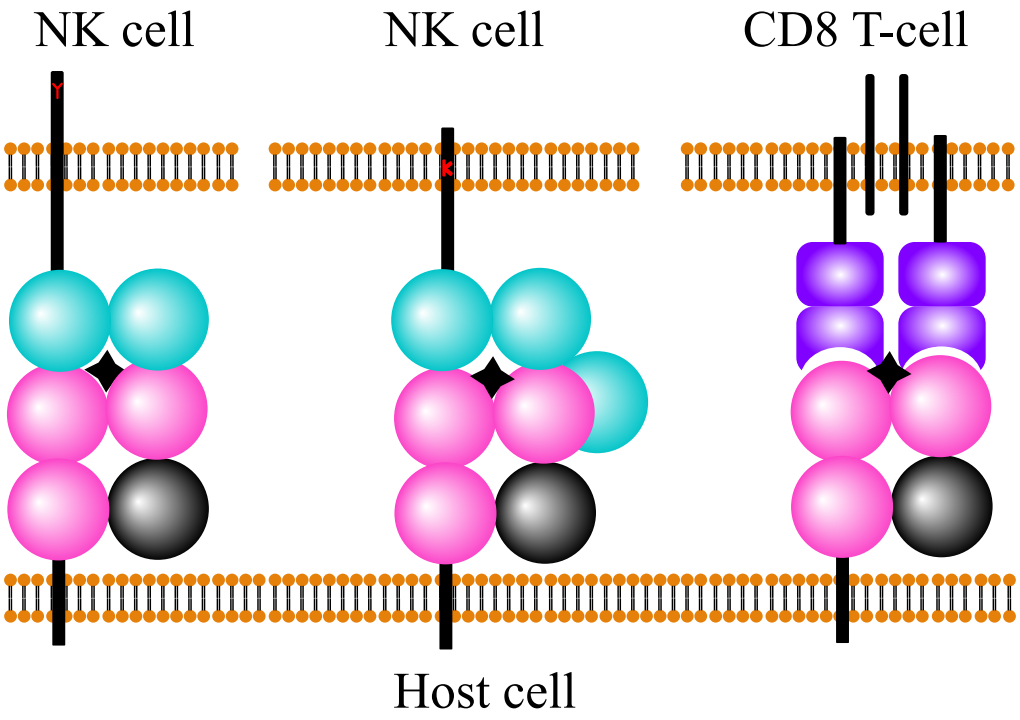


Figure 4: Diagram of a host cell MHC class I molecule interactions with NK cell KIR and T-cell TCR receptors. The pink domains are the MHC heavy chain proteins and the black domain represents the β_2-microglobulin protein. The light blue represents the KIR Ig domains, the left most is a two domain inhibitory receptor and the middle represents a three domain activating receptor. The purple domains represent the TCR receptor that is specific to the MHC class I molecule and the peptide, shown as a black cross.

and prevent viral evasion. There is significant polymorphism within both Ly49H and m157 gene sequences that has resulted in variations in the ability of the virus to evade NK cell detection [35], in what appears to be a continuation of the evolutionary arms race between the host and the virus.

A further example has been characterised within humans and another herpes virus, human cytomegolovirus (HCMV) [63, 113, 144]. The CD94-NKG2A heterodimer recognises the non-classical MHC class Ib protein HLA-E. The HLA-E molecule expresses the leader sequence peptides from the expression of other MHC molecules. This provides an indication of the level MHC class I expression by a target cell and viral suppression of MHC molecules is detected by CD94-NKG2A bound NK cells. MCMV encodes an MHC analogue, UL18, and loads HLA-E with a virally derived sequence peptide, UL40, that in tandem act as decoy for MHC suppression. UL18 complexes with the β_2-microglobulin and is recognised by the inhibitory leukocyte immunoglobulin-like receptor (LILR) B1 [24, 36]. Therefore, the expression of UL18 by the virus prevents detection of "missing-self" by the surveying NK cells. The production of the UL40 signal peptide analogue causes expression of HLA-E to prevent detection of MHC suppression by CD94-NKG2A [113]. Therefore, HCMV has evolved multiple mechanisms to subjugate NK cells in humans.

There are no examples describing viral subversion of the KIR receptors, however predictions based on the level of polymorphisms within the *KIR* sequences have indicated that it is occurring [20]. The job of recognising "missing-self" by detecting suppressed MHC by an inhibitory receptor can be achieved using monomorphic receptors that recognise the conserved regions of the MHC molecule. This has been proven with computational modelling [20], showing viral clearance with monomorphic receptors that recognise self MHC. However, this enables the possibility of decoy proteins mimicking the conserved regions of the MHC molecule to subvert the inhibitory receptors on the NK cells. Therefore, to detect suppressed MHC, yet not be fooled by decoy proteins, the KIR receptors are highly specific to their ligands. It is this specificity that drives diversity within the *KIR* sequences to generate a "heterozgous advantage" [20]. Diverse and heterozygous receptors are more likely to generate NK cells capable of specifically detecting MHC ligands thus generating a licensed NK cell population. Therefore a KIR repertoire that is only required to recognise self is achievable with degenerate receptors. However, under viral-decoy protein mediated selection pressures, MHC specific KIR repertoires are favourable. This indicates that the highly polymorphic and selective human KIR repertoire has evolved under pathogen decoy selection pressures. Furthermore, the inhibitory KIR MHC ligands have

largely been elucidated, however the specificity of the activating receptors remains largely unknown [98]. This suggests the human activating KIR ligands could be viral decoy proteins that are yet to be identified. Furthermore, it has been suggested that HCMV infection has elicited a self-specific inhibitory KIR and greater activating KIR response from NK cells, which is indicative of herpes viral decoy proteins [9]. This has been detected by the phenotypic characteristics of NK cell sub-populations generated in individuals infected with HCMV.

1.3.2 Recurrent evolution of activating receptors

Within the different species known to have expanded the *KLRA* and *KIR* genes, the recurring formation and subsequent deletion of short tailed activating genes has occurred [1]. It is believed that the inhibitory receptors are ancestral and the activating receptors are derived from them. A recombination event or point mutations within the transmembrane and cytoplasmic domains results in a short tailed activating receptor with the same ligand specificity as the inhibitory receptor. It is believed this process is driven by viral subversion of inhibitory receptors, with the expression of MHC decoy proteins to subvert the NK cells. Therefore, there is a pathogen selection pressure to recognise the decoy protein with an activating receptor instead of an inhibitory receptor. Therefore, a major role of activating NK receptors is for fighting infection by detecting viral decoy proteins. These activating NK receptors have a lower specificity than their inhibitory counterparts, however, they potentially maintain the capability to recognise self MHC. Whilst the viral infection is prevalent, maintaining functional decoy specific activating receptors provides a selective advantage. However, once viral infection has subsided at the population level, maintaining functional decoy specific activating receptors becomes a selective disadvantage. This is due to the potential for the generation of autoimmune NK cells resulting from self recognising activating receptors. Therefore, activating NK cell receptors are short lived, becoming null-alleles or pseudogenes after their function has been served.

1.3.3 Cattle MHC class I genes

The expansion of cattle *KIR* is believed to mirror the polymorphic and structurally variable MHC class I within cattle. It is predicted that cattle KIR also recognise MHC class I molecules however, this has not yet been shown. Cattle have six MHC class I loci with a maximum of three genes per haplotype, only certain MHC class I genes are found on the same haplotype [11, 32], Figure 6. Therefore, cattle have higher MHC class I structural diversity than humans and

this potentially increased ligand diversity in cattle may have driven the expansion of the *KIR* genes. The MHC class I genes in cattle are also highly polymorhic although the limit of this polymorphism has not been characterised to the same extent as the human MHC genes. Furthermore polymorphism may have been reduced during domestication of cattle due to genetic bottlenecks, founder effects and inbreeding. The MHC genes in cattle are found on chromosme 23 and the *KIR* and NKC genes map to chromosomes 19 and 5 respectively. Therefore there is no linkage between the NK cell receptors genes and their predicted ligand genes. This means the cattle NK cells have the potential to recognise six independent MHC class I molecules, far greater than the number human NK cells are required to recognise. Furthermore there is no constant MHC class I gene that is on all of the haplotypes. Therefore the cattle *KIR* haplotype must encode specificity for at least more than one gene. This suggests that cattle MHC class I expansion has driven cattle KIR evolution akin to that seen in primate *KIR*/MHC class I evolution. Furthermore, the increased structural diversity of the cattle MHC class I genes may have had a greater impact on cattle *KIR* genes. This leads to the hypothesis that pathogen selection pressures acting indirectly through structurally diverse MHC I ligands has impacted the evolution of the cattle *KIR* genes.

To test this hypothesis and determine the expansion of cattle *KIR*, a complete cattle *KIR* haplotype must be sequenced. A sequenced *KIR* haplotype will determine the extent of *KIR* expansion and will be informative to which *KIR* families have expanded and which genes have become pseudogenes. With the knowledge gained from sequencing a *KIR* haplotype it will then be possible to determine the extent of polymorphism and study the role of KIR receptors during NK cell function. Cattle have evolved from shared ruminant ancestors with other domesticated animals such as sheep and goats. These animals may also encode similar *KIR* and *MHC* genes that may also have been affected by domestication. Therefore evolution and domestication of cattle should be considered when studying the *KIR* and *MHC* genes.

1.4 Evolution and domestication of ruminant species

The first ruminant species evolved approximately 50 million years ago (mya) [65]. Speciation of ruminants approximately 32 mya resulted with the *Cervidae* (deer) clade of species and the *Bovidae* (cattle, sheep, goats, buffalo, bison) clade. The *Bovinae* species which includes cattle as well as bison, water buffalo and American buffalo split from the other *Bovinae* species, including *Caprinae* (sheep and goats) approximately 25.4 mya [65]. The *Bos* and *Bison* species diverged

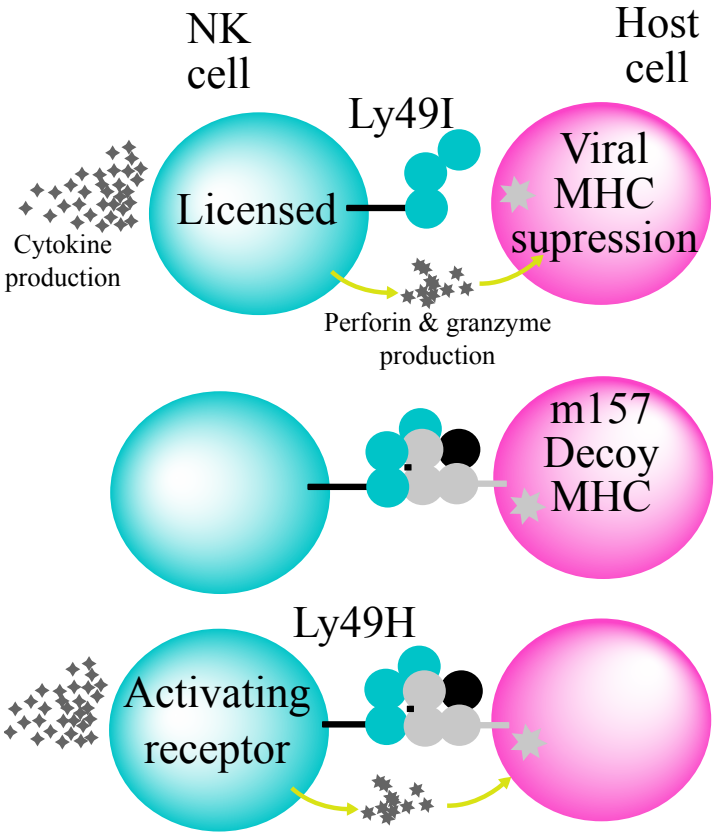


Figure 5: Diagram of NK cell subjugation. Licensed mouse NK cells expressing the inhibitory Ly49I receptor detect missing self after viral MHC suppression and kill the infected target cell. HCMV expresses the MHC decoy protein, m157 that Ly49 recognises as self and prevents killing. Switching the function of the m157 recognising receptor to an activating signal via Ly49H causes NK mediated killing of the decoy expressing virally infect host cell.

from the *Bubalus* species approximately 17 mya, with *Bos* and *Bison* subsequently splitting approximately 5.8 mya. The two most abundant agricultural *Bos* species, the taurine (*Bos taurus*) and indicine (*Bos indicus*) cattle, are believed to have diverged between 610,000 and 850,000 years ago to generate two species of ancient aurochsen, the *Bos taurus primigenius* and *Bos indicus primigenius* [90]. The most common breeds of cattle used for agriculture in Europe and north America are all of the *Bos taurus* species, which was domesticated from the ancient wild aurochs cattle (*Bos taurus primigenoius*) approximately 10,000 years ago [8, 43]. Within the last 10,000 years humans have domesticated the *Bos* species to generate thousands of specialised breeds used for milk production, meat, leather, transportation and dual purposes. Within Europe and North America, the Holstein-Friesian, or slight variants thereof, have become the highest milk yielding breed after centuries of intensive artificial selection and inbreeding [122].

The domestication of cattle for agricultural use has focused on the production traits such as milk yield and muscle growth. This has resulted in cattle that greatly outperform the productivity of animals used in agriculture 50 years ago. However, the focus on productivity may have overlooked health traits such as disease resistance and reproduction [51]. The inbreeding associated with domestication has resulted from back breeding of cattle to retain specific traits as well as the use of small numbers of bulls to sire hundreds of herds within a country. Therefore, domestication may have affected the *KIR*, and many other immune genes, within the cattle genome. There may be a lack of diversity within the *KIR* haplotypes and within the *KIR* sequences. Founder effect and genetic bottlenecks may cause a limitation in the number of *KIR* haplotypes within modern cattle as well as reducing the number of alleles for each *KIR* locus. Therefore, the effects of domestication within cattle should be considered. The cattle *KIR* complex structure may have been heavily affected by artificial selection during domestication and the level of polymorphisms within cattle may be lower than expected.

1.5 Known *KIR* in the cattle genome

The first cattle *KIR* genes to be described were *BotaKIR3DL1* and *BotaKIR2DL1* by McQueen et al in 2002 [94] and named following the human and primate nomenclature system. Storset et al described the next two *KIR* genes in 2003 [129] that included the activating receptors *BotaKIR3DS1* and *BotaKIR2DS1.* Several more cattle *KIR* genes were described in 2007 by Dobromylskyj et al [41], finding *BotaKIR3DL2* and *BotaKIR3DL3* genes along with several alleles of

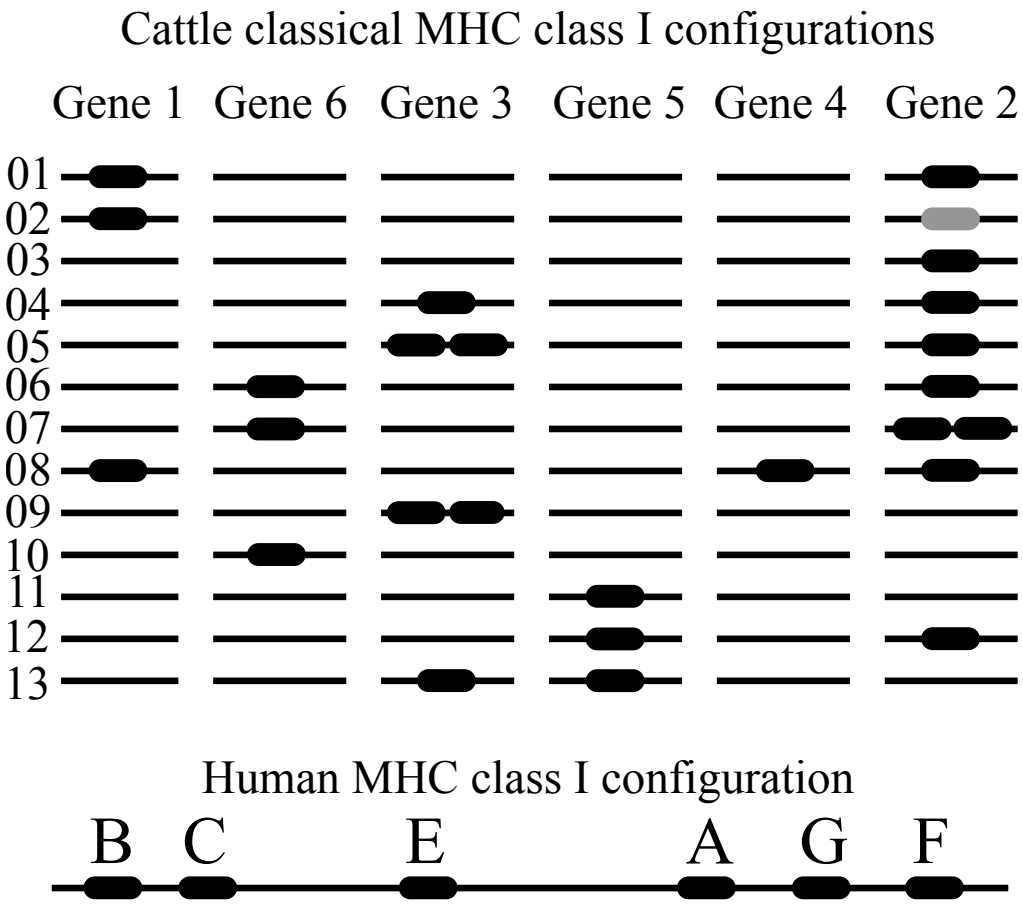


Figure 6: Diagram of cattle and human MHC class I haplotype structures. MHC class I genes are shown as black oblongs. Pseudogenes are shown as grey oblongs. The 13 different cattle haplotype forms only show the predicted orders of the classical class I genes. Non-classical genes exist but are not shown here. The exact order has not been confirmed by fully sequencing the haplotypes. The humans MHC class I haplotype region shows the gene structure seen in all humans including classical and non-classical MHC class I genes. Cattle MHC haplotypes diagram has been adapted from Ellis et al, 2014 [44]

each gene. Guethlein et al [57] went on to describe how the cattle *KIR* have evolved from the same lineage as *KIR3DX1* in humans, whilst the cattle gene *BotaKIR2DL1* is from the same lineage that has expanded in primates. Unpublished analysis of the sequences collected by these groups has determined *BotaKIR3DL3* to be a pseudogene and was therefore renamed *BotaKIR3DL1p*. *BotaKIR3DL1p* clades with *BotaKIR2DS1* and by considering identical intron sequences is potentially an allele, therefore presenting a possible example of variable haplotypes at the allelic level.

There are six *KIR* sequences that have been identified within the cattle genome, with four different molecule forms, Figure 7. However, the full genomic sequence has not been sequenced within any of these projects meaning the number of *KIR* within cattle is unknown. The cattle genome project has attempted to *de novo* assemble all of the chromosome, including chromosome 18 where the cattle LRC is located.

1.6 The LRC within the cattle genome project

Due to the mechanisms of *KIR* evolution, *KIR* within a species are very similar in sequence (over 90% sequence identity in humans) making distinguishing between the sequencing reads of different *KIR* loci difficult. Sequencing of expanded *KIR* haplotypes in different species is hard to complete. This is because the assembly of the reads is complicated when the reads only differ slightly (less than 1%) which is exacerbated with low coverage or short sequences. A cattle *KIR* haplotype has not been characterised despite the 7x read coverage produced by the cattle genome project and the advanced assembly techniques applied by the University of Maryland, there is not a fully assembled LRC [45, 151]. In the cattle genome, only one *KIR* is placed on chromosome 18 which contains the LRC, whilst several more *KIR*-like sequences are unplaced. The *KIR* haplotype will need to be sequenced again in order to assemble a complete and accurate sequence.

1.7 Advances in sequencing technology

The ability to sequence and assemble genomes or target gene complexes has become feasible due to advances in sequencing technologies and analysis techniques. The work in this thesis has benefited from these advances and has used a variety of sequencing technologies suited to the various questions asked. This has ranged from targeting individual gene exons to targeted genome enrichment to raw genome sequence analysis using traditional and new technologies.

1.7.1 Sanger sequencing

The traditional and most widely used method of sequencing until recently was developed by Frederick Sanger and colleagues in 1977 [119]. This method uses oligonucleotide primer sequences to initiate the incorporation of deoxynucleoside-triphosphates (dNTPs) onto an elongating complementary strand of DNA. The dNTPs are a mix of the four nucleotide bases and are incorporated complementary to the target DNA strand. Strand elongation of bases is terminated by the random incorporation of a di-deoxynucleosidetriphosphate (ddNTP) which doesn't contain the 3′ hydroxyl group needed to make a phosphodiester bond, but does contain a fluorescent dye. The process generates DNA fragments of differing lengths which can be differentiated using capillary electrophoresis whilst the fluorescent dye is detected to determine the terminating base of the fragment. The terminating bases from the differing fragment lengths are sequentially detected from the light wavelength emitted to generate a chromatogram of wavelengths representative of a sequence of bases.

The major current providers of this technology and sequencing chemistry, Lifesciences' ABi BigDye 3.1, are capable of generating sequence lengths just over 1,000 bases (1 kilobase or kb), which is limited by the ability to differentiate between fragment lengths over 1 kb. The throughput is relatively low with each sequencing run producing 96 x 1 kb reads, resulting in approximately 96 kilobases (kb) of sequence data per run. This low throughput limits the use of Sanger sequencing for large scale projects such as genomes and BAC clones. Traditional Sanger sequencing still provides a method to quickly and cheaply target a gene sequence, however for projects over larger regions, newer generation technology is better suited.

1.7.2 Second generation sequencing

The second generation of sequencing technologies has dramatically increased the throughput of each sequencing run, generating billions of bases but at the expense of read length. The second wave of sequencing technologies has consisted of Roche's 454, Illumina's Solexa, ABi's SOliD and finally Lifescience's ION torrent that could each sequence individual reads in a massively paralleled fashion. These platforms provide differing approaches and variable efficacies for generating sequences from DNA and RNA templates. All of the technologies included a polymerase chain reaction (PCR) step to amplify the target sequence and required the final template to be fragmented. However, each technology has different mechanisms of sequencing the DNA fragments.

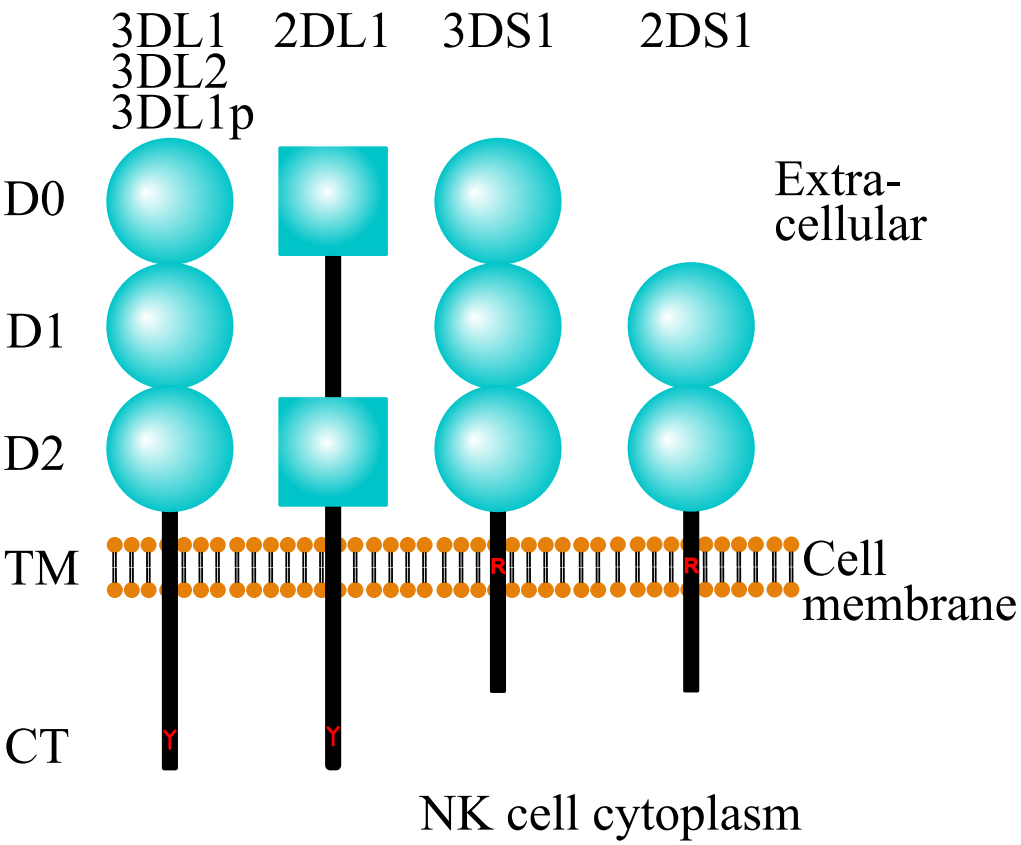


Figure 7: Diagram of known cattle KIR. Here are the predicted molecular forms of the cattle *KIR* genes previously published. Circles represent Ig domains from the X-lineage of *KIR* and squares represent the L-lineage. The short tail receptors encode an Argenine (R) residue akin to that seen in human 2DL4. Cattle 2DL1 has a D0-D2 structure whilst 2DS1 has a D1-D2 structure. The gene structure of the haplotype is unknown.

Roche 454 uses the pyrosequencing method of "sequencing by synthesis" by detecting the light from pyrophosphate release after nucleotide incorporation. Pyrophosphate labelled nucleotides are sequentially added during the run and are only incorporated if they are complementary to the next base in the target DNA strand. Incorporation is detected by the emission of light and the intensity is proportional to the number of nucleotides incorporated. This method enables 400 to 600 bp reads to be sequenced in parallel. However, the technology suffers where single nucleotides are repeated causing what is known as the "homopolymer repeat issue". The proportion of light intensity generated from multiply incorporated nucleotides diminishes with number. This becomes difficult to detect and errors are introduced. The Ion torrent system works on the same principle but detects the hydrogen ion (proton) release.

Illumina sequencing, previously called Solexa, also sequences by synthesis but uses clusters of clonal DNA template molecules seconded to a flow cell with fluorescent labelled nucleotide bases sequentially washed over. Light emission can be detected from each cluster after the bases have been sequentially incorporated. Illumina sequencing has upgraded the maximum read length of its sequencers from 25 bp to 150 bp from the HiSeq platform and 200 bp to 300 bp on the MiSeq platform. The process enables sequencing from both ends of the molecule to generate paired reads, where the first read is the sequence from the 5′ end and the second sequence is from the 3′ end. The distance between the two reads is the DNA molecule fragment size and often referred to as the library insert size. The Illumina sequencers offer unparalleled throughput with the highest accuracy, however the read lengths can be inhibitory when interrogating highly repetitive regions and the PCR steps required can introduce errors.

1.7.3 Third generation sequencing

The second wave of sequencing technology enabled entire genomes to be sequenced on a single machine with one run. However for targeted regions and genome finishing, the short reads are problematic. The third wave of sequencing technologies has addressed this by generating the longest reads, however throughput is lower. The only 3rd generation sequencing technology commercially available is Pacific Biosciences (PacBio) single molecule real time sequencing (SMRTcell sequencing) technology. However, Oxford Nanopore's potentially disruptive new minion sequencers are undergoing public testing and are an exciting prospect for sequencing large repetitive immune gene complexes like the LRC.

The PacBio SMRTcell sequencer utilises a polymerase molecule seconded to

the bottom of a nanometre wide well called a zero mode waveguide (ZMW). The ZMW is small enough that nucleotide incorporation into a DNA strand can be detected. The template strand of DNA is circularised and passes through the seconded polymerase whilst bases attached with four different fluorescent dyes are incorporated. Incorporations results in the cleavage of the dye and an emission of light which is detected as the specific base is incorporated. This single molecule sequencing process produces very long read lengths up to 25 kb, however the raw quality per read is relatively low at around 86% accuracy. There is also a considerable number of erroneous insertions that result from the polymerase "stuttering" on certain bases. The PacBio sequences are useful for providing a basic backbone sequence for a repetitive region of the genome. However, considerable bioinformatic developments have been required to use these reads and account for the high error rate.

1.8 Sequence analysis

The deluge of sequencing data generated from these technologies has required the advancement of methods to process, assemble/map and analyse the data. Numerous open-source tools have utilised mathematical and bioinformatic algorithms to answer biological questions from DNA sequence fragments and full genomic sequences. Deciding on which software to use depends on the algorithms utilised and whether they are appropriate to the data, this applies importantly to *de novo* assembly of sequencing reads.

1.8.1 Overlap Layout Consensus

The original method for the *de novo* assembly of sequencing read fragments using computers was to simply attempt to determine all of the possible overlaps between all of the fragments. Like putting the pieces of a jigsaw together, this approach attempted to put all the pairs of pieces that fitted (overlapped) together, using a graph method where each fragment is represented by a node. The overlapping fragments are laid out to generate a contiguous alignment (contig) of reads from which a consensus could be established. This overlap-layout-consensus (OLC) method worked well with the relatively long reads and low throughput of Sanger sequencing but fails to cope with the quantity of data generated by 2nd generation platforms.

1.8.2 De Bruijn graph

To tackle the quantity of data generated by short read sequencing, de Bruijn graphs have been utilised to reduce the computing power required for *de novo* assembly. The reads are reduced to *k-mers* which is representative of a DNA string of n base pairs length. Repeated *k-mers* throughout the raw sequences are reduced to the single *k-mer* sequence therefore eliminating redundancy and the calculations required. Overlapping *k-mers* form nodes with the reverse complement forming a twin node, therefore nodes are representive of *k-mers* from multiple reads and not a single read as in OLC methods. Nodes are connected by overlap from the first and last *k-mers* between the nodes. Shorter *k-mers* result in more overlaps and greater connectivity within the graph at the expense of sensitivity. Therefore a balance is required for correct and efficient assembly when choosing the *k-mer* size. Errors, repeats and paralogous sequences make more than one connection between nodes that complicate the graph and slow down or break the assembly. The raw reads are reused to generate a path along the graph that utilises the full read length of the sequences and is not compromised by the kmer length. However, the short read sequences are sometimes still too short to span the paralogous gene and repeat sequences within genomes. Therefore, de Bruijn graph assemblies can efficiently handle large datasets but are limited by errors and repeats and the read lengths used.

1.9 Aims of the project

The field of human and mouse NK biology has grown rapidly over the last 30 years and there are still many questions remaining. Understanding and exploiting cattle NK cells has huge potential for improving animal health through host genetics and vaccination. However, the major expanded NK receptor genes have not been characterised fully. To understand the mechanisms of cattle NK cell actions and potentially exploit them, this project aims to determine the genetic mechanisms responsible for generating diversity within the cattle NK cell receptor genes.

The first aim of this project is to sequence and assemble a cattle *KIR* haplotype. This will reveal the number of *KIR* genes within the genome and will enable prediction of the functional receptors cattle encode. Generating a reference sequence for the first *KIR* haplotype will provide a backbone for each of the genes to be studied; enabling polymorphism and gene presence/absence interrogation of cattle populations. Interrogating the cattle *KIR* sequences will also provide a novel insight into the expansion of the genes outside of humans

and primates. It will provide another example of NK receptor genes that have expanded and diversified through evolution and will therefore contribute to the study of evolutionary immunogenetics.

The second aim of the project is to determine the extent that domestication has impacted on the evolution of the cattle *KIR* complex. The effects of cattle domestication over the last 10,000 years will be interogated by comparing the modern cattle *KIR* complex to that of the extinct aurochs.

The third aim of this project is to sequence and assemble the sheep *KIR* complex. Comparisons can then be made between cattle and sheep *KIR* complexes. This will indicate the effects of evolution on the ruminant *KIR* complexes over the last 25 million years in much the same way humans have been compared to other primates. This will give an indication of the rate of expansion and diversification of the *KIR* genes that has occurred since the last common ancestor between sheep and cattle approximately 25 mya.

The fourth aim of this project is to determine in what species the cattle *KIR* haplotype structure has remained the same. This will also reveal to what extent cattle have variable gene presence/absence haplotypes. This will enable genotyping strategies to be targeted only at species which have the same gene structure.

The final aim of this project is to define the polymorphic regions of the cattle *KIR* haplotype. This will facilitate further high throughput genotyping projects by determining the conserved regions that primers and probes can be designed to target. It will indicate the most variable genes and alleles which could be under viral or ligand-mediated selection pressures.

2 Chapter 2. Sequence and assembly of a Cattle *KIR* haplotype

2.1 Introduction

Cattle are known to have multiple *KIR* genes [41, 56, 94, 129]. However, the extent of cattle *KIR* gene expansion and diversification is poorly understood. The current cattle genome build (as of writing this UMD 3.1) [45] is unfinished with regards to the LRC with several *KIR* genes placed on the X chromosome or on unmapped contigs. Therefore, to study the evolution and function of the cattle *KIR* genes the *KIR* complex needs to be sequenced and assembled to elucidate the gene characteristics and numbers.

Previously found cattle *KIR* genes from other projects*KIR* genes [41, 56, 94, 129] and the cattle genome, have been named following the convention set by the human and primate *KIR* gene nomenclature. However, as many cattle *KIR* genes have expanded from a different ancestral *KIR* gene, *3DX1* [56], genes of this lineage are substantially differentiated enough to have warranted renaming, thus preventing confusion with the human and primate expanded *KIR*. Therefore the previously discovered *KIR3DX1* lineage genes have been renamed with an "X" between the "D" and the "L" or "S", as have all the genes found throughout this project. The cattle genes that belong to the human and primate expanded "L" lineage follow the same nomenclature as human and primates and do not contain an "X". Table 1 contains the previously discovered genes and their new names.

To sequence and assemble the cattle *KIR* haplotype, Holstein Friesian BAC clones were isolated from an in house BAC library and sequenced with 1st and 2nd generation sequencing technologies. The sequences were assembled to provide a single contiguous consensus sequence that contains the cattle *KIR* complex.

2.2 Methods

2.2.1 BAC library screening for *KIR* positive clones

This first subsection (subsection 2.2.1) of these methods was carried out prior to the start of my project, I have included them to improve continuity and enable repetition of this work. A previously created BAC library [40] was made PCR screen-able to enable *KIR* specific primers to be used to find *KIR* positive BAC clones. Aliquots of 5 μl from each well were incubated with 5 μl water for 10 minutes at 96 °C. This template was used in PCR reactions containing cattle *KIR* specific primers, designed from previously published sequences [94] [129] [41] [56] and *KIR* containing contigs from the cattle genome build [45] (L-lineage: 3DL_ex2_S1 5′CAK AGS ATC TGG GCA CAAG′3, 3DL_ex3_AS3 5′GAA TAT GAT GCC CTG GAG CTC′3, X-lineage: 3DX_ex3_S 5′GTC TCT CSC TGT GTT TTC CAG′3, 3DX_ex4_AS 5′ATG ACG ATG TCC ACA GGA TCA′3). PCR was performed using GoTaq (Promega, UK) with optimised cycling conditions (95°C 1 min, (95°C 20s, 62°C 20s, 72°C 2.5min) x32, 72°C 5min). Initially, whole-plate templates were pooled so that full plates could be screened in one reaction; once *KIR* positive plates were identified their individual rows and columns were pooled and screened. If corresponding rows and columns shared a positive PCR result, a *KIR* containing BAC clone could be identified from the well that the row and column cross.

2.2.2 BAC plasmid DNA extraction and 454 sequencing

DNA from four *KIR* positive BAC clones was extracted using a large construct kit (Qiagen, UK) following the manufacturers protocol. Purified plasmid DNA was produced after digestion of the bacterial chromosome DNA. The super-coiled plasmids were purified through a silica column which filtered through the fragmented chromosomal DNA, allowing plasmid DNA to be eluted out separately. This pure plasmid DNA from BAC clones 095G05, 335H08, 032G11 and 068F04 was sequenced using the Roche 454 platform and titanium sequencing chemistry at the Stanford Genome Technology Centre (California, USA). The BAC clone 303D02 was sequenced using a 3 kb paired end library at the Liverpool Center for Genomic Research; the samples were multiplexed using molecular identifier tags (MID) and loaded onto one-quarter of a pico-titre plate. The BAC clone inserts were predicted to be between 110 kb and 200 kb based on previously conducted restriction digests and the average insert size of the library.

2.2.3 *De novo* assembly of 454 seuqneces

Roche 454 pyrosequences were extracted from SFF files using the SSF_extract python script written by Jose Blanca and provided in the MIRA package [27,28]. Raw sequence run statistics such as numbers, average lengths and read distribution histograms were produced using a bespoke python script (available in the appendix 9.1.4). Extracted 454 reads were screened for vector sequences using the SSAHA2 [101] program using the pBeloBAC11 vetor sequence (available from the CHORI website). The screened 454 pyrosequences were *de novo* assembled using the MIRA assembler [26, 29]. The MIRA setings used were accurate, genome, *de novo* with vector screening on.

2.2.4 Checking and editing of BAC sequence assemblies

Assemblies were checked for premature contig breaks caused by homopolymer repeats and incorrect contig joins caused by reduced read coverage using Gap4 and Gap5 from the Staden package [13]. Contigs were broken when read coverage was less than four 454 reads and/or the overlap was less than 10 bp.

2.2.5 Contig joining PCR, cloning and sequencing

Primers were designed within Gap4 at the ends of each contig. PCR using contig end primer pairs (primers shown in Table S1) was conducted using GoTaq (promega UK) and BAC clone template extracted using the CHORI BACPAC resources DNA isolation protocol (website: bacpac.chori.org/bacpacmini.htm). The thermal cycling profile was as follows: 95°C 1 min, (95°C 20s, 54°C 20s, 72°C 1min) x26, 72°C 10s. PCR products were seperated by gel electrophoresis on 1% agarose gels with 1 μl of 0.5 μg/ml concentration ethidium bromide per 100 ml gel volume. Positive PCR reactions, yielding products of expected size were excised from the gel over UV light using a scalpel. PCR products were purified from the agarose gel using qiaquick gel extraction kits (Qiagen, UK) then ligated into pGem-T easy vectors (Promega) both following the manufacturers guidelines. Ligated PCR-product vector constructs were transformed into in-house competent JM105 *E.coli* cells. Transformed cells were spread onto LB-agar ampicillin plates and grown overnight for 16 hours at 37°C; three positively transformed colonies per transformation were selected using blue/white selection and grown in culture overnight for 16 hours at 37°C in 3 ml of LB-broth with ampicillin.

Plasmid DNA from the cultures were extracted using Qiaprep spin miniprep kits (Qiagen, UK) following the manufacturers protocol. Sanger sequencing was

performed using ABI BigDye® terminator 3.1 (Life technologies, UK) following the manufacturers guidelines with either M13 forward or reverse primers. The BigDye reactions were run on ABI 3730 capillary sequencing machine at the University of Oxford (UK) Department of Zoology, resulting in three Sanger sequences per direction for each contig joining PCR reaction *i.e.* six Sanger sequences per join.

2.2.6 Hybrid assembly of Sanger and 454 sequencing reads

Sanger sequences were processed and manually edited for errors using the pregap4 module from the Staden package [127]. Edited sequence files were loaded into MIRA for hybrid *de novo* assembly with 454 sequences. Upon hybrid assembly, databases were checked and manually edited using the same criteria used for the initial assemblies with 454 reads alone. If the hybrid assembly was still unfinished *i.e.* split into contigs after editing, the process of assembly, primer design, PCR and introduction of more Sanger sequences was repeated until a single contig remained.

2.2.7 Error checking with Illumina sequencing

BAC clone DNA was prepared using the Qiagen large construct kit as described in section 2.2.2 and sequenced using the Illumina HiSeq platform and 100 bp paired end with 500 bp insert sizes at ARK Genomics, Roslin Institute, The University of Edinburgh (UK).

The raw fastq sequences were aligned to the reference sequence using bwa aln [84] then converted into sam format with bwa sampe. Once manipulated into a a sorted bam file using samtools [85] it was interrogated for SNPs using Varscan2 [79]. SNP effects such as residue changes and the exons they belong to were determined using a bespoke python script (available in appendix 9.1.3). The structure and positions of the gene sequences was confirmed by using the relative positions of the paired end reads. Gene order was confirmed by filtering reads that had greater inferred insert sizes than 1 kb. This was carried out with a set of bespoke python scripts (see appendix script 9.1.5).

2.2.8 Gene identification and annotation

Gene positions and structure were determined by BLAT searches [74] using all previously found cattle and human LRC genes against the BAC assembled consensus sequence. *KIR* genes were split into domains that enabled successful

BLAT hits of all of the *KIR* exons including the transmembrane and cytoplasmic tails regions which are not identified using whole gene sequences. BLAT hits were visualised with the artemis genome browser [117]. ITIM and transmembrane functional motifs were characterised using manual searches of the translated sequence. ITIM sequences were searched using the canonical sequence VxYxxL and slight variants thereof. Exact gene coordinates were calculated by eye. Exon positions were confirmed based on alignments of previously determined cattle cDNA sequences and splice junction donor acceptor sites were honoured based on the GT-AG motif [17].

2.2.9 Gene comparisons using phylogenetic, dot plot and sliding window analyses

KIR gene, exon or domain sequences were aligned using MAFFT [72] on automatic settings and manually corrected using Bioedit [58]. Neighbour joining phylogenetic trees were constructed using MEGA 5 [133] with 500 bootstrap replicates. Either the P-distance or Tamura-Nei algorithm was used depending on the comparison needed. Sliding window of average base sequence identity between sequences of certain window sizes was conducted using aligned sequences and a bespoke python script (available in appendix 9.1.2), the chart was generated using the matplotlib package [68] used within python. Dot plot analysis was performed using dotter [126] and edited using Inkscape (urlhttp://inkscape.org/en/).

2.3 Results

The results from this chapter and the next have been submitted for peer review to Plos Genetics. The sequence files have been submitted to genbank and IPD but will not be released until publication. Therefore I have uploaded the sequence files to Gitgub (`https://github.com/nick297/thesis_scripts/tree/master/data_files`) where the data should be obvious and clearly labelled. These sequences were too big for the appendix.

2.3.1 BAC clone DNA was successfully sequenced with a mixture of single end and paired end Roche 454 pyrosequences

To sequence and assemble the highly repetitive cattle LRC region, BAC clones were sequenced using Roche 454 pyrosequencing technology. Roche 454 sequencing was used because it has greater read length than Illumina technologies. Illumina was limited to 75 bp at the time of sequencing (early 2010) compared to the 500-600 bp produced by 454, and a significantly higher throughput than Sanger sequencing. The sequencing produced tens of thousands of pyrosequences per BAC clone, Table 2, although there is sequence number and length variation between the BAC clones sequenced. Clones 068F04 and 335H08 yielded the fewest pyrosequences, however the average read lengths and median read lengths are roughly similar.

A further BAC clone, 303D02, was sequenced using a 3 kb insert library to produce paired end reads, at the Liverpool Centre for Genomic research (UK). This BAC clone produced more sequences (193,149 Table 2), however the average and median read length is substantially lower than the other BAC clones resulting in the total bases number being comparable with 095G08 and 032G11. The read length distribution histogram for 303D02, shown in Supplementary Figure S1e, shows a two peak distribution. One peak represents the intended library preparation size peaking at just over 500 bp, the other peak is indicative of fragmented reads caused by the library insert preparation. Therefore there are pyrosequences of useful length and the overall average read length is misleading in this respect. In order to generate a complete reference sequence for the *KIR* complex, these sequences needed to be *de novo* assembled into a complete contiguous assembly.

2.3.2 BAC clone pyrosequences were partially assembled with the MIRA assembler

For *de novo* assembly of the raw pyrosequences, the open-source MIRA assembler was used [27] [28]. This was chosen because other assemblers designed specifically

for 2nd generation technology such as velvet [148] and SOAPdenovo [89] primarily use a de Bruijn graph method. These are better suited for shorter Illumina and SOLiD datasets and unsuitable for the longer read lengths of 454. Other Overlap-Layout-Consensus (OLC) assemblers such as "Newbler" that contains Roche's specific 454 sequence assembly and mapping software were not used because MIRA has higher accuracy at assembling repetitive regions due to its iterative assembly steps. MIRA is also capable of combining different sequencing technology datasets in order to produce hybrid assemblies.

The MIRA assembler produced assemblies of the raw 454 pyrosequneces that were split into contigs due to the highly repetitive nature of the haplotype. The 500 bp read length was too short to span the repeat regions which resulted in the assembly terminating into contigs. MIRA could also not resolve some of the pyrosequencing homopolymer repeat issues that resulted in assembly termination and contig formation. To finish the assembly, contigs were manually joined at breaks caused by the homopolymer repeat issue. However, to span the breaks caused by repetitive sequence, longer reads were required at these positions.

2.3.3 Assemblies required finishing with PCR and Sanger sequencing

BAC assemblies were completed by the addition long Sanger sequence reads spanning the unfinished regions. PCR primers were designed from within 500 bp of the ends of each contig and were used for PCR reactions using all possible combinations of primer pairs. The successful primer pair combinations are shown in Table S1.

Using the MIRA assembler, Sanger sequences were hybrid *de novo* assembled with the 454 sequences to create contigs containing both Sanger and 454 sequencing technologies. It is possible to insert the longer bridging Sanger sequences directly into the previously assembled BAC clone databases, then manually join the contigs. However, by taking the hybrid assembly approach, further 454 sequences are assembled with the spanning Sanger sequences, adding greater assembly confidence to the spanning region and removing reads that may have assembled incorrectly elsewhere or formed a small contig. Hybrid assemblies were targeted with two to four further rounds of PCR and Sanger sequencing depending on the BAC clone until complete assemblies were achieved.

2.3.4 Hybrid assembly of the sequenced BAC clones produced a complete cattle *KIR* haplotype

Three BAC clones, 095G08, 335H08 and 303D02, were sequenced and assembled using Roche 454 and Sanger technologies to produce a complete cattle *KIR* complex with flanking *FCAR*, *NCR1* and *LILR* genes forming a single cattle *KIR* haplotype (CKH), Figure 8. Of the three BAC clones, raw sequences from two (095G08 and 335H08) were merged together in order to generate a composite assembly as they overlapped. As BAC clone 095G08 was successfully assembled with a single round of Sanger sequencing, this facilitated in the assembly of the 335H08 within the overlapping region. By combining the two BAC clones for composite assembly this guaranteed *de novo* assembly of the 095G08 region which left the unassembled region of 335H08, which contains fewer sequences, to be finished by further targeted PCR and Sanger sequencing.

The BAC clone 303D02 provided the remainder of CKH 1 sequence. This was successfully *de novo* assembled from paired-end 454 pyrosequences with a 3 kb insert size. However, 303D02 did not overlap with 095G08. Therefore, to complete the haplotype, PCR spanning the BAC clone positions was conducted with the genomic DNA of the BAC library animal and BAC clone 369B10 which spans the two BAC clones but was not sequenced. The PCR yielded a 2 kb PCR product spanning the 1186 bp gap between the two BAC clones.

The two assemblies of 303D02 and 095G08/335H08 were combined and pregap4 [127] was used to assemble the spanning sequenced PCR products. 303D2 was determined the same haplotype as 095G08/335H08 despite no overlap because of the BAC clone 032G11. This BAC clone was *de novo* assembled and overlapped with 303D02 and 095G08. 032G11 is allelic to both 0950G8 and 303D02, and therefore represents a different haplotype. The *de novo* assembly of the first cattle *KIR* produced a complete assembly. However, before characterisation and annotation the haplotype sequence and structure needed to be verified.

2.3.5 A second *KIR* haplotype was partially sequenced

The second CKH was defined using a combination of Sanger, 454 and Illumina sequencing of two BAC clones. The BAC clone 032G11 was *de novo* assembled using 454 and targeted Sanger sequencing to create a single contig. There is a high level of sequence identity from 130,000 kb to 265,000 kb between 032G11 and CKH 1, there is also another block of high sequence identity that corresponds to block A of CKH 1, which is not contained in 032G11, Figure 9a. The 032G11

consensus sequence has high sequence identity to CKH 1, with the majority greater than 95% identity compared to CKH 1, Figure 9b. There are three points of reduced sequence identity. However, they are relatively short sequence stretches and may represent slight structural difference or pockets of sequence variation. The partial second *KIR* haplotype was characterised in parallel with the first *KIR* haplotype.

Alongside 032G11 another haplotype 2 BAC clone was identified, 068F04. This BAC clone was sequenced with 454 and Illumina but could not be *de novo* assembled despite targeted PCR sequencing attempts and therefore was mapped to haplotype 1, shown in Figure 8. This revealed a similar sequence to CKH 1 at the 5′ end of the complex. There were no further CKH 2 BAC clones identified and therefore this haplotype was only partially sequenced.

2.3.6 Haplotype 1 sequence and structure was verified with further Illumina sequencing

To verify the haplotype sequence and gene order, three BAC clones (335H08, 303D02 and not-previously sequenced 369B10) were sequenced using the Illumina HiSeq platform and 100 bp paired end chemistry at ARK Genomics, Roslin Institute, The University of Edinburgh (UK). The sequencing produced between 18 and 48 million reads per BAC clone, Table 3. Attempts to *de novo* assemble this dataset using MIRA, velvet and SOAPdenovo failed to produce contigs of usable length, this was most likely because the 2x100 bp read length was too short to span repetitive elements. This Illumuna dataset was used to confirm the sequence and structure of haplotype 1 by mapping to the haplotype 1 reference sequence produced by 454/Sanger sequencing.

The structure of the haplotype was interrogated by examining the relational distance between paired-reads *i.e.* how far apart the read pairs are from each other and does that indicate structural variation? The library preparation resulted in an insert size average ranging from 530 bp to 700 bp. Populations of reads that map away from their respective read pairs so that their inferred insert size is outside of the expected size of the library preparation could be indicative of structural variation. Artificial breaks and rearrangements in the reference sequence were generated to make simulated structural irregularities creating examples to compare against.

No SNPs were found from the Illumina mapping data and no structural irregularities could be determined by looking at the paired-end read information. However, as only one library size was used, and this library size was relatively short this verification method has some limitations. A more robust method for

Previous allele Name	New gene name	KIR lineage	allele no	Breed	Accession	reference
BotaKIR2DL1	*BotaKIR2DL1*	3DL	01	Unknown	AY075102.1	[94]
BotaKIR2DL1	*BotaKIR2DL1*	3DL	01	Holstein-Friesian	AF490399.1	[129]
BotaKIR2DS1	*BotaKIR2DXS1*	3DX	01	Holstein-Friesian	AF490400.1	[129]
BotaKIR3DL1	*BotaKIR3DXL1*	3DX	01	Holstein-Friesian	AF490402	[129]
BotaKIR3DL2-001	*BotaKIR3DXL4*	3DX	01	Holstein-Friesian	EF197118	[41]
BotaKIR3DL1N	*BotaKIR3DXL6*	3DX	01N	Unknown	AY075103.1	[94]
BotaKIR3DL3	*BotaKIR3DXL6*	3DX	02	Holstein-Friesian	EF197119	[41]
BotaKIR3DS1	*BotaKIR3DXS1*	3DX	01	Holstein-Friesian	AF490401	[129]
BotaKIR3DS1-002	*BotaKIR3DXS1*	3DX	02	Holstein-Friesian	EF197120.1	[41]

Table 1: Previous and replaced *KIR* gene names. Previous gene names that have been defined in older studies have been renamed based on L or X-lineage origins. The new names are shown here along with the designated allele.

BAC clones	No. Seqs	Total bases	Ave. read length	Median read length
095G08	77,223	44,579,184	577	555
335H08	42,937	24,715,050	575	553
032G11	83,551	48,368,449	578	555
068F04	35,423	20,651,656	583	562
303D02	193,149	43,181,999	223	192

Table 2: Raw 454 sequence numerical data. Breakdown of numbers, total bases, average read lengths and median read lengths of 454 pyrosequences for each BAC clone sequenced.

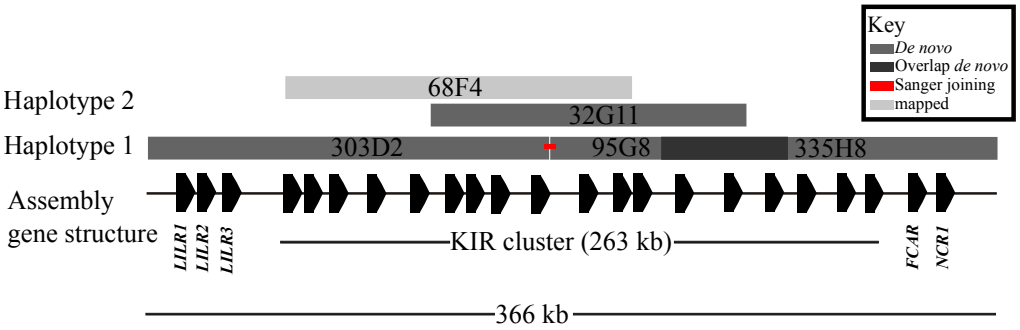


Figure 8: Overview of BAC clones used in the assembly of the haplotype sequence. Genes are represented as unequal pentagons and BAC clones are shown as long rectangles. The overlap between 095G08 and 335H08 is shown as the darkest grey and the joining PCR sequence is shown in red. The light grey 068F04 BAC clone has been mapped and not *de novo* assembled

future verifications of *de novo* assembled regions would utilise a larger insert size or several different library sizes. After verifying the haplotype sequence and structure it could be confidently characterised and annotated.

2.3.7 The cattle *KIR* haplotype was characterised by bioinformatic and manual sequence analysis

To define the genes within the assembled raw haplotype consensus sequence, the sequence needed to be characterised and annotated. CKH 1 was characterised with a combination of blat [74] alignments and manual sequence searches revealing a total of 18 discrete *KIR* loci including; 8 predicted functional *KIR* genes, 6 *KIR* null-alleles, 4 *KIR* pseudo-gene fragments as well as 3 potential *LILR* genes, an *FCAR* gene and an *NCR1* gene, Figure 10. The CKH 2 was characterised using the same methods to reveal the same structure and gene content as CKH 1, Figure 10. For CKH 2, full length gene sequences were extracted from only the BAC clone 032G11 sequence which was fully *de novo* assembled.

The gene order in the CKH (Figure 10) shows a mixture of X and L-lineage *KIR* interspersed through the complex. There are 8 predicted functional *KIR* in the CKH which is comparable to human haplotypes, with 7 genes in human haplotype A and up to 12 genes in human haplotype B. The human *KIR* haplotypes contain null-alleles on different haplotypes and two pseudogenes. Six human *KIR* encode variants that are null-alleles but also maintain functional copies. The CKH has a greater number of non-functional genes, 10 in total, two of which have no intact signalling exons however 7 out of the remaining 8 non-functioning loci are predicted to have been activating genes. For a higher resolution perspective of *KIR* sequence relationships, phylogenies were inferred from the extracted gene sequences.

2.3.8 Cattle have expanded both *KIR* lineages

To compare the cattle *KIR* genes to those previously found in other mammals, a neighbour-joining tree was constructed to infer phylogenetics. Extracted gene sequences were aligned with other species *KIR* genes using mafft [72]. Phylogeny was then inferred using MEGA5 [134] generating a neighour joining tree using exon3-intron4-exon4 of the *KIR* gene sequences. This region has been established as containing the most divergent sequence between *KIR* lineages [56] (Figure 11). Two distinct lineages clearly segregate amongst the *KIR* genes. This is highlighted by the vertical lines denoted 3DL and 3DX-lineages. The cattle genes are highlighted with a grey background and are split between X and L-lineages,

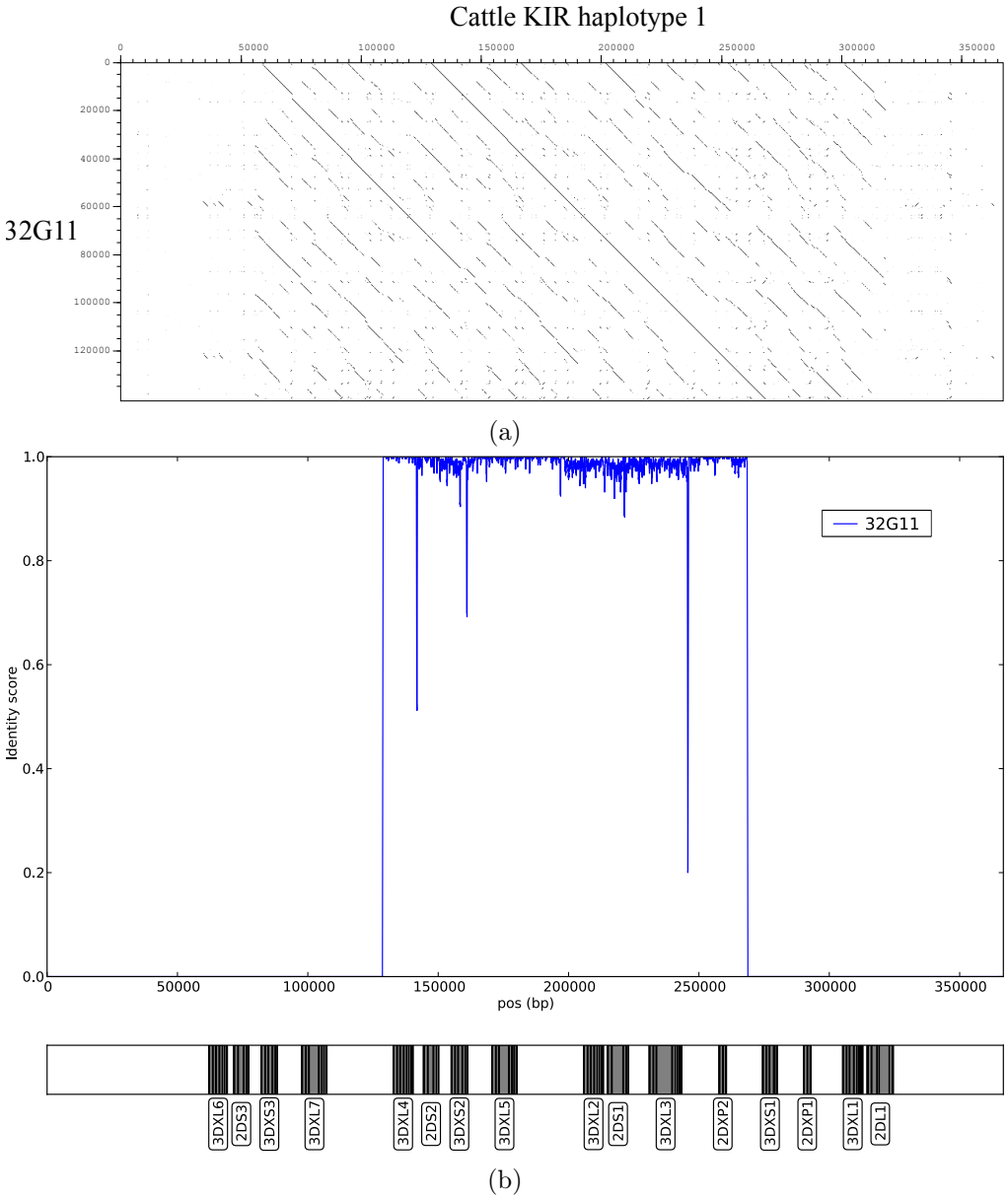


Figure 9: Sequence comparison of BAC clone 32G11 against CKH 1. Dot plot with BAC clone 032G11 sequence compared to the CKH 1 sequence, dots represent 150 bp sequence identity, lines are contiguous dots (a). A 300 bp sliding window sequence comparison of the 032G11 compared against the CKH 1 reference sequence, the line represents the mean sequence identity and *KIR* genes are indicated along the X-axis (b).

Figure 11. The cattle L-lineage genes form an outgroup with horse and pig *KIR* genes but have known no orthologs within any other species. The X-lineage cattle genes form a clade with 3DX1 genes from chimp, orang-utan, human, macaque and gibbon, with a further X-lineage gene from an African elephant. Again there are no direct orthologous genes related to any of cattle X-lineage genes although the other species are not closely related to cattle. To find potentially orthologous genes to the cattle *KIR* genes, other ruminant species will need to be studied. There is considerable diversity within the cattle X-lineage genes however, several of the genes group together to form gene groups.

2.3.9 Cattle X-lineage genes cluster into related groups

It is evident from the exon3-intron4-exon4 sequences that the cattle *KIR* genes cluster into groups of related genes, Figure 11. The cattle X-lineage gene names have been coloured in the tree to represent this grouping. There are five different groups of cattle *KIR* genes, including the L-lineage as a group, with either two, three or four separate loci within the CKH. Two groups, group IV and group III had not previously been identified along with the group V pseudogenes *BotaKIR2DXP1* and *BotaKIR2DXP2*. Each gene group has loci that had not previously been defined. There are several activating genes that had previously not been defined, this may be because they appear not to be functional.

The gene groups have been named based on the phylogenetic relationships of the ectodomain sequences. This is in contrast to the system used by the Parham lab, that groups KIR based on their molecular forms. Therefore the cattle *KIR* gene groups can contain different forms of the genes including 2 and 3 domains, and short and long tailed receptors. The groups are numbered from 5′ to 3′ based on the order each group is located. The group I *KIR* consists of three loci of three Ig domain long tailed inhibitory receptors. *BotaKIR3DL2* is very similar to the previously published *BotaKIR3DXL4*, both genes are predicted to be functional and contain highly similar sequence identities. The predicted null-allele *BotaKIR3DXL6* has a deletion within the D2 causing a frame-shift mutation and premature stop codon within the transmembrane domain. The group III *KIR* consist of two loci containing three Ig domain short tailed null-alleles. The group III clade together with the group I and group V genes and away from the other group IV genes that form an out group within the X-lineage clade. The group IV genes are formed of three inhibitory genes all encoding three Ig-domain domain genes predicted to be functional. This gene group shares high sequence similarity with the exception of a long repeat sequence insert within intron 4 of *BotaKIR3DXL3*. The group IV genes contain a characteristic five

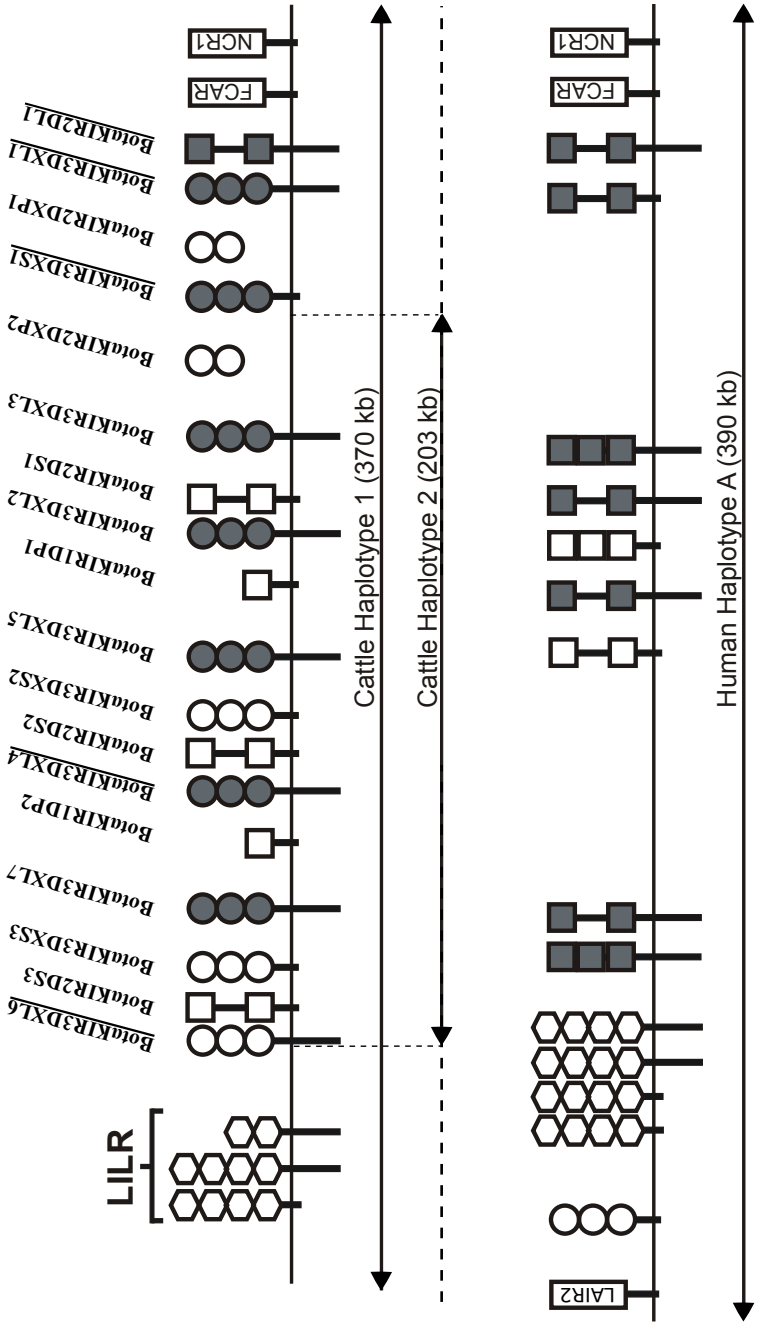


Figure 10: Cattle and human *KIR* haplotype structure comparison diagram (to scale), circles represent the X lineage *KIR* domains and squares represent the L-lineage. Open shapes represent null-alleles or pseudogenes, and filled shapes represent predicted functional. Vertical lines represent signalling tails, longer lines are inhibitory and shorter lines are activating. Arrows represent size limits of haplotype in diagram. The entire cattle structure is based on haplotype 1. Previously found *KIR* gene names are underlined. Dashed horizontal lines represent predicted haplotype. Human haplotype A is taken from the human genome project build release 104.

residue insertion within the D0 domain that is not within any of the other genes.

The group V genes contain two functional and two non-functional genes. The functional, *BotaKIR3DXL1* and *BotaKIR3DXS1*, genes have alternate signalling domains but very similar extracellular domain sequences. The pseudogenes *BotaKIR2DXP1* and *BotaKIR2DXP2* encode disrupted signal peptides to D2 domains. The pseudogene *BotaKIR2DXP1* has characteristics unlike any of the other X-lineage genes and clades away from all the other X-lineage groups when using just the D0 domain, Figure 11.

2.3.10 Cattle L-lineage genes have also expanded

The L-lineage cattle *KIR* genes, group II, have expanded alongside the X-lineage, contrary to what was previously predicted. There are four genes within group II, three of which are predicted null-alleles of short tail receptors, including *BotaKIR2DS1*, *BotaKIR2DS2* and *BotaKIR2DS3*. The only predicted group II gene is the previously defined *BotaKIR2DL1*. It is now apparent that the reason the group II were predicted to be a single gene in cattle is because only *BotaKIR2DL1* has an intact coding sequence and all previous attempts used transcription methods to detect the *KIR*.

2.3.11 Serial inactivation of short-tail genes by terminating mutations

Six *KIR* loci within CKH 1 have been identified as null-alleles because of a premature stop codon or miss-sense mutation. Table 4 shows that all but one of the genes, *BotaKIR3DXL6*01N*, are activating. These short tailed gene groups, group II and group III, have independently mutated stop codons within exon 3. Genes in the same group share the same disabling mutation, suggesting that the mutations arose prior to duplication. The *BotaKIR2DS1*02N* allele found on CKH 2 has a frame shift mutation prior to the stop codon in *BotaKIR2DS1*01N*. This prevents the premature termination of translation in *BotaKIR2DS1*02N* until an alternate stop codon at residue 218. This suggests that *BotaKIR2DS1*02N* could be a secreted form of the molecule with only two Ig domains and no signalling domains. The only non-functional long-tail gene, *BotaKIR3DXL6*02*, has a single nucleotide deletion at mRNA position 780 within the domain 2 exon that causes a frameshift which introduces a stop codon within the transmembrane domain.

BAC clone	Insert size	Read 1	Read 2	Read length (bp)	Million reads	Total bases
335H08	683	8,992,937	8,992,937	100	18	1,798,587,400
303D02	702	15,597,673	15,597,673	100	31.2	3,119,534,600
369B10	674	24,197,574	24,197,574	100	48.4	4,839,514,800
068F04	532	12,318,130	12,318,130	100	24.6	2,463,626,000

Table 3: Details of Illumina sequencing run statistics from sequenced cattle BAC clones. DNA from cattle BAC clones were sequenced with 2 x 100 bp Illumina for sequencing error correction and detection of structural problems with the *de novo* hybrid 454/Sanger assembled reads.

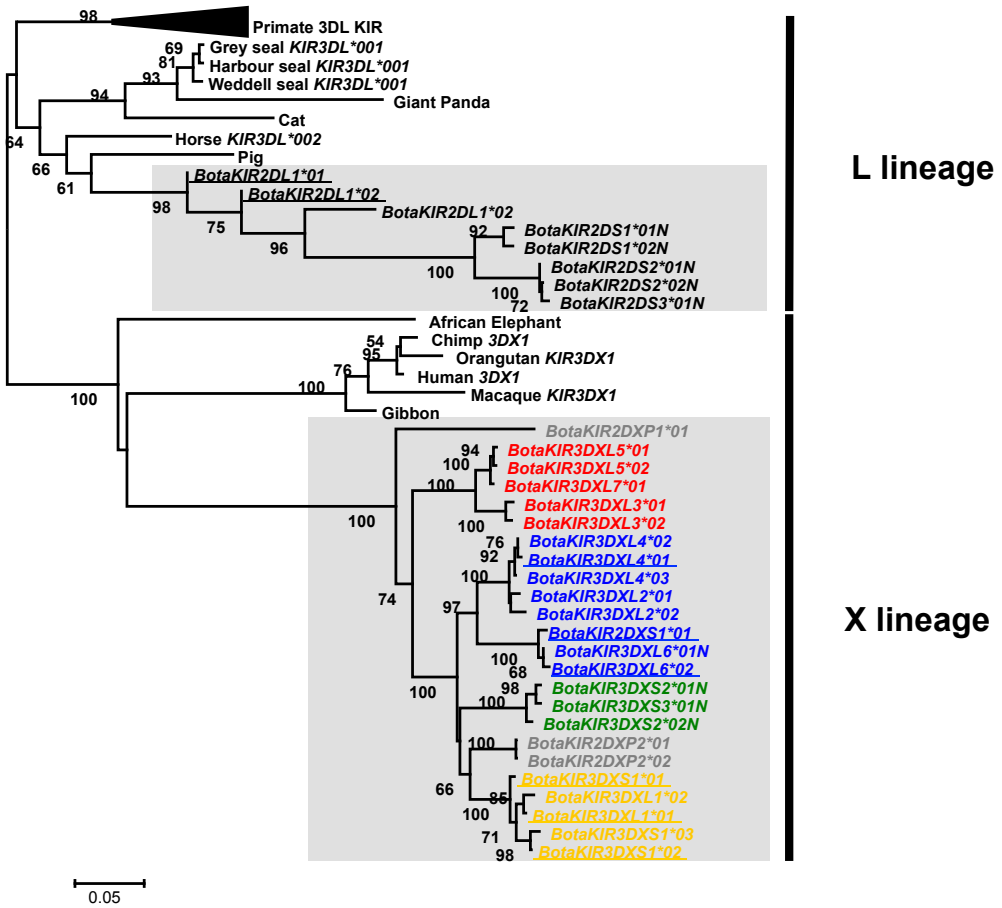


Figure 11: Neighbour Joining Phylogenetic tree of cattle *KIR* sequences from both CKH 1 and 2 and representative *KIR* sequences from several species. X and L-lineages have been labelled separately. Cattle genes are highlighted with grey background. Cattle X-lineage genes clade together into groups. These groups of genes have been coloured. Previously defined genes from before the start of this project are underlined. The Primate (including human) L-lineage node has been collapsed to aid the tree visualization. Allele numbers represent CKH 1 and 2, with the higher number denoting the CKH 2 allele.

2.3.12 Cattle *KIR* maintain the same domain structure seen in other species

Human and other primate KIR receptors have two or three Ig-like domains that bind MHC class I ligands. Except for human *KIR2DL4* which has the D0-D2 domain arrangement, two domain KIR receptors bind via the D1 and D2 contacting the face and peptide of the ligand. The D0 of the three Ig domain receptor *KIR3DL1* anchors to the conserved side of the ligand [141]. The two forms of KIR bind ligand in similar ways but vary within the D0 domain. Similarity in domain order between human and cattle *KIR* is indicative of analogous function. Each individual Ig domain exon sequence was extracted from every gene in the CKH and aligned to compare the Ig domain composition, Figure 12. The Ig-domain exon sequences segregate into groups of D0, D1 or D2 domains consistently with their order within the gene. The three domain genes maintain the D0-D1-D2 structure whilst the two domain genes have a D0-D2 structure. This is contrary to the majority of human *KIR* which have a D1-D2 structure, except *KIR2DL4* which also contains this domain order.

2.3.13 Cattle cytoplasmic tail sequences have likely originated from X-lineage genes except in *BotaKIR2DL1*

The introduction and propagation of activating function within multiple cattle *KIR* genes has occurred via recombinations with activating receptors. The 5′ end of *KIR* group sequences including the Ig domains clade together when both inhibitory and activating genes are encoded, Figure 13a. *BotaKIR3DLXL1* and *BotaKIR3DXS1* form a group that contains both inhibitory and activating tails but share highly similar Ig domain. However, these genes segregate when using the 3′ region of the gene displaying complete separation of inhibitory and activating domains, Figure 13b. All of the inhibitory and activating tails clade together away from the other mammalian *KIR* tails, except for *BotaKIR2DL1.* Therefore, there is likley two tail origins in the CKH, one for *BotaKIR2DL1* and one for the other cattle *KIR* genes. All the activating tails segregate away from the inhibitory tails. This suggests the cattle *KIR* activating tail sequence has evolved once and recombined several times with the domain sequences from inhibitory genes to generate novel activating genes.

	Domain 0				Domain 1						Domain 2				Transmembrane			
	mut.	pos. (res.)	mut.	pos. (res.)	mut.	pos. (res.)	mut.	pos. (res.)	mut.	pos. (res.)	mut.	pos. (res.)	mut.	pos. (res.)	mut.	pos. (res.)	mut.	pos. (res.)
BotaKIR2DS1*01N	Stp	106	Stp	116							Stp	146	Stp	150	Stp	220		
BotaKIR2DS1*02N	mis.	104													Stp	241		
BotaKIR2DS2*01N	Stp	106											Stp	159	Stp	236		
BotaKIR2DS2*02N	Stp	106											Stp	159				
BotaKIR2DS2*02N	Stp	106											Stp	159	Stp	236		
BotaKIR2DS3*01N	Stp	106											Stp	159				
BotaKIR3DXL6*01N											mis.	252			Stp	323	Stp	349
BotaKIR3DXS2*01N	Stp	34			Stp	134			Stp	171	Stp	243	Stp	303				
BotaKIR3DXS2*02N	Stp	34			Stp	134	ins.	164	Stp	172	Stp	244	Stp	304	Stp	345		
BotaKIR3DXS3*01N	Stp	34			Stp	134			Stp	171	Stp	243	Stp	303	Stp	343		

Table 4: Details of mutations causing null-alleles and pseudogenes within the cattle *KIR*. Stop codons are labelled Stp, indels causing frameshift mutations are labelled mis. Numbers are the residue position within the mature protein, without the signalling domains. Allele numbers represent CKH 1 and 2, with the higher number denoting the CKH 2 allele.

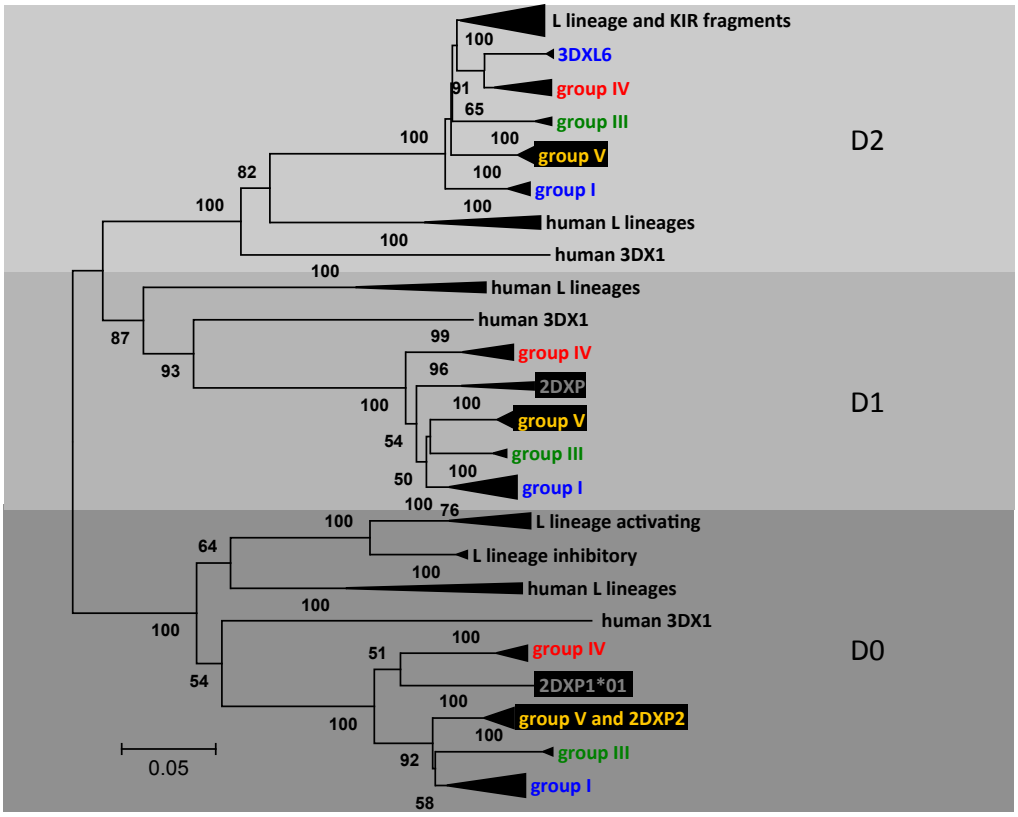


Figure 12: Phylogentic trees of cattle *KIR* Ig domain sequences. Each Ig domain sequence has been individually extracted and aligned together alongside the human L and X-lineage *KIR* Ig sequences. Similar sequences from the same group have had their common nodes collapsed to reduced visual complexity of the tree. The tree was constructed using neighbour joining and the p-distance algorithm with 100 bootstraps.

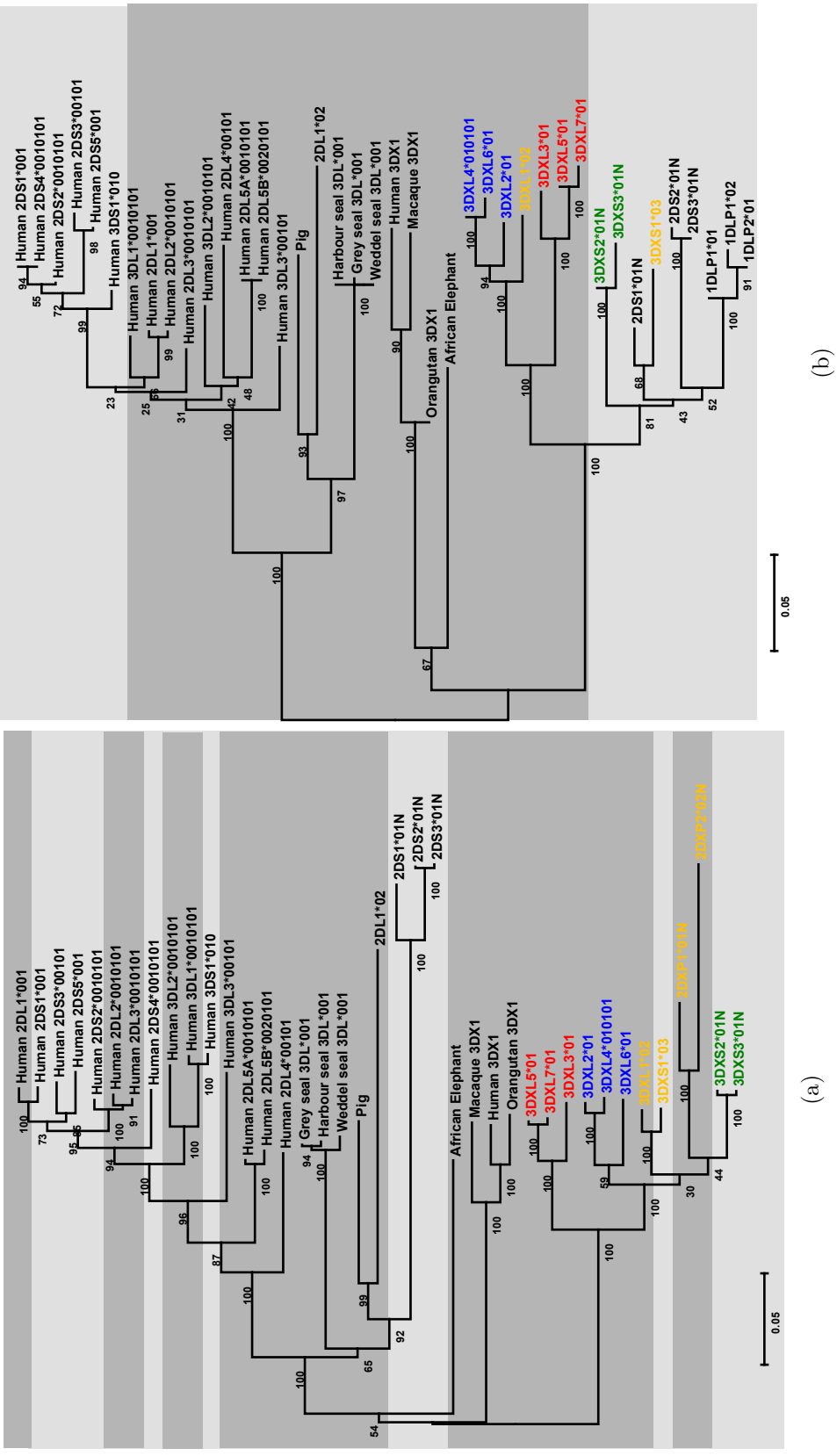


Figure 13: Functional recombination within *KIR* sequences. *KIR* 5′ sequence including signal peptides-D0-D1-D2 (a). *KIR* 3′ sequence from stem to cytoplasmic tail (b); Trees are neighbour joining using the p-distance algorithm and 100 bootstaps. The groups are colour coordinated and the background shading represents the inhibitory or activating function of the gene (or predicted function for null-alleles).

2.3.14 Cattle activating *KIR* signal through the Fcγ adapter protein rather than DAP10 or DAP12

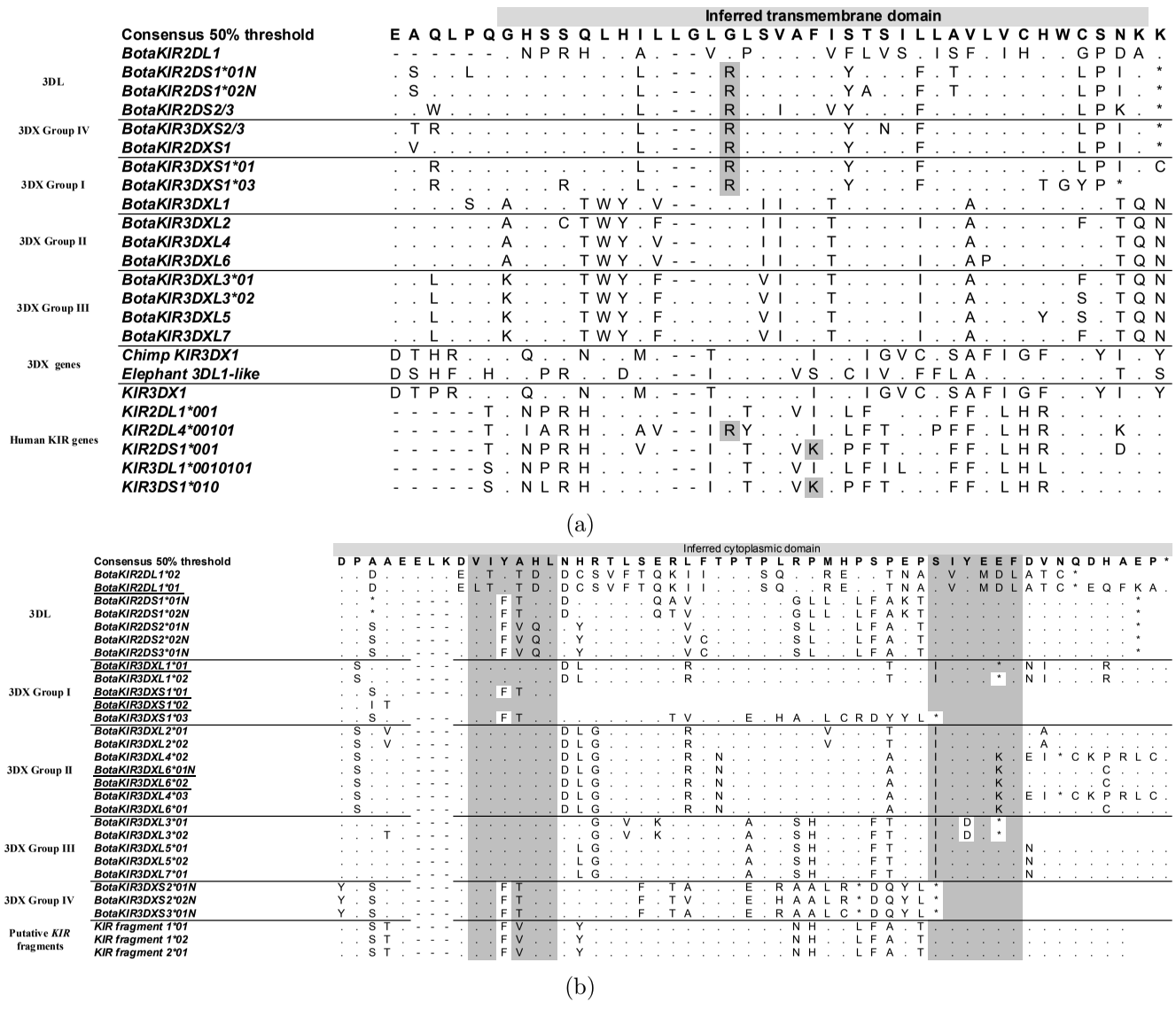
The cattle short tailed *KIR* genes and null-alleles all contain an arginine residue within the transmembrane domain at the corresponding position as the KIR2DL4 arginine residue in humans. This is contrary to the other human and primate short tailed KIR receptors that signal through a lysine residue interacting with DAP10/12 [18], Figure 14a. This alignment and the distinct groups formed in the phylogenetic tree in Figure 13b demonstrate that the codon for arginine has spontaneously mutated and not been inherited from a shared ancestor of *KIR2DL4* and cattle short tail *KIR*. The cause of the cattle *KIR* genes utilising the arginine-Fcγ pathway is unlcear as both DAP10 and DAP12 are present and functional in the cattle genome [50].

The cytoplasmic tail regions of the activating genes are not translated due to stop codons at the end of the transmembrane domian. The remains of the cytoplasmic tail regions within the genomic DNA show that the activating genes have been disrupted within the first ITIM motif, Figure 14b. Within the first ITIM motif the functional tyrosine residue has changed to a phenylalanine, which has similar electrochemical properties but has been shown to affect binding of SHP-1/2 and reduce inhibition [128]. The second ITIM has been disrupted within the X-lineage activating genes by the introduction of stop codons.

The inhibitory genes encode a conserved **VIYAHL** first ITIM motif and slightly variable **(S/I)IY(E/K)F** second ITIM motif. The variation is contained between the groups of genes, with each group showing little variation within the signalling domains. Therefore, the gene groups likely shared common ancestors before duplication and/or are constrained by their signalling adaptors. The mechanism that has produced these gene groups has resulted in very similar genes that likely share similar functional properties.

2.3.15 Gene groups were forged by the block duplication of unrelated genes, resulting in highly related genes dispersed over the length of the haplotype

Using the exon and intron sequences of the 5′ region of the gene from the signal peptides to the stem, the alleles found within haplotype 1 and haplotype 2 segregate into groups, Figure 15. Within the X-lineage *KIR* genes, groups I, III and V, form a group segregating away from the group IV genes. This could indicate the group IV genes share a more ancient ancestral gene to the other X-lineage groups, which appear to be derived from the same but more recent ancestral



(a)

(b)

Figure 14: Residue alignments of predicted transmembrane (a) and cytoplasmic domains (b). Dots represent residues identical to the reference sequence which is the top bold line of sequence. Dashes represent sequence absence. Grey shading highlights the functional motifs for each domain. Allele numbers represent CKH 1 and 2, with the higher number denoting the CKH 2 allele.

gene. The group II appears split, with *BotaKIR2DS1* segregating away from the rest of the group. This is contrary to the D0 portion of Figure 12 which shows *BotaKIR2DL1* grouping away from the other group II genes. The genes within each group are closely related, yet they are spread throughout the haplotype. The proximally closest genes within the same group are the group V genes, that are wholly contained at the 3′ end. The other gene groups are separated by either two or three genes from alternative gene groups. Each non-group V gene belongs to an alternate block to the other genes within that group.

The blocks are evident in the dot plot shown in Figure 16. The blocks have been designated block A to E, with block A the most 5′ and block E the most 3′. There is high sequence identity between block A and block B, Figure 16, with a slight break in the line of identity at the position of *BotaKIR3DXL6*. Block A and B show similarity with block C but not to the same extent. Blocks D and E share significant sequence identity, which could be a result of group V gene duplication *BotaKIR3DXL1* and *BotaKIR2DXP1* in block E to form *BotaKIR3DXS1* and *BotaKIR2DXP2* of block D. There is inevitable sequence similarity between all the blocks that is an artefact of *KIR* sequence comparison as there will always be similarities between *KIR* gene sequences. However, from the gene groups shown in Figure 15 and the identity between the blocks shown in Figure 16, two sets of blocks are defined; Set 1 comprising blocks A, B and C and set 2 containing blocks D and E. Set 1 blocks include the gene groups I - VI. Set 2 blocks consist exclusively of group V genes and the group II gene *BotaKIR2DL1*, which is notably different to the other group II genes.

To facilitate comprehension of the gene, group, block and set nomenclature hierarchy Figure 18 is a diagrammatic representation of the haplotype with the genes, groups, blocks and sets labelled. Although Figure 16 clearly shows the formation of the blocks, it only shows similarity greater than 150 bp. To accurately compare and contrast the different blocks, a higher resolution approach is required.

Sequence identity was plotted over alignments of the different blocks. Blocks within each set were aligned using MAFFT then using a sliding window approach, average sequencing identity was calculated over a window size of 500 bp for each individual base pair, Figure 19. Block B and C have very high sequence similarity to block A, Figure 19a. Block B (blue line) has largely identical sequence to block A with the exception of the first 10 kb. This includes the gene *BotaKIR3DXL6* and a segment of sequence unique to block B. Block C (green line) also has high sequence identity for the majority of the block relative to block A, albeit not as high as the block B identity. However, block C does not contain a group III

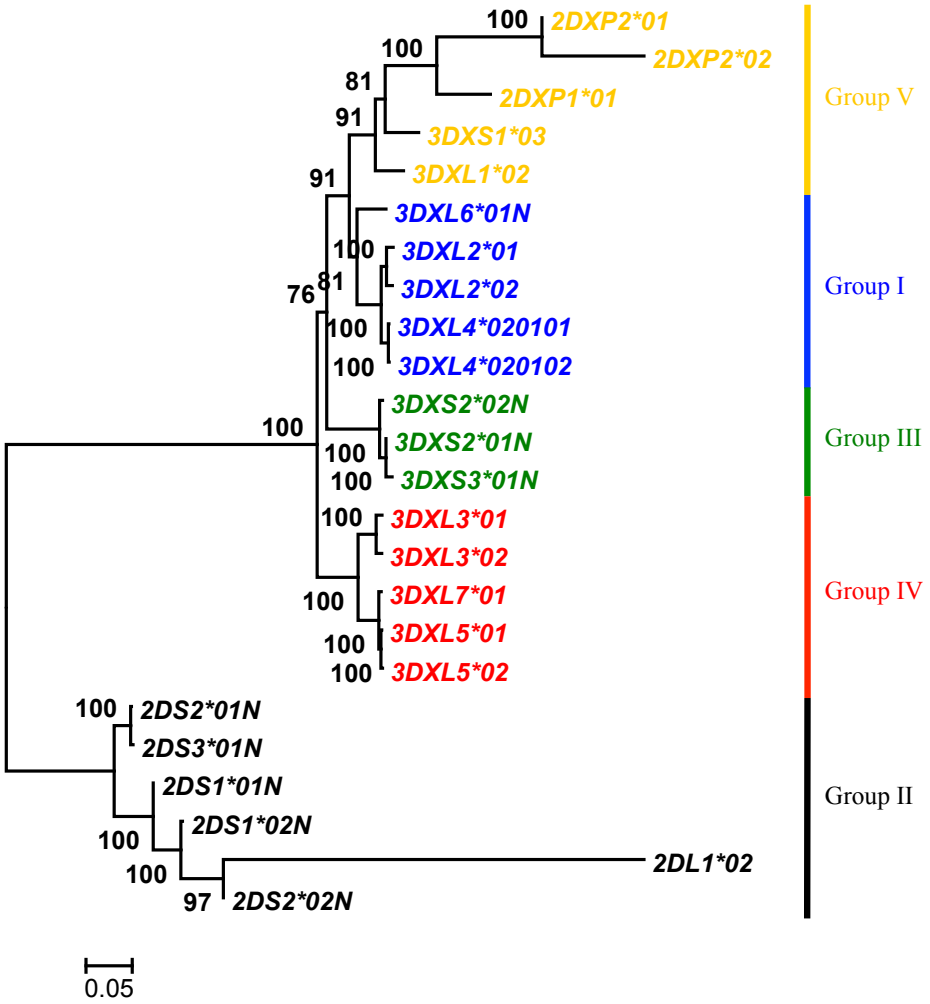


Figure 15: Neighbour joining phylogenetic tree of *KIR* gene sequences from the *de novo* assembled BAC clones. The tree is constructed using the full length genomic DNA sequence, 500-bootstrap replicates were used with the Tamura-Nei algorithm. Gene groups are colour coded.

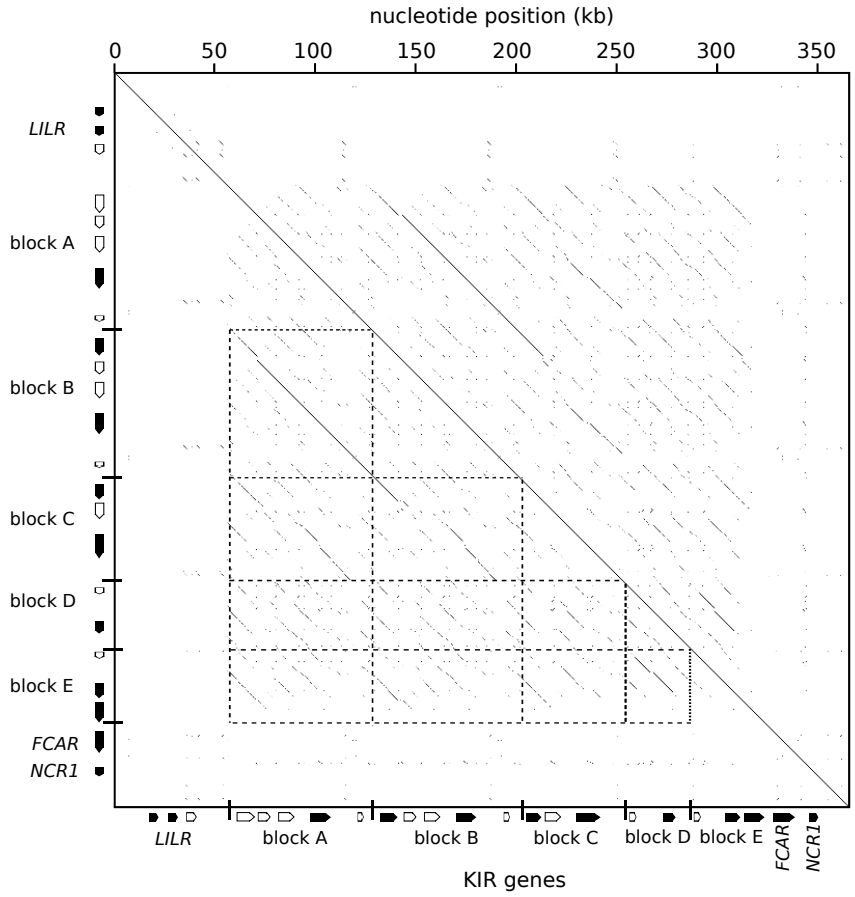


Figure 16: Dot plot of 150 bp showing regions of high sequence identity between blocks of genes. Blocks and genes are annotated along the axes, comparisons between blocks are surrounded by dashed line boxes. Dots represent exact sequence identity over 150 bp and lines represent several dots over consecutive sequence. Therefore, the diagonal line from the top left corner to the bottom right corner is identical sequence as the CKH has been compared to itself. Dots and lines outside of this diagonal represent sequence identity within the haplotype which come from a result of sequence duplication or repetitive elements. Block positions and gene content is based on gene group relationships shown in Figure 15. The dashed lines outline comparisons between different blocks within the haplotype.

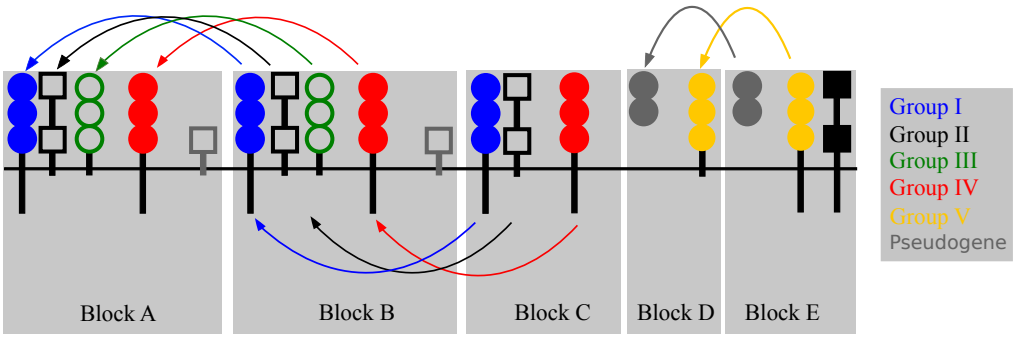


Figure 17: Cartoon representation of block duplication. Arrows represent duplication of genes and grey background represents the block. Genes are represented as they are in Figure 10 but colour coded by group.

gene, which is shown by the reduction in sequence identity midway through *2DS3* and subsequent increase in sequence identity midway through *3DXS3*. This is consistent with Figure 13b which indicates the tail end of *BotaKIR2DS1* from block A has greater similarity to the tail regions of the group V and III genes than the other group II genes. There are two regions of unique sequences specific to block C within the group IV genes. This can be seen within the intron 4 of *BotaKIR3DXL7* and is likely inserted sequence found in *BotaKIR3DXL3*.

Figure 19b shows the similarity between the set II blocks lies within the group V genes. Block D (blue line) has high sequence identity to block E for the majority of the two blocks however the *2DXP1/2* genes have reduced identity. There is also block D specific sequence between the two group V genes that is not present in block E. There is high sequence identity between the blocks at the *3DXL1* locus. However there is a drop in identity score between the exon 5 (domain 2) and exon 6 (stem domain) that signifies a breakpoint between signalling domains of *BotaKIR3DXL1* and *BotaKIR3DXS1*. This is where recombination has taken place introducing an activating tail to *BotaKIR3DXS1* in block D.

Blocks A and B may have formed from block duplication of block C, which itself may have formed through gene and haplotype rearrangements. As blocks A and B are more similar to each other it is predicted that they duplicated more recently with one of them being the product of a block duplication and subsequent group III gene insertion (or deletion) in block C. This is shown in a cartoon representation of the predicted block duplication, which shows the predicted path of gene duplication over the evolution of the CKH (Figure 17). Furthermore, there has been block duplication and gene conversion from block E to block D with *BotaKIR3DXL1* recombining with an activating tail to form *BotaKIR3DXS1*. There has likely been further duplications and gene recombinations that cannot be implied from this one full haplotype sequence. However, the model suggested here is of block duplication occurring in the two sets separately and limited to sequence still present in the haplotype. For evidence to support the order of block duplication in the model shown in Figure 17, it is hypothesized that there would be a SNP frequency gradient, with fewer total SNP numbers in each block relative to its neighbouring 3′ block based on the time of duplication. This hypothesis can be tested by looking at the SNP positions within the second cattle *KIR* haplotype.

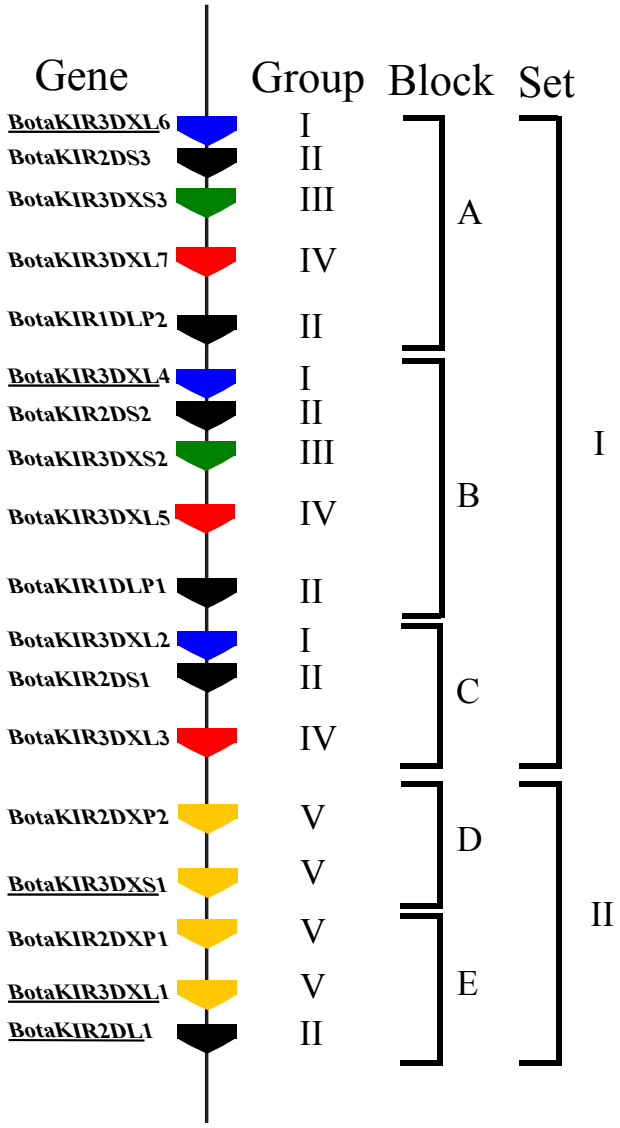
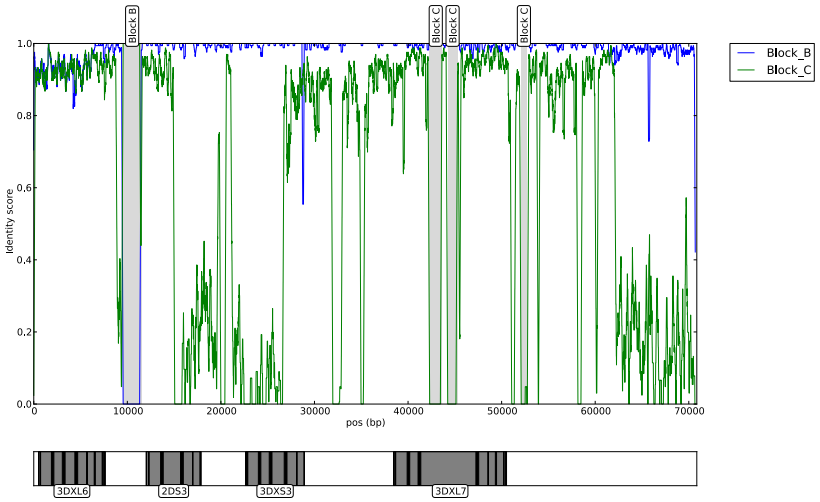
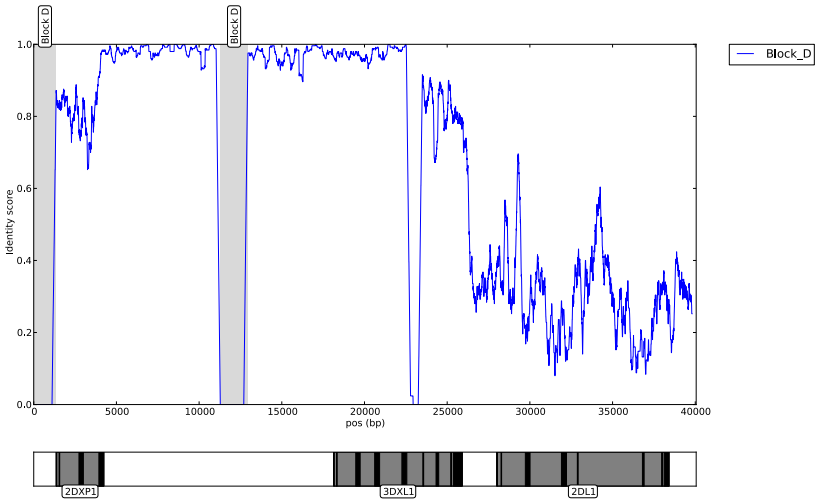


Figure 18: Hierarchy of groups, blocks and sets within the cattle KIR haplotype. Genes are colour coded depending on group.



(a)



(b)

Figure 19: Sliding window sequence identity analysis of aligned blocks. (a) 500 bp sliding window of blocks B and C compared to the block A reference. The blue and green lines represent the relative sequence identity of blocks B and C respectively to block A. X axis represents base pair position within the alignment. Y-axis represents the average identity within the sliding window for each base pair along the x-axis, 1 is identical and 0 is no identity, alignment pads are counted as 0. Vertical shaded grey columns represent unique sequence belonging to either block B or block C, labelled above the column. The genes of block A are annotated along the bottom to scale, grey rectangles show the area of the gene and black lines are the exons within the gene.(b) 500 bp sliding window of block D (blue line) to reference sequence block E using same criteria as (ABC).

2.3.16 A second CKH (CKH2) shows identical gene structure and polymorphic sites

The second haplotype two BAC clone, 68F4, could not be successfully *de novo* assembled due to poor sequencing quality and the highly repetitive nature of the region. In an attempt to fully sequence this clone it was sequenced again with the Illumina HiSeq platform yielding 24.6 million reads at 2 x 100 bp, however this also did not enable *de novo* assembly despite several different assembly techniques with differing combinations of 454, Sanger and Illumina datasets. Therefore in order to determine the polymorphic positions for the second haplotype over the entire sequence available both raw sequence data from both BAC clones was aligned to the CKH 1 reference sequence. Variable positions were then called using Varscan2 and were kept subject to conservative criteria implemented to filter out false SNPs. The criteria were based on whether the position had Illumina and or 454 coverage and whether they agreed. Due to the higher accuracy of Illumina chemistry, disagreements between Illumina and 454 data were awarded to Illumina so long as the Illumina read coverage was greater than 500.

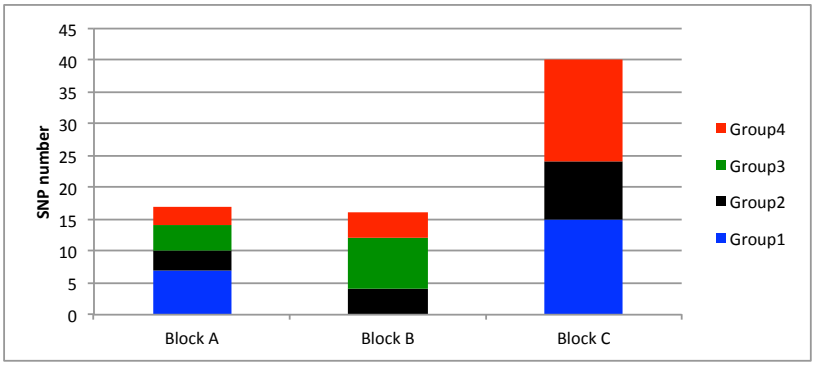
The final SNPs called are shown in Table 5, the effects of variable positions within exons of functional genes have been determined and are shown in the table. For the region that has been sequenced, haplotype 2 has identical gene structure to haplotype 1, however there are polymorphic positions that create allelic variation between the two haplotypes within the *KIR* exons. The variable positions in haplotype 2 are concentrated in the Ig domains of the genes (Figure 20b). This suggests that the variation has been driven by selection pressures resulting from the interactions between the Ig domains and their MHC class 1 (or class 1-like) ligands.

There is a greater number of SNPs between the 3′ end genes on these haplotypes. Genes within block C have greater numbers of SNPs between haplotypes when compared to their corresponding group genes in block A and block B, shown in Figure 20a. This suggests that block A and B were a more recent duplication than block C which has received greater mutation numbers over time as a consequence of selection pressure and/or genetic drift. Conversely the blocks A and B could be constrained by their ligands with further mutations within these genes resulting in a loss of fitness as ligand specificity may be reduced. Block A has slightly higher number of SNPs compared to block C. This can be accounted for by *BotaKIR3DXL6* which may have been introduced into block A by recombination after block A duplication from block B as shown in Figure 19a. Interestingly no SNPs were found in *BotaKIR3DXL4*, this could point to

gene	group	domain	CDS pos.	codon pos	S/NS	residue	haplotype 1	haplotype 2
3DXL6	1	Domain 1	375	3	S	104	A	A
	I		528	3	N	155	W	*
	I	Domain 2	804	3	S	247	A	A
	I		882	3	N	273	H	Q
	I		887	2	N	275	P	R
	I		894	3	N	277	F	L
	I		921	3	N	286	L	F
2DS3	2	Domain 0	268			69		
	II		287			75		
	II		314			84		
3DXS3	3	Domain 1	410			116		
	III		570			169		
	III	Domain 2	651			196		
	III	Stem	962			300		
3DXL7	4	Domain 0	133	1	N	24	V	M
	IV		208	1	N	49	E	K
	IV	Domain 1	507	3	S	148	T	T
2DS2	2	Domain 0	269			69		
	II		287			75		
	II		314			84		
	II	Stem	657			198		
3DXS2	3	Domain 0	196			45		
	III	Domain 1	411			116		
	III		571			170		
	III		626			188		
	III	Domain 2	656			198		
	III		674			204		
	III		768			235		
	III	Stem	963			300		
3DXL5	4	Domain 0	133	1	N	24	V	M
	IV		249	3	S	62	P	P
	IV	Domain 1	393	3	S	110	L	L
	IV		397	1	N	112	G	R
3DXL2	2	Domain 0	80	2	N	6	V	E
	I		111	3	S	16	P	P
	I		298	1	N	79	W	R
	I		299	2	N	79	W	*
	I		308	2	N	82	H	R
	I	Domain 1	360	3	S	99	I	I
	I		381	3	S	106	L	L
	I		400	1	N	113	V	M
	I		482	2	N	140	R	H
	I		628	1	N	189	I	V
	I	Domain 2	679	1	N	206	V	M
	I		781	1	N	240	A	S
	I		896	2	N	278	S	L
	I	Stem	982	1	N	307	S	T
	I	Trans	1068	3	S	335	I	I
2DS1	2	Domain 0	92			10		
	II		123			20		
	II		237			58		
	II		240			59		
	II		282			73		
	II	Domain 2	393			110		
	II		404			114		
	II	Trans	716			218		
	II		775			238		
3DXL3	4	Signal 2	36	3	S	-9	R	R
	IV	Domain 0	87	3	S	8	L	L
	IV		153	3	S	30	F	F
	IV	Domain 1	397	1	N	112	S	R
	IV		453	3	S	130	I	I
	IV		509	2	N	149	H	R
	IV		544	1	N	161	M	V
	IV	Domain 2	661	1	N	200	P	S
	IV		662	2	N	200	P	R
	IV		666	3	S	201	S	S
	IV		691	1	N	210	V	L
	IV		730	1	N	223	K	E
	IV		832	1	N	257	R	C
	IV		892	1	N	277	H	Y
	IV	Trans	1106	2	N	348	F	S
	IV	Tail 2	1183	1	N	374	A	T

Table 5: Table of haplotype 2 SNPs showing positions and residue changes between CKH 1 and 2 within the *KIR* gene exons.

functional importance of *BotaKIR3DXL4* as it may be constrained to binding it's ligand. Although synonymous substitutions would be suspected they are not seen, there are however variable positions within the intron sequences proving this gene has alleles.



(a)

Domain	SNPs
Signal peptide 2	1
Domain 0	23
Domain 1	19
Domain 2	21
Stem	4
Transmembrane domain	4
Cytolplasmic tail domain 2	1

(b)

Figure 20: SNP comparison between CKH 1 and 2. 20a Bar chart of total SNPs within exons for gene blocks A, B and C. Bars are broken down to gene groups by colour. 20b Table of SNP numbers by exon within the *KIR* genes.

2.4 Discussion

The first complete cattle *KIR* haplotype sequence has been assembled from a Holstein-Friesian BAC library using Roche 454 pyrosequencing and finished with targeted Sanger sequencing. A partial second haplotype from the same animal has also been assembled, demonstrating identical gene content and order but with considerable polymorphic diversity within the complex. Further BAC clones from the second haplotype could not be found within the BAC library. Therefore, to fully genotype this animal further sequencing of genomic DNA would be required.

The results in this chapter show the cattle *KIR* complex has a total of 18 discrete loci over 266 kb of sequence. Eight of these loci encode predicted functional genes, six encode null-alleles and four encode pseudogene fragments. Of the eight functional genes, only one loci encodes an activating KIR, the other seven encode intact and predicted functional inhibitory receptors. However, all but one of the null-alleles encoded activating genes, demonstrating a significant bias toward inhibitory receptors within the cattle *KIR* haplotype. The cattle *KIR* complex contains both X and L-lineage genes. Cattle are the only species known to have expanded the X-lineage *KIR* genes, which remains a single gene within humans and non-human primates. Contrary to previous theories, cattle have also expanded the L-lineage genes (group II cattle *KIR*). However, only a single L-lineage gene, *BotaKIR2DL1*, is predicted to be functional within the cattle *KIR* complex, the other L-lineage genes are all null-alleles. Therefore the major functional genes within the cattle *KIR* haplotype are X-lineage, inhibitory genes with three Ig domains.

Sequencing of a partial second *KIR* haplotype has revealed the same structure and gene content but with notable polymorphisms concentrated within the Ig domain exons sequences. The *KIR* complex is flanked by the *LILR* genes at the 5′ end and the *FCAR* gene at the 3′ to create an LRC structure syntenic with other mammalian species.

2.4.1 Gene and block duplication mechanisms and models

This cattle *KIR* haplotype has expanded through a series of block duplications and subsequent rearrangements. This has resulted in 5 discrete blocks of *KIR* genes. Block duplication is likely to have occurred through non-allelic homologous recombination (NAHR) during meiosis and is depicted in Figure 17. Of the five discrete blocks, three of the blocks; blocks A, B and C, contain three of the same gene groups; I, II and IV, present in the same order. Blocks A, B and C have very high sequence identity and were likely generated from two NAHR

events. The first event gave rise to block B from block C, and subsequently block B generated block A. Before duplication of block B to form block A and after duplication of block C to form block B, a group III gene was inserted into block B. Group III genes are exclusive to blocks A and B and are likely derived from recombination between group I or V gene and the tail of an activating gene. This likely occurred prior to the duplication and diversification of the groups I and V which share a recent common ancestral gene. Although Figure 13a suggests that the group III and V share similar Ig-like domains, Figure 12 shows group III has Ig domains similar to both group I and group V. Therefore group III likely evolved before the formation of the blocks that can be seen now and has since been inserted into block B before duplication to form block A.

I predict that block D and E arose through a separate event with block D forming from block E. This would have included a gene recombination event between the extra-cellular domains of the group V ancestral gene and the trans-membrane domain of a group II activating gene generating *BotaKIR3DXS1*. The order of the gene duplication was predicted from the similarity between genes and blocks which is highest between block A and B than A and C or B and C. Furthermore, the number of SNPs in the second haplotype is higher in block C than B or A. However, this is a hypothesis based on two haplotypes and needs further evidence. The disrupted sequence of *BotaKIR3DXL6* is substantially different to the other group I genes *BotaKIR3DXL2* and *BotaKIR3DXL4* and may be a product of further recombination that is not accounted for in the model described.

There were inevitably further block duplications and gene recombination events that cannot be elucidated from the current cattle *KIR* haplotype sequenced. These further unpredicted events have been deleted from the current haplotype by more recent events. However, fragments of previous genes remain, such as the *BotaKIR1DP* and *BotaKIR2DXP* pseudogenes. These pseudogenes could be artefacts of the gene and block duplication events that occurred prior to the blocks seen now. To further understand the evolution of the cattle *KIR* haplotype and the origins of the *KIR* gene fragments, sequence analysis of other ruminant *KIR* haplotypes would need to be performed. This could highlight any shared genes that have been disrupted during the evolution of the cattle *KIR* haplotype, and would therefore have been present in the common ancestral haplotype.

The block duplication has generated five disparate groups of cattle *KIR* genes, each represented at two to four discrete loci within the *KIR* complex. The high levels of sequence similarity is likely to have prevented the identification of individual gene members between members of each gene group as discrete loci.

It is probable that the gene groups would have been classified as alleles without this completed reference sequence. It would also be prudent to consider any future alleles as potentially new discrete loci that have been generated through further block duplications and copy number variations, *i.e.* we cannot rule out the possibility that alleles sequenced independently are actually new genes elsewhere on the chromosome. To correctly genotype the *KIR* content of diploid cattle genomes, a quantitative approach would be required. This could take the form of qPCR or digital droplet PCR (ddPCR). However, to robustly design the probes required, the level of polymorphism within the *KIR* genes would need to be gauged.

2.4.2 Inhibitory *KIR* haplotype and the effect on NK cells

The cattle *KIR* haplotype sequenced in this project has 18 discrete loci. However, four loci contain pseudogenes and a further six loci encode null-alleles. Of the remaining eight functional genes, only one encodes a short activating receptor. Therefore, the cattle *KIR* haplotype has an inhibitory receptor bias akin to that seen in the human A haplotype. The inhibitory bias of the A haplotype in humans is postulated to generate a more potent NK cell response to combating infections [110].

The human A haplotype is found in nearly all human populations alongside the B haplotype that is believed to play a role in NK cell remodelling of the spiral arteries during trophoblast implantation in pregnancy [110]. The two haplotypes are believed to be maintained within human populations through balancing selection; haplotype A providing a selective advantage for fighting infections and haplotype B providing a selective advantage preventing pre-eclampsia [66].

It could be suggested that the inhibitory receptor bias seen in the cattle *KIR* haplotype results in a potent NK cell response to pathogens, generated via natural selection through generations of co-evolution with infectious diseases. However, cattle have been extensively bred and domesticated by humans which may have influenced their *KIR* genes, selecting individuals and breeding from them may have introduced a bias to one *KIR* haplotype. This potential founder effect can be investigated by interrogating the genome of ancestral cattle, the aurochs, which pre-dates the domestication of cattle [14, 43]. There is currently no evidence of a corresponding "B" haplotype in cattle however only a two haplotypes have been sequenced. Furthermore, it may be possible that the null-alleles encoding short receptors could be functional in other animals. This will become evident with further sequencing to gauge the levels of polymorphisms with cattle *KIR*.

2.4.3 Functional ablation of activating receptors

Of the six null-allele loci in the cattle *KIR* complex, we predict five loci would have encoded short tail activating receptors. A further two disrupted pseudogenes, *1DP1* and *1DP2*, would also encode short tail activating receptors. Therefore, seven discrete loci containing activating receptors have become non-functional which is in contrast to the single non-functional inhibitory receptor, *BotaKIR3DXL6*01N*, within the complex. This supports the model outlined by Laurent Abi-Rached and Peter Parham [1], whereby activating receptors are derived from inhibitory receptors, then subsequently become non-functional. Viral subversion of inhibitory receptors, utilising ligand homologues, generates selection pressures forming activating receptors from inhibitory genes. The genetic mechanism involved is either point mutations generating a lysine within the transmembrane domain and a stop codon before the ITIM motifs or through recombination of the entire transmembrane domain with another activating gene. Short tailed activating genes are short lived because they can be detrimental to host fitness due to recognition of self ligands leading to autoimmunity.

The short tailed null-alleles and pseudogenes within the cattle *KIR* complex from a common gene group have likely duplicated after mutation created premature stop codons. As seen in Table 4, each group of genes each have stop codons in the same position. Therefore these null-alleles have been carried over from block duplication after they became non-functional. Nonetheless, the process of inactivating short tailed receptors has occurred independently three times during the evolution of the cattle *KIR* haplotype, with short-tailed genes from groups II and III, and the gene fragments *1DP1* and *1DP2* all becoming non-functional. Therefore, it is possible that during the evolution of the CKH, these three gene groups had inhibitory ancestral receptors that were subverted by viral infections, creating a selection pressure to generate activating genes. Subsequently, the activating receptors became detrimental to the host in the absence of viral selection pressures disappeared and the genes became non-functional as a result of negative selection pressures.

2.4.4 Evolution of activating *KIR* receptors in cattle

The activating *KIR* gene short tail exon sequence in cattle, including the arginine containing transmembrane domain and disrupted cytoplasmic tail, has evolved once prior to duplicating and recombining throughout the haplotype. This is shown in Figure 13b demonstrating that all of the tail segments clade together within their functional groups, including both X and L-lineages. Therefore, the

most parsimonious explanation is that the activating tail was inherited from a single activating receptor and has duplicated throughout the haplotype via recombination with inhibitory genes. This creates activating receptors that recognise the same ligands of the inhibitory receptors but translate these interactions to an activating signal. This has happened between the group I/V ancestral gene and the group III genes that have similar ectodomains (Figure 13a) but different signalling domains (Figure 13b). This has occured in the L-lineage group II genes and potentially more recently in the group V genes, where *BotaKIR3DXL1* and *BotaKIR3DXS1* have nearly identical Ig domain sequence but different signalling domains.

The group V genes are an example of paired-receptors, proteins that share nearly identical ligand binding extracellular domains but have different signalling domains. These paired-receptors have been found in several gene families and include an inhibitory allele or locus together with an activating allele or locus [2]. Selection pressures during host-pathogen co-evolution are believed to be the driving force behind the evolution of the paired-receptors. One notable example are the Ly49I and Ly49H recepors in mice, which both recognise the m157 protein produced by murine cytomegalovirus (MCMV) [4]. This decoy protein subverts the inhibitory receptor Ly49I in MCMV susceptible mice. However, the activating paired-receptor Ly49H binds the protein and causes resistance to the infection (reviewed in detail here [4]). It is thought that *BotaKIR3DXL1* and *BotaKIR3DXS1* have undergone similar selection pressures to generate paired-receptors. These group V genes are found encoded on the same haplotype and therefore, if both are expressed, could produce opposing signals to the same target cell. The translation and cell surface expression of these genes may be controlled by further factors such as intron silencing to prevent contradictory signals.

The evolutionary origin of the activating tail in cattle is unknown and has evolved independently to primate activating tails. This is because primate activating tails transduce a signal through a charged lysine residue within the transmembrane domain, whereas cattle utilise a charged arginine. These residues are at slightly different relative positions within the transmembrane domain and interact with different signalling adapters. The human and primate KIR2DL4 receptor has an arginine residue within the same position as the cattle activating receptors. However, there is very low sequence identity and this functional similarity is likely a result of convergent evolution to utilise the same signalling adapters than of shared ancestral sequence.

2.4.5 Conclusions from the first cattle *KIR* haplotype sequence

In this chapter we have sequenced and assembled the first cattle *KIR* haplotype using massively parallel sequencing and finished with targeted standard sequencing. The finished haplotype took several cycles of targeted sequencing, hybrid assembly and manual editing which demonstrated the difficultly in producing a reliable *KIR* complex from the whole genome assembly attempts. This first cattle *KIR* haplotype displays the importance of *KIR* genes within the cattle genome. With several functional loci, it likely evolved through years of pathogen selection pressures meaning the KIR receptors are crucial to the health and fitness of cattle. Now with the capabilities afforded by complete genomic sequences for each of the *KIR* loci, it is now possible to study each gene individually. This will enable the extent of polymorphisms and gene presence or absence to be studied for each *KIR* gene. Furthermore the complete haplotype reference sequences will facilitate polymorphism and copy number variation investigation by short read mapping studies. From this, the function and importance of the cattle *KIR* genes can be interrogated.

3 Chapter 3. *KIR* in the ancient cattle genome

3.1 Introduction

The animal used to sequence and assemble the *KIR* complex in chapter 2 was a Holstein-Friesian (HF), a breed almost exclusively used for dairy production in Europe and north America. HF have undergone intensive artificial selection to make it the highest milk producing breed in the world. The intensive selection for production traits has been prioritised over selection for health based traits i.e. disease resistance [52]. This process may have affected immune gene complexes such as the *KIR* complex, resulting in haplotypes that do not reflect evolution by natural selection. There may also be a lack of diversity within the *KIR* complex caused by inbreeding, reliance on restricted number of sires and the founder effect.

It is important to understand the differences between the *KIR* complex in modern domesticated cattle and the *KIR* complex that evolved before domestication. This is because the *KIR* complex of cattle prior to domestication has evolved through natural selection pressures. Differences between the wild and domesticated complex may be an indication of loss of function. There is evidence that cattle NK cells play a much reduced role in pregnancy compared to humans, so the role of *KIR* is likely to be involved in host-pathogen recognition and immune surveillance [106]. This could be useful for future breeding programmes where the *KIR* alongside other immune genes are considered for improving animal health.

To gain an indication of the influence domestication has had on the cattle *KIR* complex, the ancestral cattle genome was interrogated for *KIR* sequences. Modern taurine (*Bos taurus*) cattle, including the Holstein-Friesian breed, originated from the aurochsen species of cattle. This species, *Bos primigenious*, has been shown to be the ancestor to taurine cattle and pre-dates human domestication of livestock [43]. The aurochs sequenced here has been radio carbon dated to be approximately 6,700 years old and originated from the Derbyshire area [43]. Although cattle are predicted to have been domesticated approximately 10,000 years ago, this bone pre-dates the arrival of the human populations responsible for domesticating livestock to the British Isles. Therefore the aurochs genome studied here is believed to have been wild and not a product of artificial selection via domestication. To genotype the aurochs for *KIR* the raw sequencing reads were obtained and mapped to the cattle *KIR* complex. This enabled *KIR* gene presence/absence to be determined however the fragmented ancient DNA molecules caused reduced read length that impacted the resolution of *KIR* genotyping.

3.2 Methods

3.2.1 Custom cattle genome construction

A custom cattle genome reference sequence was generated utilising the UMD build 3.1 (ID GCA_000003055.3). Blat searches revealed the locations of the *KIR* sequences within the genome build and a bespoke python script was used to generate a new genome with the *KIR* sequences omitted, described in the appendix (section 9.2.4). From this, two reference sequences were generated; one with each cattle *KIR* sequence from the haplotype 1 sequence in chapter 2 inserted, and another with the full length *KIR* complex inserted. The first reference sequence was used to determine group aligning reads and the second was used for uniquely aligning reads and coverage breadth and depth along the haplotype.

3.2.2 Sequence alignment bioinformatic pipeline of aurochs genome Illumina reads

The aurochs genome has been sequenced, but not published, based on DNA extracted from the proximal half of a humerus bone, previously used to determine the mitochondiral genome sequence [43]. The DNA was sequenced using the Illumina HiSeq 2000 platform with 37 or 74 cycles.. The Illumina reads had been trimmed for adapter sequence and quality by the providers at Trinity College Dublin. All sequences from each sequencing run library were aligned to the *KIR* complex sequence, assembled in chapter 2. Reads were aligned using the BWA aln algorithm [84] with default settings, aligned reads and their read partner were extracted using a bespoke python script described in the appendix (section 9.2.2). The extracted reads were aligned to the custom cattle genome build using BWA aln with default settings. Alignments were filtered for uniquely mapping reads using a combination of samtools and grep commands described in the appendix (section 9.2.3). Reads that alternatively aligned to genes from other groups were extracted using a bespoke python script described in the appendix (section 9.2.3).

3.2.3 Sequencing coverage breadth and depth calculation

Coverage depth was calculated using bedtools coverageBed program and plotted with a sliding window (line smoothing) value of 300 bp using a bespoke python script (section 9.2.6). The bed format coverage depth file produced was used to calculate the coverage breath using a bespoke python script (section 9.2.6). Positions mapped with more that 1x aurochs or simulated read were calculated,

positions with more than 1000 x BAC DNA reads were calculated to generate a total percentage coverage for the entire complex as well as each intron, exon and gene. BAC sequencing required a higher threshold due to the higher sequencing depth generated and the spill over reads from other BAC clones.

3.2.4 Simulated dataset creation and analysis

Simulated datasets were generated from the haplotype 1 cattle *KIR* complex reference sequence assembled in chapter 2. A bespoke python script (appendix section 9.2.5) was used to generate artificial fastq sequences of varying lengths from the *KIR* complex sequence. The script cut the sequence with varying overlap dependent on the specified sequence length using the normal and reverse complemented sequence. Each base position was assigned a quality score of Q30 in order to represent good quality Illumina sequencing. Each simulated dataset was aligned using the pipeline described in section 3.2.2, however BWA mem [83] was used due to its higher accuracy with longer read lengths (over 100 bp). Accuracy was checked at shorter read lengths and was identical to BWA aln (data not shown). Increments of 15 bp were used from 15 bp to 2000 bp DNA fragment lengths. The coverage depth and breadth was calculated using the methods described in section 3.2.3. Coverage breadth over fragment length was plotted using a bepsoke python script for all the different simulated fragment sizes generated (appendix section 9.2.5).

3.2.5 BAC clone DNA sequencing

BAC clone DNA was previously produced in section 2.2.7. Raw reads were artifially cut down to a read length of 35 bp in order to simulate the aurochs genome sequencing.

3.2.6 High resolution loci defining SNP analysis

Files containing the samtools pileup format were generated for each *KIR* sequence based on the group aligning reads from the aurochs genome. Sequences from each group of *KIR* were aligned using mafft and manually corrected using seaview to generate an alignment in fasta format. Each difference between pairs of sequences (AKA loci defining positions) within this alignment was interrogated within the pileup file to determine aurochs genome sequence concordance with the loci defining position. This was performed using a bespoke python script and pipeline described in section 2.2.7 in the appendix. The total number of loci defining positions (LDPs) for each gene pair was calculated along with the

number of LDPs covered by aurochs mapped sequencing reads. The number of reads that were discordant with the reference genes LDP was calculated to define genes that were not represented. The percentage of reads corresponding to the LDP was determined at 50%, 75% and 100% thresholds.

3.2.7 SNP calling

SNPs were called from the group read filtered alignments using the same methods described in section 2.2.7. Residue changes for aurochs were compiled based on *KIR* group as reads may have alternatively mapped to the other group genes.

3.3 Results

3.3.1 The aurochs raw genome sequencing reads were aligned to the cattle *KIR* complex

The aurochs raw genome sequences were supplied by collaborators at Trinity College Dublin and have not been published yet. The data represents the genome of an individual, radio carbon dated at approximately 6,700 years old. The ancient DNA was isolated from a bone artefact and sequenced with the Illumina platform, as previously described by the group after sequencing the mitochondrial DNA [43]. The ancient DNA was heavily fragmented and therefore average read length is 37 bp making *de novo* assembly impossible, Table 6. To determine the *KIR* content and any *KIR* sequence variation within the aurochs genome, the raw reads needed to be aligned to the *KIR* complex sequence assembled in chapter 2.

Mapping an entire genome to just the *KIR* complex results in a surplus of aligned reads that should map to other regions of the genome. However, aligning all of the raw aurochs sequencing reads to the entire cattle genome and the *KIR* complex requires a lot of computational time and resources. Therefore, to reduce the quantity of the raw reads to align but also remove reads that align to other positions within the genome, a custom pipeline was established. Firstly, the raw reads were aligned to the *KIR* complex reference sequence, then the reads that aligned, and their sequencing pair (if available), were subsequently extracted from the original pool of aurochs genome sequences. These extracted reads were then aligned to a custom genome build which has had all *KIR*-like sequences removed and the *KIR* complex reference sequence inserted as a standalone chromosome.

Alongside the aurochs genome, raw sequencing reads from human and dog genomes were aligned to the cattle *KIR* complex as negative controls, Table 6. Humans have expanded the *KIR* but the sequences vary significantly, and dogs have no *KIR* genes within their genome [60]. Although these two genomes were sequenced with 100 bp reads, they were artificially reduced to 37 bp read lengths to simulate the aurochs genome sequences. Alignment of reads from these genomes to the cattle *KIR* complex would highlight non-specific regions of the cattle *KIR* complex.

The raw sequencing reads from each genome were aligned to the *KIR* complex then subsequently the aligning reads were extracted, Table 7. No reads from the negative control genomes aligned to the cattle *KIR* complex. This suggests that the cattle *KIR* complex is specific to cattle and not unrelated mammals.

From a total of just under 2.7 million reads from the aurochs genome, 88,952 reads aligned to the cattle *KIR* complex which is representative of 0.00332%

of the total genome reads, Table 7. However, when these reads were extracted and aligned to the *KIR* complex and custom genome build, only 13.44% of the extracted reads aligned to the *KIR* complex, Table 8. Therefore, a large portion of the *KIR* complex reads extracted are from repeat regions in the genome. There are several different repeat sequences within the cattle *KIR* complex that are repeated across the genome. Therefore, all the repeat reads from the genome will have aligned to the *KIR* complex sequence. This may also be true for gene sequences from other regions of the genome that are similar to the LRC gene sequences. By aligning the extracted reads to the custom genome, these excess reads have have been prevented from aligning to the *KIR* complex.

3.3.2 Simulated data reveals the limitations of short read genomic alignments to cattle *KIR* complex

Before the short read aligned data could be interpreted, the limitations of the read lengths were explored. The aurochs genome raw sequences have an average read length of 37 bp; to understand the quantity of the LRC that could be accurately mapped using these short reads, simulated datasets were generated.

The percentage of read coverage breadth increases with fragment size when aligned to the cattle *KIR* complex, Figure 21. Uniquely mapping reads need a greater fragment size in order to completely cover the complex. Therefore to align reads to the complex uniquely, longer reads are required to span a greater number of loci defining positions. The percent breadth of coverage over just the uniquely mapping *KIR* exon sequence is lower than the percent breadth of coverage over the uniquely mapped entire *KIR* complex up to a fragment size of 290 bp. A fragment size over 290 bp results in higher percentage breadth of coverage in the *KIR* exon regions than the entire complex. To uniquely map reads to just the *KIR* exons, a read fragment length of at least 550 bp is required. This is beyond the capabilities of current Illumina technology which is used for the majority of genome resequencing projects. To uniquely align reads to the entire complex, a fragment length of over 1.2 kb is required. This simulated dataset suggests that the breadth of coverage for the aurochs raw genome reads will only be around 50-60% of the *KIR* complex due to the limited read length of 37 bp. This 37 bp read length is inadequate for the repetitive region of the genome targeted here. Therefore, a lack of aurochs sequence coverage is not indicative of gene absence within the aurochs genome. Further quantitative comparison between the aurochs breadth of coverage and positive control datasets is needed to determine gene presence or absence. Furthermore determining sequence variation within the *KIR* will not be possible for all of the genes.

Animal	Specis	Accession	Technology	Samples	No. Reads	Av length	Bases (Gb)
Aurochs	*Bos primigenious*	not released	Illumina	1	2,680,282,530	37.00	99.73
Korean Jindo Dog	*Canis lupus familiaris*	DRP000492	Illumina	1	1,075,574,618	100.00	107.56
Human	*Homo sapiens*	SRP002509	Illumina	1	1,269,435,784	100.00	126.94

Table 6: Table of raw whole genome sequencing details from the aurochs and two negative control animals, human and korean dog, genome sequencing projects. Data provided from NCBI SRA with the accession numbers shown. Aurochs genome data was provided by David McHugh *et al.* from University College Dublin (publication in preparation). Average read length (Av. length) is shown in base pairs. Bases is a calculation of total number of base pairs for each genome, despite having more reads, the lower read length of the aurochs genome sequencing has few total bases than the positive controls.

Animal	Haplotype				KIR exons				mapped reads	mapped reads
	N reads	total bases	Av length	Av cov.	N reads	total bases	Av length	Av cov.	% (haplotype)	% (KIR exons)
Auroch	88,952	2,943,238	33.09	7.95	4847	161,379	33.29	8.97	0.00332	0.00018
Korean Jindo Dog	0	0	0.00	0.00	0	0	0.00	0.00	0.00000	0.00000
Human	0	0	0.00	0.00	0	0	0.00	0.00	0.00000	0.00000

Table 7: Table showing details of extracted LRC reads from the aurochs raw genome sequencing. This is the details of the raw reads mapping to the *KIR* haplotype sequence and just the *KIR* exon sequences. The mapped read % is the quantity of the total raw reads that map to the haplotype or the *KIR* exons. The reads that mapped to the haplotype were extracted to be mapped to the custom genome build.

N reads	Total bases	Av length	% of extracted
11,951	441,851	37.0	13.44

Table 8: Table of statistics from aurochs extracted LRC reads mapped to the LRC in the custom genome. Number of reads, total bases and the average length in base pairs of the extracted aurochs reads that mapped to the *KIR* haplotype within the custom genome build. The custom genome build has all *KIR* sequences removed and the *KIR* haplotype from chapter 2 has been inserted as a standalone chromosome.

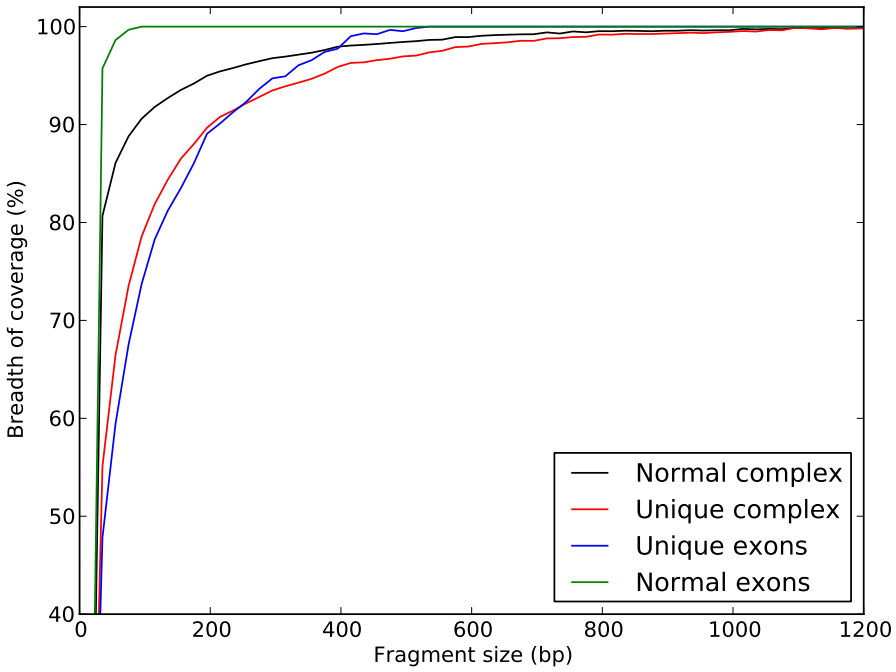


Figure 21: Chart showing percentage of read coverage breadth over simulated fragment length. Simulated sequence fragments were artificially generated from the *KIR* haplotype then aligned back to the *KIR* haplotype. Read coverage breadth was calculated for positions with read coverage of 1x or above over the entire haplotype and just the exons. This is displayed as a percentage where 100% would be complete coverage of the haplotype or *KIR* exons. BWA mem was used for all alignments and Unique represents uniquely mapping reads, Normal represents reads unfiltered. The four lines show the total unfiltered *KIR* complex (black) and *KIR* exons (green), as well as the uniquely aligned *KIR* complex (red) and *KIR* exons (blue).

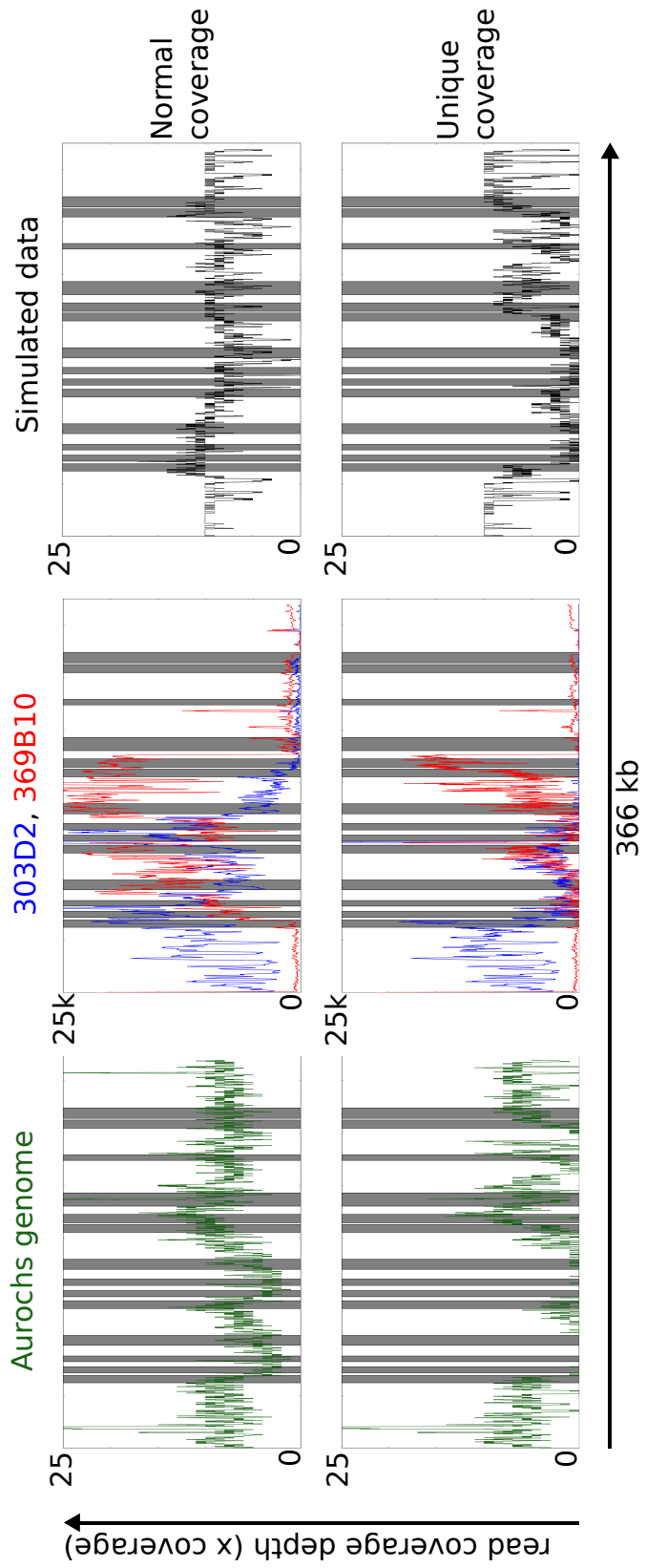


Figure 22: Read fragment coverage depth histograms of aurochs (green), BAC clones (303D02 blue, 369B10 red) and simulated data (black). X-axis represents position along the *KIR* complex and the Y-axis represents the read coverage depth in base pairs. Vertical grey columns represent the *KIR* positions along the complex. The read depths coverage line has been smoothed with a sliding window of 300 bp. The three top charts represent normal read coverage where reads have not been filtered. The bottom three charts represent unique coverage where reads have been filtered for fragments that only map to one position.

3.3.3 Uniquely mapped read coverage depth and breadth is reduced in repetitive areas of the *KIR* complex

The aurochs genome normal coverage shows that all *KIR* loci are represented by raw sequencing reads, Figure 22. However, when the ambiguously mapping reads are filtered out leaving only uniquely mapping reads, read coverage is reduced. There are very few uniquely mapping reads over *BotaKIR2DS2/3* and *BotaKIR3DXS2/3* within the blocks A and B of the *KIR* complex. To prove this is an artefact of the repetitive nature of the complex, and not the lack of these genes from the aurochs genome, the same alignment pipeline was repeated using positive control datasets.

The BAC clones (303D02 and 369B10) were used to assemble and verify the first *KIR* complex sequence and therefore contain identical sequence to the assembled complex used as the reference sequence. The Illumina sequences from the BAC clones were artificially reduced to 35 bp lengths. Therefore, the BAC reads are comparable to the raw aurochs genome reads. Simulated reads as described in subsection 3.3.2 were used with a read length of 35 bp to be comparable with the aurochs genome reads. Both the BAC 35 bp reads and the simulated data show that after filtering for uniquley mapping reads, there is a significant reduction in read coverage over the *KIR* sequences within block A and B, Figure 22, which confirms the lack of read coverage is a consequence of insufficient read length and not genomic structural variation. There is also a reduction in normal read depth coverage over the block A and B *KIR* genes in the aurochs genome. The coverage depth is roughly half of the rest of the complex. This suggests that the aurochs genome may be heterozygous, with one haplotype representing the complex assembled in chapter 2 and another a reduced gene number haplotype.

Read coverage breadth was calculated as a percentage of sequence coverage over the X-axis. Coverage along the X-axis of the haplotype is the amount of sequence accounted for by mapped raw reads. Read breadth coverage of 100% indicates that the entire gene is accounted for by aligned sequencing reads. The breadth of coverage was calculated for each gene in the cattle *KIR* complex from the aurochs, 35 bp BAC and simulated data alignments, Table 9. The coverage is based on uniquely mapped reads and is equivalent to the data shown in Figure 22. There is reduced read coverage breadth for the *2DS2/3*, *3DXS2/3* and *3DXL5/7* *KIR* in all of the datasets, table 9. This further demonstrates that the reduction in read coverage is the result of inadequate read length and not absence of the genes.

	Aurochs	BAC	Simulated
3DXL6	68.3	80.6	87.7
2DS3	7.1	17.8	15.6
3DXS3	5.5	0.8	17.2
3DXL7	6.8	19.8	18.3
3DXL4	28.7	51.9	47.3
2DS2	3.9	19.0	17.9
3DXS2	6.9	13.4	16.5
3DXL5	13.5	21.9	17.2
3DXL2	51.7	62.1	52.0
2DS1	68.3	75.0	74.1
3DXL3	65.2	4.3	75.0
3DXS1	58.7	66.9	61.4
3DXL1	55.2	72.1	68.0
2DL1	83.3	67.6	95.4

Table 9: Table of aurochs *KIR* total breadth of sequence coverage. Each number is a percentage of the total gene length that is covered by at least one sequence (1000x coverage required for the BAC clones). Numbers have been shaded with lower read breadth of coverage a darker shade of grey. *BotaKIR3DXL3* is very low within the BAC clones because it was not covered by the BACs sequenced with Illumina data and is therefore not represented.

3.3.4 High resolution analysis of the loci defining single nucleotide positions predicted *KIR* gene presence within the aurochs genome

The uniquely mapping reads only align to the cattle *KIR* complex if they cannot be aligned elsewhere to the reference sequence. Therefore the uniquely mapping reads align to loci defining positions. These are positions that define the *KIR* sequence from other *KIR* within the same group. Within the more divergent *KIR* groups there are more loci defining positions that enable reads to align uniquely to those loci. However, within the least divergent groups such as the short tailed group II and group III genes, there are fewer loci defining positions for the reads to align to uniquely. Therefore, these genes cannot be confirmed with as much confidence as the more divergent genes. To enable the distinction of group II and III genes a strategy using more sequences and comparing the loci defining positions within each group was used.

For this high resolution loci defining position analysis, normal read alignments were used and only fragments that alternatively mapped to the same gene groups were kept. Sequencing fragments that alternatively aligned to genes from other gene groups were discarded. This provided greater read coverage over the less divergent genes than using just the uniquely mapping reads. However, many reads alternatively aligned to two or more different loci from the same gene group. To distinguish between the loci and determine the likelihood of a gene being represented by the aligned raw sequences, the single nucleotide differences between the loci were analysed. These loci defining positions are the single nucleotide positions that distinguish one loci from the others in the same gene group. All of the loci defining positions for each gene were compared to calculate gene representation. Both haplotype 1 and 2 alleles were used for loci defining positions, however, it is likely further alleles exist and therefore this approach will not account for these. Furthermore, novel genes that have not been defined yet will not be detected using this approach.

For each gene, the number of possible loci defining positions were calculated. The total number of comparable loci defining positions varies between genes; the comparison of *3DXL3* against *3DXL5* has significantly more total comparable positions than *2DS2* against *2DS3*. The actual number of compared positions was dependent on the coverage breadth of sequencing. Because the positions are loci defining, there is greater potential for unambiguous sequence alignment over these positions. Therefore, absence of aligned sequence over the loci defining positions may be an indication of gene absence from the genome. At positions where sequence had aligned from the raw genome reads, concordance between the

loci defining positions and the genome were counted. This was done by counting the individual reads at the position with the same base. Heterozygosity within the aligned genomes was considered and concordance was measured at 100%, 75% and 50% of the aligned reads.

The aurochs genome contains each *KIR* gene locus, however *2DS2* is missing notably more sequence than the other genes, Figure 23. Despite the lack of sequence aligning to *2DS2* there is still over 50% of the locus defining positions represented and consistent with the reference genes. Therefore, this locus is likely represented within the aurochs genome. These findings suggest that the aurochs genome contains the same *KIR* genes found in the HF *KIR* haplotype sequenced in chapter 2.

This analysis robustly confirms the presence of the HF *KIR* within the aurochs genome. However, the presence of novel cattle *KIR* sequences in the aurochs genome cannot be predicted. The presence of all the HF *KIR* within the aurochs genome confirms the structure of the haplotype had formed before the domestication of *Bos taurus* cattle breeds. This is significant as it proves the HF *KIR* complex evolved through natural selection prior to domestication and has been unaffected by artificial selection pressures.

3.3.5 The *KIR* sequences have remained functionally unchanged within the aurochs genome

It was not possible to determine SNPs within all of the aurochs *KIR* sequences due to the lack of coverage caused by the inadequate read length. Each gene was instead interrogated for SNPs in the context of its own group by using the group aligning reads described in subsection 3.3.4. Therefore, the SNPs reported are compiled from all the genes in each group. There was no evidence of any mutation resulting in a null-allele or pseudo gene encoding a functioning allele. The predicted functional *KIR* within the aurochs genome encode residue variations compared to the HF *KIR*, Table 10. Each gene group encodes several residue changes that may have an effect on the receptor function, but that has not been established yet. There was no evidence of the predicted functional *KIR* in the aurochs genome encoding null-allele variants.

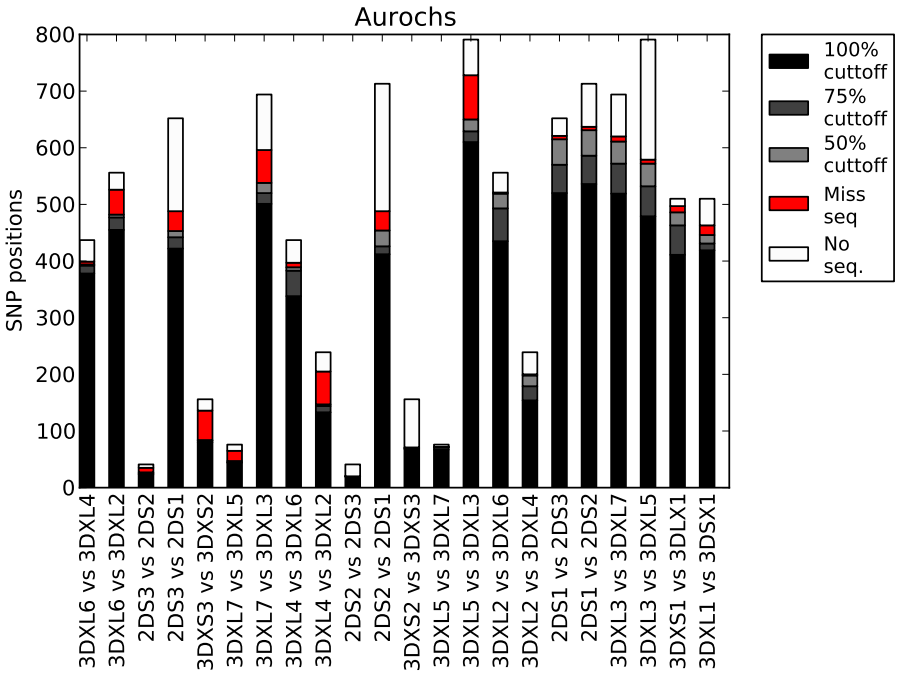


Figure 23: High resolution SNP analysis of *Bos primigenius*. Comparison of gene defining SNP positions between gene group loci. Each bar is representative of the total unique positions between two genes from the same group. Aurochs reads that aligned to the *KIR* haplotype within the custom genome build were filtered out if they alternatively mapped to genes from a different gene group. From this filtered alignment, the number of positions that corresponded to the *KIR* locus and not the alternative group *KIR* locus was calculated. This was conducted using different thresholds of sequences corresponding to the reference base. Black represents 100% of the reads were consistant with the base, dark grey was 75% and light grey was 50%. Red represents reads that were not consistent with the reference base and may be the result of reads mapping to the wrong gene within the *KIR* group. White represents the number of loci defining positions that were not covered by reads.

3.4 Discussion

To genotype the *KIR* within the aurochs genome, raw sequencing reads were aligned to the cattle *KIR* complex. Through different analyses of the aligned reads it has been established that the aurochs genome represented the same *KIR* as the Holstein-Friesian (HF) reference complex. However, the sequencing read length of 37 bp is limiting, preventing SNP calling from individual genes and the prediction of novel genes not found within the HF. Nonetheless, the aurochs *KIR* complex appears to have the same gene content as the HF complex with detectable residue changes between the *KIR* groups containing functionally intact genes.

3.4.1 Cattle *KIR* evolved through natural selection

The fact that all of the HF *KIR* are represented within the aurochs genome confirms that they evolved before human intervention and domestication of taurine cattle. Therefore, the cattle *KIR* complex has evolved through natural selection and has not been altered by artificial selection pressures or founder effect. It is thought NK cells play a much reduced role in cattle pregnancy than in humans [106], therefore it is hypothesised that cattle NK cells are primarily involved in detecting infected and transformed cells. The function of KIR is predicted to be licensing NK cells and surveying MHC class I for presence/absence or viral peptides. Therefore, the major selection pressures that have acted on the *KIR* during cattle evolution are predicted to have been pathogen and MHC class 1 mediated. This would confirm the genes have a significant relevance in the study of animal health. The cattle *KIR* genes have likely evolved to play a significant role in fighting disease and therefore likely still do. This chapter has indicated their importance and reinforced the need to study the diversity and function of the receptors further.

3.4.2 Cattle *KIR* null-alleles were deactivated from mediation by selection pressures occurring before domestication

The process of cattle *KIR* evolution has resulted in the rise and fall of short tailed activating receptors. Based on models from other species, these genes are likely to have changed function from inhibitory to activating, to fight the subversion of viruses using MHC class 1-like decoy proteins. Once the viral threat has subsided from the cattle populations the requirement for the short tailed *KIR* receptors disappears, this has left five null-allele activating *KIR* within the genome. The aurochs genome shows no evidence of these genes being functional.

Therefore, the evolutionary process leading to their inactivation has happened before domestication of taurine cattle. This indicates the non-functional genes were not a result of founder effect and the cattle *KIR* complex is a relevant example of an innate immune complex that has expanded and diversified as a result of pathogen selection pressures.

3.4.3 Variable MHC leads to non-variable *KIR*?

There are a total of six different cattle MHC genes, with a maximum of three genes on a haplotype [61]. These gene variable haplotypes are considerably different to the constant three gene classical class 1 haplotypes seen in humans and primates. Humans and primates have gene presence absence variation within their *KIR* complexes but do not have this variation within their classical class I complexes. Could the reversal of this trend in cattle be an indication of receptor-ligand restriction, where only one of the complexes can be gene variable? Therefore, as cattle have gene variable MHC class I complexes, they require gene constant *KIR* complexes to recognise at least one of the MHC class I ligands (assuming MHC class I are the KIR ligands in cattle). The evidence so far suggests that cattle have gene constant *KIR* complexes, however more information is needed to confirm this.

3.4.4 In what ancestral species did the current cattle *KIR* gene complex form

The aurochs genome studied in this chapter has been radio carbon dated to approximately 6,700 years old. The cattle *KIR* complex has therefore formed sometime before this. We know the human and simian primates, that shared a common ancestor with cattle approximately 60 mya [105], have expanded a different lineage of *KIR* gene, from the L-lineage, however they contain a single copy X-lineage gene, *KIR3DX1*. We know that the two *KIR* lineages diverged approximately 135 mya [56]. Therefore the cattle *KIR* complex has formed between 135 mya and 7,000 years ago. This is a large time frame with many different speciation events, to determine in which species the cattle *KIR* complex formed, other ruminants genomes should first be interrogated. It is not known how similar other ruminants that diverged from cattle approximately 26 mya [65], such as sheep and goats, *KIR* complexes are to cattle. Understanding the similarities and differences may indicate functionally important gene groups or families that have been maintained independently within each species. It will also provide an insight into the rate of *KIR* expansion and diversification in non-primate species.

Domain	D0								D1						D2				STM
Residue	31	42	45	47	54	56	58	79	113	121	127	156	158	163	199	225	235	279	307
Consensus	R	T	R	R	H	F	N	W	Q	R	V	A	F	M	S	E	R	H	M
BotaKIR3DXL1		R/T	K/E	H	D/H					R	L/M							C/R	
BotaKIR3DXS1		T	K	R/C	H					R	L							C/R	
aurochs		R/T	K/E	H/R	D/H					R/H	L/M							C/R	
BotaKIR3DXL2	R			G		F	N	W/Q				A	F			E	M		
BotaKIR3DXL4	R			G		F	N	W				A	F			E	M		
BotaKIR3DXL6	R			R		V	N	W				A	F			E	R		
aurochs	R/H			R/H		F/L	N/K	W/R				A/V	F/L			E/K	M/I		
BotaKIR3DXL3									S/R	C				M/V	Y				M
BotaKIR3DXL5									G/R	R				M	S				M
BotaKIR3DXL7									R	R				M	S				M
aurochs									S/R	C/R				M/V	Y/S				M/T

Table 10: Table showing the variable residues within the compiled aurochs *KIR* groups compared to the Holstein-Friesian reference sequence. SNPs were called from the filtered alignment of group aligning reads described in Figure 23. Genes from each group have been compiled to a single representative gene because the individual genes cannot be distinguished from this dataset. Shaded residues are predicted to have been under positive selection (based on PAML prediction). Red residues are unique to aurochs.

4 Chapter 4. *KIR* in the sheep genome

4.1 Introduction

Cattle have expanded and diversified the *KIR* genes independently to humans. Sheep, another *Bovidae* species, shared a common ancestor to cattle approximately 25.4 mya [65]. Sequencing and assembling the sheep *KIR* haplotype is key to understanding the evolution of the KIR receptors in ruminants. Understanding the similarities and differences between the cattle and sheep haplotypes may indicate functionally important gene groups. It will also highlight the extent of *KIR* expansion in the two species over the last 25.4 million years. This will give an indication of the *KIRs* importance within the immune system. As described in chapter 3, the cattle *KIR* expansion likely occurred through natural selection and not domestication; the extent of sheep *KIR* expansion from its common ancestor with cattle could also have been the result of natural selection pressures. However, as sheep have also undergone domestication, artificial selection cannot be discounted from influencing the sheep *KIR* gene complex. Predicted functional similarities between the two species may be an indication of common pathogens, such as foot and mouth disease virus (FMDV), bluetongue virus (BTV) and Schmallenberg virus. These pathogens or closely related pathogens have infected both sheep and cattle during their evolution. This may have shaped their immune systems via an evolutionary arms race between pathogen and host, as the pathogen adapts to hide from, or subvert the immune system, the host has to generate new receptors and mechanisms to detect the pathogen.

Though the sheep *KIR* have not been characterised in any previous publications, there is an assembled *KIR* haplotype within the 3.1 build of the sheep genome [5]. However, this region contains several scaffolded sequences resulting in unfinished regions of the sheep *KIR* complex and reducing confidence in the annotated assembly. However, the combination of second generation sequence technologies used to generate this *KIR* complex is likely to have had greater assembly success than the traditional Sanger sequencing utilised by the cattle genome project. Nevertheless, as shown in chapter 2, the assembly of an expanded *KIR* complex sequence may need finishing by the addition of longer sequencing reads to span repeat regions. Therefore, this haplotype sequence needs to be confirmed before full annotation and analysis.

In order to finish the sheep *KIR* haplotype, a similar strategy to the assembly of cattle *KIR* haplotype (chapter 2) was employed. BAC clones from the texel breed of sheep were sequenced with Pacific Biosciences smrt cell sequencing. This BAC library was also used in the sheep genome project and is therefore directly

comparable to the sheep *KIR* haplotype assembled as part of the whole genome attempt.

4.2 Methods

4.2.1 BAC clone DNA preparation and sequencing

BAC clone DNA was prepared using the same method described in section 2.2.2. BAC clone 263M01 DNA was re-suspended in TE whilst 422J05 was re-suspended in water. BAC clone DNA was sequenced using the PacBio RS II system with either version 1.3.3 or 2.0.3 SMRT cell (Pacific Biosciences INC, California, USA). The DNA library preparation and sequencing run was conducted by GATC biotech (GATC-biotech AG, Konstanz, Germany). A single movie was recorded in order to generate long reads.

4.2.2 Assembly of PacBio sequence data

Vector sequence was screened from the raw sequence files by aligning the pTarBAC1.3 vector sequence to the raw fastq sequences using BLASR [22]. A bespoke python script was used to generate a list of reads that do not contain vector sequence (script shown in section 9.3.1). This was added to the HGAP assembly process [30] XML file and run via the SMRTpipe pipline (Pacific Bioscience smrtportal analysis pipeline).

4.2.3 Sequence characterisation and annotation

Consensus sequences were characterised using the same methods described in section 2.2.8.

4.2.4 Sequence and gene analysis

Phylogenetic tree construction, dot plots and sliding window analysis were conducted using the same methods described in section 2.2.9.

4.2.5 BAC clone DNA Illumina sequencing

BAC clone DNA was sequenced using the same methods described in section 2.2.7. Error checking was conducted using the same methods described in section 2.2.7.

4.2.6 Sheep genome characterisation

The *KIR* and *LILR* region of the sheep genome build 3.1 was extracted from chromosome 14. Sequence was extracted from position 59,546,707 to 60,686,753 then

reverse complemented. This extracted genome build sequence was characterised using the same methods described in section 2.2.8.

4.3 Results

4.3.1 PacBio sequencing yielded long reads that fully assembled using HGAP

To sequence and assemble the sheep *KIR* haplotype, two BAC clones were sequenced using the Pacific Biosciences RS II platform. This process uses template DNA and polymerase molecule complexes seconded to the bottom of wells called zero-mode waveguides (ZMWs). Phospholinked nucleotides are introduced to the ZMW chambers and are sequentially incorporated into the complementary strand of DNA by the polymerase molecule. Each incorporated nucleotide base emits a different light that is detected within the ZMWs to generate a contiguous base sequence. Each ZMW containing a polymerase-DNA complex yields a read containing adapter and insert DNA sequence. These long reads known as polymerase reads are split into sub-reads that only contain the DNA insert sequence.

The PacBio sequencing of the two BAC clones yielded over 75,000 polymerase reads per BAC clone, as shown in Table 11. The BAC clone 263M01 was sequenced with the version 1.3.3 SMRT cell. Therefore, 263M01 yielded half as many polymerase reads as 422J05, which was sequenced with the version 2.0.2 SMRT cell. As part of the HGAP assembly process the polymerase reads were filtered, removing reads less than 50 bp long and with a quality of less than 0.75. The PacBio sequencing process yielded reads averaging between 4.2 and 4.7 kb, relatively long compared to Illumina (100-250 bp) and 454 (approximately 500 bp). However, the base quality of the reads was notably lower at 0.2-0.3 pre-filtered and 0.85 post-filter; compared to 0.999 capable from Illumina sequencing. Before assembly of the reads, vector sequence was screened using BLASR [22]. The subreads containing vector were blacklisted preventing vector sequence contamination in the final assembly.

PacBio reads were processed and assembled using HGAP [30], available within the smrtportal analysis pipeline. The pre-assembly process generates long seed reads with an average length over 6.1 kb, compared to the 4.2 kb average subread length, Table 12. The longest reads from the SMRT sequencing are utilised as pre-assembly seed reads. The sub-reads are subsequently aligned to the pre-assembly reads sequence using BLASR, correcting sequencing errors. These seed reads are then assembled using the Celera assembler [100].

By using HGAP, both BAC clones were successfully assembled from pre-assembly reads into one complete contig per BAC clone. The sub-reads were re-mapped again onto the assembled consensus sequences, correcting any errors

BAC clone	Pre-filter reads				Post-filter			
	Number	Quality	Av. Length (bp)	Bases	Number	Quality	Av. Length (bp)	Bases
422J05	150,292	0.223	1,086	163,214,084	32,344	0.853	4,240	137,123,579
263M01	75,153	0.316	1,881	141,349,264	18,636	0.853	4,728	88,107,160

Table 11: Table of PacBio sequence details. Pre-filtered reads are the raw polymerase reads produced from the sequencing process. Post-filtered reads have been filtered for reads length and quality. Removing reads with a length shorter than 50 bp and a quality score below 0.75.

BAC Name	Pre-assembled reads				Subreads mapping					Contig Length
	Yield	Number	Av. Length	N50	Number	Av. Length	Bases	Accuracy	Coverage	
422J05	0.403	3,472	6,193	7,518	36,293	3,406	123,621,426	87.15	479.5	249,849
263M01	0.377	2,673	6,357	7,752	19,744	4,031	79,588,205	84.57	378.24	198,520

Table 12: Table of PacBio HGAP assembly details. Lengths are in base pairs (bp). Pre-assembled reads are the longest reads generated from post-filtered PacBio reads. Subreads mapping are all of the post-filtered reads aligned to the pre-assembled reads to improve accuracy. Each BAC clone was assembled to a single contig. Accuracy is the average accuracy of the aligned reads. Av. Length is the average lengths of the reads in base pairs. N50 is a length value (bp) whereby 50% of the assembled sequence is within contigs equal or larger than the N50 size. Yield is a proportion of reads used for pre-assembly.

in the pre-assembled read sequences. This resulted in two highly accurate BAC clone consensus sequences of approximately 250 kb and 200 kb for 422J05 and 263M01 respectively, Table 12. The sub-read mapping coverage, 479 reads and 378 reads for 422J05 and 263M03 respectively, is high enough to overwhelm the inherent PacBio error rate.

To confirm the accuracy of the consensus sequences, BAC clones 422J05 and 263M01 were sequenced using the Illumina HiSeq 2000 platform. These high quality reads were aligned to the consensus sequence to check for sequence and structural errors using the same methods described in section 2.2.7. The BAC clone consensus sequences were verified as accurate and no structural recombinations were identified.

The consensus sequences for each BAC clone were aligned together manually to generate a complete assembly consensus sequence. The two BAC clones have an overlap of 151,910 bp, Figure 24. There were no differences between the BAC clones; therefore, they almost certainly belong to the same haplotype. The PabBio sequence and assembly process resulted in a complete haplotype sequence of 253,918 bp that could then be characterised and annotated.

4.3.2 Characterisation of the assembled BAC clones revealed an expanded sheep *KIR* gene haplotype

The assembled haplotype sequence was characterised using a combination of BLAT and manual sequence searches. This revealed a total of 14 *KIR* loci spread over 197,340 bp, Figure 24. The *KIR* are flanked by *FCAR* and *NCR1* at the 3′ end of the assembly which is syntenic to other mammalian LRC regions. The assembly does not contain any flanking *LILR* at the 5′ end. Therefore, this haplotype could contain more *KIR* that have not been sequenced in these two BAC clones and this haplotype is considered incomplete.

The start and end positions for each sheep *KIR* gene sequence were determined based on the blat searches, known *KIR* gene structures, splice donor/acceptor sites and stop codon positions, Table 13. The full sheep *KIR* gene sequences were extracted and aligned along with the cattle *KIR* sequences to determine the exact intron/exon boundaries. The exon positions for each sheep *KIR* gene was determined based on this alignment, the exons sequenced for each gene were extracted to provide the predicted mRNA coding sequences. Translation of predicted mRNA sequence from the sheep *KIR* exons enabled prediction of functionally intact genes by searching for premature stop codons within the Ig, stem and transmembrane domains, Table 13. There are a total of eight intact sheep *KIR* genes and six pseudogenes or null-alleles based on the open reading

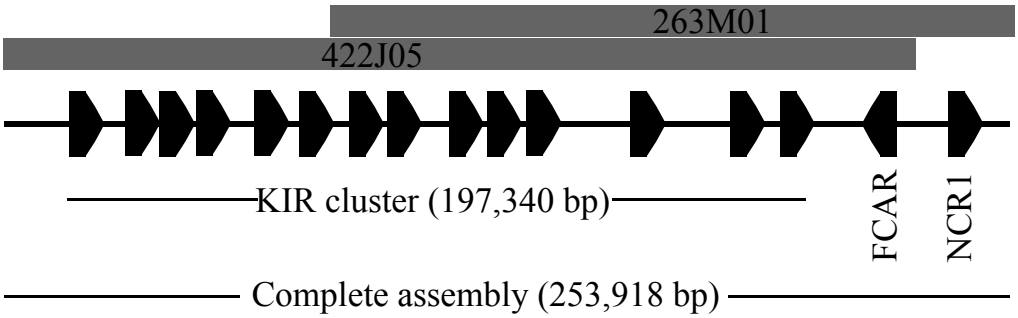


Figure 24: Overview of BAC positions used in the assembled haplotype sequence. Genes are represented as unequal pentagons, orientation of the gene is indicated by the direction of the pentagon. BAC clones are shown as long rectangles with BAC clones names shown within the rectangles. The *KIR* loci positions have been identified within a complex of almost 200 kb.

frames extracted from the assemble sheep *KIR* haplotype sequence, Table 13. Sheep *KIR* 01 and *KIR* 08 sequences contain signal peptides and two Ig domains but do not contain a stem, transmembrane domain or cytoplasmic tail. Sheep *KIR* 04 sequence only contains a single Ig domain but does not encode signal peptides. Sheep *KIR* 02 contains four Ig domains that appear to be fully intact and encode an open reading frame from signal peptide the stop codon at the end of the cytoplasmic tail.

4.3.3 Sheep have independently expanded L and X lineage genes

To indicate the lineages of the sheep *KIR* genes and their relationships to the cattle *KIR* groups, phylogenetic analysis of the exon 3 sequences was used. Sheep *KIR* sequences from the first assembled haplotype were extracted then aligned with cattle and a selection of other mammalian *KIR* sequences. A phylogentic tree was constructed using the D0 domain (exon 3) to distinguish the relationship of the sheep *KIR* sequences to the other mammalian *KIR*. The sheep *KIR* sequences are most similar to the cattle *KIR* sequences within both the X and L-lineages. However, none of the individual sheep *KIR* are more related to a single cattle *KIR* than they are to another sheep *KIR* meaning there are no obvious orthologous *KIR* between the two species, Figure 25. To aid comprehension, gene names have been retrospectively renamed within Figure 25 based on relationships with the cattle *KIR* gene sequences, Table 14.

The sheep *KIR* sequences clade together to form groups, Figure 26. All of the sheep *KIR* genes are related to cattle *KIR* groups, meaning they expanded from the same original genes. X-lineage *KIR* expansion in sheep has undergone an alternative route to cattle. Both sheep and cattle have expanded the group IV *KIR* to a similar extent, three loci in cattle and at least four loci in sheep. The sheep have expanded *KIR* related to the *Bota2DXP1* pseudogene, which is a single locus in cattle but has expanded to at least four loci in sheep. Conversely cattle have expanded the Group I, III and IV *KIR*, which share a common ancestral gene with the group VII sheep *KIR*. This group has only expanded to two loci in sheep compared to the eight loci in cattle. Therefore sheep and cattle have undergone species specific *KIR* expansion as well as combined expansion of the group IV genes. This suggests an important pan-species role for the group IV *KIR* that could have been driven by the impact of a related ligand or pathogen between cattle and sheep.

KIR	start pos	Gene (bp)	mRNA (bp)	Full ORF	Exon 01	Exon 02	Exon 03	Exon 04	Exon 05	Exon 06	Exon 07	Exon 08	Exon 09	Exon 10
01	12334	2832	641	No	34	26	281	300						
02	22764	11392	1633	Yes	35	36	279	300	300	300	51	123	53	156
03	36177	7111	1090	No	35	35	280	302	51	118	53	216		
04	49123	3308	726	No	300	50	121	52	203					
05	60051	7144	1318	Yes	35	36	279	300	303	51	123	53	138	
06	69247	5902	846	No	36	35	342	260	51	122				
07	85965	7316	1333	Yes	35	36	273	300	300	51	126	56	156	
08	101687	2828	638	No	35	36	267	300						
09	114400	7142	1352	Yes	35	36	279	300	300	51	126	51	174	
10	123560	5903	737	Yes	36	35	209	283	51	123				
11	138671	9391	1294	Yes	35	36	279	300	300	51	126	50	117	
12	158035	15790	1340	No	35	36	278	300	180	124	51	123	56	157
13	188705	6067	1112	Yes	35	36	267	300	300	51	123			
14	204466	7090	1325	Yes	35	36	267	300	300	51	126	53	157	

Table 13: Table of sheep gene positions. *KIR* are numbered from 5′ to 3′ from 01 to 14. The start position is the left most base from the start codon of the first exon which is the signal peptide for all except KIR 04. The gene column represents the total length of the gene in base pairs (bp) including introns. The mRNA column is the predicted mRNA sequence length based on blat alignments and splice donor acceptor sites. The Full ORF column indicates whether the open reading frame is intact for the full length of the predicted mRNA sequence, "No" represents a stop codon before the end of the gene sequence. The length in bp for each exons is shown for each gene.

KIR	Gene name	allele number	Group
01	2DXP2	01	VI
02	4DXL1	01	IV
03	2DS3	01*N	II
04	1DLP1	01	II
05	3DXL4	01	VI
06	2DS2	01*N	II
07	3DXL2	01	VI
08	2DXP1	01	IV
09	3DXL5	01	IV
10	2DS1	01	II
11	3DXL3	01	IV
12	3DXP1	01	VI
13	3DXS1	01	VII
14	3DXL1	01	VII

Table 14: Table of sheep gene names assigned after characterisation. *KIR* names were assigned based on cattle *KIR* nomenclature discussed previously. The groups were assigned based on the relationships to cattle *KIR* gene groups. Sheep *KIR* gene names in Figure 25 were retrospectively renamed and coloured based on this table.

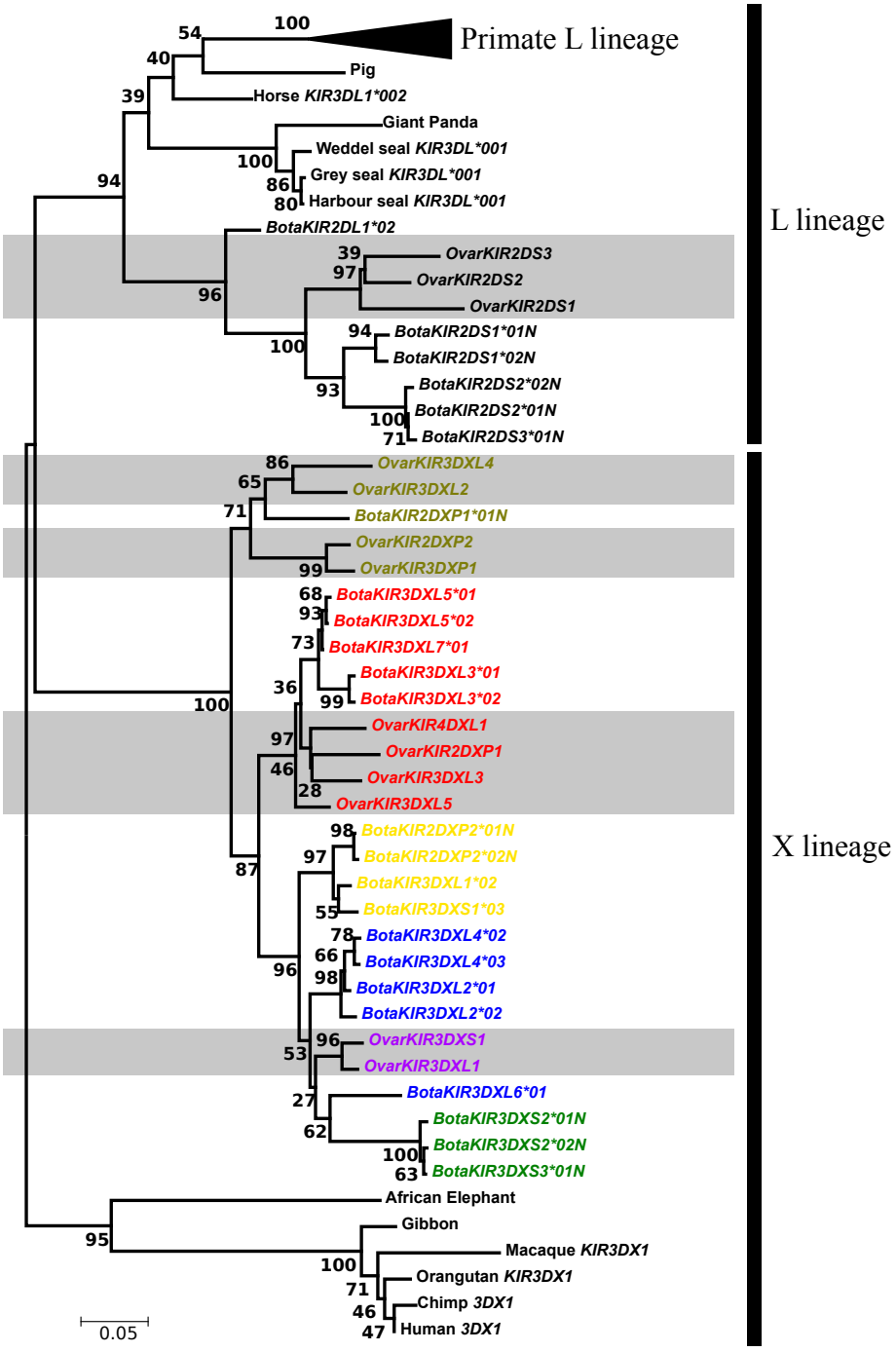


Figure 25: Neighbour-joining phylogenetic tree of selected mammalian species *KIR* genes, using only the exon 3 sequence. 500 bootstrap replicates and Tamura-Nei algorithm was used. Gene groups have been colour coded based on relation to cattle *KIR* groups and novel sheep *KIR* groups. The sheep *KIR* sequences have grey box backgrounds. X and L lineages have been annotated. Sheep *KIR* names have been assigned based on further characterisation of predicted functionality.

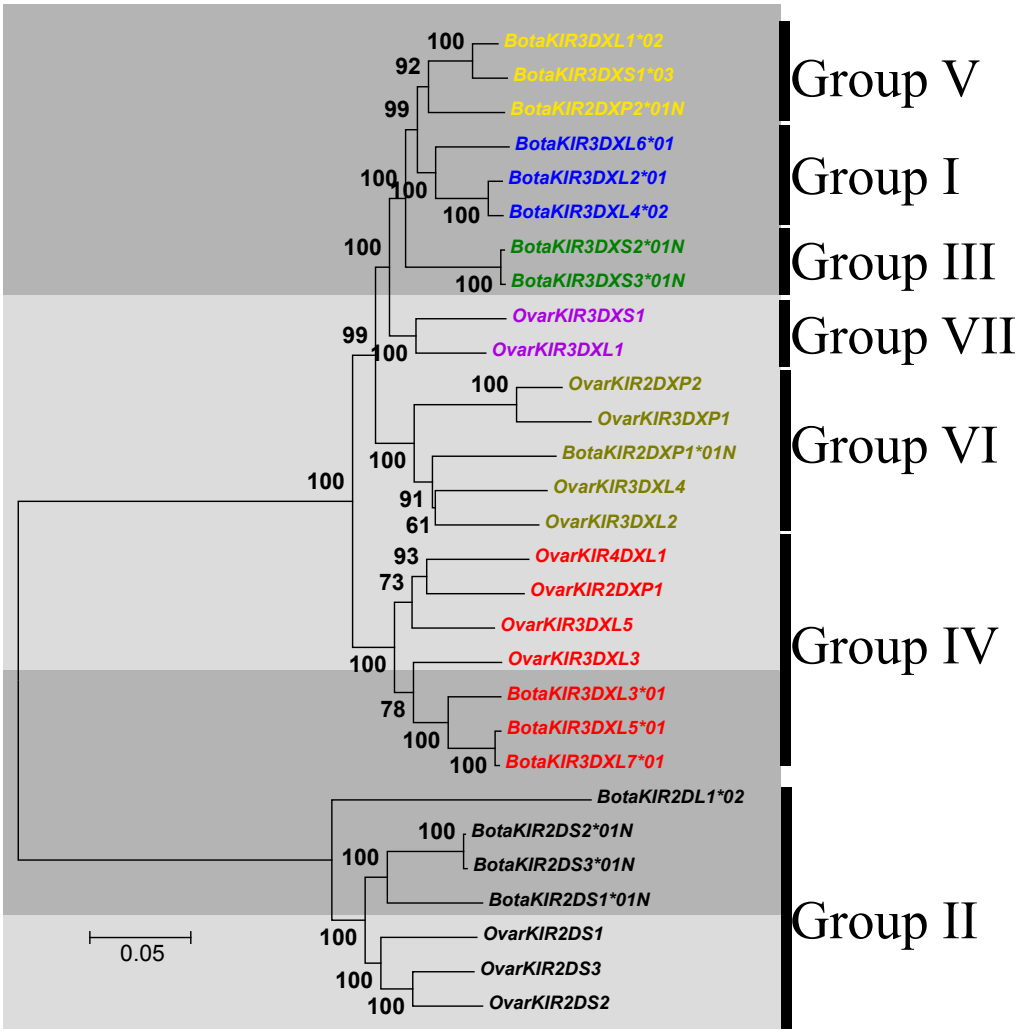


Figure 26: Neighbour-joining phylogenetic tree of selected mammalian species *KIR* genes, using only the extracellular Domains (signal to stem) with intronic sequence. 500 bootstrap replicates and p-distance algorithm was used. Gene groups have been colour coded retrospectively and sheep *KIR* genes have light grey box backgrounds. Gene groups have been annotated.

4.3.4 Sheep *KIR* domain order is consistent with cattle *KIR* genes

Sheep *KIR* Ig domain sequences were individually extracted and aligned with the cattle *KIR* Ig domain sequences. A neighbour joining phylogenetic tree was constructed from the resulting alignment. Sheep *KIR*, like cattle *KIR*, have maintained the Ig domain order, D0-D1-D2, for three Ig KIR, and D0-D2 for two domain KIR, Figure 27. Except *KIR2DL4*, this is different to the human two domain KIR that encode the D1-D2 form of the receptor

One of the Sheep *KIR* sequences contains a four Ig-domain gene, which belongs to the group IV. This gene, named *BotaKIR4DXL1*, has the domain structure D0-D1-D1-D2. The second D1 domain has originated from a group VI gene. This domain has been inserted via an unknown recombination event. The result is a four Ig-domain long tail gene with intact coding sequence. This is the first X-lineage gene with four domains to be characterised.

4.3.5 The *KIR* activating tail sequence evolved before *Bovinae* speciation

The evolution of activating *KIR* transmembrane sequences occurs from a number of mechanisms including one or all of the following: Inactivation of the ITIM motifs by substitution of the tyrosine residue, stop codon introduction at the end of the transmembrane domain and the introduction of a charged residue within the transmebrane domain. This may occur only once, with the activating gene sequence propagating within the immune complex by gene recombination. To determine whether the sheep and cattle *KIR* activating sequence evolved multiple times, or once and propagated through recombination, the signalling regions of the genes were compared using phylogenetics. The transmembrane domain sequence of the activating genes is conserved resulting in low divergence between the genes, Figure 28. The node support between the short tail sequences is low and there is no segregation between species or between L and X lineages. The inhibitory tail regions show greater concordance between segregating sequences and group affiliations. However, the inhibitory genes have greater sequence length including the cytoplasmic domains to contribute to the phylogeny. Therefore, as the activating tails only include the transmembrane domain, less sequence is available for phylogeny construction. This could contribute to the low node support seen between the activating genes.

The sheep short tailed *KIR* genes contain a transmembrane domain arginine residue in the same position as the cattle short tail *KIR* genes (Figure 29). The sheep group II L-lineage gene, *OvarKIR2DS3*, contains a lysine residue at

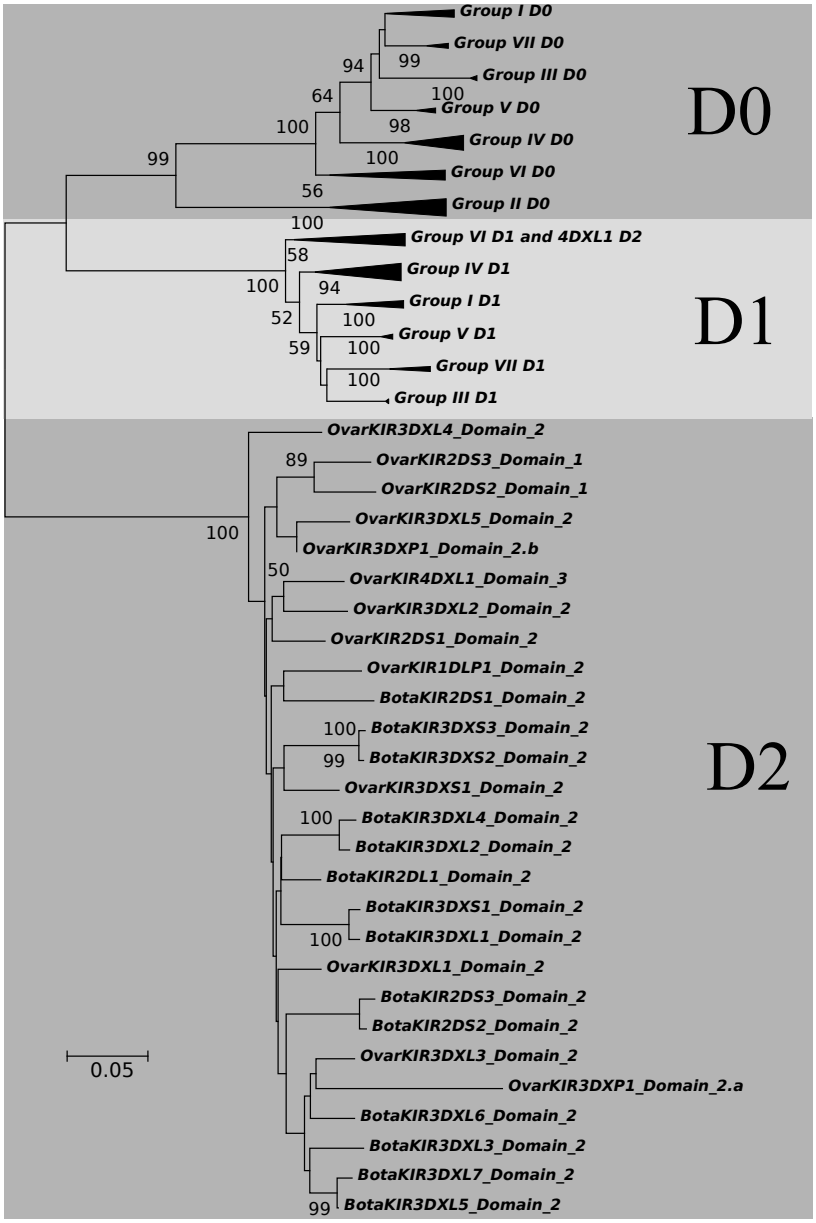


Figure 27: Neighbour-joining phylogenetic tree of Ig domain exon sequences from sheep (*OvarKIR*) and cattle (*BotaKIR*). Groups that clade together have had nodes collapsed to reduce the visual complexity. Gene group nodes that have been collapsed are represented by triangles. Node support scores have been removed if less than 50%. The groups of sequences have been shaded based on D0, D1 or D2 groups.

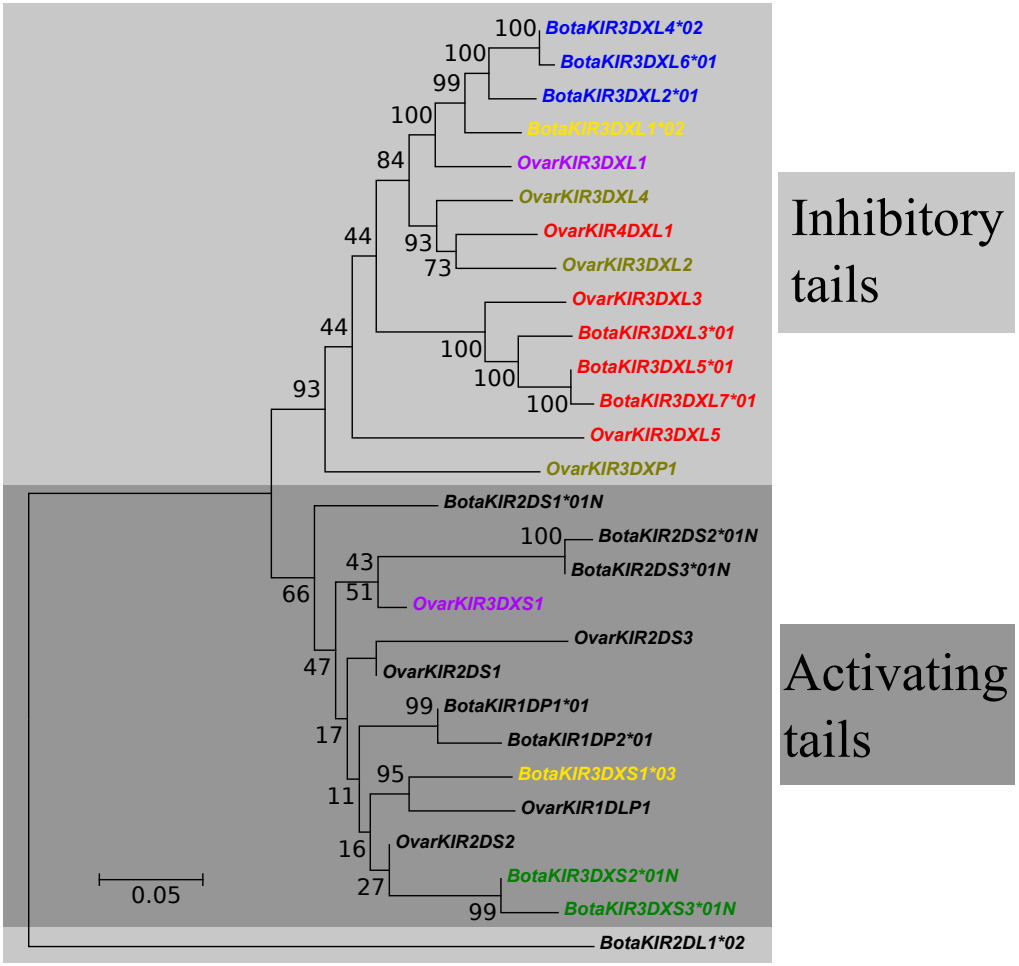


Figure 28: Neighbour joining phylogenetic tree of tail region sequences from cattle and sheep *KIR*. Uses the p-distance algorithm and 500 bootstrap replicates. Based on genomic DNA from the transmembrane domain to the end of the cytoplasmic domain. Genes have been colour coded based on gene groups and backgrounds have been shaded based on activating and inhibitory signalling.

the same position as the arginine residues in the other short tail genes. This is the only lysine encoding *KIR* transmembrane domain characterised within a ruminant species. Lysine is the active residue encoded within short tail human L-lineage *KIR*, however, the human *KIR* lysine is in a different location to the *OvarKIR2DS3*. Therefore, the *OvarKIR2DS3* lysine has evolved independently to the human *KIR* lysine, however this gene is a null-allele within the haplotype sequence here.

The sheep long tail *KIR* genes encode two functional ITIM motifs. With the exception of *OvarKIR3DXL3* which terminated before the second ITIM, the sheep genes encode ITIMs in the same relative positions as the cattle *KIR* genes (Figure 30). The ITIM motifs maintain the same canonical sequence as cattle with VxYxxL for the first ITIM and IxYxxF for the second. There is minor variation in *OvarKIR4DXL1* in ITIM 1, and *OvarKIR3DXP1* and *OvarKIR3DXL5* in ITIM 2. These variations should not impact the signalling of the genes as the residue substitutions have similar biochemical properties. The variation within *OvarKIR3DXLP1* may have affected signalling from ITIM 2 by the introduction of the polar amino acid arginine which has been shown to affect binding of SHP-1/2 and reduce inhibition [128]. The similarities between the cattle and sheep ITIM and transmembrane motifs suggests they signal through the same adapter molecules. The rate of evolution within the adapter molecules is likely to be lower as they are constrained by other signalling receptors. This constraint also constrains the *KIR* signalling domains which remain conserved between species so they can interact with the signalling adapter molecules.

4.3.6 Sheep *FCAR* gene is inverted compared to cattle

The sheep LRC contains an inverted *FCAR* gene when compared to cattle and all other mammals, the coding orientation of the gene is on the negative strand compared to the *KIR* genes on the positve DNA strand, Figure 31. The inversion involves sequence until the end of the most 3′ *KIR* gene. The other flanking gene, *NCR1*, is in the same coding orientation as cattle *KIR*, thus demonstrating the inversion is limited to the *FCAR* gene.

4.3.7 Summary of the Sheep *KIR* haplotype structure

The *KIR* gene sequences were characterised for features including long or short tail, number of domains, intact coding sequence and X or L lineage. The genes were assigned names bases upon these characteristics and have been shown in Figure 32. The sheep LRC contains at least fourteen discrete *KIR* loci, and

Figure 29: Amino acid alignment of sheep and cattle transmembrane domain. Signalling residues have shaded grey backgrounds. The consensus sequence is based on a 50% threshold. Dots represent same as the consensus and white space represents no sequence.

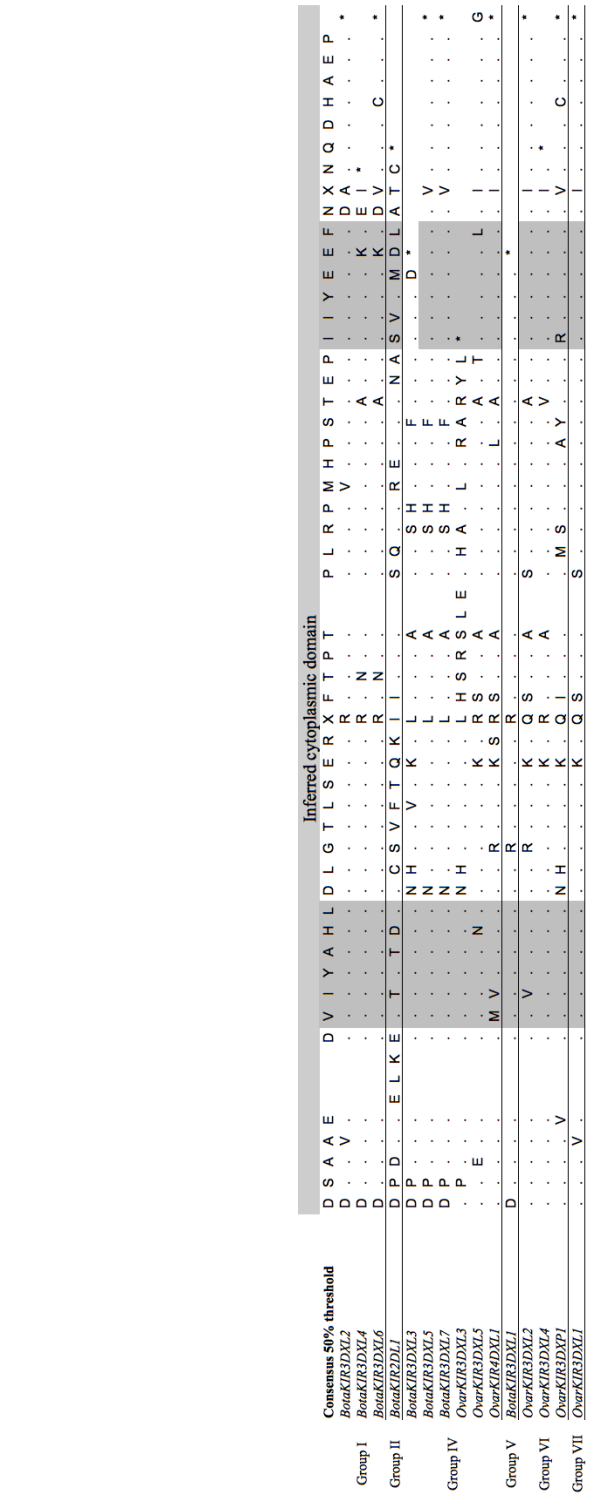


Figure 30: Amino acid alignment of sheep and cattle cytoplasmic domains. ITIM motifs have shaded grey backgrounds. The consensus sequence is based on a 50% threshold. Dots represent same as the consensus and white space represents no sequence.

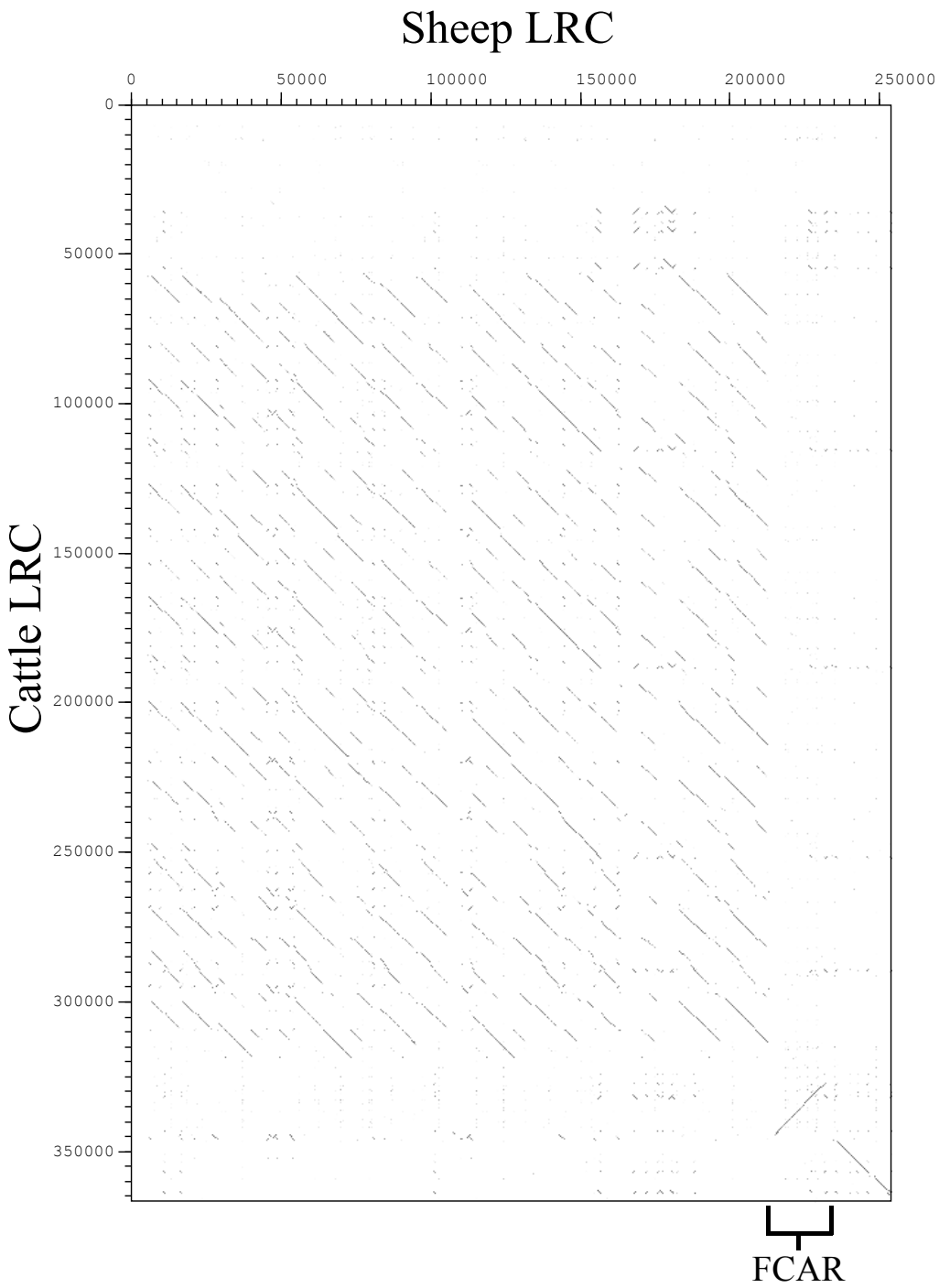


Figure 31: 150 bp dot plot of cattle LRC against the sheep LRC. Dots represent sequence identity over 150 bp between cattle and sheep *KIR* complexes. Lines are contiguous dots representative of larger regions of sequence identity. The position of the inverted *FCAR* gene is annotated on the diagram. Distances on X and Y axis are in base pairs (bp).

singular *FCAR* and *NCR1* genes, Figures 24 and 32. There are eight predicted functional *KIR* loci, four pseudogenes and two potential null-alleles. The pseudogenes were determined to be completely non-functional due to large sections of missing sequence. Of the fourteen *KIR* loci, ten are from the X-lineage and four are of L-lineage origin. All L-lineage loci are short tailed including a single predicted functional gene, *OvarKIR2DS1*01*.

4.3.8 Sheep genome LRC is partially correct but poorly annotated

The sheep genome project [5], although not producing a complete genome yet, have assembled a partial LRC within the Oar_v3.1 build and annotation release 100. The sequence for the LRC was extracted, reverse complemented, and compared against the BAC clone assembly using a 150 bp dot plot, Figure 33. The dot plot shows a region of high sequence identity at the 3′ end from approximately 110 kb to the end of the BAC clone assembly. This region is interrupted by breaks in the continuous diagonal line, which represents areas of no sequence similarity. These are a result of scaffolds within the genome assembly. Scaffolds occur where contigs have been joined utilising paired-end read placement information, however the sequence between the contigs is unknown and replaced with Ns. The sequence to the 5′ of the 110 kb position shows no continuous sequence similarity between the BAC clone assembly and the genome build. Therefore, this portion of the assembly differs between the BAC assembled sequence and the genome build. This is despite the sheep genome project using the same BAC library used here.

The sheep genome build LRC shows high sequence identity to the 3′ half of the BAC assembly. This is from the *3DXL5* gene to the end of the BAC assembly, Figure 34. The highly similar half of the alignment spans all the *KIR* genes to the 3′ of *3DXL5*. This region is highly similar with the majority of the sequence having 0.95 to 1.0 identity scores. This is interspersed by occasional drops in identity, which is a result of short (1-50 bp) indels and scaffold sequence Ns reducing the identity score. This 3′ half of the genome build corresponds to the BAC assembly, however there is very little similarity within the 5′ half.

The 5′ half of the BAC assembly, encompassing *2DXP2* to *2DXP1* has a relatively lower identity score compared to the genome build LRC. There are three discrete peaks of higher sequence identity score in the 5′ half of the complex, Figure 34. However, they are likely *KIR* sequences from the genome build that have aligned and are not exact matches. The identity scores never exceed 0.92 and therefore are not considered allelic or in the same gene group. To delineate the extent of the similarities and difference between the two builds, the genes

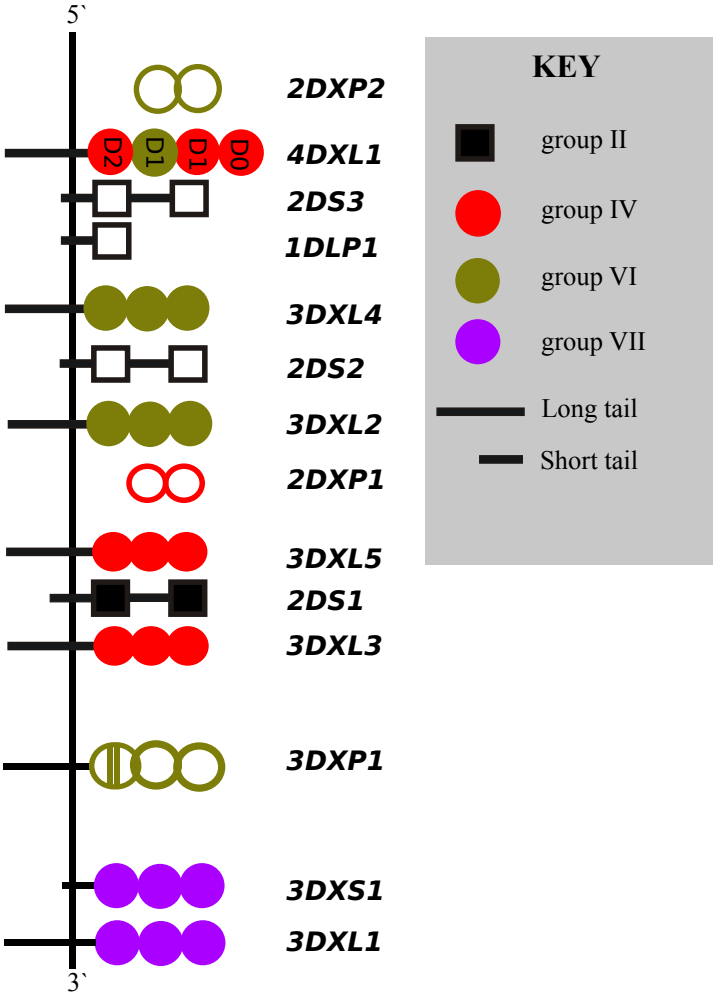


Figure 32: Overview of gene positions within the assembled sheep haplotype. The *KIR* complex is 197,340 bp in length. The figure shows X-lineage Ig-domains as circles and L-lineage Ig-domains as squares. Functional genes are full coloured squares or circles and non-functional genes are just borders. Long and short tails are shown as long and short lines. Broken domain is representative of an insertion within the domain. Where domains are shown and no tail, sequence is missing for the transmembrane and cytoplasmic exons. Colours represent gene groups based on phylogenetic sequence comparison with cattle *KIR*.

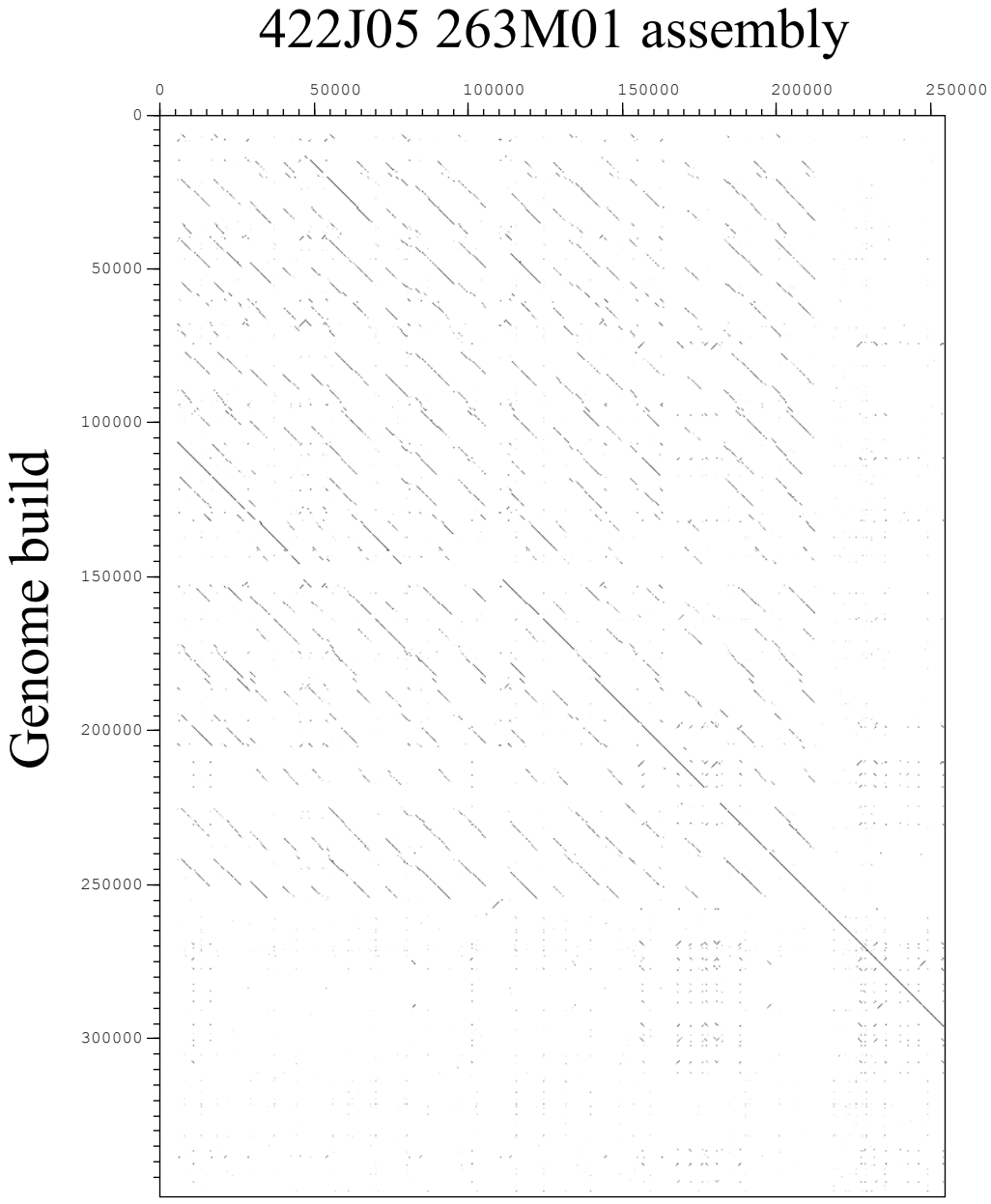


Figure 33: 150 bp dot plot of the two LRC assemblies. Using the sheep BAC clone assembly consensus sequence against the sheep genome build. Dots represent sequence identity over 150 bp between the two *KIR* haplotypes. Lines are contiguous dots representative of larger regions of sequence identity.

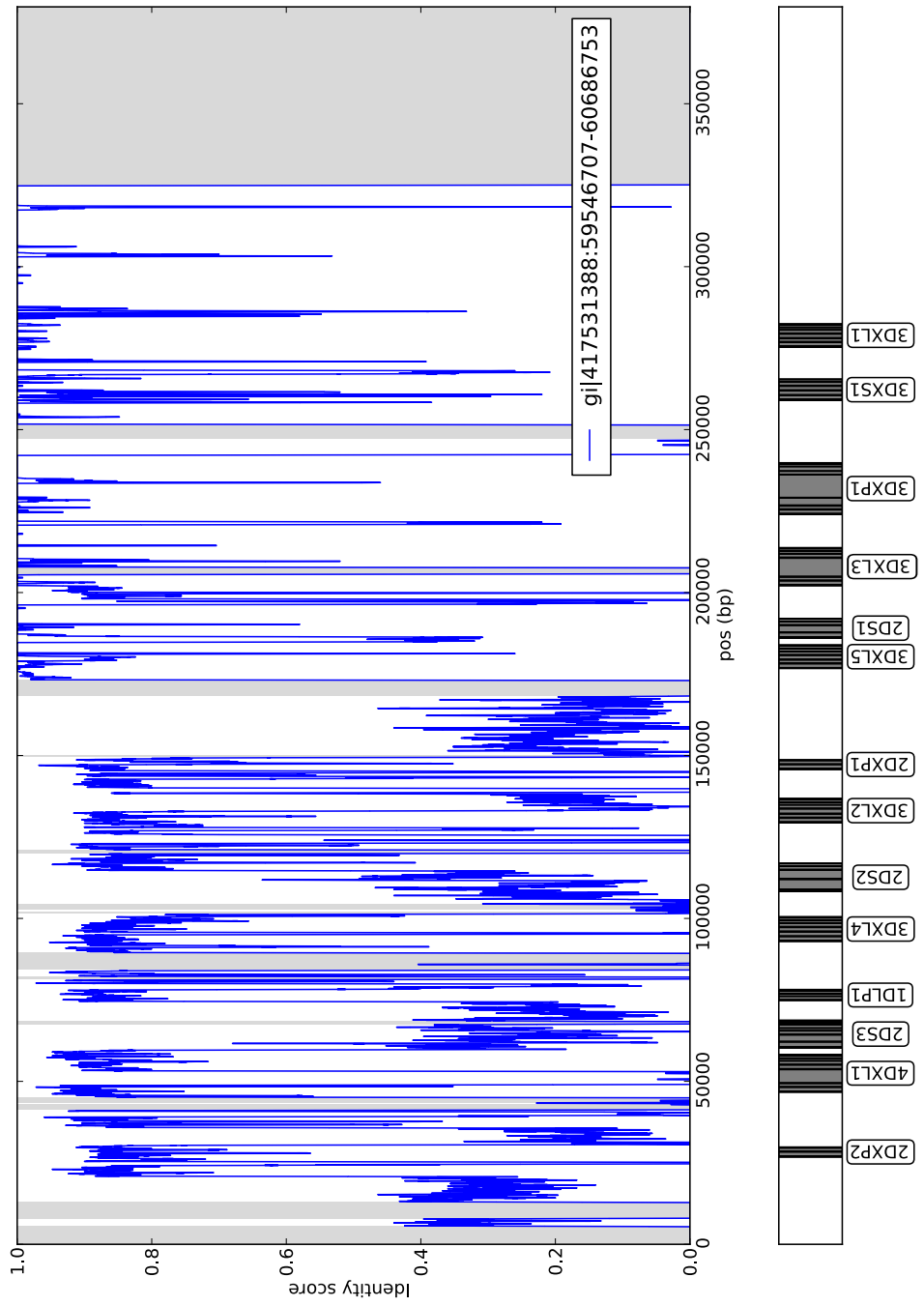


Figure 34: 500 bp sliding window analysis of the sheep genome build (blue line) shown against the BAC clone assembly consensus sequence. The blue line in the figure shows average sequence similarity of a sliding window in the genome build LRC when compared to the BAC assembly, which is the reference sequence. Vertical light grey shading represents unique sequence in the genome assembly. The track at the bottom represents the genes within the reference BAC assembly sequences. Grey rectangles are the genes and the black lines represent exons.

were compared.

The annotation of the genome was largely inaccurate due to the automated GNOMON process. Instead, the raw genome build was annotated using the same methods employed to annotate the BAC assembly, Figure 35. Three genes from the genome annotation were each composed of two individual genes, shown in the custom annotation as long genes. The custom annotated *KIR* were labelled KIR one to thirteen based on their position, from the 5′ to the 3′ of the reverse complemented genome assembly build LRC.

The sheep genome *KIR* sequences from the custom annotation were extracted and aligned with the BAC assembled *KIR* genes, revealing that sheep genome *KIR* 9 corresponded to *OvarKIR3DXL5* and sheep genome *KIR* 13 corresponded to *OvarKIR3DXL1*, Figure 36. Between these, genes correspond as expected from the dot plot and sliding window analysis. Except for *BotaKIR2DS1*01*, which was not represented within the genome sequence. Figures 33 and 34 suggest a break at this position in the genome, omitting the *OvarKIR2DS1* gene. The rest of the 3′ genome genes correspond to the BAC assembly in the same order. Therefore, the 3′ half of the two haplotypes are highly identical at the gene level.

Of the eight *KIR* genes from the 5′ half, five show no identity to any of the BAC assembly *KIR* genes. Sheep genome *KIR* 1, sheep genome *KIR* 6 and sheep genome *KIR* 8 correspond to *OvarKIR3DXL4*, *OvarKIR2DXP2* and *OvarKIR2DS2* respectively, Figure 36. The synteny between the two different haplotype assemblies has been summarised in Figure 37. The three genes within the 5′ end that correspond between the assemblies are not in the same order. This explains the three peaks with reduced (sub 0.95) sequence identity in the 5′ end of Figure 34; the three genes have aligned but not to syntenic genes, resulting in reduced identity scores. These genes may be the result of variable haplotype gene structure, or the effects of an incorrectly assembled haplotype.

4.3.9 Illumina sequenced BAC clones were aligned to the different *KIR* assemblies and confirmed alternate haplotype structures

Two further BAC clones, 179E01 and 127N14, were sequenced alongside the two assembled BAC clones, 263M01 and 422J05, on the Illumina HiSeq platform, Table 15. The reads were aligned to both the BAC assembly consensus sequence and the genome build, Figure 38. Each BAC clone read depth histogram is coloured so that individual BAC clone raw sequence alignments can be distinguished. The two clones used in the BAC assembly, 263M01 (green) and 422J05 (blue), resequenced and aligned back onto the assembly consensus sequence with

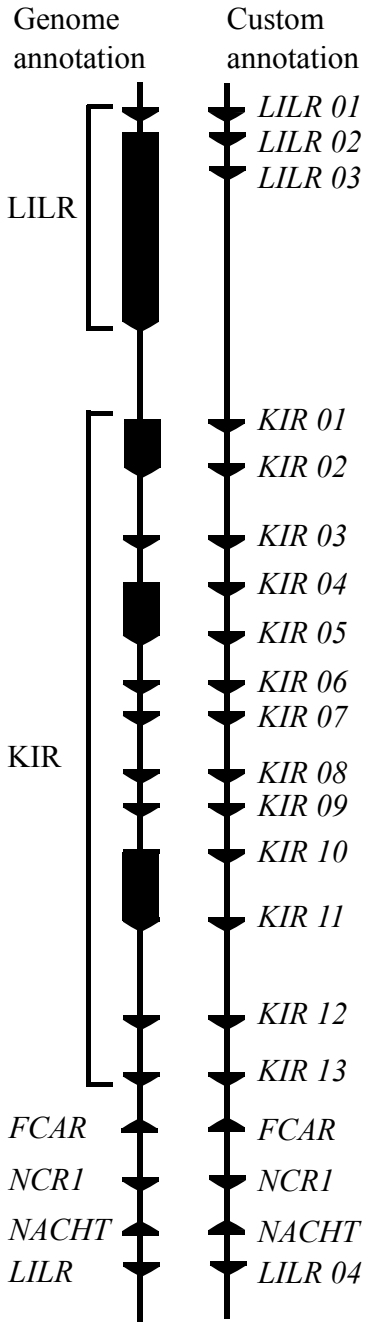


Figure 35: Diagrammatic representation of the sheep genome build. Both the genome annotation and our custom annotation are shown with uneven pentagons representing genes, elongated pentagons represent genes in the genome annotation that contain several genes. Custom annotation was done using a combination of blat search results and manual sequence searching.

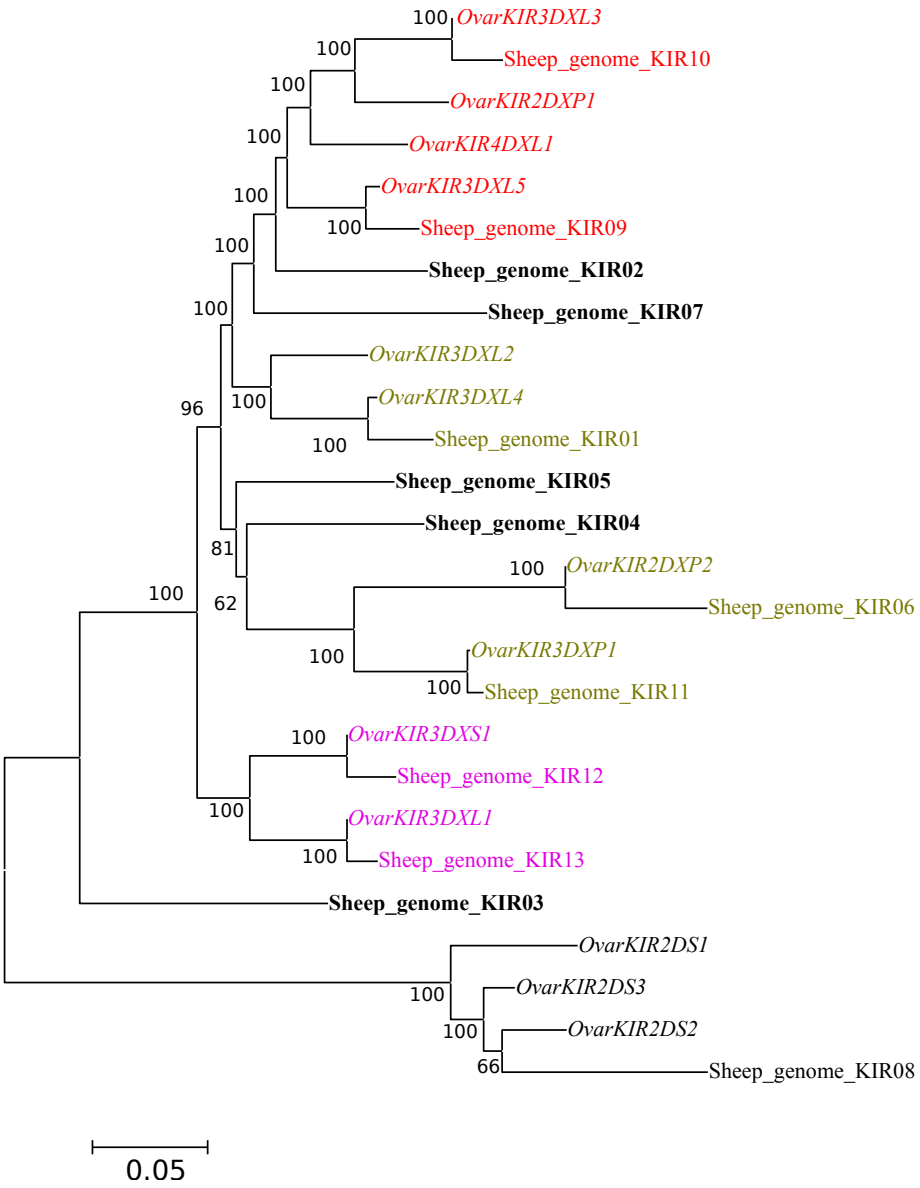


Figure 36: Neighbour joining phylogenetic tree of genome build *KIR* genes and the BAC assembly *KIR* genes. P-distance algorithm and 500 boostrap replicates used. Genome genes are from custom annotation and not the genes annotated by the genome project. Genes have been colour coded based on groups. The genes in bold typeset are not seen in the BAC assembly and therefore potentially different or new genes.

even and consistent coverage depth, Figure 38a. A further BAC clone, 179E01 (red), was aligned which maps to the 3′ end of the assembly.

These three BAC clones alongside 127N14 were also aligned to the genome assembly build, Figure 38b. The mapping of 422J05 to the genome build highlights the lack of corresponding sequence between the two assemblies. The 3′ of 422J05 is mapped even and consistently. However, the 5′ end of the 422J05 read coverage histogram is minimal, except for the *KIR* 1 position. From the results shown in section 4.3.8 this higher read coverage can be explained by the translocation of *BotaKIR3DXL4* in the sheep genome build, shown in Figure 37. As the reads from 422J05 do not map to the inconsistent 5′ region of the genome assembly, the 5′ end is the result of structural variation and not miss-assembly. If the reads from 422J05 had mapped to the 5′ region, then an alternate assembly could have been possible. This alternate assembly would have generated the genome build that differs to the BAC assembly. Instead it is predicted that the genome build has an alternate *KIR* complex structure.

BAC clone mapping confirms the genome build placement of a *LILR* gene at the 3′ of the LRC. The BAC clones 263M01 and 179E01, both map to the BAC assembly and the genome assembly with consistent and even read depth coverage. This is consistent with the findings from section 4.3.8. This suggests the 3′ end of the assembly is consistent between the two builds. The BAC clone 179E01 spans both the BAC assembly and the *LILR* gene within the genome build at the 3′ end of the LRC. This confirms that there is a *LILR* gene at the 3′ end of the LRC.

Further *LILR* genes are confirmed at the 5′ end of the genome build by BAC clone mapping. The BAC clone 127N14 (light blue) maps to the 5′ region of the genome build but does not overlap with any of the *KIR* containing BAC clones. This BAC clone contains *LILR* genes but is separated from the BAC assembly build. Therefore the 5′ end of the assembly cannot be confirmed as complete. The flanking region containing *LILR* is not connected to the BAC assembly, therefore, there may be more *KIR* genes between the BAC assembled region and the 127N14 BAC clone.

Mapping the BAC illumina reads to both *KIR* assemblies has revealed that the *KIR* complex is flanked by *LILR* genes on both sides. It has also confirmed that the genome *KIR* haplotype is the result of structural variation and not incorrect assembly.

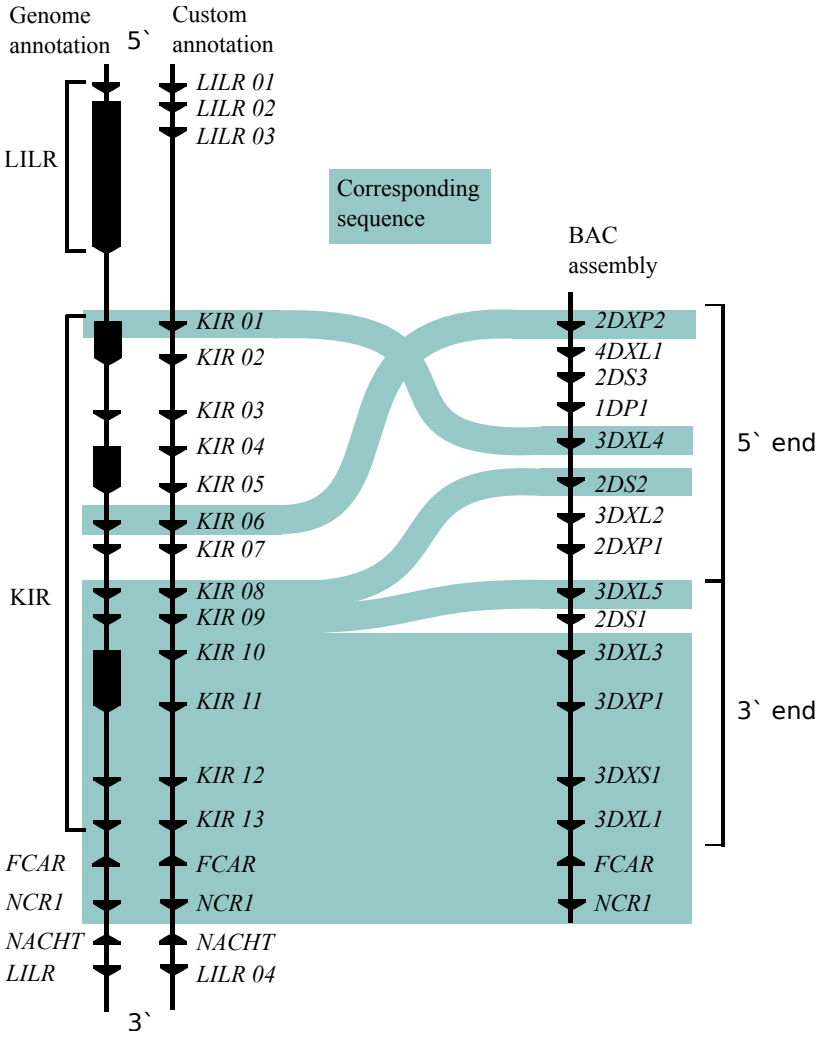


Figure 37: Diagram showing sheep genome *KIR* gene sequence similarity with the BAC assembled *KIR* gene sequences. Both the genome annotation and the custom annotation are shown with uneven pentagons representing genes, elongated pentagons represent genes in the genome annotation that contain several genes. Light blue lines represent corresponding sequence between the genome and the BAC assembled LRC

BAC clone	Insert size	Read 1	Read 2	RL (bp)	Million reads	Total bases
127N14	705	13,082,137	13,082,137	100	26.2	2,616,427,400
179E01	721	7,552,360	7,552,360	100	15.1	1,510,472,000
422J05	532	18,425,097	18,425,097	100	36.9	3,685,019,400
263M01	703	9,359,342	9,359,342	100	18.7	1,871,868,400

Table 15: Table of sheep BAC Illumina sequencing details. Insert size is the length of the fragmented DNA used for library prep and sequencing. Read 1 is the forward read, first to be sequenced and Read 2 is the reverse read, second to be sequenced. The average read length (RL) is 100 bp for all the samples from 2x100 bp sequencing with 200 cycles.

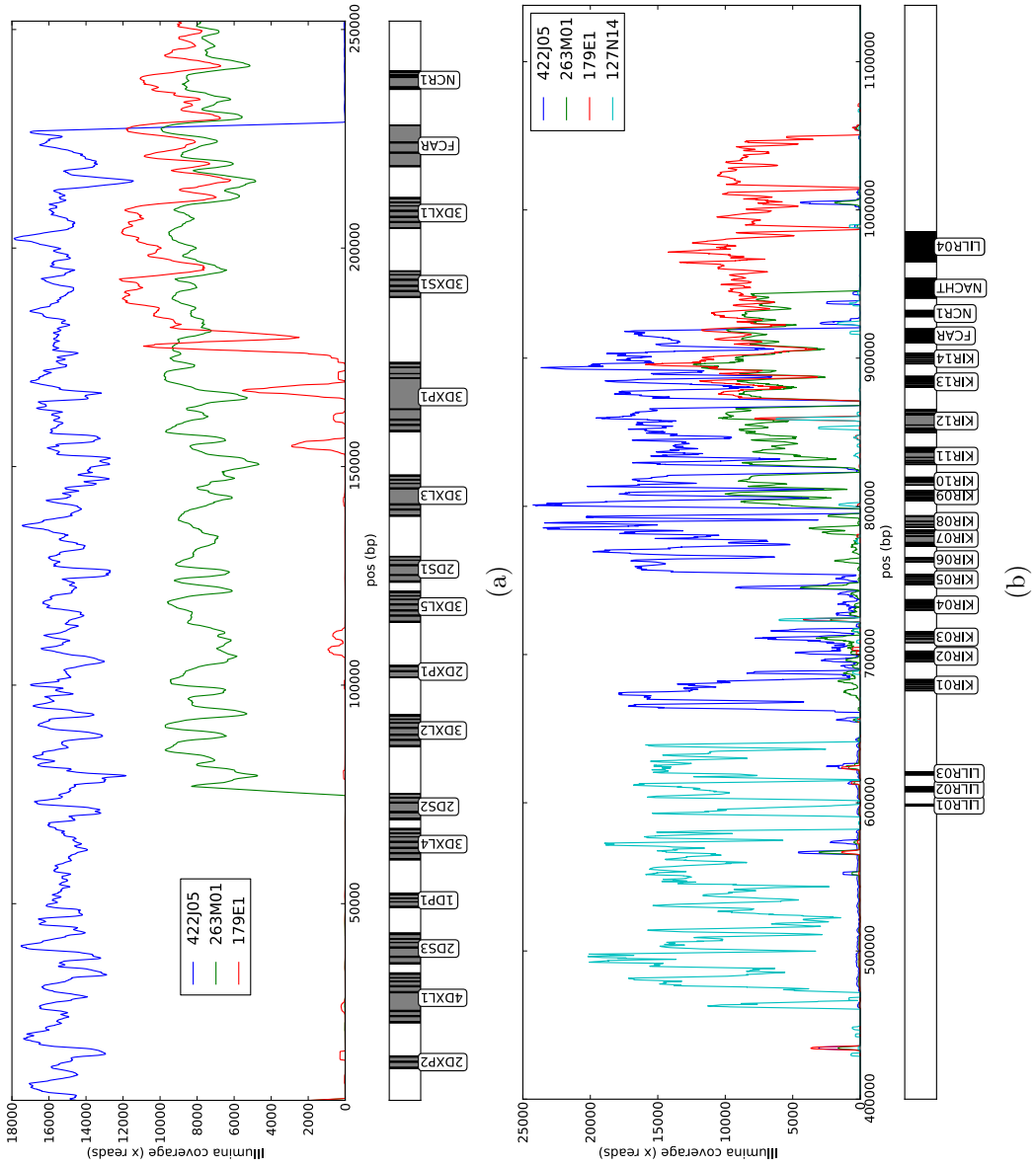


Figure 38: Read depth coverage histogram of sheep BAC clone Illumina reads mapped to the BAC assembly (a), and the genome assembly (b). Read depth is shown along the Y axis and is coloured depending on the BAC clone. Lines are smoothed using a mean value from a 1 kb sliding window. The X-axis starts at 40 kb in (b), longer sequence was used to ensure sequence mapping didn't occur adjacent to the *KIR* and *LILR* complexes. No mapping was observed and therefore not shown in this figure.

4.3.10 The last common ancestor of sheep and cattle likely contained at least five *KIR* genes

By comparing the common gene families between the sheep and cattle *KIR* haplotype, an ancestral *KIR* haplotype has been predicted. The ancestral haplotype, includes five genes, three X-lineage and two L-lineage, Figure 39. As the genes within the same groups between the two species are unrelated, the gene group expansion has likely occurred after the sheep-cattle speciation event. The sheep-cattle common ancestor haplotype must have contained, but is not limited to, five genes that have expanded and diversified independently within the sheep and cattle genomes.

The genes predicted to be in the sheep cattle common ancestor genome are, a group IV gene, which has expanded to create seven discrete loci within both genomes. A group VI gene, that has expanded within sheep to form four discrete loci yet has remained as a single disrupted loci in cattle. Two group II genes, a long tailed ancestor of *BotaKIR2DL1* that has been deleted from the sheep genome, and a short tail gene that has expanded to form nine separate loci in both genomes. This gene has also likely donated the activating tail domain sequence to other gene groups resulting in more activating genes. The group 0 gene in the common ancestor has expanded to form the group I, III, and V genes in cattle and the VII genes in sheep. This gene group is an ancestral gene group that has significantly diversified through cattle and sheep evolution. From this predicted ancestral *KIR* haplotype, it can be predicted that other ruminant species such as goats have expanded *KIR* similar to the *KIR* expanded in cattle and sheep.

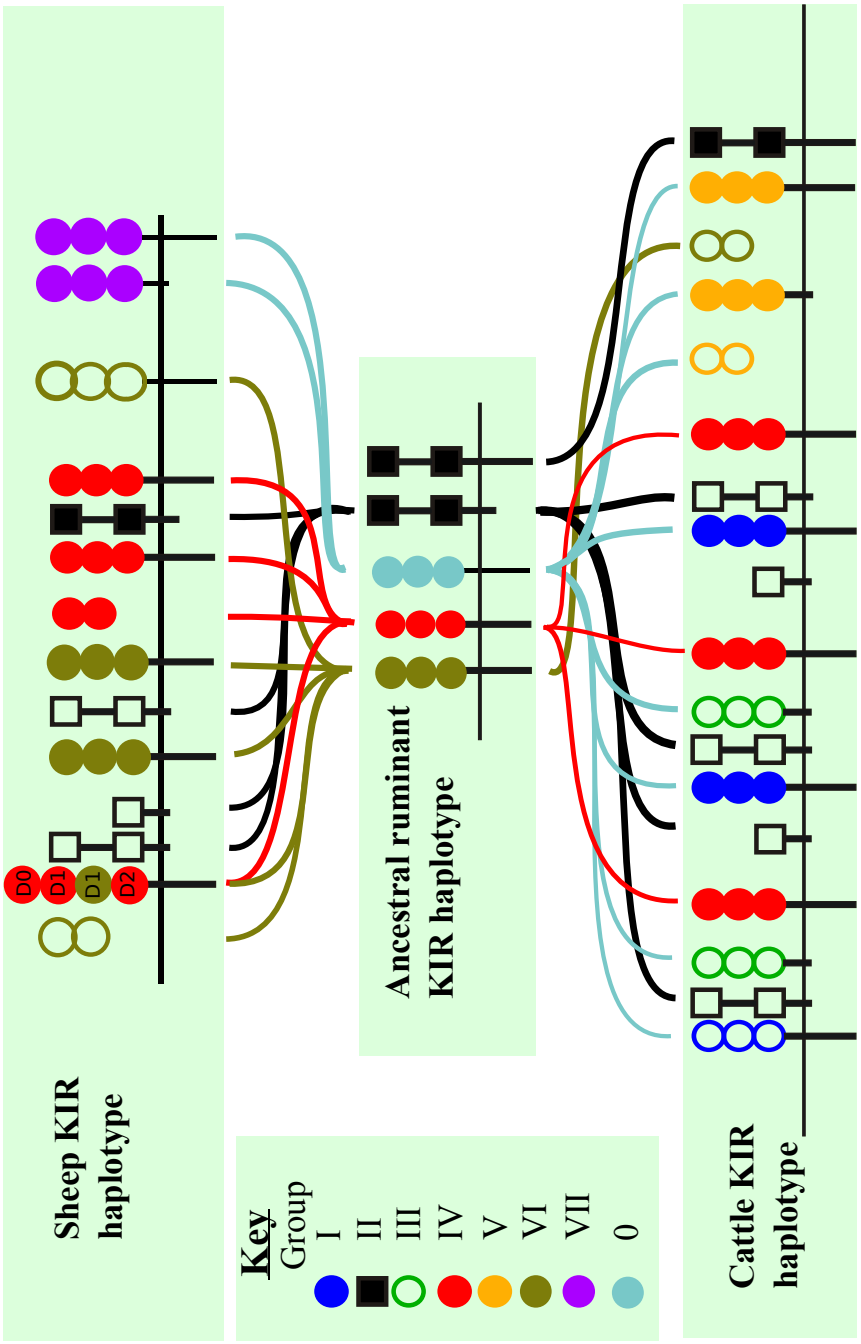


Figure 39: Diagram predicting ancestral ruminant haplotype. Lines represent origins from the predicted ancestral haplotype. Genes and lines have been colour coded based on group. The light blue colour represents the ancestral X group 0 from which groups I, III, V and VII have originated.

4.4 Discussion

A sheep *KIR* complex has been partially assembled from two overlapping BAC clones. The 3′ end of the complex is anchored to the flanking gene *FCAR* and corresponds with the genome build. The 5′ end of the haplotype is not anchored to the flanking *LILR* genes and it does not correspond to the genome build. Because the 5′ region does not anchor to any flanking genes, the haplotype is considered incomplete as further uncharacterised *KIR* may exist.

4.4.1 Sheep *KIR* genes have not undergone block duplication

Unlike the expansion of the cattle *KIR* haplotype, the sheep *KIR* complex has not evolved through block duplication. Although the mechanisms of sheep *KIR* expansion are unknown, the genes appear to have duplicated individually. This has resulted in localised pockets of gene groups. With the exception of *4DXL1*, *3DXP1* and the group II genes, the group III, VI and VII genes are localised to the middle, left and right of the haplotype respectively.

4.4.2 Sheep have expanded an ancient X lineage gene group

Sheep have expanded an ancient X-lineage gene group that has not expanded in cattle. The gene *BotaKIR2DXP1* is the only group VI gene in the cattle genome but has expanded to four loci in the sheep genome. This suggests that the functional group VI genes fulfil a niche function that is specific to the sheep, such as a sheep specific infection. Alternatively the group VI genes could be fulfilling the niche left by the limited expansion of the group 0 genes in the sheep genomes. In cattle the group 0 genes have expanded into the group I, III and V genes occupying seven loci. However, sheep have only expanded the group 0 genes into the group VII genes, occupying two loci.

4.4.3 There is no *Bota2DL1* orthologue within the sheep *KIR* complex

Despite the expansion of the short tail group II genes within the sheep genome, an orthologue of *BotaKIR2DL1* has not been found in the sheep genome. The most parsimonious explanation is that this gene was inherited from the ancestral ruminant *KIR* haplotype, but was lost after the inversion of the *FCAR* gene. In the cattle *KIR* haplotype, the *BotaKIR2DL1* gene is located adjacent to the *FCAR* gene. Therefore, assuming the same order was inherited within the sheep genome, it is reasonable to predict the recombination event that inverted the FCAR gene also deleted the long tailed group II gene.

4.4.4 The genome build may represent a second haplotype with structural variation and different gene content

The variation between the genome build and the BAC assembled *KIR* complex is the result of structural variation between haplotypes and not misassembly. The was indicated by alignment of raw illumina sequencing reads from BAC clone 422J05 to the genome build. The genome build could be the result of misassembly of these raw reads. However, the reads do not map and therefore could not be rearranged and misassembled to generate this genome build. I predict the genome build represents a second haplotype, albeit unfinished with scaffold sequence.

4.4.5 Conclusions from the sheep *KIR* haplotype

The sheep *KIR* haplotype has expanded from a five *KIR* loci ancestral haplotype to a multi-gene family complex with fourteen discrete loci. There are similarities between the sheep haplotype and the cattle haplotype, such as the dominance of three Ig-domain X-lineage inhibitory genes and the functional ablation of short tailed receptors. However, the sheep and cattle have expanded their *KIR* independently via alternate mechanisms, resulting in two haplotypes with no synteny. Nonetheless this has resulted in two distinct haplotypes between the species with similar characteristics. This indicates that both cattle and sheep have undergone similar selection pressures that have impacted on the evolution of their *KIR* genes.

5 Chapter 5. *KIR* and different *Bovidae* genomes

5.1 Introduction

From the Holstein-Friesian (HF) and aurochs *KIR* complexes studied in chapters 2 and 3 it is predicted the cattle *KIR* complex has no variation in gene content, and that the cattle *KIR* complex has evolved before domestication approximately 6,700 years ago. Unlike the *KIR* in primates and *KLRA* in mice, there is no evidence of gene presence/absence variation in cattle. It is now theorised that the sheep *KIR* complex, although evolving from a common ancestor that contained at least five of the same *KIR*, is significantly different to the cattle *KIR* complex. Therefore the cattle *KIR* haplotype has formed within the last 25 million years. In this chapter, the genomes of other *Bovinae* species are interrogated for *KIR* presence/absence with the intention of determining an indication to which the *KIR* complex structure is consistent with the *KIR* complex sequenced from the HF BAC library.

The cattle genome project has provided a whole genome reference sequence that has enabled a flurry of resequencing projects in the hope of identifying genetic variations responsible for production, health and general physiological traits. These resequencing projects have utilised the major advances in sequencing throughput afforded by second generation sequencing technology. Therefore, full genomes are sequenced without the intention of *de novo* assembly, resulting in cheaper and faster projects. The short reads are aligned to the cattle reference genome to provide an alignment that can be interrogated for gene presence/absence and SNP/indel detection. As the *KIR* region within the the cattle genome reference sequence is unfinished, this process does not yield usable results for the *KIR* complex.

The importance and benefits of open access datasets has been well received within the genomics community resulting in the provision of raw short read sequencing from such resequencing projects. These datasets have become freely available with the intention that further analysis can contribute to and enhance the projects. Therefore this chapter has focused on utilising the freely available *bovinae* genome raw sequence datasets for characterising *KIR* complex structures. The aim of this chapter is to indicate which *bovinae* species maintain the same gene structure as the HF *KIR* complex. A further aim of this chapter is to discover any potentially gene variable *KIR* complexes.

The bioinformatics pipeline developed in chapter 3 has provided a robust method to genotype animals based on whole genome short read sequences. Unfortunately due to the repetitive nature of the *KIR* complex and the inadequate

read length of the resequencing projects, SNP positions cannot be confirmed within individual genes using these datasets. However, by downloading the short read archive datasets from these genome resequencing projects it has been possible to genotype seven *bovinae* species. The short read data from the sheep genome project has also been used to gauge the stringency of this pipeline to distinguish between ruminant species *KIR* sequences.

The aim of this chapter is to gauge the extent to which the cattle *KIR* complex remains similar to the HF reference within other *Bovinae* species. This will enable the accurate targeting of genotyping projects to only use the breeds and species that are predicted to be compatible with the HF *KIR* complex. The eight genome sequences extensively studied in this chapter are representative of different speciation events during cattle evolution. The *Bos taurus* genomes from Angus and Fleckvieh animals are closely related to HF and therefore will indicate structural variation within the species. These breeds are used for different purposes than HF, the Fleckvieh is a dual purpose European breed similar to Simental and the Angus is used worldwide for beef production. Differences between these animals and the HF *KIR* complex may be the product of founder effect or the different breeding process used to generate these breeds. The Nellore and Sahiwal cattle are both *Bos indicus* breeds which split between 610,000 and 850,000 years ago [90]. The speciation of these animals and *Bos taurus* pre-date the aurochs characterised in chapter 3. Therefore, there is a greater chance of *KIR* complex sequence diversity between *Bos indicus* and *Bos taurus*. The Sahiwal data used here has been pooled from 22 individual animals and provides a cross-representation of the breed. The Nellore cattle breed originates from India but has become a major beef breed in Brazil after being transported there in the 19th century. The Kuchinoshima-Ushi (KU) is an isolated breed of cattle native to Japan, it is believed to retain much of phenotypic traits described of the ancient native Japanese cattle [73]. Utilising the KU genome sequences provides a unique insight into the *KIR* complex of a disparate and isolated island breed that has not had veterinary intervention. The Yak genome represents another *Bos* species outside of *Bos taurine* that will indicate the level of variation within the *Bos* species. The water buffalo genome will provide an indication of *KIR* complex conservation within the wider *Bovinae* species. It has been shown in chapter 4 that the sheep *KIR* complex is different to the HF therefore *Bovidae* *KIR* complexes are not all the same. The sheep genome sequences used in this study provides a gauge for the accuracy and limit of this analysis.

5.2 Methods

5.2.1 KU gDNA PCR for *KIR* genes

Genomic DNA from a KU breed of cattle was provided by the NODAI genome research center at the Tokyo university of agriculture. The sample was from the animal used in a SNP discovery resequencing project [73]. The genomic DNA was whole genome amplified using a QIAGEN REPLI-g Mini Kit (QIAGEN, UK) following the manufacturers guidelines. Primers were designed within the conserved intron sequence flanking the D0 and D2 domains. Primer pair sequences were as follows: 2DS23 int3 S ATGAAACTGCCTCTCCTCCTTCC and 2DS23 int4 AS GGTTTCATTGAGTTACACAAGCCC, 2DS23 int2 S ATTGGGTCACAAGAGTCAGATATGG and 2DS23 int3 AS GGAGCACTTCCTGTCGTTTTGAC, group2 int2 S AGCCCACCCACGAGAGCA and 3DX int3 AS CTCTGGAGACATTCCTGGGACTC, 3DXS23 int2 S GGTTAGCCCAGGTTTGGACTTG and 3DXS23 int3 AS TCCCTGGTTCCGTGGTGG. The optimised thermocyler conditions were as follows (95°C 1 min, (95°C 1.5 min, 62°C 30 s, 72°C 2 min) x32, 72°C 5 min). Predicted band sizes were calculated based on distance between primer pairs within the HF *KIR* gene sequences. PCR product band sizes were measured using electrophoresis on a 1% agarose gel. The *KIR* genes targeted were were confirmed by direct PCR product Sanger sequencing. PCR bands were excised then products were extracted and cleaned up using QIAGEN qiaquick gel extraction kits. PCR products were sequenced using the same PCR primers by Source BioSciences (Nottingham, UK) using ABi BigDye 3.1 and read using an ABi 3730 (Applied Biosystems).

5.2.2 Boinformatics pipelines

All of the sequence analysis methods conducted in this chapter including alignments, filtering, coverage depth and breath calculations and loci defining position analysis have been described within chapter 3 section 3.2.

5.3 Results

5.3.1 Non-Illumina sequenced genomes had disproportionate alignment statistics and were removed from further analysis

The raw genome sequences of ten different ruminant species were aligned to the cattle *KIR* complex using the same pipeline described in chapter 3. The majority of genomes were sequenced using the Illumina platform producing read lengths of 37 bp to 100 bp, Table 16. One animal, the Goldwyn bull (*Bos taurus*) was sequenced with the ABi SOLiD platform [130]. The Hereford genome was sequenced using Sanger technology and was used to assemble the cattle genome [45]. The buffalo [135], sheep [5] and yak [114] genomes which were sequenced as part of *de novo* assembly projects. The other genomes used, including the Angus, Fleckvieh, KU, Nellore and Sahiwal, were part of resequencing projects for SNP discovery. The angus (not published, same project as the nellore) and sahiwal [103] genomes were pooled from 18 and 22 individuals respectively.

Cattle *KIR* complex reads were pulled from the raw genome sequences using the bespoke pipeline used in chapter 3. This reduced the overall number of reads to just the reads that aligned to the complex. The proportion of reads extracted ranged from 0.002% to 1.2% of the total genome reads, Table 17. This range varied greatly depending on the technology used. With the non-Illumina sequenced genomes mapping a disproportionate quantity of reads to the complex relative to the Illumina sequenced genomes. The Illumina sequenced genomes ranged between 0.002 and and 0.01% of the genome reads. The variation in proportion of reads mapping to the complex is likely a result of the number of repeat regions sequenced.

The percentage of extracted reads from the genomes that subsequently aligned to the *KIR* complex after re-mapping ranges between 1.96% and 27.16%, Table 18. Therefore, at least 72% of the reads initially extracted from each genome are not originally from the *KIR* complex and align to other parts of the genomes. The sequence alignment of the Goldwyn animal was omitted from the further analysis of the *KIR* complex. This is because to align the colourspace reads a different pipeline was used and the average read length is very low at 27 bp. Therefore, the results are not comparable with the Illumina datasets. The Sanger sequenced Hereford genome also had to be aligned and interrogated using different techniques; the Sanger sequences also have lower base quality, combined with reduced read coverage could potentially introduce error. To enable accurate comparison between the genomes, the Hereford alignment was also dropped from further analysis. Therefore, the analysed genomes were all sequenced exclusively

Animal	Specis	Accession	Technology	Samples	No. Reads	Av length	Bases (Gb)
Angus	*Bos taurus*	SRP015694	Illumina	18	693,759,538	75.00	52.03
Buffalo	*Bubalus bubalis*	SRP001574	Illumina	1	1,906,168,360	61.63	117.49
Fleckvieh	*Bos taurus*	ERP000015	Illumina	1	1,201,868,938	36.00	43.27
Goldwyn (Holstein)	*Bos taurus*	SRP016124	SOLiD	1	1,540,371,084	36.75	56.60
Hereford	*Bos taurus*	AAFC0000000	Sanger	1	38,222,472	988.00	37.77
Kuchinoshima	*Bos taurus*	DRP000172	Illumina	1	1,005,754,450	70.26	70.67
Nellore	*Bos indicus*	SRP015694	Illumina	1	624,886,204	75.00	46.87
Sahiwal	*Bos indicus*	ERP000443	Illumina	22	601,247,607	91.00	55.08
Yak	*Bos grunniens*	SRP009062	Illumina	1	4,391,373,218	81.21	356.63
Human	*Homo sapiens*	SRP002509	Illumina	1	1,269,435,784	100.00	126.94

Table 16: Table showing details of the raw whole genome sequences. The common name of the genome (Animal) and species name are shown for each raw genome used. Each species has several different breeds which is denoted by the animals name. The accession number is the short read archive reference identifier. Not all of the genomes have been published. Samples is the number of individuals pooled for the genome sequencing.

Animal	Haplotype				KIR exons				mapped reads	mapped reads
	N reads	total bases	Av length	Av cov.	N reads	total bases	Av length	Av cov.	% (haplotype)	% (KIR exons)
Angus	56,413	4,230,975	75.00	11.43	3162	237,150	75.00	13.18	0.00813	0.00046
Buffalo	53,376	3,919,056	73.42	10.59	3,298	244,168	74.04	13.57	0.00280	0.00017
Fleckvieh	101,486	3,653,496	36.00	9.87	6,281	226,116	36.00	12.56	0.00844	0.00052
Goldwyn (Holstein)	18,504,323	500,846,500	27.07	1,353.22	666,617	18,732,950	28.10	1040.95	1.20129	0.04328
Hereford	98,461	99,889,656	1,014.51	269.89	1,253	1,304,805	1,041.34	72.51	0.25760	0.00328
Kuchinoshima	105,267	7,314,529	69.49	19.76	7,508	527,808	70.30	29.33	0.01047	0.00075
Nellore	56,201	4,215,075	75.00	11.39	3,306	247,950	75.00	13.78	0.00899	0.00053
Sahiwal	45,758	3,785,522	82.73	10.23	3,859	325,087	84.24	18.06	0.00761	0.00064
Yak	512,536	40,276,963	78.58	108.82	40,482	3,259,301	80.51	181.11	0.01167	0.00092
Sheep	43,363	2,445,456	56.39	6.61	2,137	133,939	62.68	7.44	0.00618	0.00034

Table 17: Table showing details of the extracted LRC reads. Details are of the reads aligned to the haplotype in order to pull out *KIR* reads to be realigned to the haplotype and the custom genome build. Haplotype represents reads that aligned to the full length *KIR* haplotype. Exons represents reads that aligned to just the *KIR* exon sequences. The mapped read % is the proportion of the raw genome sequencing reads that aligned to the haplotype or exon sequences. Average coverage (Av. cov.) has been calculated as the mean read depth over the reference sequence. Average length has been recalculated based on the reads that have aligned to the haplotypes.

using Illumina technology and the results were comparable.

Between the animals studied there is variability in the proportion of reads mapping to the complex. This is caused by a number of factors including the total number of reads sequenced, sequencing library preparation variation, read length and potentially the number or variation of *KIR* genes within the genome. To study the presence or absence of *KIR* genes within each genome, the same approaches described in chapter 3 were taken.

5.3.2 Read coverage depth indicates *KIR* gene presence absence variation

Except for one, all of *Bos* species genomes show expected sequence coverage along the *KIR* complex, which is exemplified by the Angus genome in Figure 40. The other *Bos* genome read depth coverage profiles are shown in the appendix (section 9.4.1). The read coverage is similar to that seen within the aurochs genome and positive control BAC and simulated datasets, Chapter 3 Figure 22, with reduced uniquely mapping read depth coverage over the *BotaKIR2DS2/3* and *BotaKIR3DXS2/3* loci. Therefore it is indicated from these alignments that the *Bos* species all encode gene identical *KIR* complexes. However, the Kuchinosima-Ushi genome has reduced unfiltered normal read depth coverage over the *BotaKIR2DS2/3* and *BotaKIR3DXS2/3* loci, Figure 41. Within all the other genomes studied, these genes have lower read coverage after filtering for uniquely mapping reads, however in the KU the read coverage is low before filtering. Therefore this may be an indication that these genes are not present within the KU genome.

The water buffalo genome shows far greater uneven read depth coverage, suggesting a potentially different *KIR* complex to cattle, Figure 42. This species is more related to *Bos* than to sheep and therefore more likely to have similar *KIR* complex to cattle than sheep. The sheep read depth coverage shows that reads map over the majority of the haplotype but not in a consistent or even pattern, Figure 43. It has been shown in chapter 4 that the sheep *KIR* complex contains similar genes but a substantially different haplotype sequence and structure. The sheep genome read coverage depth profile suggests similarity within some genes but no uniform coverage. Therefore by using the sheep genome as a negative control it shows that read coverage depth is a good indicator of gene presence or absence within the cattle *KIR* complex. To assess the difference in *KIR* representation between genes and species at a quantitative level, read breadth coverage was calculated.

5.3.3 Read coverage breadth reveals *KIR* gene presence absence variation in the KU and Nellore *Bos* species

Read coverage breadth was calculated as a percentage of sequence coverage over the X-axis and is based on uniquely mapped reads, this was conducted using the same methods described in chapter 3. Data from chapter 3 has been included in this chapter for reference. Read length has a positive correlation with coverage breadth as longer reads span more loci defining positions. Therefore, the genomes with longer read lengths are likely to have greater coverage breadth than genomes with shorter read lengths.

There is a pattern of coverage breadth shared between the different animals, Table 19. The genes *BotaKIR2DS2/3* and *BotaKIR3DXS2/3* have low read coverage breadth in all the animals. As the read coverage breadth is low in the positive control BAC and simulated datasets, this is a result of inadequate read length incapable of aligning uniquely to one of the two very similar loci and not structural variation as discussed in chapter 3. The sheep genome reads, used here as a negative control, show reduced read breadth coverage in all of the genes. The sheep does not have greater than 30.1% read coverage in any of the genes. The sheep genome alignment is therefore a useful indication of the level of coverage breadth required for the gene to be present.

The gene *BotaKIR3DXS1* has reduced breadth coverage in the nellore breed and the yak. This may be an indication of that genes absence in those genomes. The Kuchinomshima has greater reduction of *BotaKIR2DS2/3* and *BotaKIR3DXS2/3* read coverage breadth than the other *Bos* species. Although the read breadth coverage is relatively low for these genes in all the species analysed, the KU has particularly low read breadth coverage. This is despite the relatively longer read length of the sequenced KU genome. This further indicates *BotaKIR2DS2/3* and *BotaKIR3DXS2/3* are absent from the KU genome. The coverage depth and breadth analysis has shown that the *KIR* complex gene content is maintained in the *Bos* species. It has been indicated that the KU genome has a lack of *BotaKIR2DS2/3* and *BotaKIR3DXS2/3* genes and that the buffalo and sheep genomes do not contain the same *KIR* genes as cattle.

5.3.4 High resolution analysis of the loci defining positions predicts *KIR* presence and absence

Due to the low sequence coverage resulting from inadequate read length and reduced mapability, an alternative but complementary approach to determine gene presence or absence was required. The high resolution approach to calculate

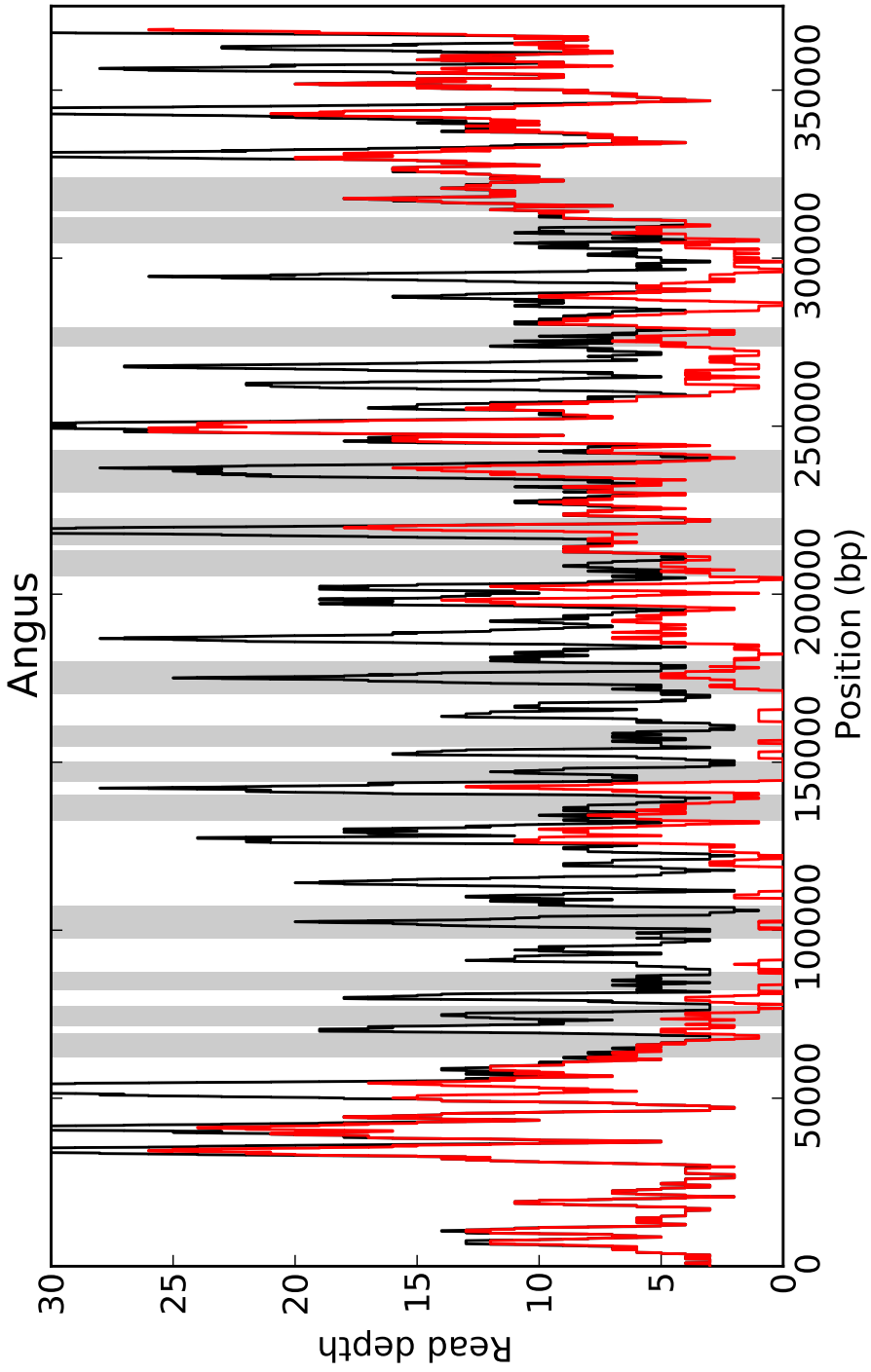


Figure 40: Read depth coverage of the Angus genome over the *KIR* complex. The black line represents the normal unfiltered read coverage and the red line represents reads filtered for unique alignment. The read coverage is a mean of a 300 bp sliding window. For reference the grey columns represent *KIR* loci positions and from left to right are: *3DXL6*, *2DS3*, *3DXS3*, *3DXL7*, *3DXL4*, *2DS2*, *3DXS2*, *3DXL5*, *3DXL2*, *2DS1*, *3DXL3*, *3DXS1*, *3DXL1* and *2DL1*

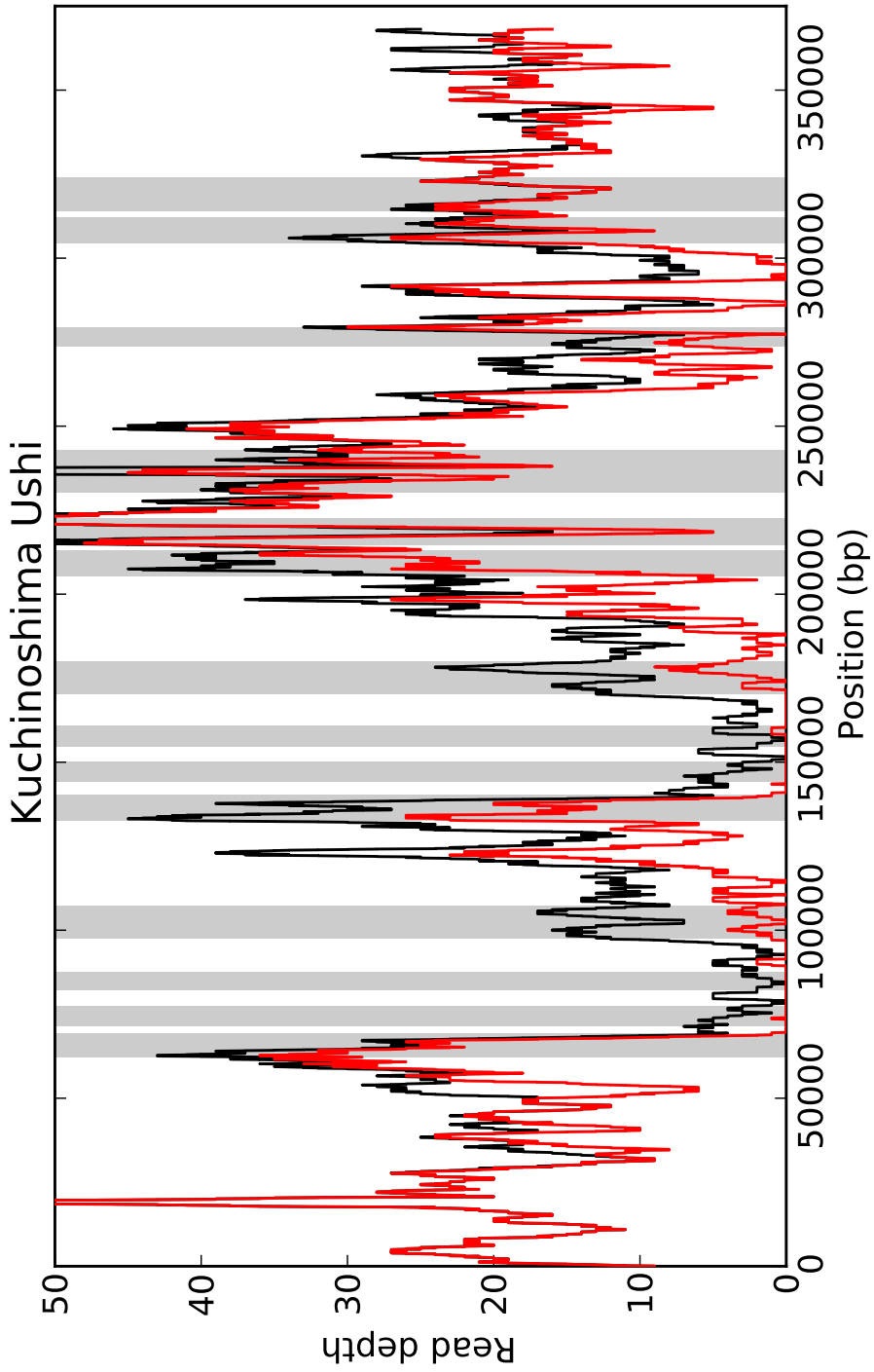


Figure 41: Read depth coverage of the KU genome over the *KIR* complex. The black line represents the normal unfiltered read coverage and the red line represents reads filtered for unique alignment. The read coverage is a mean of a 300 bp sliding window. For reference the grey columns represent *KIR* loci positions and from left to right are: *3DXL6*, *2DS3*, *3DXS3*, *3DXL7*, *3DXL4*, *2DS2*, *3DXS2*, *3DXL5*, *3DXL2*, *2DS1*, *3DXL3*, *3DXS1*, *3DXL1* and *2DL1*

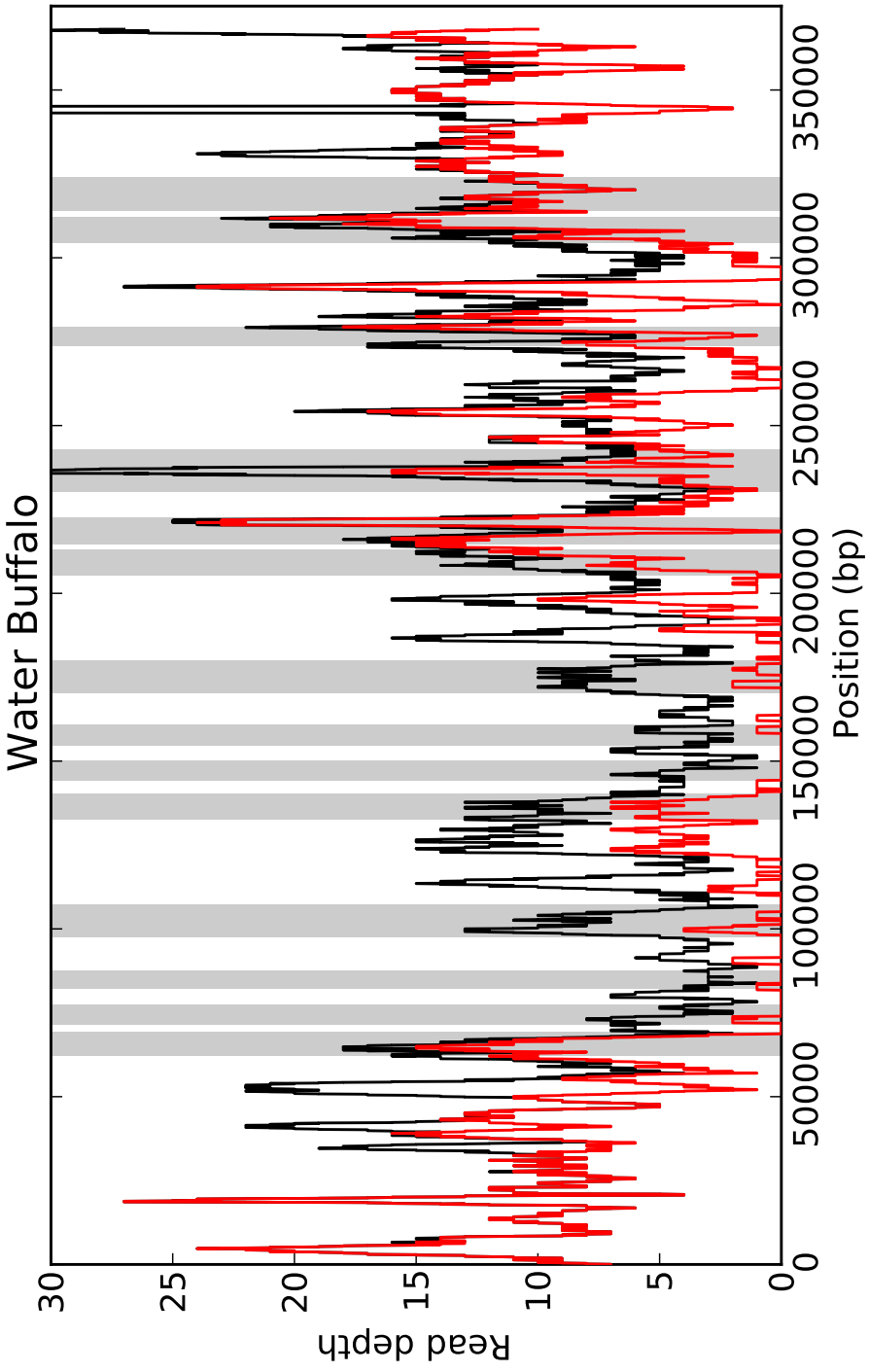


Figure 42: Read depth coverage of the water buffalo genome over the *KIR* complex. The black line represents the normal unfiltered read coverage and the red line represents reads filtered for unique alignment. The read coverage is a mean of a 300 bp sliding window. For reference the grey columns represent *KIR* loci positions and from left to right are: *3DXL6*, *2DS3*, *3DXS3*, *3DXL7*, *3DXL4*, *2DS2*, *3DXS2*, *3DXL5*, *3DXL2*, *2DS1*, *3DXL3*, *3DXS1*, *3DXL1* and *2DL1*

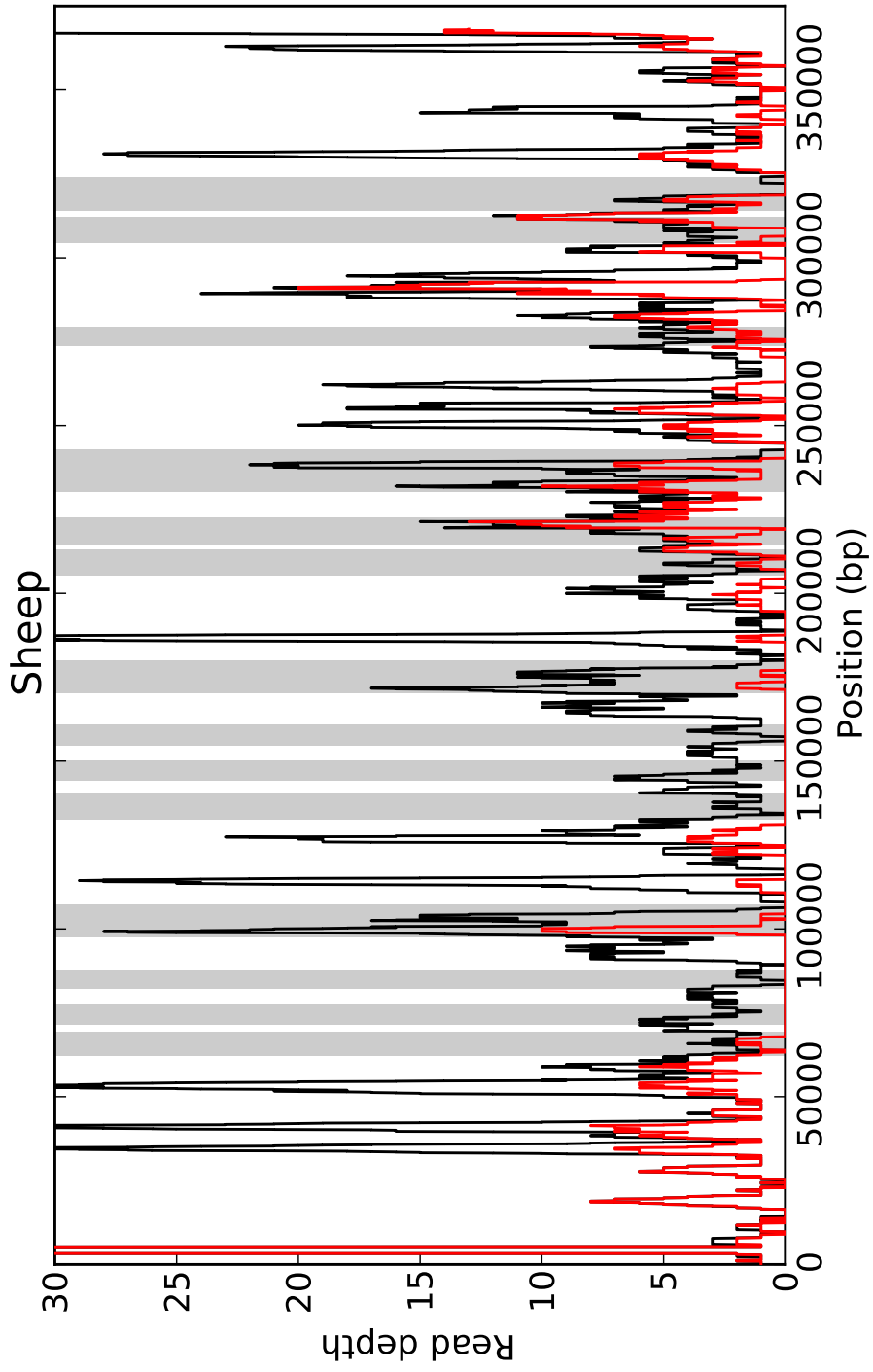


Figure 43: Read depth coverage of the sheep genome over the *KIR* complex. The black line represents the normal unfiltered read coverage and the red line represents reads filtered for unique alignment. The read coverage is a mean of a 300 bp sliding window. For reference the grey columns represent *KIR* loci positions and from left to right are: *3DXL6*, *2DS3*, *3DXS3*, *3DXL7*, *3DXL4*, *2DS2*, *3DXS2*, *3DXL5*, *3DXL2*, *2DS1*, *3DXL3*, *3DXS1*, *3DXL1* and *2DL1*

Animal	**N reads**	**Total bases**	**Av length**	**% of extracted**
Angus	8,786	658,950	75.0	15.57
Buffalo	10,694	795,584	74.4	20.04
Fleckvieh	15,945	574,020	36.0	15.71
Goldwyn (Holstein)	2,290,998	62,339,725	27.2	12.38
Hereford	1,932	1,978,209	1,023.9	1.96
Kuchinoshima	27,262	1,943,328	71.3	25.90
Nellore	9,188	689,100	75.0	16.35
Sahiwal	12,429	1,059,292	85.2	27.16
Yak	114,895	9,411,289	81.9	22.42
Sheep	4,295	286,014	66.6	9.90

Table 18: Table showing reads mapping to LRC within custom genome. The details are of the extracted raw reads described in Table 17 that have aligned to the *KIR* haplotype reference sequence embedded within the custom genome build. The % of extracted is the proportion of extracted reads that subsequently re-aligned to the *KIR* haplotype sequence.

Gene	BAC	Simulated	Aurochs	Angus	Buffalo	FleckVieh	Kuchinoshima	Nellore	Sahiwal	Sheep	Yak
3DXL6	80.6	87.7	68.3	78.9	76.1	84.9	95.4	67.9	93.9	30.1	95.9
2DS3	17.8	15.6	7.1	22.2	7.8	11.8	5.3	16.9	16.6	2.9	26.6
3DXS3	0.8	17.2	5.5	18.9	7.7	15.6	3.8	19.8	28.8	2.4	41.9
3DXL7	19.8	18.3	6.8	12.3	14.6	9.7	18.7	24.4	22.2	12	27.4
3DXL4	51.9	47.3	28.7	56.6	42.8	43.5	61	64.3	76.7	8.5	84.9
2DS2	19	17.9	3.9	16.7	4.6	14.1	4.9	22.8	24.1	4	36.9
3DXS2	13.4	16.5	6.9	15.4	6.8	12.7	3.4	16.8	25.5	3.9	37.2
3DXL5	21.9	17.2	13.5	29.6	10.2	13	26.8	14.3	38.9	10.2	50
3DXL2	62.1	52	51.7	52.8	45.7	39.7	63.1	36.9	63.5	12	68.2
2DS1	75	74.1	68.3	80.5	71.1	69.7	81.7	69.2	77	27.8	90.2
3DXL3	4.3	75	65.2	71.7	44.1	63.1	90.9	82.8	95.2	28.5	98.7
3DXS1	66.9	61.4	58.7	52.9	35.2	48.8	46.8	8.7	59.8	16.1	34.4
3DXL1	72.1	68	55.2	64.6	61.9	63.3	89.4	63.8	89.2	19.8	96.1
2DL1	67.6	95.4	83.3	89.4	83.7	92.4	99.6	89.5	97.7	19.9	100

Table 19: Read coverage breadth of various *Bovidae* species. Numbers are percentages of coverage for each gene, based on uniquely mapping reads. Blue data bars visually represent the coverage breadth percentage.

concordance within loci defining positions used in Chapter 3 section 3.3.4, was repeated with each genome aligned to the *KIR* complex. This high resolution loci defining position analysis shows that the majority of positions are represented within the *Bos taurus* species, Figure 44. Each gene has over half the loci defining positions represented in 100% of the reads aligned. The level of discordance between the angus and fleckvieh genomes and the loci defining positions is very low. The majority of the genes are well represented by sequenced genomes, except for *3DXL3*, which has the highest proportion of missing sequence.

The *Bos indicus* genomes have comparable gene content to the *Bos taurus* species, Figure 45. The sahiwal analysis shows greater heterozygosity with a larger proportion of gene positions represented by 50% to 75% of reads. This may be an artefact of pooling 22 animals for whole genome sequencing, as some genomes may have sequence or structural variation. The sahiwal genomes that are missing *KIR* genes or have polymorphisms in the loci defining positions would reduce the number of reads that are consistent with the reference loci defining positions. Therefore, a number of the sahiwal genomes analysed in the pooled sample may not have the same *KIR* gene structure or sequence as *Bos taurus*. However, the composite pool of genomes analysed here shows representation of all the *KIR* genes. Therefore, within the sahiwal breed, all the *KIR* genes analysed are present. The Nellore genome shows a reduction in the number of loci defining positions represented by reads in the *3DXS1* sequence, Figure 45a, this reduction is significant compared to the other species studied. This corresponds with the read breadth coverage analysis suggesting that *3DXS1* is absent from the Nellore genome.

The Kuchinomishima-Ushi genome shows a large proportion of missing sequence in the four genes *2DS2/3* and *3DXS2/3*, Figure 46. This analysis corresponds to the read breadth and depth coverage from the normal and uniquely mapped sequences. This is further evidence that *2DS2/3* and *3DXS2/3* genes are not represented within this genome and that the KU has a different *KIR* haplotype structure to HF, with a reduced gene number.

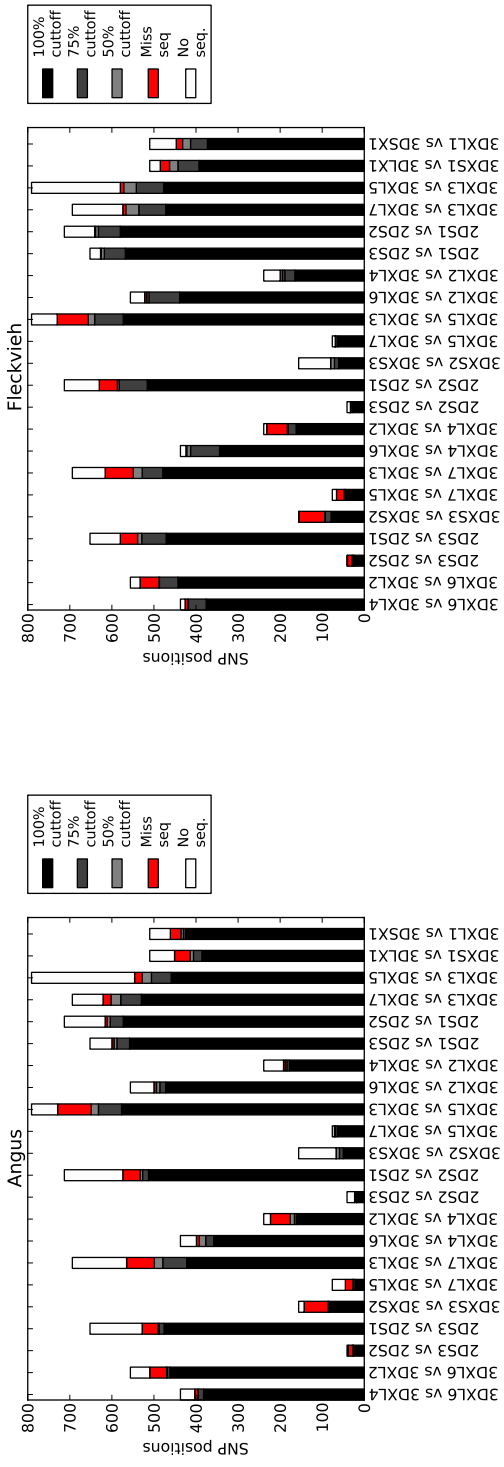
All of the *KIR* loci are represented within the yak genome, Supplementary Figure S6. Like the Sahiwal genomes, the yak genome shows greater heterozygosity, with the majority of positions corresponding to greater than 75% of the reads. Unlike the Sahiwal analysis, the yak genome is from a single animal and therefore this variation likely consists within a heterozygous *KIR* complex of the yak genome, or copy number variation of certain genes. This analysis predicts the presence or absence of a gene but it does not predict novel genes. Therefore the variation within the yak *KIR* complex may be the presence of new *KIR* genes

that have not yet been characterised.

The water buffalo and sheep genome analysis demonstrate the sensitivity of this characterisation, Figure 47. The water buffalo is more related to *Bos taurus* than sheep, therefore there is greater potential to share a similar *KIR* complex structure. The water buffalo genome represents moderate similarity to the majority of *KIR* loci, Figure 47a, however it is clear that there is significant diversity between the *Bos* and *Bubalus* species within the *KIR* complex. Although supporting greater than 50% of the loci defining positions in several of the genes, a number of genes have very little or no support. There are low levels of correspondence between the cattle and sheep but clearly there are no identical *KIR* genes, Figure 47b. The sheep *KIR* complex, as shown in chapter 4, has similar *KIR* genes and structure but no *KIR* genes are the same. Therefore, the sheep genome alignments supports this analysis pipeline and the interpretation of results.

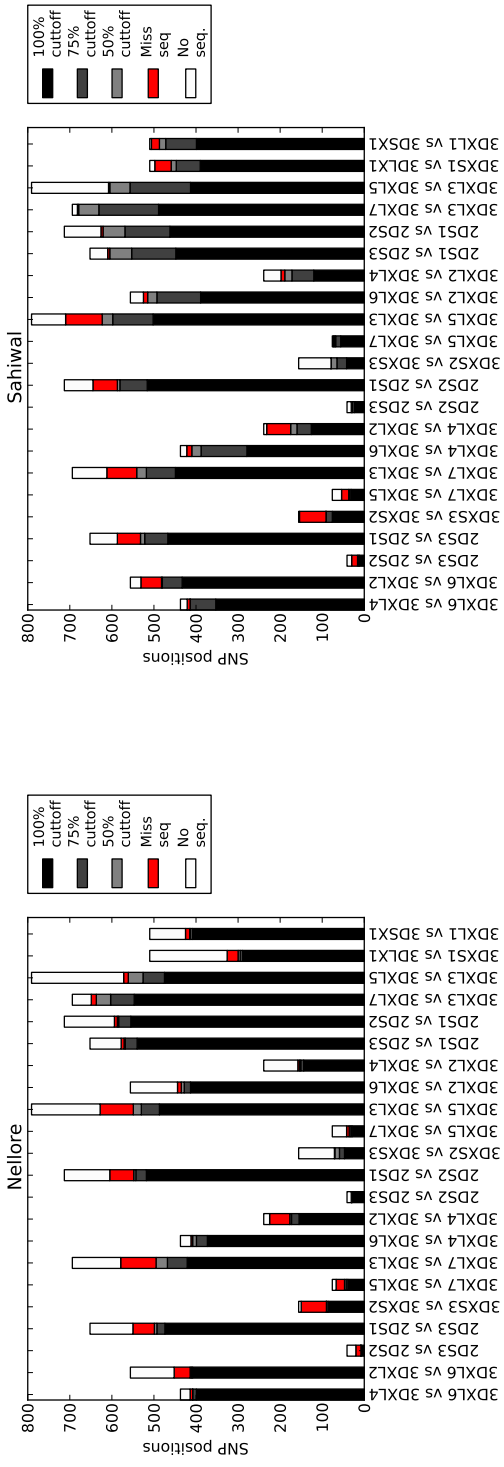
5.3.5 PCR of KU gDNA confirms absence of four *KIR* genes

To confirm the absence of the *2DS2/3* and *3DXS2/3* genes within the KU genome, PCR was conducted with group II and group III specific primers. Low concentration genomic DNA template was provided by the NODAI genome research center at the Tokyo university of agriculture. This sample was expanded to a usable quantity by whole genome amplification. PCR primers specific to *2DS2/3* and *3DXS2/3* genes were designed within areas of known sequence conservation. The PCR reaction was also performed on several DNA templates as positive and negative controls including the BAC DNA used to assemble the haplotype, several related Holstein-Freisian cattle and a yak. The results clearly showed that the these primers did not amplify any product from the KU template DNA, Table 20. Positive control primers designed to amplify group II genes worked as expected.



(a) (b)

Figure 44: High resolution SNP analysis of *Bos taurus*. Comparison of gene defining SNP positions between gene group loci. The columns represent the total number of variable nucleotide positions between the two genes. The black columns represent the number of positions that correspond to the first gene within 100% of the aligned reads. The grey columns are the same for 75% and 50% of the reads. The red columns represent the total number of positions that are discordant between the aligned reads and the first gene. The white columns are representative of the total number of positions that are not covered by sequence.



(a) (b)

Figure 45: High resolution SNP analysis of *Bos indicus*. Comparison of gene defining SNP positions between gene group loci. The columns represent the total number of variable nucleotide positions between the two genes. The black columns represent the number of positions that correspond to the first gene within 100% of the aligned reads. The grey columns are the same for 75% and 50% of the reads. The red columns represent the total number of positions that are discordant between the aligned reads and the first gene. The white columns are representative of the total number of positions that are not covered by sequence.

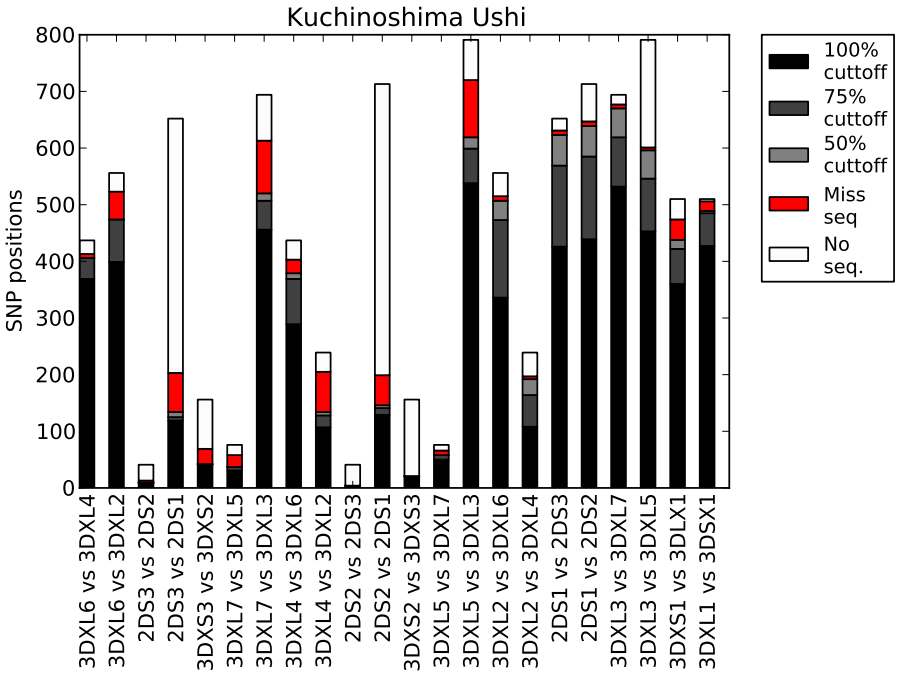
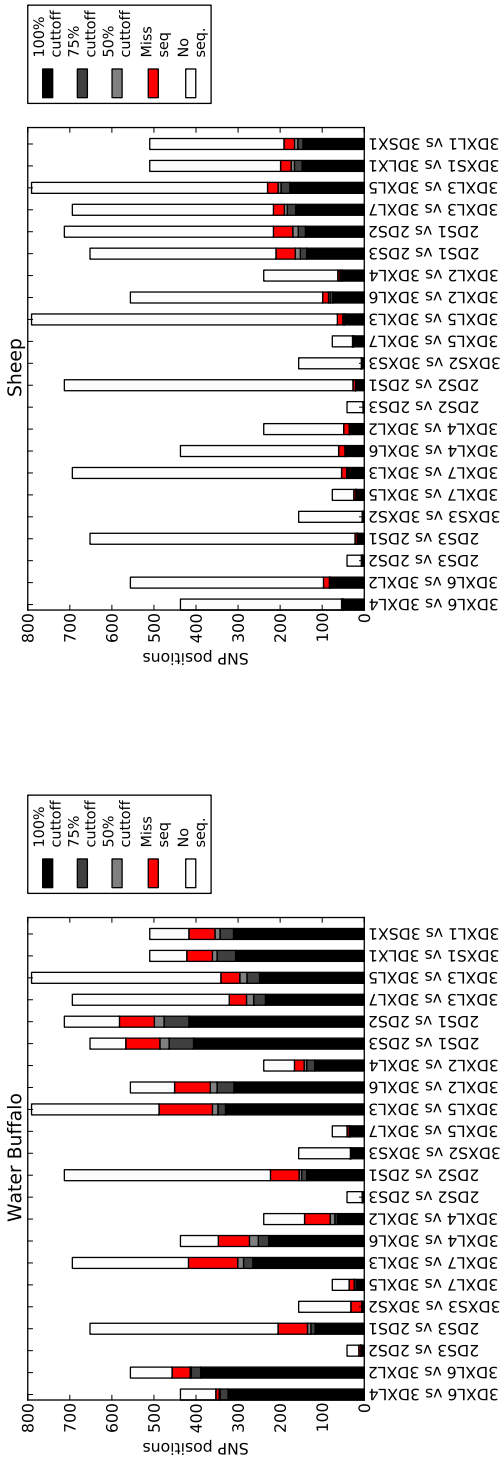


Figure 46: High resolution SNP analysis of other Kuchinoshima-Ushi cattle. Comparison of gene defining SNP positions between gene group loci. The columns represent the total number of variable nucleotide positions between the two genes. The black columns represent the number of positions that correspond to the first gene within 100% of the aligned reads. The grey columns are the same for 75% and 50% of the reads. The red columns represent the total number of positions that are discordant between the aligned reads and the first gene. The white columns are representative of the total number of positions that are not covered by sequence.

Target	Pr. band size	335H08	917	206	Yak	145	KU
2DS2/3 D0	482 bp	X	✓	✓	✓	✓	X
2DS2/3 D2	465 bp	X	✓	✓	✓	✓	X
3DXS2/3 D0	1201 bp	X	✓	✓	✓	✓	X
Group II D0	547 bp	X	✓	✓	✓	✓	✓

Table 20: Table of results from *2DS2/3* and *3DXS2/3* targeted PCR. The target is the gene and domain exon sequence targeted by primers designed within the flanking intron sequence, the predicted band sizes are shown (Pr. band size). Numbers and names along the top row are templates used for PCR amplification. 335H08 is BAC clone DNA that does not contain any of the genes targeted. 917, 206 and 145 are all gDNA templates from Holstein-Friesian cattle. The yak is gDNA from a *Bos grunniens* sample. The Kuchinohsima is the whole genome amplified gDNA. Numbers are representative of PCR band sizes in base pairs. Ticks (✓) represent correct band sizes after electrophoresis, crosses (X) represent no product.



(a) (b)

Figure 47: High resolution SNP analysis of other bovinae species. Comparison of gene defining SNP positions between gene group loci. The columns represent the total number of variable nucleotide positions between the two genes. The black columns represent the number of positions that correspond to the first gene within 100% of the aligned reads. The grey columns are the same for 75% and 50% of the reads. The red columns represent the total number of positions that are discordant between the aligned reads and the first gene. The white columns are representative of the total number of positions that are not covered by sequence.

5.4 Discussion

In this chapter whole genome sequencing reads from within the *Bos* species and beyond have been aligned to the HF *KIR* complex to determine gene presence or absence. This bioinformatic analysis has shown that although *Bos* species have the same *KIR* complex as the HF reference, there is gene presence/absence variation within *Bovinae*, Figure 48. This has been verified with PCR screening of the KU genome.

5.4.1 Evolution of the KU *KIR* complex

It is predicted from this analysis that the KU cattle have a reduced gene number *KIR* complex, Figure 48. The genomic DNA for the KU cattle individual used in this analysis does not contain four null-allele *KIR* that have been identified within all the other *Bos* species studied. Assuming the animal in this study is representative of the KU species, the species has either lost the genes, or never had the genes to begin with.

One explanation is that the species has truncated the complex removing the genes that are no longer required. In HF these *KIR* are non-functional, therefore in KU they may have been removed from the complex by non-allelic homologous recombination (NAHR) with no selection pressures to maintain the genes. Alternatively these *KIR* may have been functional within the KU; mirroring the process of inactivation that was seen in HF the KU may have removed the genes using a different mechanism, deleting them completely as opposed to silencing them through premature stop codons.

An alternative explanation is that these *KIR* were not present within the founding cattle that were brought to the Kuchinoshima island. This would mean that the *KIR* complex did not become truncated but was the result of founder effect. This explanation goes against the block duplication theory described in chapter 2. The other *KIR*, *3DXL4/6* and *3DXL5/7*, from the same blocks, block A and block B, as the missing *KIR* are present within the KU genome. Therefore if the KU and it's ancestors never had the missing genes, then the blocks in HF must have evolved by a different mechanism. Because this explanation is less likely, it is believed that the KU previously had these *KIR* and subsequently lost them during their evolution.

Furthermore this truncated complex may be seen within other species. Until further complexes have been sequenced it cannot be discounted that this complex is a common form that is seen in disparate cattle populations. It could therefore have been a complex that was in the founding cattle to be transported to Kuchi-

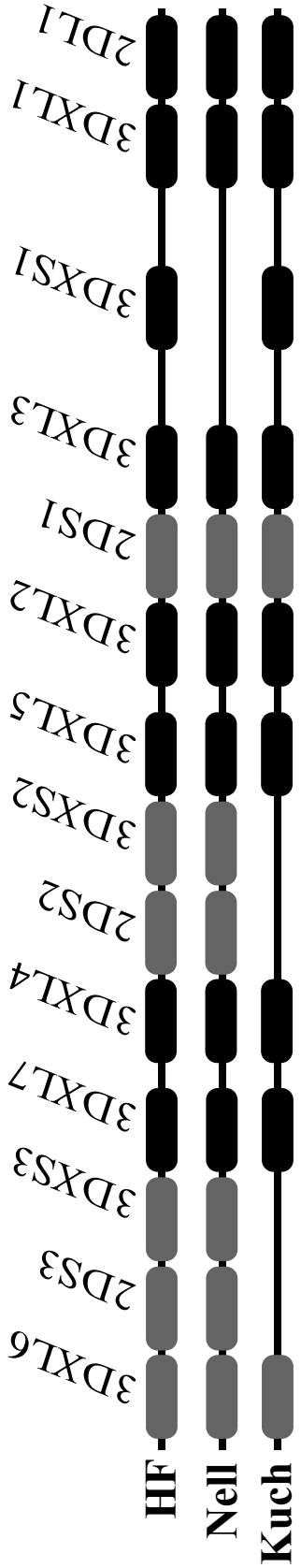


Figure 48: Diagram of variable gene *KIR* haplotypes. Black oblongs represent predicted functional genes, grey oblongs represent predicted null-alleles. *KIR* names are shown along the top and animal names are shown along the left side. HF represents the classical Holstein-Friesian haplotype sequence that is seen within all of the *Bos* species except the Nellore (Nell) and KU (Kuch).

noshima. If the complex is the dominant form within the KU population, it may either have been the only form inherited or it has been selected for because of an advantageous trait.

Unpublished work in the Hammond lab has shown that the null-alleles, specifically *BotaKIR3DXS2* and *BotaKIR3DXS2* are transcribed. However, these transcripts contain the premature stop codon that prevents full length protein translation and are not potentially functional versions of the null-alleles. Could these short translated sequences or transcripts of null alleles be interfering with the normal function of the cattle KIR? And therfore is there a benefit from deleting these null-alleles? Alternatively could the presence of the null-allele sequences enable hybrid-composite genes to form? This would occur by the polymerase skipping certain stop-codon containing domains to produce a novel *KIR* transcript based on exons from different genes. Such genes have never been found but only few full length exon *KIR* sequences from cattle lymphocytes cDNA have been sequenced.

5.4.2 The evolution of 3DXS1 has occurred relatively recently within the *KIR* complex

The analysis within this chapter has indicated that the genome of Nellore species of cattle does not contain *BotaKIR3DXS1*. Although the study has released the raw genome sequences for the Angus and Nellore cattle to the EBI and NCBI short read archives (accession SRP015694), the results have not been published. Therefore, there is very little information regarding this animal, meaning the results for the Nellore and Angus should be considered with caution. Nonetheless, if it is assumed that the animal used was a Nellore cattle (*Bos indicus*) and that this animal is representative of its species, then *BotaKIR3DXS1* has either been deleted or it had not evolved within this animal. Further investigation into this breed and this locus are on-ongoing within the Hammond lab.

BotaKIR3DXS1 is believed to have duplicated from the Ig domain exon sequences of *BotaKIR3DXL1* and the transmembrane domain exon sequences of an activating *KIR*. The result is an activating KIR with the same ligand specificity as an inhibitory KIR. These paired receptors are hypothesised to have evolved in response to viral subversion of an inhibitory KIR ligand. The virus produces a ligand homologue protein that interacts with the KIR receptors thus masking down regulation of MHC. The host responds after the spontaneous recombination of the paired receptors to generate the activating variant KIR that targets the cells expressing the decoy protein. Therefore the virally infected cells are killed by NK cells naturally within the host resulting in eradication of the virus from

the population. The activating KIR recognises self and is therefore dangerous to the host. It is therefore likely to be short lived within evolutionary terms. *BotaKIR3DXS1* is predicted to be functional within HF cattle, it is hypothesised it remains functional and has not been deactivated due to the persistence of the pathogen it evolved to target.

The *Bos indicus* species split with *Bos taurus* before domestication and both species were domesticated independently, however admixture between the two sub species has occurred. Nonetheless, it is conceivable that *BotaKIR3DXS1* did not exist before this split then formed within *Bos taurus primigenious* aurochsen. However, the Yak genome has strong evidence for containing *BotaKIR3DXS1*. Therefore, as the Yak diverged before taurine and indicine cattle, the most parsimonious explanation is that the nellore has lost *BotaKIR3DXS1* during its evolution. It could be hypothesised that the nellore cattle has lost *BotaKIR3DXS1* as a result of lack of pathogen selection pressure. Unlike the other species studied, the nellore may have become isolated from the pathogen that promoted *BotaKIR3DXS1* evolution. It is predicted BotaKIR3DXS1 recognises the same ligand as BotaKIR3DXL1, therefore expressing BotaKIR3DXS1 could be detrimental to the host causing autoimmune problems. Therefore, with the lack of pathogen selection pressures, its removal from the genome was beneficial to the host.

5.4.3 The cattle *KIR* complex has evolved in *Bos* species

This chapter has shown that the *KIR* complex is largely consistent within the *Bos* species. The *KIR* complex is shown to have the at least same gene content in the disparate *Bos* species of *Bos taurus*, *Bos taurus primigenious*, *Bos indicus* and *Bos grunnienes*. However, the larger *Bovinae* clade does not share this structure as the water buffalo genome has shown evidence of a different *KIR* complex. Therefore it is believed that the HF *KIR* complex has evolved within the *Bos* species during the last 5.4 million years. To confirm this further, other *Bos* species such as *Bos javanicus*, *Bos gaurus* and *Bos sauveli* could be sequenced.

6 Chapter 6. Variation in the cattle *KIR* complex

6.1 Introduction

The assembly of the cattle *KIR* haplotype has enabled the systematic analysis of each of the cattle *KIR* genes. Before this can begin the extent of polymorphismic variation within the cattle *KIR* sequences needs to be determined. Primer and probe design for real time assays and genotyping studies are likely to be sub optimal without knowledge of the polymorphic positions within the *KIR* sequences. Determining the most polymorphic *KIR* and the most variable positions within each gene will indicate genes under selection and enable modelling of ligand binding variation.

By utilising publicly available short read sequences it has been possible to identify that the *KIR* complex largely has remained unchanged during its evolution within the *Bos* species. These short reads alignments have however indicated that the cattle *KIR* complex, like in other species, is highly polymorphic. However, the inadequate read length of these studies has not allowed confident use of this polymorphism data. Therefore to determine the extent of polymorphism within the *KIR* complex, further sequencing is required to provide longer reads capable of aligning to the complex with high confidence.

First attempts involved designing primers within the intron sequences surrounding each *KIR* exon sequence and sequencing directly from the PCR product. This initial pilot project was intended to determine the most conserved sequences for further generic primer design that would amplify all of the *KIR* exon sequences for NGS sequencing. After designing and implementing the pilot phase of this project the results suggested that the polymorphic nature of the *KIR* complex and the secondary structures that were forming prevented detection all of the *KIR* loci targeted.

Instead a targeted sequence capture and enrichment approach was taken which enabled the whole *KIR* complex to be sequenced. A bespoke Nimblegen EZ capture developer assay was created using the *KIR* complex as a reference sequence. Tiling probes were created that are complementary to the *KIR* complex sequence, Figure 49. After fragmentation and sequencing adapter ligation of the of the target animal genomic DNA, tiling probes bind to the fragmented target region enabling amplification of the *KIR* complex via ligation-mediated PCR (lmPCR). The enriched genomes can then be sequenced using an Illumina MiSeq to produce relatively long reads of 2x250 bp compared to the 2x100 bp produced by the genome resequencing projects. These reads can be artificially joined to create a theoretical maximum read of 490 bp. At this read length, most

of the problems associated with inadequate read length for unique alignment to the *KIR* complex exons are negated, as discussed in chapter 3.

The cattle *KIR* haplotype sequenced within this project was from a HF animal. The HF breed of cattle has undergone intensive domestication to improve productivity, this may have had an effect on the level of polymorphisms with the breed. The MHC herd in Compton, West Berks, UK is owned by The Pirbright Institute and consists of back-bred HF cattle with homozygous MHC haplotypes. The MHC herd of cattle will be intensively interrogated for the role of NK cells in various cattle diseases, however the lack of detail for the level of variation within the *KIR* complex means the KIR cannot be robustly interrogated yet. To study the extent of *KIR* sequence polymorphisms within the the HF breed, DNA from fourteen MHC herd animals was enriched and sequenced. Animals from different MHC genotypes were sequenced to get an indication of diversity across the entire herd. This provided complete *KIR* exon sequence data for an unprecedented 28 HF cattle *KIR* haplotypes. To determine the rate of error in the process DNA from one of the animals was enriched and sequenced twice. To gain a wider indication of *KIR* variation within the UK dairy herds, two breeding British Friesian bulls were sequenced, Blackisle Glen Grant and Nerewater Tiptop. These bulls have sired hundreds of offspring within the UK and worldwide dairy herds, each containing a *KIR* haplotype sequenced here. Therefore this is a powerful indicator of SNP positions within the *KIR* complex.

To confirm the Kuchinoshima-Ushi *KIR* complex results described in chapter 5, the whole genome amplification DNA sample was also enriched and sequenced. This enabled the polymorphic positions to be interrogated to look for specific SNPs and indels to the KU. To begin to determine the *KIR* complex gene structure and extent of polymorphisms within the genetically isolated Chillingham breed of cattle, Chillingham bullock and heifer DNA samples were enriched and sequenced. The Chillingham cattle, which appear phenotypically similar to the white park breed of cattle have lived within an enclosed park at Chillingham castle, Northumberland. They have not had any veterinary intervention and are believed to be the best representation of native wild UK cattle, therefore pre-date the intensive domestication of the last two centuries. The species has undergone genetic bottlenecks that has reduced sequence variation between the animals. Four *Bos indicus* cattle, two Sahiwal and two Nellore were sequenced to determine the extent of genetic diversity within a more distant cattle sub species. Sequencing these animals provides information on variable positions and genes that might not be present within the the HF *KIR* haplotypes. The data produced from four rounds of capture, enrichment and sequencing provides a rich source

of information about the diversity of the cattle *KIR* sequences. The variable positions will be used to enable design of genotyping assays as both conserved and allele specific positions were previously unknown.

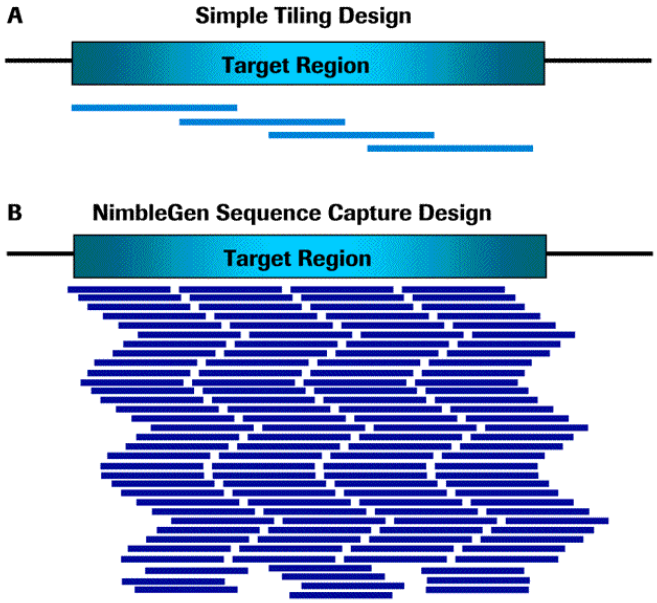


Figure 49: Diagram of tiling probes over the target region. Blue lines represent probes designed to complement the target region. Simple tiling design (a) uses fewer 120 bp length probes than the Nimblegen design (b) that uses 50-105 bp length probes. Image taken from the Nimblegen support website.

6.2 Methods

The data in this chapter is part of an ongoing Immunogentics lab project to characterise the genetic diversity within the cattle *KIR* and NKC complexes. The wet-lab work has been conducted by Doctor Mark Gibson within the Immunogenetics lab, Pirbright Institute. All of the analysis and some of the DNA preparation was conducted by myself.

6.2.1 DNA preparation

The Kuchinoshima-Ushi genomic DNA is the same whole genome amplified sample used within chapter 5. The Chillingham bull DNA has been whole genome amplified using the same methods described in chapter 5 section 5.2.1. The dairy cattle samples have been extracted from PBMCs using TRIzol® reagent (Life technologies, UK) following the manufacturers guidelines.

DNA was fragmented using a Covaris focused ultrasonicator (Covaris, inc. Massachusetts, USA) to produce fragment sizes of approximately 420 bp, measured on an Agilent bioanalyzer (Agilent Technologies, Inc. California, USA). Illumina TruSeq LT (Illumina, UK) library adapters were ligated to the fragments before genome enrichment.

6.2.2 Targeted genome enrichment of the *KIR* complex sequence

The NKC and *KIR* complex target regions were enriched using a Roche Nimblegen SeqCap EZ developer library (Roche, Basel Switzerland) following the manufacturers guidelines for short read Illumina sequencing. Custom probes were designed based on the *KIR* complex. The probe design was an iterative process that relied on the custom genome build with the assembled *KIR* haplotype with input from both Nimblegen and myself for the final probe design.

6.2.3 DNA sequencing

Enrichment capture library preparations were sequenced using an Illumina MiSeq at the Pirbright Institute. The Illumina TruSeq LT library preparation was used with either 460 or 500 cycles to produce either 2x230 bp or 2x250 bp reads respectively. A total of four capture experiments and sequencing runs were conducted, with four animals in the first run (HF504805, HF505183, HF504882 and HF104766), six animals in the second (HF404818, HF598, HF4222, HF505204, HF705206, HF204375) and third runs (HF982, HF766, HF405, HF159, Chillingham bull 250b, Kuchinoshima-Ushi), and nine animals in the fourth run (Blackisle, Chillingham3, HF252, HF652, Nellore NE14, Nellore NE43, Nerewater, Sahi-

wal SW2, Sahiwal SW3). All of the samples designated with the HF have been sourced from the Pirbright Institute MHC herd in Compton, West Berkshire and have characterised MHC class I haplotypes.

6.2.4 Sequence analysis and variant detection

The raw fastq files were downloaded from the MiSeq and had adapter sequences cut using cutadapt [91]. Cut sequences were aligned to the custom cattle genome build described within chapter 3, section 3.2.1, with BWA-MEM and Bowtie2 [81]. Reads were uniquley aligned with BWA-MEM by using the C1 option and the other default settings. Reads were uniquely aligned with bowtie2 by filtering out reads that aligned to other positions with an equal map score using a bespoke python script described in the appendix section 9.5.1. SNPs and indels were called using Varscan2 [79] with 20% or higher proportion of supporting reads. SNPs were called from both BWA and Bowtie2 alignments and kept if they were called by both. The positions and potential residue changes of the SNPs were annotated using the same pipeline described in chapter 2, section 2.2.7. Shared SNPs statistics were determined using MySQL.

Copy number variation of the cattle *KIR* genes was predicted by calculating the fold increase or decrease of read coverage depth compared to a baseline reference sequence provided by the HF4222 read depth coverage. The ratio change in read depth coverage over the 3′ region of the KIR complex encompassing the non-variable *FCAR* and *NCR1* genes was calculated between the baseline HF4222 and a target animal coverage depths. This ratio was used to compare the relative read depth coverage between the two animals over the entire haplotype. Each base position was compared between the two animals and calculated as a change from the relative baseline. All the relative depths for each base position were calculated for each *KIR* exon to give an indication of CNV per gene. Details of the script used are in the appendix within section 9.5.2.

The number of SNPs each animal contains that are different or not seen within in another animal was calculated using MySQL. This was repeated for all the combinations of animals in a pairwise fashion to generate a matrix of SNP differences. The matrix was used to generate a dendrogram and therefore infer phylogenetics using the shared SNP positions between all of the animals. The entire haplotype sequence was used excluding the LILR regions (start at 60 kb into the CKH reference sequence). A bespoke python script was written to generate the dendrogram, details are in the appendix section 9.5.3.

6.3 Results

6.3.1 DNA fragment length biased sample sequence distribution

The Illumina MiSeq generated up to 2.5 gigabases (Gb) of nucleotide sequence data for each of the animals used in the enrichment captures, Figure 50. There was considerable differences in total bases generated between the animals in the study with HF705206 producing nearly 4x as many reads as HF982. This was partly due to the number of animals used in each sequencing run. There were four runs, the first run had four animals whilst the second and third run both had six animals each, the final run contained nine animals. Therefore there were fewer reads per animal in the later runs which would have reduced the number of bases per animal. This was partially counteracted by the decision to use 500 cycles for the second, third and fourth runs instead of 460 cycles that was used in the first run. Therefore, the second, third and fourth runs produced 2x250 bp reads compared to the 2x230 bp reads from the first run. This is effectively a nucleotide base increase of 8% per animal in the second, third and fourth runs. The library preparation for Nellore NE43 was sub-optimal and produced the lowest total bases, it is unknown what caused this and the results from this animal cannot be used.

The biggest differences in total bases yielded were caused by the library sizes. The HF705206 and HF204375 libraries were prepared with the animals in the first capture enrichment but were run on the MiSeq alongside the animals from the second capture enrichment. The first capture enrichment had a lower median fragment size than the second, third and fourth, Figure 51. Therefore, when the first capture animal library preparations were sequenced alongside the second capture animals, the shorter fragments of the first capture animals were biased by the sequencing process. This bias resulted in many more sequences for HF705206 and HF204375 compared to the other animals sequenced in the second capture.

6.3.2 Alignment of raw sequences revealed 50% non-specific enrichment

The raw sequencing reads from each capture enrichment were aligned to the custom cattle genome described in chapter 3, section 3.2.1. This genome build has had all *KIR* sequences removed and the assembled *KIR* complex sequence inserted as a standalone chromosome. The number of bases aligned to each chromosome was calculated revealing that approximately 50% of the reads were not specific to the *KIR* complex or the NKC, Figure 52. These reads are the result of non-specific probe binding and enrichment that has amplified unwanted

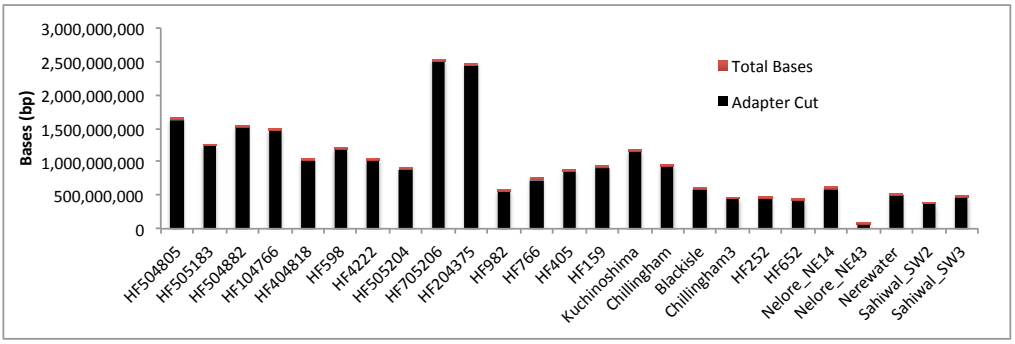


Figure 50: Total bases produced from Illumina MiSeq sequencing for each capture experiment. Total bases are shown (red), which is slightly greater than the quantity of bases after adapters have been cut (black). The animals are in order of capture experiment, the first four were from capture 1, the second and third six were from capture 2 and 3 respectively, the final 9 are from capture 4.

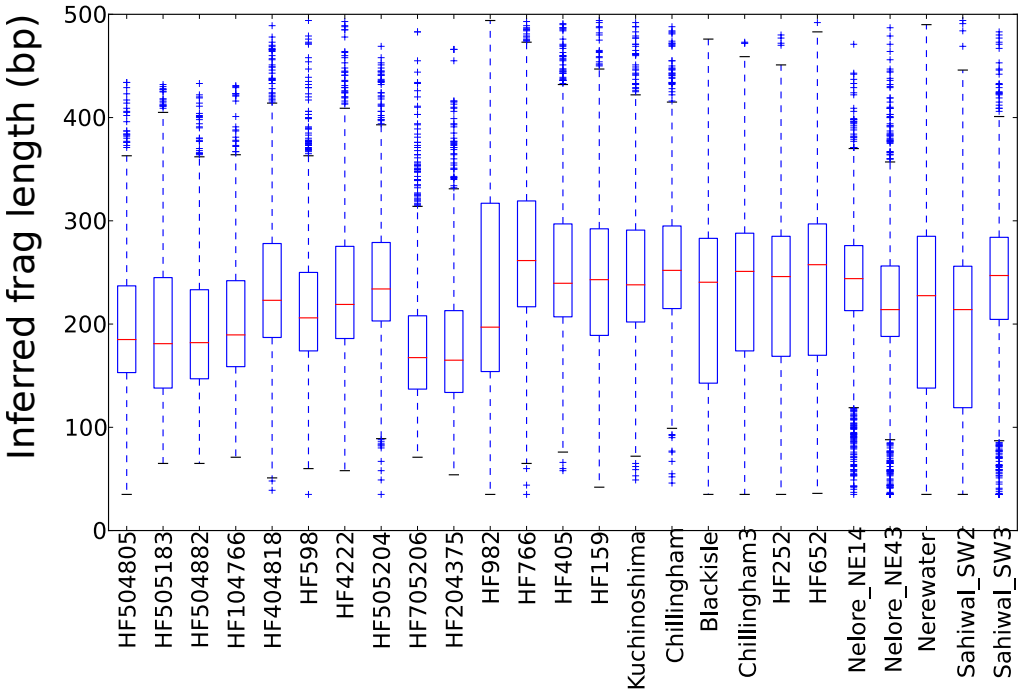


Figure 51: Box plots showing the inferred DNA fragment size distributions. Fragment sizes are inferred by joined read pairs. Box plot edges represent the 25th and 75th percentiles from the distribution curve. Whiskers represent the 5th an 95th percentile ranges and the outliers are represented by crosses. Data is representative of 10,000 random fragments per animal. This data follows the same trend as the full data set (not shown). The animals are in order of capture experiment, the first four were from capture 1, the second and third six were from capture 2 and 3 respectively, the final 9 are from capture 4.

regions of the genome. The number of non-specific read bases aligning to each chromosome is proportional to the size of the chromosome, with more non-specific reads aligning to the larger chromosomes, Figure 52. This is in line with what was predicted for the capture and enrichment technology.

6.3.3 Fragment length has no effect on probe specificity

For each capture experiment the fragment size was increased in an effort to increase coverage over the repetitive regions of the *KIR* complex. The difficulties of aligning short reads to the *KIR* complex and the inadequacies of short read lengths were discussed in chapter 3, section 3.3.2. The effect of increasing the fragment size may have affected the specificity of the probes resulting in greater non-specific binding. To interrogate this effect the fragment sizes were calculated and compared. The fragment lengths of the reads aligning to the *KIR* complex and NKC, which were the intended targets for this project, were inferred by calculating read pair mapping distance. The fragment lengths for the reads aligning to the non-specific chromosomes were also calculated. The actual sizes of the fragments for each enrichment capture library sequencing run was calculated by joining the Illumina read pairs to create one artificial long read from a pair of short reads. A comparison of fragments sizes between the capture targets (*KIR* complex and NKC), non-specific chromosomes and actual sizes revealed no differences, Figure 53. This was repeated on all of the the animals sequenced to determine if the increase in fragment sizes used had an effect on the specificity of probe binding (data shown in the appendix section 9.5.4). There was no correlation between fragment size and probe specificity.

The library sizes were calculated to be significantly shorter than the fragment sizes determined by the bioanalyser. This was an unknown consequence of the enrichment and sequencing process and may have impacted on sequence breadth coverage. Due to this shorter effective fragment size, the reads were aligned individually as opposed to alignment after artificial joining. This is because more sequences were available with this approach. A comparison between aligning reads independently and with joined reads revealed no difference in sequence coverage breadth.

6.3.4 Read depth coverage confirmed the presence of *KIR* genes within the HF and confirmed the reduced *KIR* complex in KU

The coverage depth of the aligned reads over the *KIR* complex for each of the animals sequenced showed that each *KIR* locus had sequence coverage, Figure 54.

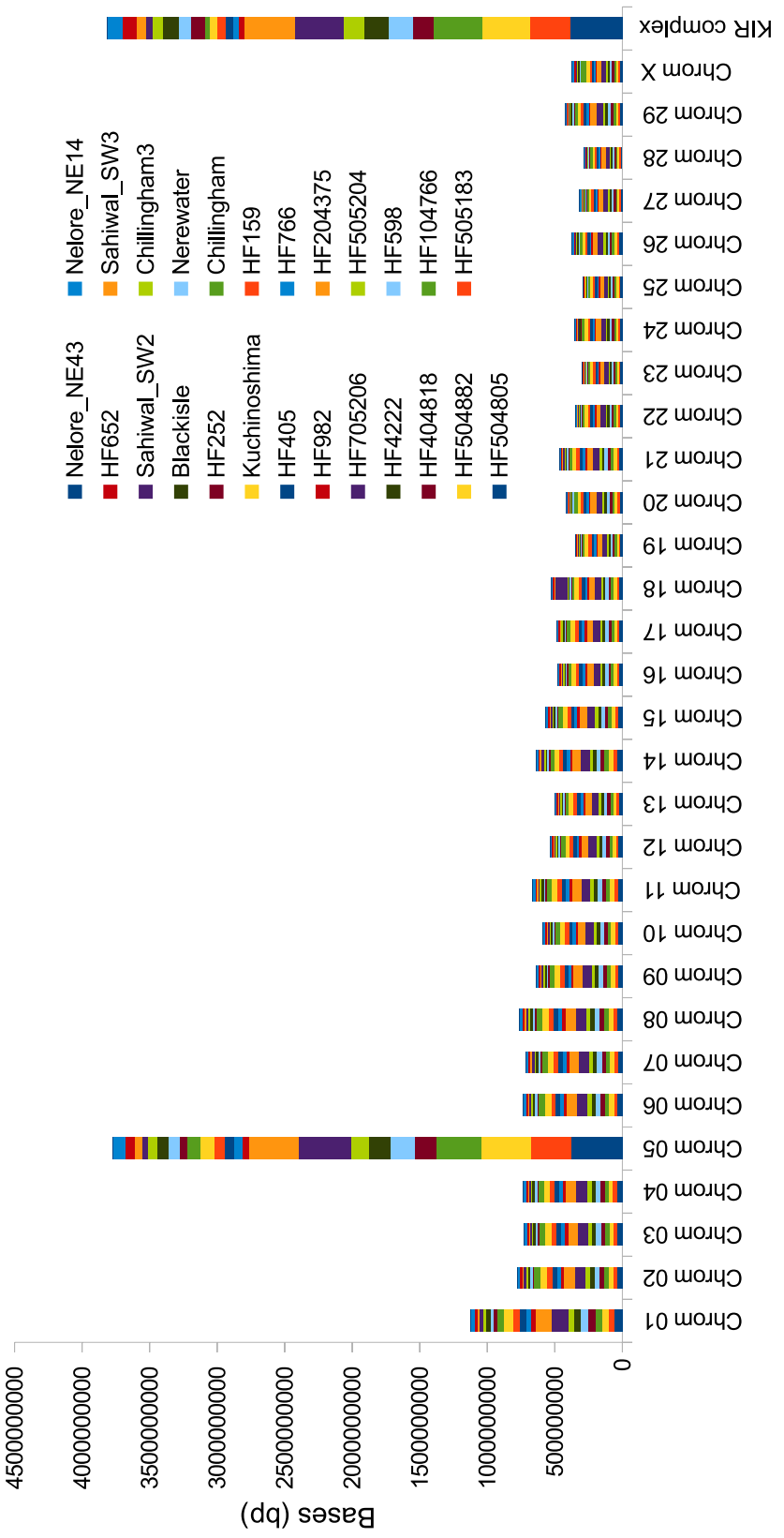


Figure 52: Number of bases aligning per chromosome. Each colour is representative of the number of bases aligning to that chromosome from a particular animal. Chromosomes are ordered by size. The *KIR* complex and chromosome 05 are the target regions. Chromosome 05 contains the target sequences and non-specifically bound sequences outside of the NKC.

This has been exemplified by the animal HF4222 which was the animal used to sequence the *KIR* complex and thus the read coverage represents normal diploid *KIR* content, Figure 54a. The read coverage for HF4222 is uneven through the complex which is a result of sequencing bias and the unequal number of probes used along the complex. The HF4222 read depth coverage can be used as a baseline for comparison of the other animals to. The Kuchinoshima-Ushi read coverage confirms the conclusions of chapter 5, that the *BotaKIR2DS2/3* and *BotaKIR3DXS/3* positions are absent within the Kuchinoshima-Ushi genome, Figure 54b.

6.3.5 Read depth coverage revealed CNV within the cattle *KIR* haplotype

Copy number variation of the cattle *KIR* genes was predicted by calculating the fold increase or decrease of read coverage depth compared to a baseline reference sequence provided by the HF4222 read depth coverage. This further confirmed the absence of *2DS2/3* and *3DXS2/3* loci from the KU *KIR* complex, Figure 55. The process was repeated for all the animals (appendix section 9.5.6) and indicated potential CNV of other *KIR* genes, Figure 56 and Table 21. Animals HF159 and Sahiwal SW2 are predicted to be lacking *BotaKIR3DXS1*, Figure 56 and Table 21, which has been confirmed by expression analysis within HF159 in a separate project within the lab. One of the Sahiwal animals sequenced, SW3, is predicted to encode a heterozygous *KIR* genotype with a full *KIR* content haplotype similar to the HF4222 and a truncated *KIR* haplotype akin to that sequence within the KU.

The Nellore NE14 animal has increased relative read depth coverage over the *2DS1* and *3DXL1* genes, suggesting CNV of these genes within the genome. The HF504805 and HF504882 animals have reduced read coverage over the *2DS3* to *3DXS2* loci which may be a result of a truncated haplotype. It has not been possible to confirm these different haplotype structures yet; either within the lab or through analysis of SNP heterozygosity within the "intact" haplotype.

6.3.6 Different aligners produced different results, thus a combination of aligners were used for SNP detection

To determine the most accurate method of aligning the reads to the highly repetitive *KIR* complex, the alignment was repeated several times with a range of different parameters. No two aligners produced the same results, Figure 57. Aligners were chosen based on capability to align 250 bp reads accurately and the ability

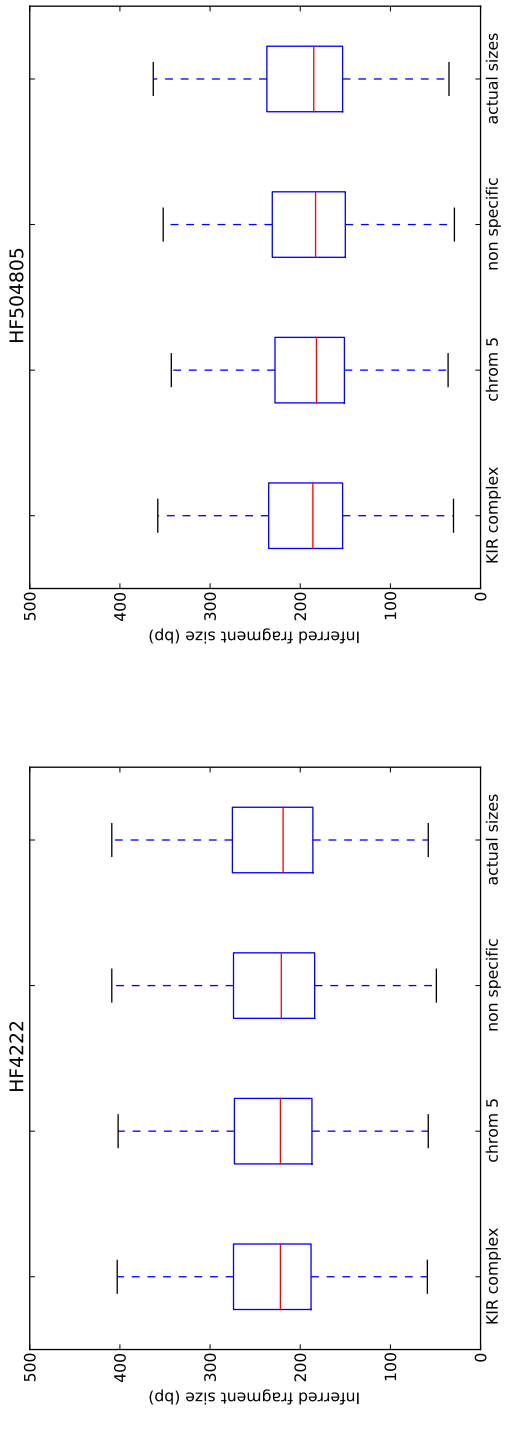
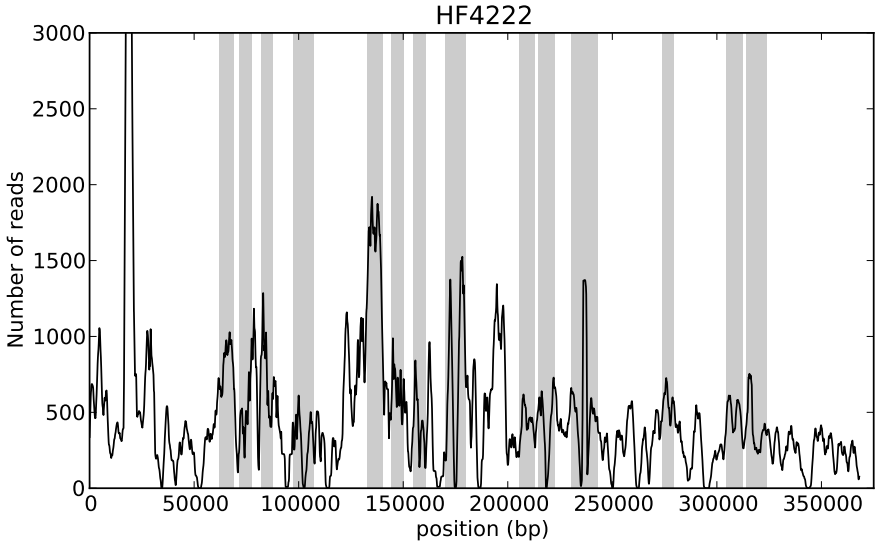
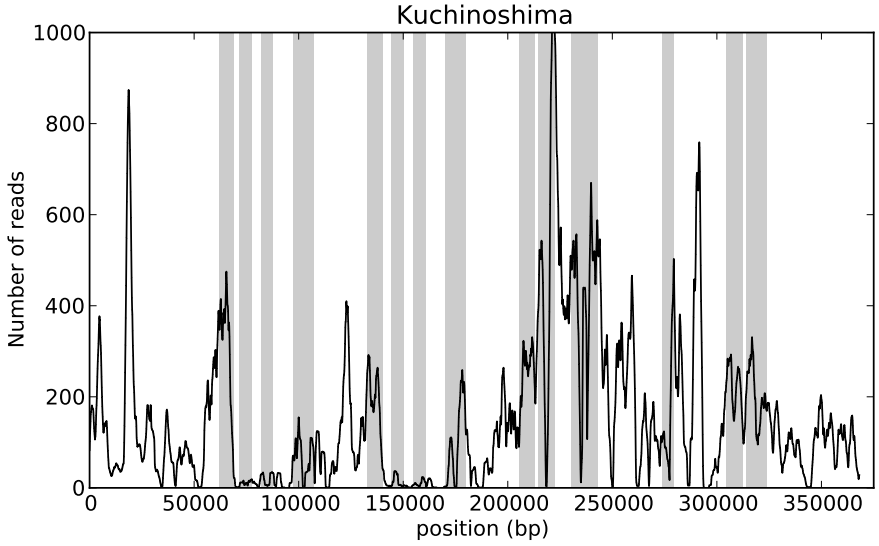


Figure 53: Inferred fragment sizes for the target regions *KIR* complex and NKC and the non-specific regions. Actual sizes have been calculated from joining read pairs for all the reads for that animal. Fragment sizes for animal HF4222 (53a), which is the same animal used to generate the *KIR* complex has higher overall fragment size than animal HF504805 (53b) which was sequenced in the first capture experiment. Further box plots for each animal are shown in the appendix section 9.5.4. No significant differences were found between the the actual sizes and the inferred fragment sizes.



(a)



(b)

Figure 54: Read coverage depth of animals HF4222 (a) and Kuchinoshima-Ushi (b) enrichment capture sequences over the *KIR* complex. The rest are shown in the appendix section 9.5.5 and follow the same pattern exemplified by HF4222. The black line represents a sliding window (1 kb) average of read coverage depth. Read coverage reduced from the different capture experiments leading to lower peak coverage in the Kuchinoshima compared to HF4222 cattle. Grey columns are for reference and represent the *KIR* positions. Positions of *KIR* from left to right are: *3DXL6*, *2DS3*, *3DXS3*, *3DXL7*, *3DXL4*, *2DS2*, *3DXS2*, *3DXL5*, *3DXL2*, *2DS1*, *3DXL3*, *3DXS1*, *3DXL1* and *2DL1*

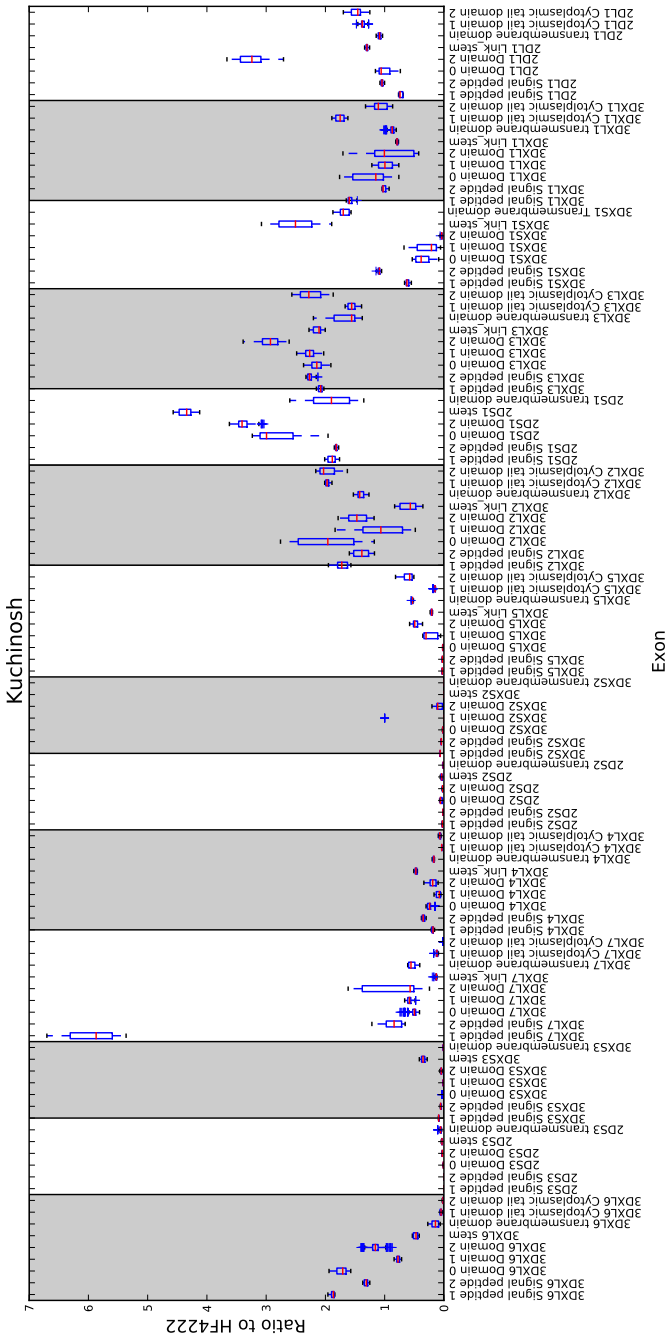


Figure 55: CNV exon prediction of KU. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation. The ratio is relative to the diploid heterozygous 4222 animal which the first haplotypes were sequenced from therefore 1 is indicative of two copies at that locus. Grey and blank columns distinguish genes and exons are labelled along the X-axis. Each box plot contains all of the ratio values for an exon. The ratio has been calculated using the *3′* region of the sequenced haplotype containing NCR1, FCAR including introns and intragenic regions.

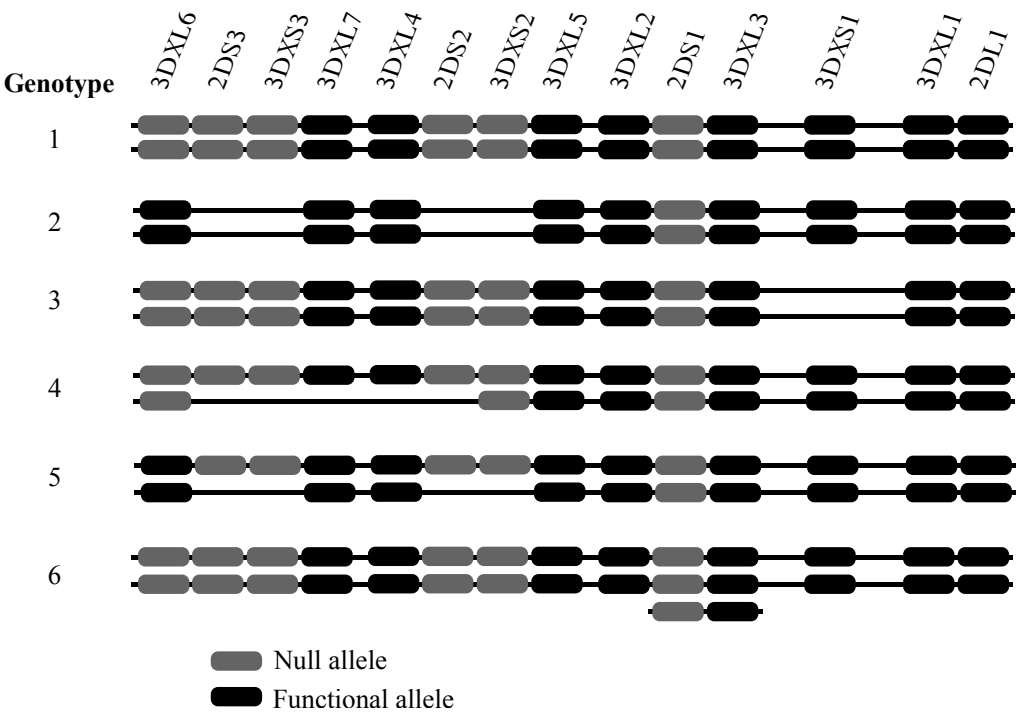


Figure 56: Diagram of predicted *KIR* genotype structures based on CNV analysis in section 9.5.6. Haplotypes cannot be distinguished, therefore these are predicted genotype structures and open to interpretation and verification. Black blocks are predicted functional and grey blocks are null-alleles. Genotypes have been numbered 1 to 6. The genotype of each animal is shown in Table 21.

Animal	KIR Genotype	MHC Genotype
HF504805	4	A10/A14
HF505183	1	A31
HF504882	4	A14
HF104766	1	A10/A31
HF404818	1	A18
HF598	1	A18
HF4222	1	A14
HF505204	1	A14
HF705206	1	A14
HF204375	1	A14
HF982	1	A31
HF766	1	A10/A31
HF405	1	A18
HF159	3	A31
Kuchinoshima	2	Unknown
Chillingham250	1	A10
HF252	1	A18
Nerewater	1	A11/A18
Blackisle	1	A14/A18
Chillingham3	1	A10
Sahiwal_SW2	3	Unknown
Sahiwal_SW3	5	Unknown
HF652	1	A31
Nelore_NE14	6	Unknown

Table 21: Table showing the predicted KIR genotypes from Figure 56. MHC class I genotypes are also shown. Where one haplotype is shown animal is homozygous.

to remove reads that do not uniquely align. BWA-MEM and Bowtie2 shared the highest number of SNPs. Due to the repetitive nature of the *KIR* haplotype and the uncertainties this provides, the conservative approach of using only the SNPs that were reported by both BWA-MEM and Bowtie2 was employed.

The sheer number of SNPs called here makes it impossible to show them all within this thesis. Therefore, the SNPs from within the exon sequences are shown in the appendix tables S3 to S16. Further SNPs within the introns and intergenic regions can be found online `https://github.com/nick297/thesis_scripts/tree/master/data_files`.

6.3.7 Duplicate samples revealed a low error rate (0.12%) but high prevalence of missed SNPs (8.94%)

The enrichment process involves a PCR amplification step with 24 cycles using a high-fidelity taq, which has the potential to introduce errors. Although the Illumina sequencing process produced high quality sequences, there is still the potential for errors to be introduced during sequencing. To determine the error rate of SNPs called from the enrichment and sequencing process, a duplicate sample was used. The SNPs called for each of the duplicate samples was compared to determine unique SNPs for each sample. Unique SNPs between the duplicates are likely to be spontaneous error because they have not been identified within the duplicate sample. The unique SNPs between the duplicate samples were spread evenly over the *KIR* complex, Figure 58. There is a peak in erroneous SNPs over the *LILR* region which is likely due to the increased read coverage in this area. The cattle *LILR* region, like the *KIR* region within the genome build, remains unfinished, therefore the probes have likely picked up more *LILR* genes than the three sequences within the assembled BAC clone sequence used here. Therefore, multiple *LILR* genes are being mapped to single loci generating relatively more SNPs. The duplicate samples were sequenced in the first and third capture experiments. The number of SNPs and unique (erroneous) SNPs from the first capture experiment duplicate were 1513 and 127 respectively. Alternatively, the number of SNPs and unique (erroneous) SNPs from the third capture experiment duplicate were 1762 and 166 respectively. Of these unique SNPs, only 4 from each animal were not detected within any of the other animals sequenced. Finding SNPs within other animals is mathematically very unlikely to be a false positive and more likely to be a failure within the capture experiment to detect the SNPs within one of the duplicates. All of the SNP positions missed between the duplicates had sequence coverage. Therefore, the enrichment process may have biased one haplotype between the duplicates. Alternatively slight contami-

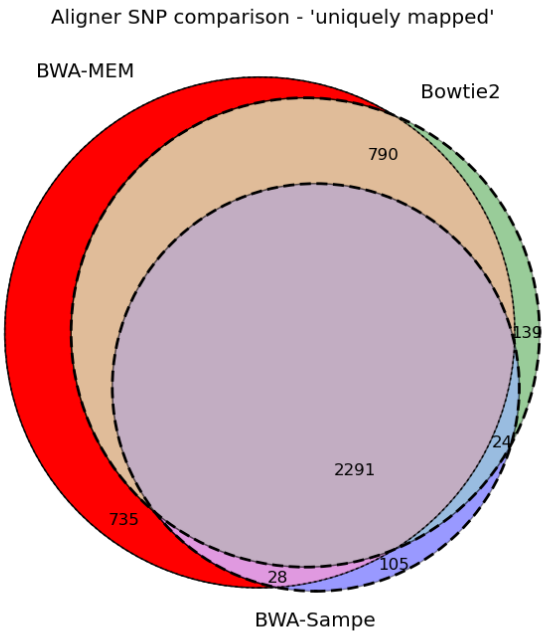


Figure 57: Venn diagram showing the three most popular aligners and the shared SNPs between them. Using the same dataset and removing reads that do not map uniquely SNPs were compared between the different aligners. SNPs unique to BWA-Mem (red), Bowtie2 (green) and BWA-sampe (purple) are shown. SNPs that are shared between two or three of the aligners are also shown within the diagram.

nation or bleed over during sequencing may have affected the results to generate missed SNPs. The error rate for false positive SNPs has been calculated at 0.12% and the enrichment and sequencing procedure has a 8.94% margin for missing SNPs.

6.3.8 Total SNP numbers per animal varied as did relative proportions of SNP numbers per intergenic, intron and exonic sequence

The number of SNPs varies between each animal, with the KU, Nelore NE14 and Sahiwal each containing over 5,000 SNPs over the *KIR* complex, Figure 59b. The Nelore N43 has the lowest number of SNPs which was a result of the low yield sequencing. Therefore, the results from this animal are unreliable. There is a trend between the number of *KIR* SNPs within an animal and their MHC genotype. The cattle with A10 and/or A18 MHC haplotypes have approximately 1,500 SNPs over the entire haplotype, Figure 59a. The cattle with A14 or A31 MHC haplotypes have over 2,000 SNPs over the entire haplotype. The exception here is 4222 (A14/A14) which has effectively a single haplotype sequenced here. This may be an indication of the back breeding process used to generate the homozygous MHC class I genotypes, which has propagated SNPs within the *KIR* complex of related animals. This may also be an indication of artificial selection increasing or decreasing diversity within the *KIR* complex.

The proportional number of SNPs within the *KIR* exons, introns and intergenic sequences varies with the total number of SNPs, Figure 59b. The animals with over 2,000 SNPs within the *KIR* complex have proportionally more SNPs within the intron and exon sequences than the animals with less than 2,000 SNPs. This further points to modulation of *KIR* sequence diversity within cattle with certain MHC genotypes, which is likely a product of back breeding and not the MHC molecules. Shared SNPs between the animals will confirm this observation.

6.3.9 Shared SNPs within the *KIR* loci between animals is likely a result of back breeding for homozygous MHC haplotypes

By using a pairwise comparison of shared SNP positons between each animal it has been possible to infer phylogenies based on called SNPs, Figure 60. Each SNP for each animal is compared to all the other animals to determine the number of SNPs within each animal that is not present within the other animals. This similarity matrix, when plotted as a dendrogram, clearly segregates the different MHC genotyped animals, Figure 60. The A10/A18 animals form a group with

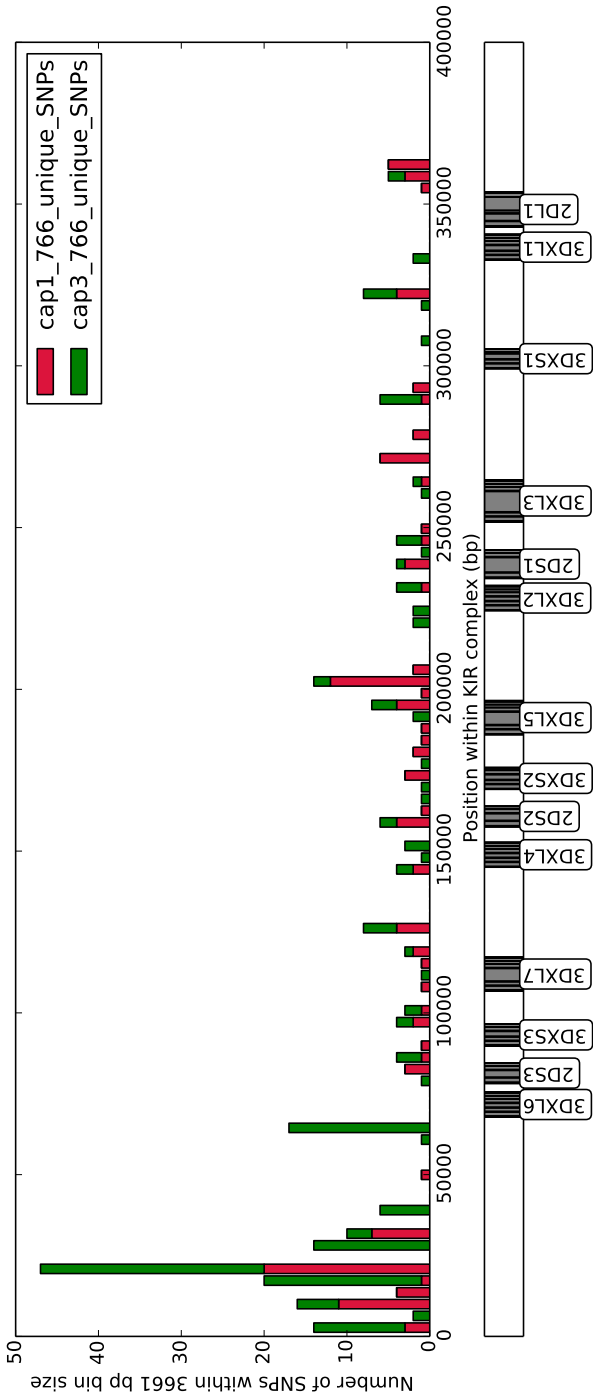
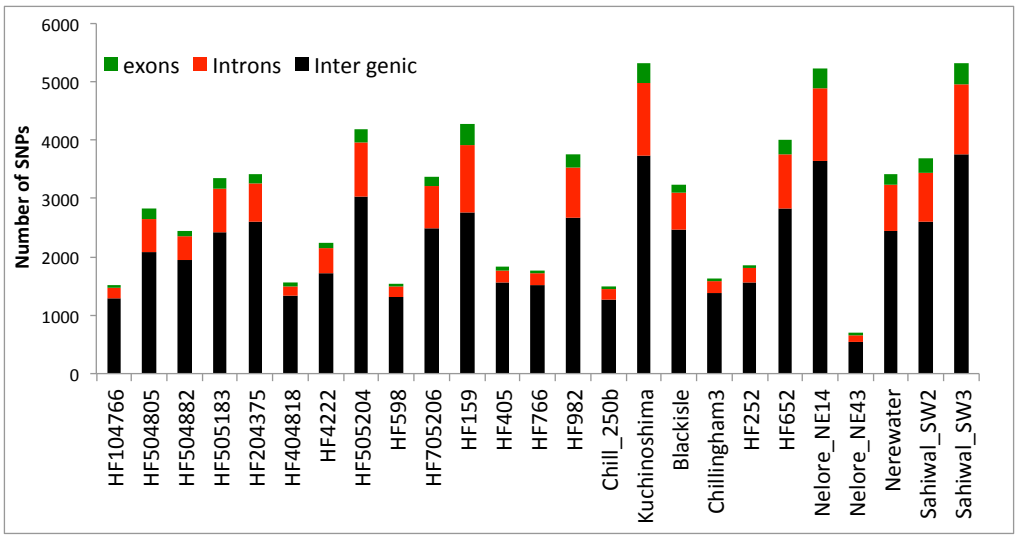
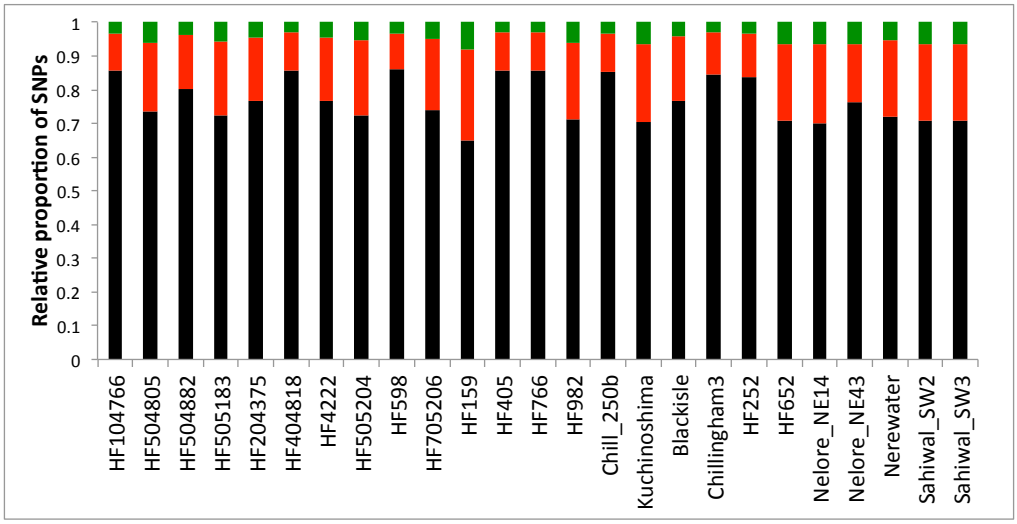


Figure 58: Histogram of missed SNPs over the cattle *KIR* complex. Red bars indicate SNPs from the capture 1 HF766 duplicate and green bars represent SNPs from the capture 2 HF766 duplicate. The 5′ (left) of the chart shows many SNPs over the region containing the *LILR*. The *KIR* positions are labelled along the X-axis. Grey represents the gene introns and black represents the exons.



(a)



(b)

Figure 59: Total numbers of SNPs called within each animal over the entire *KIR* complex (a) and proportional representation of the total numbers of SNPs within the intergenic (black, bottom), intron (red, middle) and exon (green ,top) sequences (b).

shorter branch lengths, which is a result of fewer SNPs differences and fewer SNPs within these animals. The A14/A18 are different to the A14/A31 because the A14 haplotype originate from different sources each with divergent *KIR* haplotypes associated. Analysis of the shared SNPs over the haplotype between the different animals has revealed linkage based on MHC haplotype, as well as variable levels of diversity which may be a result of the breeding programs used to generate the MHC herd. The positions of these SNPs within the *KIR* complex will reveal areas high diversity which may be an indication of selection pressures influencing *KIR* evolution.

6.3.10 Total SNP positions reveals a gradient of SNP frequency over the *KIR* complex

SNPs were called over the *KIR* complex for each of the HF animals sequenced. The number of positions over the *LILR* region at the 5′ of the *KIR* is significantly higher than the rest of the complex, Figure 61. This is likely a result of the several *LILR* sequences aligning to just the three *LILR* loci in the reference sequence. Therefore, multiple genes are aligned to the wrong position causing the level of polymorphisms to be artificially high within the *LILR* complex. The dairy cattle sequenced here appear to have a higher frequency of SNPs within and around the *KIR* genes at the 3′ of the complex. There appears to be a trend to the 5′ of the *KIR* complex which, with the exception of *3DXL6*, has fewer SNPs, Figure 61. This supports the hypotheses postulated in chapter 2 which states that the 5′ blocks have evolved more recently and thus have fewer polymorphisms. The notable exceptions are *3DXL6* and *2DL1*, which both appear to have a significantly higher number of SNPs in and around their gene sequence.

6.3.11 *BotaKIR3DXL3* contains the largest number of polymorphisms

The total number of SNPs within the functional genes collated for all of animals shows that *BotaKIR3DXL3* contains the largest number of polymorphic positions within the coding sequence, Figure 62. *BotaKIR3DXL1* has the lowest quantity of SNP positions, indicating a potentially conserved function for this gene. There are more non-synonymous than synonymous SNPs indicating greater changes in protein sequence and potentially receptor function as a result of nucleotide changes.

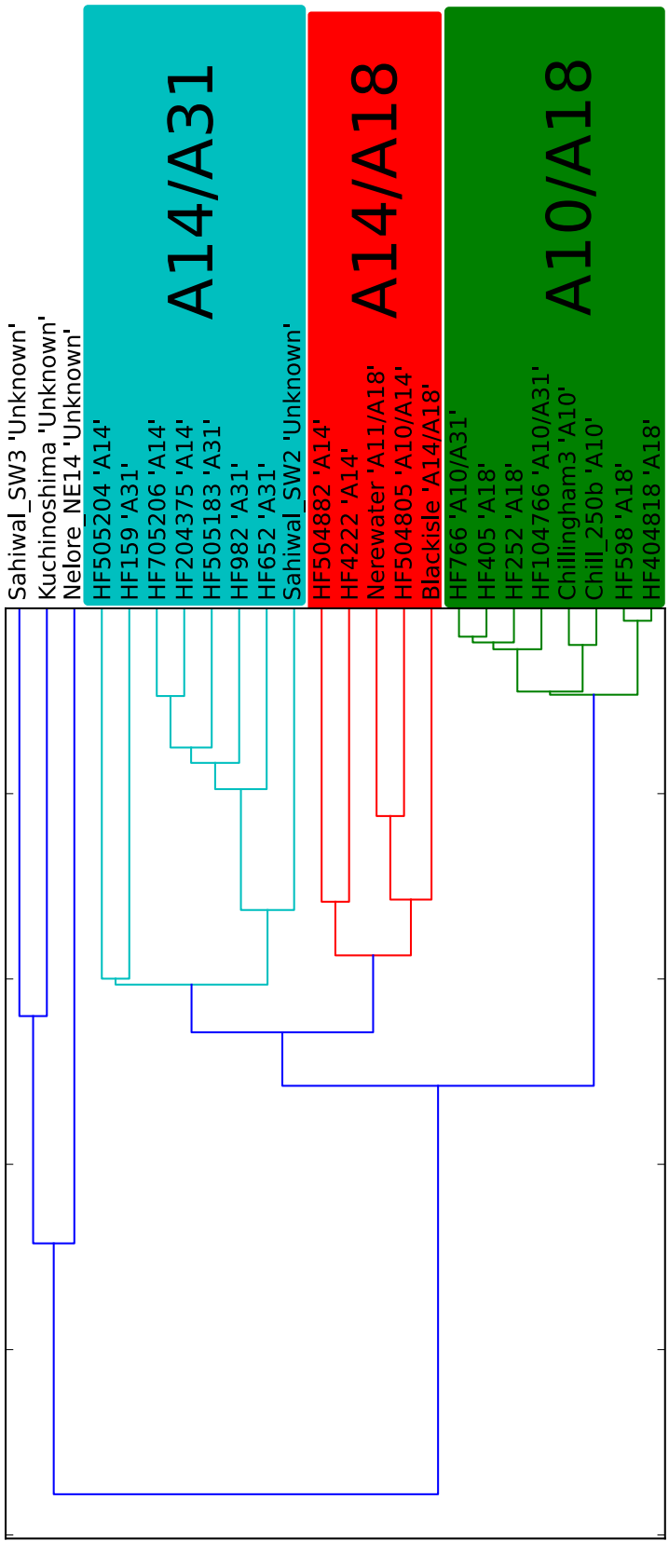


Figure 60: Dendrogram of SNP differences between animals. Animals that share the fewest SNP differences have the shortest branch lengths between nodes. The entire *KIR* complex is used without the *LILR* genes.

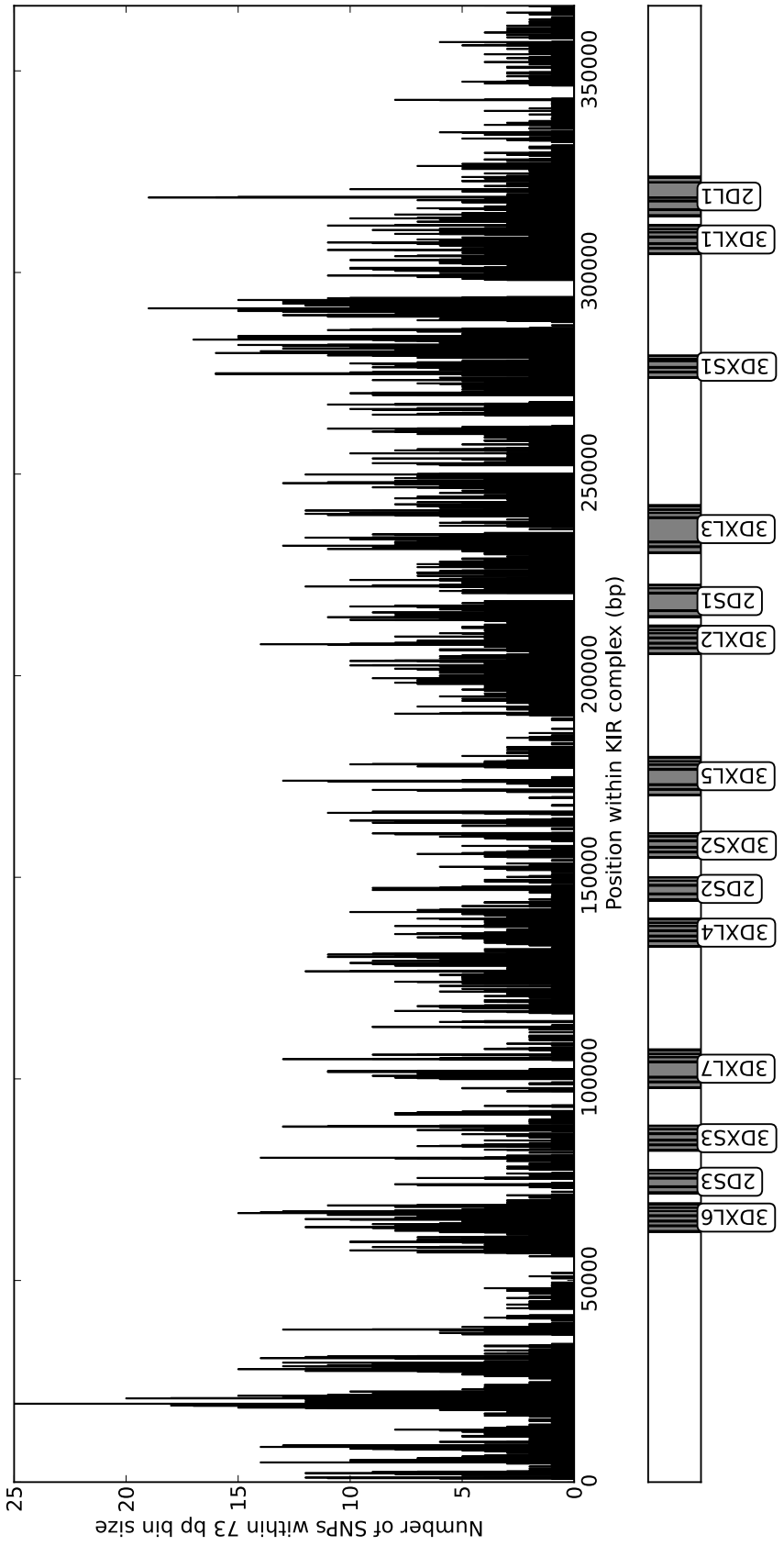


Figure 61: Histogram of dairy cattle SNPs over the entire cattle *KIR* complex. A bin size of 73 bp was used. *KIR* positions are shown along the X-axis, grey boxes represent gene positions and black boxes represent exon positions. The peaks in SNP numbers at the 5′ end are over the *LILR* genes and therefore are probably representative of several *LILR* loci mapped to just three loci.

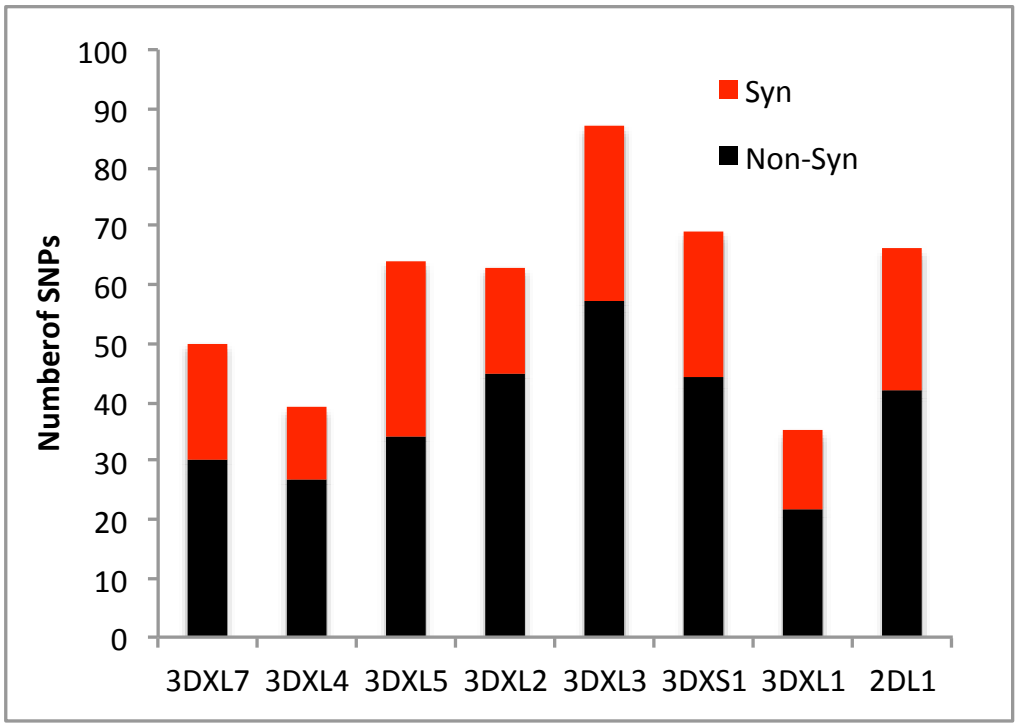


Figure 62: Stacked bar chart showing the synonymous and non-synonymous SNPs within each functional *KIR* gene coding sequence.

6.3.12 Polymorphisms are focused within the Ig domains and the transmembrane domain of *3DXS1*

The majority of collated SNPs from all of the animals are focused within the Ig domains of the functional *KIR* genes, Figure 63. This is likely to be the result of ligand mediated selection pressures influencing the evolution of the domains that interact with MHC class I. The other non-Ig domains have relatively low levels of polymorphisms with the exception of the stem and transmembrane domains of *BotaKIR3DXS1*. This is the only predicted functional *KIR* gene within the complex and may be an indication of attenuation of the activating function within this gene. The functional arginine residue (residue 332) within transmembrane domain of *3DXS1* has a non-synonymous change to glutamine within both the KU and Sahiwal SW3 animals, Supplementary Table S10. Furthermore, two more arginine residues (339 and 346) within *3DXS1* transmembrane domain have undergone residue changes to serine and glutamine respectively. This may impact the ability of the receptor to recruit the adapter molecule and activate the NK cell. Therefore, alongside the polymorphisms within the Ig domains, this may be evidence of the attenuation of the 3DXS1 receptors ability to recognise ligand, signal and activate the NK cell. It is now hypothesised that as certain genotypes do not contain the *3DXS1* gene, this gene is being actively attenuated and deleted because it has served its function.

6.3.13 Non-synonymous SNP numbers within the functional KIR genes indicates locus specific modulation of different Ig domains

The exon with the highest number of SNPs is *2DL1* Ig 2, which is dominated by SNPs from HF652, HF159 and HF505183, Figure 64, and KU and the Sahiwals Figure 66. The other domains with high number of SNPs include *3DXL6* Ig2, *3DXL5* Ig 1, *3DXL2* Ig 0 and *3DXL3* Ig 1. Interestingly there is no trend for SNPs within a particular exon for all of the *KIR*. Instead each *KIR* has a majority of SNPs within either the Ig domain 0, 1 or 2. This is true for both the MHC herd animals, Figures 64 and 65, and the non-MHC herd animals, Figures 66 and 67. A majority of SNPs are from the same animals; HF652, HF982, HF159, HF505204 and HF505183 within the MHC herd, Figures 64 and 65, and KU and Sahiwals in the non-MHC herd animals, Figures 66 and 67.

SNPs within the KU and Sahiwals are expected because of their relatively distant (geographically and evolutionary) relationships to HF cattle. However, this level of polymorphisms within HF cattle is notable, suggesting a lot of functional

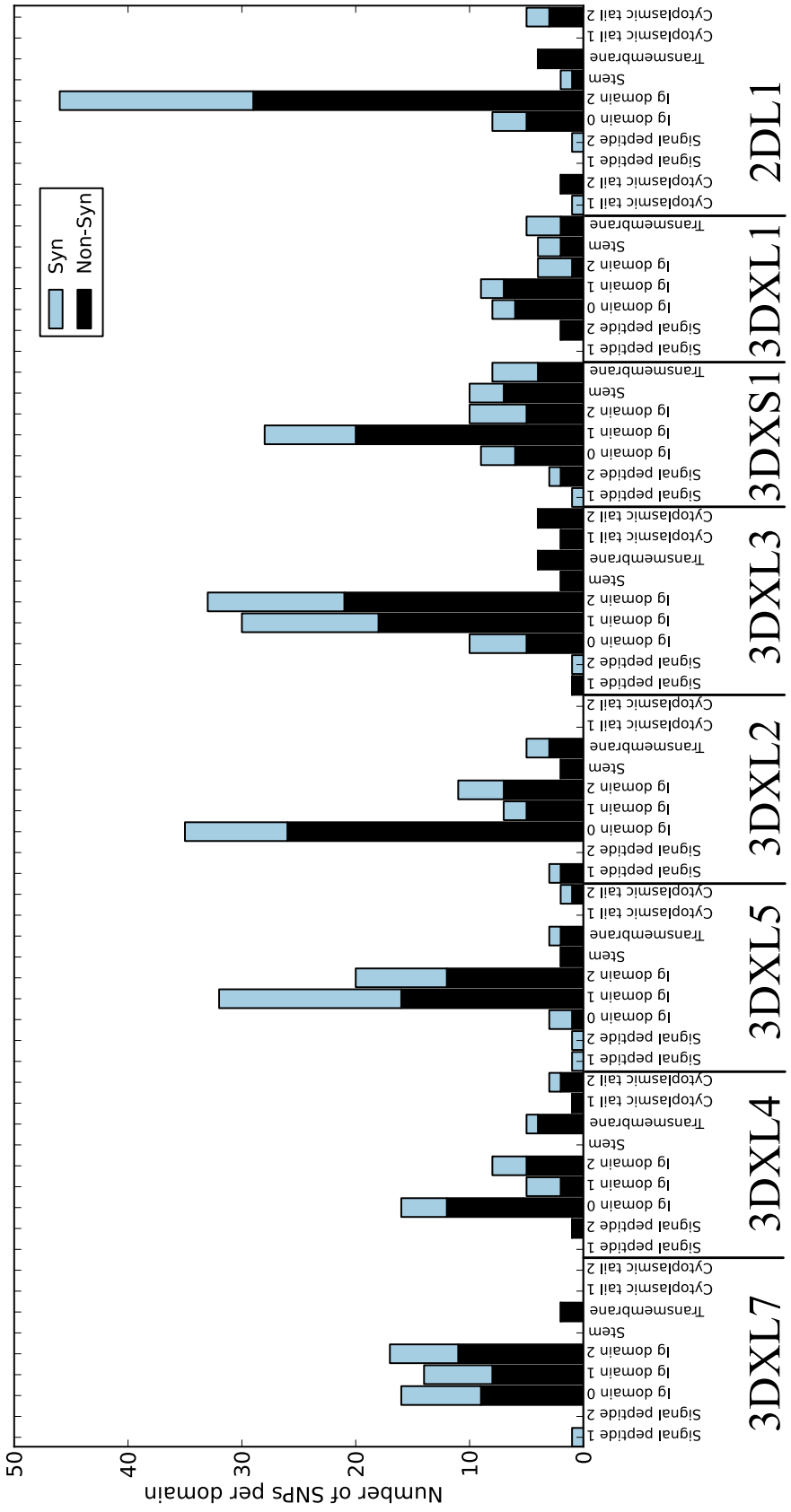


Figure 63: Stacked bar chart showing the synonymous and non-synonymous SNPs within each *KIR* exon. Each bar is representative of a domain, labelled on the X-axis.

variation within the *KIR* of the dairy breed. There is a notable lack of diversity within the *3DXL1* and *3DXS1* genes with the majority of SNPs for the latter coming from five animals. Nonetheless the majority of non-synonymous SNPs in *3DXS1* are within the D1, this could indicate a region of sequence divergence away from sharing identity with *3DXL1*. A higher resolution analysis by phasing the SNPs and determining full sequence alleles is required before conclusions can be drawn from this dataset. Although care has been taken to avoid ambiguously mapped sequencing fragments, the data cannot be fully trusted until confirmed further. The SNP data collated here will enable primer and probe design for genotyping strategies that would have otherwise failed due to the highly polymorhpic sequences.

6.3.14 The KU and Sahiwal SW2 have a predicted functional *3DXL6* allele

There is a homozygous cytosine insertion at gene position 3984 of *BotaKIR3DXL6* in the KU and Sahiwal SW2 genomes, Table 22. This insertion reverses the reference sequence frame-shift mutation that causes a premature stop codon within the gene. It is therefore believed that the KU and SW2 cattle contain a functional copy of *BotaKIR3DXL6*. There is an alternative insertion of a thymine at position 9834 which is heterozygous within HF505183, HF204375, HF505204, HF705206, HF982 and homozygous within HF159. This insertion may also produce a functioning *BotaKIR3DXL6*, however, further deletions downstream of this insertion within all of these HF cattle may counteract this effect. Therefore, it is believed that only the KU and SW2 have functioning *BotaKIR3DXL6* alleles.

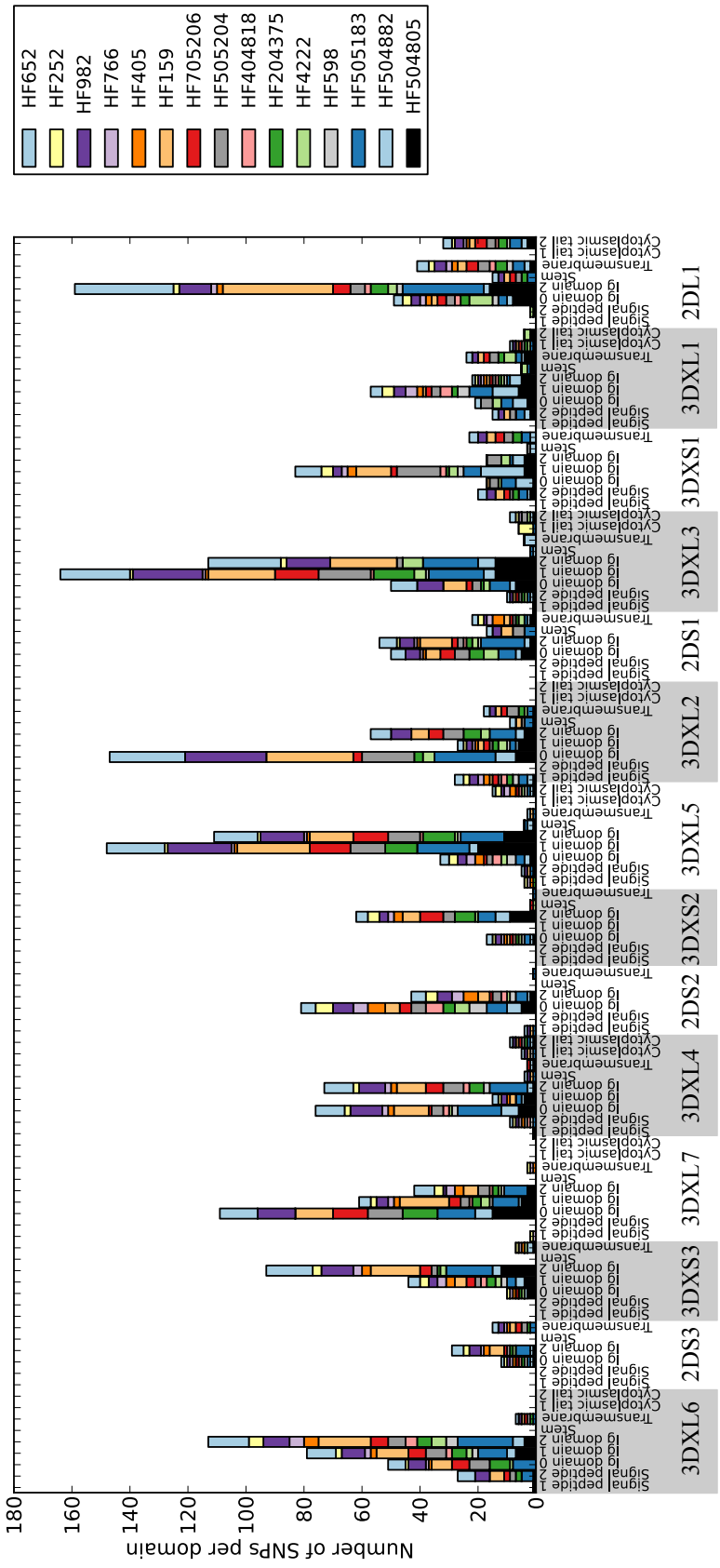


Figure 64: Stacked bar chart showing the numbers of SNPs for each capture animal within each *KIR* exon. Each bar is representative of a domain, labelled on the X-axis. Each bar is broken into stacked bars representative of each animal, each animal is colour coded and described within the legend.

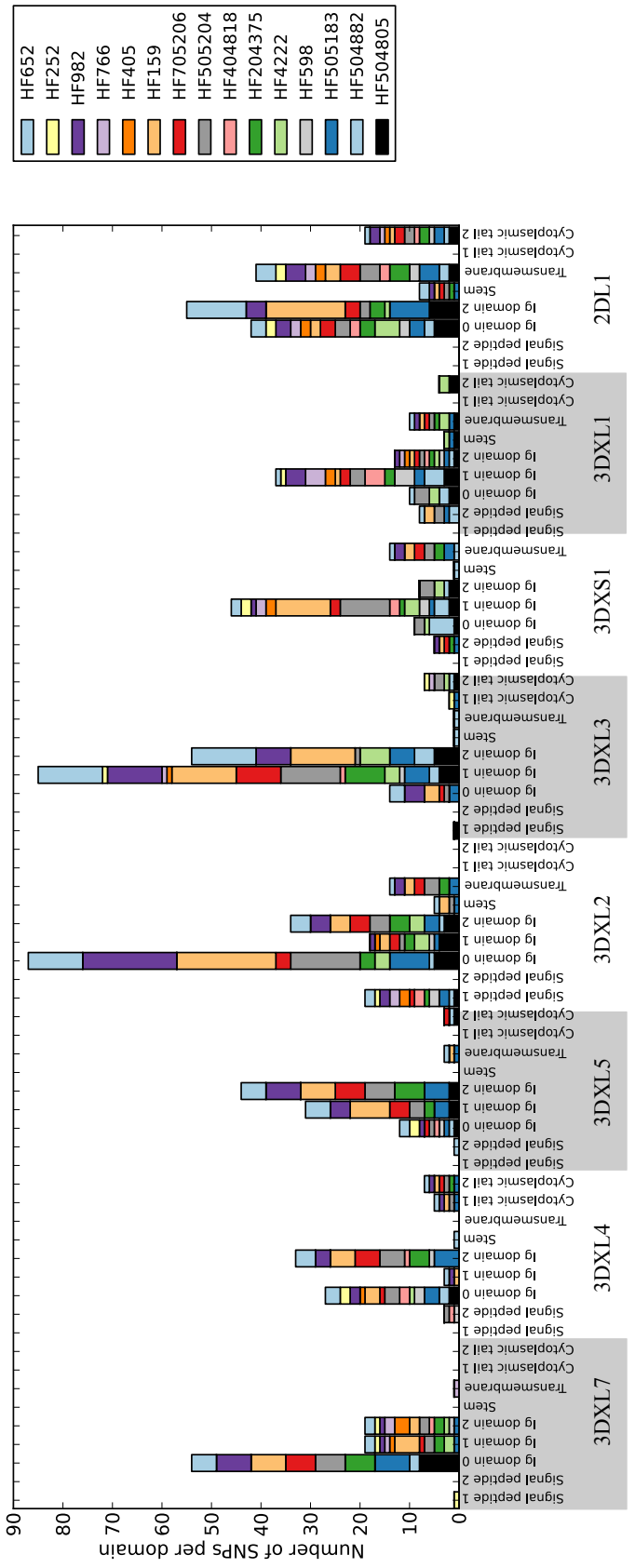


Figure 65: Stacked bar chart showing the numbers of SNPs for each capture animal within each *KIR* exon. Each bar is representative of a domain, labelled on the X-axis. Each bar is broken into stacked bars representative of each animal, each animal is colour coded and described within the legend.

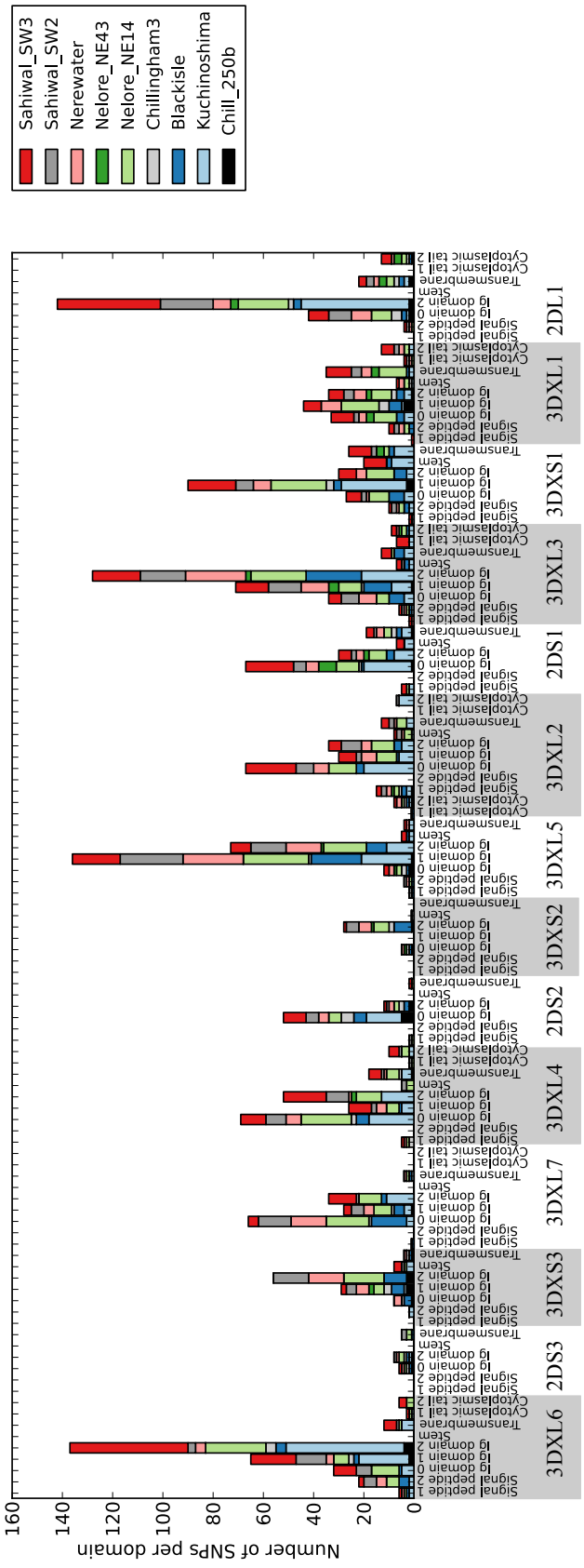


Figure 66: Stacked bar chart showing the numbers of SNPs for each capture animal within each *KIR* exon. Each bar is representative of a domain, labelled on the X-axis. Each bar is broken into stacked bars representative of each animal, each animal is colour coded and described within the legend.

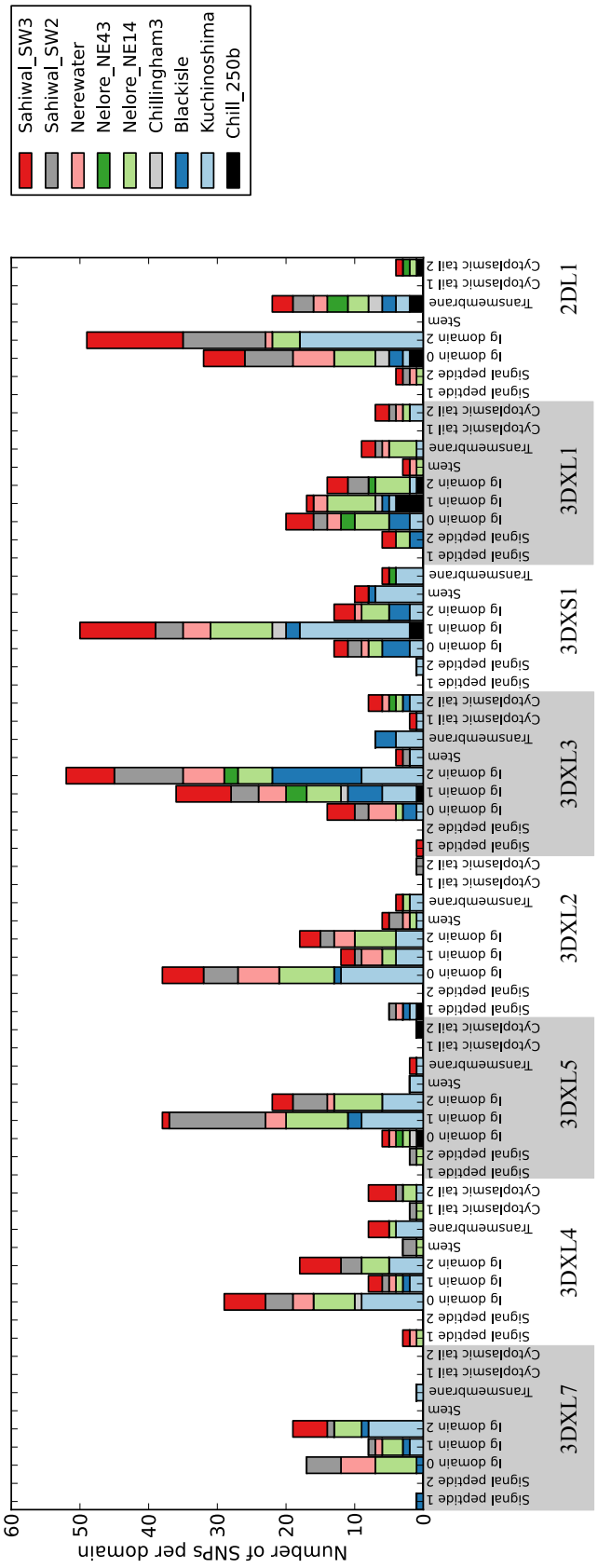


Figure 67: Stacked bar chart showing the numbers of SNPs for each capture animal within each *KIR* exon. Each bar is representative of a domain, labelled on the X-axis. Each bar is broken into stacked bars representative of each animal, each animal is colour coded and described within the legend.

HaplotYpe_pos	gene	Domain	gDNA_pos	cDNA_pos	HaplotYpe_ref	HF104766 %	HF504805 %	HF504882 %	HF505183 %	HF204375 %	HF404818 %	HF4222 %	HF505204 %	HF598 %	HF705206 %	Chillingham %	HF159 %	HF405 %	HF766 %	HF982 %	Kuchinoshima %	Blackisle %	HF252 %	Nelore_NE14 %	Nerewater %	Chillingham3 %	HF652 %	Sahiwal_SW2 %	Sahiwal_SW3 %
64386	3DXL6	D 0	1416	131	CCG				C 49	C 49			C 46		C 49		C 95			C 53				C 41			C 37	C 90	
66847	3DXL6	D 2	3877	670	GT					G 21							G 34												
66953	3DXL6	D 2	3983	776	C				CT 26	CT 49			CT 30		CT 34		CT 89			T 27	CT 42			CT 81			CT 28	CT 89	CT 54
66954	3DXL6	D 2	3984	777	T																TC 94								TC 81
67075	3DXL6	D 2	4105	898	TCAGACTC																T 26								
85718	3DXS3	D 1	2650	489	A		ATGC	ATGC 28	ATGC	ATGC		ATGC 21			ATGC 37	ATGC 48	ATGC 34					ATGC 31		ATGC 32	ATGC 41	ATGC 45		ATGC 42	
104994	3DXL7	D 2	6369	677	AG				A 24			A 27			A 22		A 90										A 27		
136314	3DXL4	D 1	2673	490	ATGC						A 22																		
136328	3DXL4	D 1	2687	504	GC		G 23				G 22										C 80			G 26					C 67
140696	3DXL4	CT 2	7055	1290	CGAAATTT																C 67								
146866	2DS2	D 0	1761	309	CT																TC 75								TC 22
146906	2DS2	D 0	1801	349	T															T 21							T 27	T 20	
158228	3DXS2	D 1	2540	380	CT	C 100			C 100								T 35							T 26			A 54	A 90	
171352	3DXL5	SP 2	179	37	TTC												A 94			A 44							T 34	T 67	T 43
215654	2DS1	SP 2	290	67	AG				A 49	A 45		T 39	A 47		A 55		T 60			T 31	T 58			T 53	T 30		T 24	T 26	
217118	2DS1	D 0	1754	304	TC		T 25		T 28	T 27		T 25	T 58		T 34		T 32	T 20	T 29	T 32		T 29	T 31		T 29				
217162	2DS1	D 0	1798	348	TC	T 31			T 35	T 40	T 31	T 25	T 36	T 33	T 25	T 25	T 32			T 21		G 48						G 33	
231499	3DXL3	SP 2	178	37	TTC			G 42	T 31	T 28			T 22		T 36		T 28												
243053	3DXL3	CT 2	11732	1186	GA												T 31										T 24	T 46	
274896	3DXS1	SP 2	178	37	TTC												A 32												
277350	3DXS1	D 1	2632	490	ATGC										G 24	G 26	G 74											G 59	
277364	3DXS1	D 1	2646	504	GT					G 27		G 29					GT 96			GT 48	GT 97	GT 39		GT 94		GT 53		GT 88	GT 95
280295	3DXS1	TM	5577	1088	G			GT 63	GT 55	GT 51		T 21	GT 97		GT 54						T 28								
316616	2DL1	D 0	1808	350	TC																								
318659	2DL1	D 2	3851	543	TCTGGGTC																								

Table 22: The indel for each animal is shown along with the percentage of reads aligning that contain the indel. Indels are based on the leftmost base of the reference (haplotype_ref). Therefore an indel of C with reference CCG is a deletion of CG after the leftmost C. Positions with the gene (gDNA pos) and predicted cDNA (cDNA pos) sequence are shown.

6.4 Discussion

Within this chapter the *KIR* complex sequences for fifteen animals within the MHC-herd at Compton and a further nine individual cattle have been enriched and sequenced using a bespoke process never used before in ruminants. Of the nine different animals, two are from the Chillingham herd in Northumbria including one bull (Chillingham 250) and a heiffer (Chillingham 3), four *Bos indicus* cattle including two Nelore and two Sahiwal cattle, although one of the Nellore (NE43) sequencing was unsuccessful, and two bulls from the British Friesian breeding stock which will account for a high proportion of the dairy herds in the UK. A duplicate sample was used from a HF cattle for error rate detection, which was determined to be approximately 8.94% of SNPs called. The sequences of *KIR* exons generated here has enabled sequence variant calling, a powerful indicator of functional variation between receptors and will enable future studies to characterise the KIR functions.

6.4.1 Limitations of the capture experiment

There are a number of limitations with the approach taken in this study. Firstly the missed SNP rate is very high. At 8.94%, one in eleven SNPs could be missed by the process. Therefore, the results within this chapter need to be confirmed by further targeted sequencing before they can be properly considered SNPs. However, the distribution of missed SNPs appears to be even over the *KIR* complex and are not associated with any *KIR* or group.

A further limitation with this approach is the bias towards only the *KIR* previously described. The probes have been designed against the *KIR* complex sequenced in chapter 2 and is therefore less likely to be capable of picking up novel *KIR* genes. The probes are designed with a specificity of approximately 80% and therefore should be capable of extracting *KIR* sequences with a sequence similarity of 80% and higher.

If the probes have picked up novel *KIR* sequences, it is unlikely the bioinformatic pipeline employed here would be capable of detecting that they were novel. The raw sequencing reads are aligned to the *KIR* complex reference sequence and therefore variation as a result of a novel gene would be detected as an allelic variation.

So far it has not been possible to link the SNPs and predict allele sequences. This may be possible by haplotype phasing however the reduced read coverage over certain repetitive regions or regions with low probe density have prevented this approach. Linking SNPs locally within exons is achievable and will be a focus

in the future. This, alongside hereditary information will enable SNP linking and full length allele sequences to be established.

6.4.2 SNPs focused within the Ig domains suggests ligand mediated selection pressures

The majority of non-synonymous SNPs identified were within the Ig domain sequences of the *KIR*. As these receptor domains interact with ligand it is likely that the SNPs are the result of ligand mediated selection pressures. These might include variations within the MHC ligands or the pathogen peptides they express. The extent of the polymorphisms shown in the Ig domains of the *KIR* genes suggests that the HF breed, despite intensive inbreeding, has maintained diversity. This may have been the effect of a large founding group of individuals used to breed the HF, or it is the result of rapid diversification within the breed. To understand whether these SNPs have evolved within the HF breed, other breeds would have to be sequenced to determine if they are breed specific.

6.4.3 No preference for Ig domain SNPs

Across the cattle studied in this chapter each *KIR* appears to have diversified the Ig domains by the introduction of non-synonymous mutations. However, there is no trend to a single Ig domain over all the *KIR*. Instead each *KIR* appears to have diversified an Ig domain independently to the other genes within the group. This could suggest that each gene within groups I and IV recognise a different ligand. Polymorphisms within specific Ig domains may improve the stability of interactions between the receptor and ligand depending on the combinations of KIR and MHC. Therefore the SNPs focused within specific Ig domains might be in response to complementary changes within the MHC molecule. However, the interaction between KIR and MHC in cattle has not been proven. The work of this project has been to enable these interactions to be interrogated with greater confidence.

6.4.4 Further variation within the *KIR* complex

It was discovered in chapter 5 that the KU *KIR* complex was truncated with the removal of four null-alleles. The results in this chapter have confirmed this discovery and have also detected a single nucleotide insertion within the D2 exon sequence of *BotaKIR3DXL6* in the KU and one of the Sahiwal genomes. This insertion is predicted to change the gene from a null-allele to a functional copy. Therefore, the KU, despite containing fewer *KIR* than the HF, encodes an extra

functional inhibitory *KIR*. It is predicted that the KU and Sahiwal diverged from the rest of *Bos taurus* before the inactivation of *BotaKIR3DXL6*. Therefore this gene may have never become a null-allele within the KU and Sahiwal. There could be a functional reason why this gene has been maintained as a functional copy in the Kuchinoshima such as an MHC ligand that remains in this population but not the HF.

6.4.5 Conclusions from determining polymorphisms within cattle *KIR*

The work in this chapter has shown that the cattle *KIR* sequences are highly variable within the Ig domains, containing many polymorphisms that may affect receptor binding to ligand. It has confirmed that the *KIR* complex has remained unchanged within the HF breed and has confirmed that the Kuchinoshima-Ushi has a truncated complex. The data from this project will provide a resource for further studies to begin to interrogate the function of the KIR. This data can now be used for genotyping cattle by designing primers and probes within the conserved regions where no or few SNPs and indels have been discovered. The effects of these SNPs could be analysed in a reverse genetics approach interrogating KIR and MHC restriction and binding.

7 Chapter 7. Discussion

7.1 Summary of findings

The aim of this project was to determine the genetic mechanisms responsible for generating diversity within the cattle NK cell receptor repertoire, the focus was on the *KIR* genes as the most expanded of the NK cell receptor families in cattle. This required fully sequencing a cattle *KIR* haplotype sequence to determine how *KIR* evolved and the level of sequence diversity within the species. Previously, several *KIR* cDNA sequences had been identified but the extent of the of their expansion was unknown as a full haplotype had not been sequenced. The extent of sequence polymorphism within cattle *KIR* genes had not been studied due to the uncertainty of how many *KIR* genes are in the haplotype. Furthermore the *KIR* gene sequences in related species had not been investigated, therefore the extent to which cattle shared *KIR* genes with related species was unknown.

In this project, a HF cattle *KIR* haplotype has been sequenced and assembled, these *KIR* genes have been identified within an ancient cattle genome, a sheep *KIR* haplotype has been sequenced and assembled for comparison with cattle, polymorphisms and gene presence absence has been determined in multiple related breeds and species to HF cattle. This work has revealed the genetic mechanisms involved in shaping the cattle *KIR* gene repertoire and has provided a platform for future investigation into the receptor functions.

7.1.1 Cattle *KIR* have expanded through block duplication

The first cattle *KIR* haplotype has been fully sequenced and assembled revealing a gene dense and repetitive immune complex. The cattle *KIR* genes have expanded within the haplotype through block duplication resulting in groups of genes sharing highly similar sequence identities. Cattle *KIR* have evolved from at least two ancestral mammalian genes that have independently expanded in the ruminant and primate genomes. Contrary to previous understanding, cattle have expanded *KIR* from both the X and L-lineages, however, only one L-lineage *KIR*, *BotaKIR2DL1*, is predicted to encode a functional receptor. The resulting expanded cattle *KIR* haplotype is dominated by predicted functional inhibitory genes and predicted non-functional activating genes. These inhibitory receptors are predicted to enable the NK cells to recognise multiple different MHC class I ligands. This suggests the *KIR* in cattle robustly enable NK cell education, generating licensed NK cells capable of killing MHC class I suppressed host cells. It is therefore predicted that KIR expressed by cattle NK cells play an important role in the ability of the innate immune system to detect and kill virally

infected host cells. This is due to the NK cells ability to recognise host cells with virally down-regulated MHC class I genes and with licensing to kill the cells. The non-functional activating receptors are predicted to have evolved as a result of selection pressures from virally derived MHC decoy proteins. This selection pressure must have subsided for the activating genes to become non-functional through the introduction of premature terminating stop codons. The paired receptor genes *KIR3DXL1* and *BotaKIR3DXS1* are predicted to be the result of this selection pressure currently acting upon the haplotype. However, further investigation is required into the function of these receptors.

The reasons for specifically expanding NK cell receptor gene families is unknown, as cattle also encode a functional *KLRA1* gene, the question remains why have cattle evolved to use diverse KIR receptors instead of the KLRA? Additionally cattle may have expanded CD94/NKG2A receptors [10], providing further complexity of cattle NK receptor gene evolution. Furthermore why some species, such as dogs, seals and bats, have not expanded any NK cell receptor gene families remains unknown. Therefore the triggers for expanding NK cell receptor gene repertoires is uncertain.

7.1.2 Cattle *KIR* have evolved through natural selection

The aurochs genome contains the same *KIR* loci as the Holstein-Freisian (HF) cattle *KIR* haplotype that has been sequenced and assembled. Novel *KIR* genes and gene order could not be confirmed using the short reads sequenced from the ancient aurochs genomic DNA. It has been possible to conclude that the HF has not gained *KIR* since domestication began from the aurochs cattle approximately 10,000 years ago. Therefore, the cattle *KIR* haplotype has evolved as a product of natural selection and not through the artificial selection of domestication. It has been suggested that the structure of the complex had been generated via centuries of domestication meaning the genes may not have evolved entirely through natural selection. Domestication may have artificially selected for production traits causing the propagation of sub-optimal *KIR* alleles within the cattle genome. This is important for the relevance of the cattle *KIR* haplotype as a model for NK cell receptor gene expansion and its potential exploitation for animal health.

The aurochs genome studied here was isolated from a bone discovered in Derbyshire that has been radio carbon dated to be approximately 6,700 years old [43]. Although cattle are believed to have been first domesticated approximately 10,000 years ago [69], this animal pre-dates the arrival of domesticated cattle to the British isles along with the first humans to domesticate livestock. The

last aurochs became extinct in 1627 [116], therefore domesticated cattle and wild aurochs coexisted for many centuries. This may have led to interbreeding between the two sub species as farmers sought to insert certain aurochs traits to their herd. This admixture may have maintained the *KIR* haplotype gene content within domesticated cattle suggesting the HF *KIR* complex has had a relatively shorter evolutionary period under artificial selection than 6,700 years.

7.1.3 Sheep *KIR* reveal the evolution of 5 ancient gene families in *Bovidae*

Sequencing the sheep *KIR* haplotype has revealed another gene dense immune complex that has similar features but is not the same as the cattle *KIR* haplotype. The two species have both expanded *KIR* from the same five gene groups suggesting a shared ancestral haplotype of at least three X-lineage and two L-lineage *KIR* genes. It is also predicted that a single activating tail sequence was inherited from the ancestral haplotype which has subsequently recombined several times throughout the cattle and sheep *KIR* haplotypes. This is based on the sequence identity and phylogenetic reconstruction of the *KIR* signalling domain sequences between species, showing they are all very highly related with no divergence between the species or *KIR* groups. The only common short tailed group between the species is the group II *KIR*, therefore the activating domain may have been inherited through this group. It is therefore predicted this activating sequence was inherited from a two domain L-lineage gene, the ancestral gene of *BotaKIR2DS1/2/3* and *OvarKIR2DS1/2/3*. Cattle and sheep have inherited the same *KIR* genes and expanded them independently to form two unique haplotypes. Sheep have diversified a *KIR* group (group VI) that has remained a single pseudogene in cattle whereas cattle have expanded a group (group 0) into three groups (groups I,III and V) that have remained as two genes in sheep (group VII). However, both cattle and sheep have expanded the group IV genes suggesting species specic roles for the other X-lineage *KIR* and a shared *Bovidae* specific role pivoting around the group IV *KIR*.

To understand the roles of the group IV *KIR* within both species, the ligands will need to be determined. Due to the sequence similarity of the ligand binding Ig domains, it is hypothesised that the group IV *KIR* in cattle and sheep recognise a similar ligand, potentially an orthologous MHC class I gene shared between the two species.

7.1.4 The cattle *KIR* complex gene content is predicted to be the same within the *Bos* species

Whole genome raw sequence analysis of the *KIR* complex has revealed that the *Bos* species, including zebu cattle (*Bos indicus*) and the Yak (*Bos gruniens*) as well as two other taurine breeds, have the same *KIR* as the HF. Analysis of the more divergent water buffalo species (*Bubalus bubalis*), which shared a last common ancestor cattle approximately 17 mya [65], revealed a potentially different *KIR* haplotype structure, however this haplotype cannot be characterised using the short read dataset as *de novo* assembly is impossible. This indicated the *KIR* haplotype is more divergent outside of the *Bos* species and that genotyping strategies should work between the different species within the *Bos* clade. Therefore, the HF cattle *KIR* haplotype structure formed within the *Bos* clade between 5.8 mya when *Bos taurus* and *bos indicus* shared a last common ancestor, and 17 mya when *Bos* and *Bubalus* shared a last common ancestor [65]. The genome of the bison has not been interrogated yet and therefore it is unknown if the haplotype formed within the larger *bos* and *bison* clade.

This WGS *KIR* alignment analysis also revealed the KU cattle has a truncated *KIR* haplotype. Missing sequence at four null-allele positions, it is predicted this animal and potentially the rest of the breed have deleted these four loci. Therefore it has been proven within this study that there are variable *KIR* haplotypes within cattle. The KU encodes a predicted functional *BotaKIR3DXL6* allele that is a null-allele within the HF and other genomes interrogated. Therefore, this animal encodes a diverse and functionally variable *KIR* haplotype compared to HF.

Furthermore, analysis of the Nellore genome suggests a lack of sequence at the *BotaKIR3DXS1* locus, however details for this dataset are ambiguous. Therefore, further investigation into this locus within the Nellore breed is required.

7.1.5 Non-synonymous SNP numbers within the functional *KIR* genes indicates locus specific modulation of different Ig domains

The *KIR* coding sequences are very polymorphic between individuals of the same breed. The SNPs are concentrated mainly within the D0, D1 and D2 domains. Therefore it is predicted that the cattle KIR have co-evolved with their ligands resulting in variable extracellular domains. Interestingly polymorphisms are focused within different Ig domains for each *KIR* gene, meaning there has been no specific modulation of a particular domain within all of the cattle *KIR*. It is hypothesised that each *KIR* gene is undergoing locus specific modulation based

on the ligand which it recognised. This could be a result of divergence in ligand specificity from the other genes within the group or an impact from the different mechanisms by which the receptors bind ligand.

7.1.6 Attenuation of *BotaKIR3DXS1* suggests a transient gene currently undergoing negative selection

The single predicted functional activating gene, *BotaKIR3DXS1*, within the cattle *KIR* complex is predicted to be a paired receptor with the inhibitory *BotaKIR3DXL1*. These two genes contain highly similar sequence within the Ig domains and the receptors are predicted to recognise, or have recognised, the same ligand. Therefore, it is predicted that *BotaKIR3DXS1* evolved as a result of gene recombination between *BotaKIR3DL1* and an activating gene, and it recognised a virally encoded decoy protein that subverted BotaKIR3XL1 expressing cattle NK cells. Therefore it is predicted that BotaKIR3DXS1 provided a functional role recognising the decoy protein and activating NK cells to kill virally infected cells.

Through sequencing of the entire *KIR* complex of 24 different cattle, its has been shown that there is gene presence absence variation as well as considerable polymorphic variation of the *BotaKIR3DXS1* locus. The gene presence absence variation, as seen in HF159 and the Sahiwal cattle, may be the result of sequencing a haplotype that never contained the gene, it may pre date the evolution of *BotaKIR3DXS1* which could have evolved within a separate haplotype not found within these animals. Alternatively this haplotype may have deleted *BotaKIR3DXS1* as it is no longer useful due to the subsidence of the potential pathogen selection pressures. Relative to *BotaKIR3DXL1*, *BotaKIR3DXS1* contains a high level polymorphic variation specifically within the D1 Ig domain. This may have reduced the BotaKIR3DXS1 receptor specificity for ligand therefore reducing its ability to activate NK cells. Furthermore the receptor is undergoing attenuation within the signalling domain with the active arginine residue changed to a glutamine within the KU and Sahiwal animals. This is another mechanism by which BotaKIR3DXS1 will be unable to activate cattle NK cells. As it is predicted that the *BotaKIR3DXS1* activating tail evolved before *BotaKIR3DXS1*, the alteration of arginine to glutamine has most likely occurred as BotaKIR3DXS1 has been functional. Therefore it is hypothesised that the *BotaKIR3DXS1* gene is being negatively selected because it is no longer required for recognising viral decoy proteins and is potentially a detriment to the host by generating autoreactive NK cells that recognise and kill host cells expressing the same ligand that BotaKIR3DXL1 binds.

7.1.7 Conclusions

In response to the aim of the title of this thesis, "determine the genetic mechanisms responsible for generating diversity within the cattle NK cell receptor repertoire", there are two mechanisms defined here.

Firstly, the cattle *KIR* gene complex has expanded via block duplication, predicted to be a result of non-allelic homologous recombination during meiosis. The cattle *KIR* genes have duplicated from at least five ancestral *KIR* shared with sheep to 18 discrete loci, each locus encoding unique sequence. Therefore, the first mechanism for generating diversity determined here has been genomic recombination generating duplicated genes.

Secondly, each cattle *KIR* sequence has subsequently undergone base substitutions differentiating it from the other loci after duplication. This has resulted in diversity between the duplicated genes but has also generated alleles multiple at each *KIR* locus. Therefore, the second mechanism for generating diversity determined here has been nucleotide substitutions generating polymorphic *KIR* sequence.

Finally there may be further mechanisms that have not been determined here but have been eluded to or hypothesised from the results obtained. The gene dense *KIR* complex may enable the utilisation of null-allele intact exon sequences for the generation of composite mRNA sequences from multiple *KIR* loci. Furthermore, splice variants may alter the receptor structures expressed. To determine if either of these are possible, further sequencing of NK cell transcripts is required, either through targeted sequencing of NK cell cDNA or on a larger scale such as RNA-seq or exome capture of NK cell receptor gene mRNA.

7.2 Future work

The results from the work in this project will enable the future study of KIR receptors within the cattle immune system. A number of prominent questions are now being asked of the KIR receptors that should be addressed next.

7.2.1 Determine the ligands for cattle KIR

The ligands for the cattle KIR are unknown and it is predicted that, like primates, the cattle MHC class I and viral class I-like molecules are recognised by cattle KIR. Sequencing of the *KIR* haplotype has provided the full length gene sequences required for determining the functions of the receptors. This will enable KIR proteins to be artificially expressed and characterised. Fusion proteins of KIR Ig domains linked to the Fc receptor of IgG can be used to detect

KIR binding to specific MHC class I transfected target cells. Therefore, differing combinations of each cattle MHC class I and *KIR* gene can be expressed to determine the receptor-ligand combinations that interact. This approach would also allow for *KIR* null-alleles to be "corrected" and expressed so that the ligands of inactivated genes can be determined.

As described in the introduction, the cattle classical class I haplotype is gene variable with a total of six different genes, Figure 6. Therefore, not all cattle MHC haplotypes will contain a single consistent MHC class I molecule. This could have driven cattle *KIR* expansion as more receptors are needed to ensure NK education via diverse host ligands. The single MHC class I gene containing haplotypes such as A18 or H2 [44] contain only a single copy of gene 5 or 6 respectively. Homozygous cattle for these haplotypes will have only a single classical MHC class I molecule expressed. If there is no KIR receptor specific for this molecule, it is predicted fewer licensed NK cells will exist. Therefore, it is predicted at least one KIR receptor is specific for MHC class I gene 5 or 6.

Conversely, animals heterozygous for haplotypes 08 and 13, such as a Boran Holstein-Friesian cross [44], have the potential to encode five different MHC class I molecules on two different haplotypes. Fewer different NK cell receptors would be required to generate licensed NK cells as more ligands would be available. To determine if host MHC genotypes affects NK cell KIR acquisition in cattle, NK cells will need to be phenotyped for receptors expressed on the cell surface. However, this requires the production of antibodies specific to each KIR receptor, something that has been difficult in all species. Currently there is a single cattle KIR antibody, for BotaKIR2DL1, however the sequences produced in this project may enable further antibodies to be raised. Alternatively, *KIR* transcription levels can be measured using qPCR of NK cell mRNA to give an indication of KIR expression on the cell surface. This work is currently being conducted within the Immunogenetics group at the Pirbright institute and has been made possible by the sequencing of the cattle *KIR* haplotype.

7.2.2 Previous role of null-allele activating receptor genes

Cattle have seen an expansion of activating KIR receptor genes with a subsequent inactivation of all but one of the genes. This has resulted in seven *KIR* loci that encode inactivated activating receptor genes. A possible cause of this has been the rise of viral decoy proteins that triggered the evolution of activating KIR to detect virally infected cells. Therefore, it is hypothesised that the cattle *KIR* haplotype activating KIR receptors have evolved from viral decoy protein selection pressures.

In humans, primates, bats and rodents, betaherpesviruses are the genus of viruses that regularly encode MHC decoy proteins to subvert the host immune system. These include MsHV in bats [150], HCMV in humans [144], CCMV in chimpanzees [37], MCMV in mice [4], RCMV in rats [140] and GPCMV in guinea pigs [120]. However, there are no betaherpesviruses known to infect livestock [112]. The known bovine herpes viruses (BHV) are either alpha (BHV-1, BHV-2, BHV-5), or gamma (BHV-4) herpesviruses, which are not known for decoy protein production. However, rodent herpesvirus Peru (RHVP), a gammaherpesvirus, encodes an MHC class 1 homologue proteins [86], furthermore RHVP also encodes a chemokine decoy [88]. Therefore, there is precedent for something similar to be present in the cattle gammaherpesvirus BHV-4 [108]. BHV-1 an alphaherpesvirus has been shown to suppress cattle MHC class I [54], however there is no evidence of a decoy protein within its genome.

There is no evidence for decoy proteins encoded by cattle herpesviruses, however there may be undiscovered cattle specific large DNA viruses in the wild that posses this functionality. Alternatively, extinct cattle betaherpesviruses may have encoded decoy proteins and subverted the cattle NK cells, therefore generating a selection pressure upon the *KIR* haplotype to generate the activating receptors. Future attempts to sequence ancient ruminant genomes could therefore also focus on identifying novel viral genomes that may have encoded decoy MHC proteins.

8 Chapter 8. Bibliography

References

[1] Laurent Abi-rached and Peter Parham. Natural selection drives recurrent formation of activating killer cell immunoglobulin-like receptor and Ly49 from inhibitory homologues. *The Journal of Immunology*, 201(8):1319–32, April 2005.

[2] Munir Akkaya and a Neil Barclay. How do pathogens drive the evolution of paired receptors? *European journal of immunology*, 43(2):303–13, February 2013.

[3] Marcus Altfeld, Lena Fadda, Davor Frleta, and Nina Bhardwaj. DCs and NK cells: critical effectors in the immune response to HIV-1. *Nature Reviews Immunology*, 11(3):176–186, March 2011.

[4] Hisashi Arase, Edward S Mocarski, Ann E Campbell, Ann B Hill, and Lewis L Lanier. Direct recognition of cytomegalovirus by activating and inhibitory NK cell receptors. *Science (New York, N.Y.)*, 296(5571):1323–6, May 2002.

[5] a L Archibald, N E Cockett, B P Dalrymple, T Faraut, J W Kijas, J F Maddox, J C McEwan, V Hutton Oddy, H W Raadsma, C Wade, J Wang, W Wang, and X Xun. The sheep genome reference sequence: a work in progress. *Animal genetics*, 41(5):449–53, October 2010.

[6] Anne Averdam, Beatrix Petersen, Cornelia Rosner, Jennifer Neff, Christian Roos, Manfred Eberle, Fabienne Aujard, Claudia Münch, Werner Schempp, Mary Carrington, Takashi Shiina, Hidetoshi Inoko, Florian Knaust, Penny Coggill, Harminder Sehra, Stephan Beck, Laurent Abi-Rached, Richard Reinhardt, and Lutz Walter. A novel system of polymorphic and diverse NK cell receptors in primates. *PLoS genetics*, 5(10):e1000688, October 2009.

[7] Alexander David Barrow and John Trowsdale. You say ITAM and I say ITIM, let's call the whole thing off: the ambiguity of immunoreceptor signalling. *European Journal of Immunology*, 36(7):1646–1653, July 2006.

[8] Albano Beja-Pereira, David Caramelli, Carles Lalueza-Fox, Cristiano Vernesi, Nuno Ferrand, Antonella Casoli, Felix Goyache, Luis J Royo, Serena Conti, Martina Lari, Andrea Martini, Lahousine Ouragh, Ayed Magid, Abdulkarim Atash, Attila Zsolnai, Paolo Boscato, Costas Triantaphylidis, Konstantoula Ploumi, Luca Sineo, Francesco Mallegni, Pierre Taberlet, Georg Erhardt, Lourdes Sampietro, Jaume Bertranpetit, Guido Barbujani, Gordon Luikart, and Giorgio Bertorelle. The origin of European cattle: evidence from modern and ancient DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 103(21):8113–8, May 2006.

[9] Vivien Béziat, James a Traherne, Lisa L Liu, Jyothi Jayaraman, Monika Enqvist, Stella Larsson, John Trowsdale, and Karl-Johan Malmberg. Influence of KIR gene copy number on natural killer cell education. *Blood*, 121(23):4703–7, June 2013.

[10] James Birch and Shirley a Ellis. Complexity in the cattle CD94/NKG2 gene families. *Immunogenetics*, 59(4):273–80, April 2007.

[11] James Birch, Lisa Murphy, N D MacHugh, and S A Ellis. Generation and maintenance of diversity in the cattle MHC class I region. *Immunogenetics*, pages 670–679, 2006.

[12] Jeroen H Blokhuis, Marit K van der Wiel, Gaby G M Doxiadis, and Ronald E Bontrop. The mosaic of KIR haplotypes in rhesus macaques. *Immunogenetics*, 62(5):295–306, May 2010.

[13] James K Bonfield and Andrew Whitwham. Gap5–editing the billion fragment sequence assembly. *Bioinformatics (Oxford, England)*, 26(14):1699–703, July 2010.

[14] Silvia Bonfiglio, Catarina Ginja, Anna De Gaetano, Alessandro Achilli, Anna Olivieri, Licia Colli, Kassahun Tesfaye, Saif Hassan Agha, Luis T Gama, Federica Cattonaro, M Cecilia T Penedo, Paolo Ajmone-Marsan, Antonio Torroni, and Luca Ferretti. Origin and spread of Bos taurus: new clues from mitochondrial genomes belonging to haplogroup T1. *PloS one*, 7(6):e38601, January 2012.

[15] J C Boyington, S a Motyka, P Schuck, a G Brooks, and P D Sun. Crystal structure of an NK cell immunoglobulin-like receptor in complex with its class I MHC ligand. *Nature*, 405(6786):537–543, June 2000.

[16] Petter Brodin, Tadepally Lakshmikanth, Sofia Johansson, Klas Kärre, Petter Höglund, K Karre, and P Hoglund. The strength of inhibitory input during education quantitatively tunes the functional responsiveness of individual natural killer cells. *Blood*, 113(11):2434–2441, March 2009.

[17] M Burset, I a Seledtsov, and V V Solovyev. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic acids research*, 28(21):4364–75, November 2000.

[18] Matthew E Call, Kai W Wucherpfennig, and James J Chou. The structural basis for intramembrane assembly of an activating immunoreceptor complex. *Nature immunology*, 11(11):1023–9, November 2010.

[19] Kerry S Campbell and Amanda K Purdy. Structure/function of human killer cell immunoglobulin-like receptors: lessons from polymorphisms, evolution, crystal structures and mutations. *Immunology*, 132(3):315–325, March 2011.

[20] Paola Carrillo-Bustamante, Can KeÅ§mir, and Rob J de Boer. Virus encoded MHC-like decoys diversify the inhibitory KIR repertoire. *PLoS computational biology*, 9(10):e1003264, January 2013.

[21] A Cerwenka and L L Lanier. Natural killer cells, viruses and cancer. *Nat Rev Immunol*, 1(1):41–49, 2001.

[22] Mark J Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *Bmc Bioinformatics*, 13(238), 2012.

[23] A Chalifour, L Scarpellino, J Back, P Brodin, E Devevre, F Gros, F Levy, G Leclercq, P Hoglund, F Beermann, and W Held. A Role for cis Interaction between the Inhibitory Ly49A Receptor and MHIC Class I for Natural Killer Cell Education. *Immunity*, 30(3):337–347, 2009.

[24] T L Chapman, a P Heikeman, and P J Bjorkman. The inhibitory receptor LIR-1 uses a common binding interaction to recognize class I MHC molecules and the viral homolog UL18. *Immunity*, 11(5):603–613, November 1999.

[25] Yong Chen, Yi Shi, Hao Cheng, Yun-Qing An, and George F Gao. Structural immunology and crystallography help immunologists see the immune system in action: how T and NK cells touch their ligands. *IUBMB life*, 61(6):579–90, June 2009.

[26] B Chevreux, T Pfisterer, B Drescher, A J Driesel, W E G Muller, T Wetter, and S Suhai. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, 14(6):1147–1159, 2004.

[27] B Chevreux, T Wetter, and S Suhai. Genome sequence assembly using trace signals and additional sequence information. 1999.

[28] Bastien Chevreux, Thomas Pfisterer, Bernd Drescher, Albert J Driesel, Werner E G Müller, Thomas Wetter, and Sándor Suhai. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Research*, pages 1147–1159, 2004.

[29] B Chevreux Wetter, T., and Suhai, S. Genome sequence assembly using trace signals and additional sequence information. *Comput. Sci. Biol.: Proc. German Conference on Bioinformatics*, (GCB âĂŹ99):pp. 45–56, 1999.

[30] Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron a Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, Stephen W Turner, and Jonas Korlach. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods*, 10(6):563–9, June 2013.

[31] P Cingolani, A Platts, M Coon, T Nguyen, L Wang, S J Land, X Lu, and D M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012.

[32] Gemma F Codner, James Birch, John a Hammond, and Shirley a Ellis. Constraints on haplotype structure and variable gene frequencies suggest a functional hierarchy within cattle MHC class I. *Immunogenetics*, 64(6):435–445, June 2012.

[33] C G Cook, G J Letchworth, and G a Splitter. Bovine naturally cytolytic cell activation against bovine herpes virus type 1-infected cells does not require late viral glycoproteins. *Immunology*, 66(4):565–569, April 1989.

[34] M A Cooper, T A Fehniger, and M A Caligiuri. The biology of human natural killer-cell subsets. *Trends Immunol*, 22(11):633–640, 2001.

[35] Alexandra J Corbett, Jerome D Coudert, Catherine a Forbes, and Anthony a Scalzo. Functional consequences of natural sequence variation of murine cytomegalovirus m157 for Ly49 receptor specificity and NK cell activation. *Journal of immunology (Baltimore, Md. : 1950)*, 186(3):1713–1722, February 2011.

[36] D Cosman, N Fanger, L Borges, M Kubin, W Chin, L Peterson, and M L Hsu. A novel immunoglobulin superfamily receptor for cellular and viral MHC class I molecules. *Immunity*, 7(2):273–82, August 1997.

[37] a. J Davison. The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome. *Journal of General Virology*, 84(1):17–28, January 2003.

[38] M L Delano and D G Brownstein. Innate resistance to lethal mousepox is genetically linked to the NK gene complex on chromosome 6 and correlates with early restriction of virus replication by cells with an NK phenotype. *J Virol*, 69(9):5875–5877, September 1995.

[39] Michel Denis, Denise L Keen, Natalie a Parlane, Anne K Storset, and Bryce M Buddle. Bovine natural killer cells restrict the replication of Mycobacterium bovis in bovine macrophages and enhance IL-12 release by infected macrophages. *Tuberculosis (Edinb)*, 87(1):53–62, January 2007.

[40] F Di Palma, S D Archibald, J R Young, and S A Ellis. A BAC contig of approximately 400 kb contains the classical class I major histocompatibility complex (MHC) genes of cattle. *European journal of immunogenetics : official journal of the British Society for Histocompatibility and Immunogenetics*, 29(1):65–68, February 2002.

[41] Melanie Dobromylskyj and Shirley Ellis. Complexity in cattle KIR genes : transcription and genome analysis. *Infection and Immunity*, pages 463–472, 2007.

[42] C Dohring and M Colonna. Human natural killer cell inhibitory receptors bind to HLA class I molecules. *Eur J Immunol*, 26(2):365–369, 1996.

[43] Ceiridwen J Edwards, David A Magee, Stephen D E Park, Paul A Mcgettigan, J Amanda, Alison Murphy, Emma K Finlay, Beth Shapiro, Andrew T Chamberlain, B Martin, Daniel G Bradley, Brendan J Loftus, and David E Machugh. A complete mitochondrial genome sequence from a mesolithic wild aurochs (Bos primigenius). *PLoS One*, 5(2):000–8, 2010.

[44] Shirley a. Ellis and John a. Hammond. The Functional Significance of Cattle Major Histocompatibility Complex Class I Genetic Diversity. *Annual Review of Animal Biosciences*, 2(1):285–306, February 2014.

[45] C G Elsik, R L Tellam, K C Worley, R A Gibbs, A R R Abatepaulo, C A Abbey, D L Adelson, J Aerts, V Ahola, L Alexander, T Alioto, I G Almeida, A F Amadio, E Anatriello, S E Antonarakis, J M Anzola, A Astashyn, S M Bahadue, C L Baldwin, W Barris, R Baxter, S N Bell, A K Bennett, G L Bennett, F H Biase, C R Boldt, D G Bradley, F S L Brinkman, C L Brinkmeyer-Langford, W C Brown, M J Brownstein, C Buhay, A R Caetano, F Camara, J A Carroll, W A Carvalho, T Casey, E P Cervelatti, J Chack, E Chacko, M M Chandrabose, J E Chapin, C E Chapple, H C Chen, L Chen, Y Cheng, Z Cheng, C P Childers, C G Chitko-McKown, R Chiu, J W Choi, J Chrast, A J Colley, T Connelley, A Cree, S Curry, B Dalrymple, M Diep Dao, C Davis, C J F de Oliveira, I K F de Miranda Santos, T A de Campos, H Deobald, E Devinoy, C M Dickens, Y Ding, H H Dinh, M De Donato, K E Donohue, R Donthu, P Dovc, S Dugan-Rocha, K J Durbin, A Eberlein, R C Edgar, A Egan, A Eggen, E E Eichler, E Elhaik, S A Ellis, L Elnitski, O Ermolaeva, E Eyras, C J Fitzsimmons, G R Fowler, A M Franzin, K Fritz, R A Gabisi, G R Garcia, J F Garcia, S Genini, D Gerlach, J B German, J G R Gilbert, C A Gill, C J Gladney, E J Glass, J Goodell, J R Grant, D Graur, M L Greaser, J A Green, R D Green, L Guan, R Guigo, D L Hadsell, D E Hagen, H A Hakimov, R Halgren, D L Hamernik, C Hamilton, G P Harhay, J L Harrow, E A Hart, N Hastings, P Havlak, C N Henrichsen, J Hernandez, M Hernandez, C T A Herzig, S G Hiendleder, S Hines, M E Hitchens, W Hlavina, M Hobbs, M Holder, R A Holt, Z L Hu, J Hume, A Iivanainen, A Ingham, T Iso-Touru, C Jamis, O Jann, K Jensen, S N Jhangiani, H Y Jiang, A J Johnson, S J M Jones, V Joshi, T Junier, D Kapetis, S M Kappes, Y Kapustin, J W Keele, M P Kent, T Kerr, S S Khalil, H Khatib, B Kiryutin, P Kitts, F Kokocinski, D Kolbehdari, C L Kovar, E V Kriventseva, C G Kumar, D Kumar, K K Lahmers, M Landrum, D M Larkin, L P L Lau, R Leach, J C M Lee, S Lee, D G Lemay, H A Lewin, L R Lewis, C X Li, S Lien, G E Liu, Y S Liu, Y Liu, K M Logan, J Lopez, R J Lozado, Y S Lutzow, D J Lynn, M D MacNeil, D Maglott, R Malinverni, N J Maqbool, E Marques, M A Marra, W F Martin, N F Martins, S R Maruyama, L K Matukumalli, R Mazza, J C McEwan, S D McKay, K L McLean, S McWilliam, J F Medrano, E Memili, C Moen, A L Molenaar, S S Moore, R Moore, D D More, B T

Moreno, M B Morgan, C T Muntean, D M Muzny, H P Nandakumar, L V Nazareth, N B Nguyen, F W Nicholas, M F G Nogueira, G O Okwuonu, I Olsaker, S D Pant, F Panzitta, R C P Pastor, B M Patel, G M Payne, M Plass, M A Poli, N Poslusny, K Pruitt, L L Pu, X Qin, S Rachagani, J M Raison, S Ranganathan, A Ratnakumar, A Razpet, J Reecy, J T Reese, Y Ren, A Reymond, P K Riggs, M Rijnkels, G Rincon, A Roberts, N Rodriguez-Osorio, S L Rodriguez-Zas, N E Romero, A Rosenwald, S J Ruiz, A Sabo, H Salih, L Sando, J Santibanez, V Sapojnikov, J E Schein, S M Schmutz, R D Schnabel, L Schook, S M Searle, S W Seo, Y F Shen, L B Shen, L Sherman, L C Skow, T Smith, W M Snelling, E Sodergren, V Solovyev, H Song, J Z Song, T S Sonstegard, B R Southey, A Souvorov, D Spurlock, D Steffen, R T Stone, P Stothard, Y Sugimoto, J V Sweedler, A Takasuga, I Tammen, M Taniguchi, J F Taylor, B P V L Telugu, C Ucla, J M Urbanski, Y T Utsunomiya, C P Van, S Vattathil, C P Verschoor, A J Waardenberg, A Walker, Z Q Wang, R Ward, J T Warren, R C Waterman, R Weikard, G M Weinstock, T H Welsh, D A Wheeler, T T Wheeler, S N White, M D Whitside, K Wilczek-Boney, J L Williams, R L Williams, L G Wilming, J Womack, R A Wright, K R Wunderlich, C Wyss, M Q Yang, J Q Yang, E M Zdobnov, J K Zhang, F Q Zhao, B Zhu, and Bovine Genome Sequencing & Analysis Consortium. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. *Science*, 324(5926):522–528, 2009.

[46] Lena Fadda, Gwenoline Borhis, Parvin Ahmed, Kuldeep Cheent, Sophie V Pageon, Angelica Cazaly, Marco A Purbhoo, and Salim I Khakoo. Peptide antagonism as a mechanism for NK cell activation. *PNAS*, 107(22):1–6, 2010.

[47] Q R Fan, E O Long, and D C Wiley. Crystal structure of the human natural killer cell inhibitory receptor KIR2DL1-HLA-Cw4 complex. *Nature Immunology*, 2(5):452–460, 2001.

[48] Z Fan and Q Zhang. Molecular mechanisms of lymphocyte-mediated cytotoxicity. *Cell Mol Immunol*, 2(4):259–264, 2005.

[49] J Feng, D Garrity, M E Call, H Moffett, and K W Wucherpfennig. Convergence on a distinctive assembly mechanism by unrelated families of activating immune receptors. *Immunity*, 22(4):427–438, 2005.

[50] Y Fikri, J Nyabenda, J Content, and K Huygen. Cloning, sequencing, and cell surface expression pattern of bovine immunoreceptor NKG2D and adaptor molecules DAP10 and DAP12. *Immunogenetics*, 59(8):653–659, 2007.

[51] M W Fisher and D J Mellor. Developing a systematic strategy incorporating ethical , animal welfare and practice principles to guide the genetic improvement of dairy cattle. *New Zealand veterinary journal*, 56(3):100–6, June 2008.

[52] M W Fisher and D J Mellor. Developing a systematic strategy incorporating ethical, animal welfare and practice principles to guide the genetic improvement of dairy cattle. *New Zealand veterinary journal*, 56(3):100–6, June 2008.

[53] P O Flores-Villanueva, E J Yunis, J C Delgado, E Vittinghoff, S Buchbinder, J Y Leung, A M Uglialoro, O P Clavijo, E S Rosenberg, S A Kalams, J D Braun, S L Boswell, B D Walker, and A E Goldfeld. Control of HIV-1 viremia and protection from AIDS are associated with HLA-Bw4 homozygosity. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5140–5145, 2001.

[54] R S Gopinath, a P N Ambagala, S Hinkley, and S Srikumaran. Effects of virion host shut-off activity of bovine herpesvirus 1 on MHC class I expression. *Viral immunology*, 15(4):595–608, January 2002.

[55] A H Greenberg and J H Playfair. Spontaneously arising cytotoxicity to the P-815-Y mastocytoma in NZB mice. *Clinical and Experimental Immunology*, 16(1):99–110, 1974.

[56] Lisbeth a Guethlein, Laurent Abi-Rached, John a Hammond, and Peter Parham. The expanded cattle KIR genes are orthologous to the conserved single-copy KIR3DX1 gene of primates. *Immunogenetics*, 59(6):517–22, June 2007.

[57] Lisbeth A Guethlein, Anastazia M Older Aguilar, Laurent Abi-Rached, Peter Parham, O Aguilar, M Anastazia, and A M O Aguilar. Evolution of killer cell Ig-like receptor (KIR) genes: definition of an orangutan KIR haplotype reveals expansion of lineage III KIR associated with the emergence of MHC-C. *The Journal of Immunology*, 179(18354292883892002637related:Td_c9yGMt_4J):491–504, 2007.

[58] T Hall. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *http://www.mbio.ncsu.edu/BioEdit/bioedit*, 2004.

[59] John a Hammond, Lisbeth a Guethlein, Laurent Abi-Rached, Achim K Moesta, and Peter Parham. Evolution and survival of marine carnivores did not require a diversity of killer cell Ig-like receptors or Ly49 NK cell receptors. *Journal of immunology (Baltimore, Md. : 1950)*, 182(6):3618–27, March 2009.

[60] John a Hammond, Lisbeth a Guethlein, Laurent Abi-Rached, Achim K Moesta, and Peter Parham. Evolution and survival of marine carnivores did not require a diversity of killer cell Ig-like receptors or Ly49 NK cell receptors. *Journal of immunology (Baltimore, Md. : 1950)*, 182(6):3618–27, March 2009.

[61] John a Hammond, Steven G E Marsh, James Robinson, Christopher J Davies, Michael J Stear, and Shirley a Ellis. Cattle MHC nomenclature: is

it possible to assign sequences to discrete class I genes? *Immunogenetics*, 64(6):475–80, June 2012.

[62] Xiao-song S He, Monia Draghi, Kutubuddin Mahmood, Tyson H Holmes, George W Kemble, Cornelia L Dekker, Ann M Arvin, Peter Parham, and Harry B Greenberg. T cell-dependent production of IFN-gamma by NK cells in response to influenza A virus. *J Clin Invest*, 114(12):1812–1819, 2004.

[63] Susan L Heatley, Gabriella Pietra, Jie Lin, Jacqueline M L Widjaja, Christopher M Harpur, Sue Lester, Jamie Rossjohn, Jeff Szer, Anthony Schwarer, Kenneth Bradstock, Peter G Bardy, Maria Cristina Mingari, Lorenzo Moretta, Lucy C Sullivan, and Andrew G Brooks. Polymorphism in human cytomegalovirus UL40 impacts on recognition of human leukocyte antigen-E (HLA-E) by natural killer cells. *The Journal of biological chemistry*, 288(12):8679–8690, March 2013.

[64] P A Henkart. Mechanism of lymphocyte-mediated cytotoxicity. *Annu Rev Immunol*, 3:31–58, 1985.

[65] Manuel Hernández Fernández and Elisabeth S Vrba. A complete estimate of the phylogenetic relationships in Ruminantia: a dated species-level supertree of the extant ruminants. *Biological reviews of the Cambridge Philosophical Society*, 80(2):269–302, May 2005.

[66] Susan E Hiby, James J Walker, Kevin M O'shaughnessy, Christopher W G Redman, Mary Carrington, John Trowsdale, and Ashley Moffett. Combinations of maternal KIR and fetal HLA-C genes influence the risk of preeclampsia and reproductive success. *The Journal of experimental medicine*, 200(8):957–65, October 2004.

[67] Petter Höglund and Petter Brodin. Current perspectives of natural killer cell education by MHC class I molecules. *Nature reviews. Immunology*, 10(10):724–34, October 2010.

[68] J D Hunter. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[69] E Isaac. On the Domestication of Cattle: Zoology and cultural history both illuminate the view that the original motive was religious, not economic. *Science (New York, N.Y.)*, 137(3525):195–204, July 1962.

[70] D Kagi, B Ledermann, K Burki, R M Zinkernagel, and H Hengartner. Molecular mechanisms of lymphocyte-mediated cytotoxicity and their role in immunological protection and pathogenesis in vivo. *Annu Rev Immunol*, 14:207–232, 1996.

[71] K Kärre, HG Ljunggren, G Piontek, and R Kiessling. Selective rejection of HâĂŞ2-deficient lymphoma variants suggests alternative immune defence strategy. *Nature*, 319(6055):675–678, 1986.

[72] K Katoh, K Misawa, K Kuma, and T Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 2002.

[73] Ryouka Kawahara-Miki, Kaoru Tsuda, Yuh Shiwa, Yuko Arai-Kichise, Takashi Matsumoto, Yu Kanesaki, Sen-ichi Oda, Shizufumi Ebihara, Shunsuke Yajima, Hirofumi Yoshikawa, and Tomohiro Kono. Whole-genome resequencing shows numerous genes with nonsynonymous SNPs in the Japanese native cattle Kuchinoshima-Ushi. *BMC genomics*, 12(1):103, January 2011.

[74] W James Kent. BLAT âĂŤ The BLAST-Like Alignment Tool. *Genome Research*, pages 656–664, 2002.

[75] S I Khakoo, R Rajalingam, B P Shum, K Weidenbach, L Flodin, D G Muir, F Canavez, S L Cooper, N M Valiante, L L Lanier, and P Parham. Rapid evolution of NK cell receptor systems demonstrated by comparison of chimpanzees and humans. *Immunity*, 12(6):687–698, 2000.

[76] Salim I Khakoo, Chloe L Thio, Maureen P Martin, Collin R Brooks, Xiaojiang J Gao, Jacquie Astemborski, Jie Cheng, James J Goedert, David Vlahov, Margaret Hilgartner, Steven Cox, Ann-Margeret M Little, Graeme J Alexander, Matthew E Cramp, Stephen J O'Brien, William M C Rosenberg, David L Thomas, and Mary Carrington. HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection. *Science*, 305(5685):872–874, August 2004.

[77] R Kiessling, E Klein, H Pross, and H Wigzell. "Natural" killer cells in the mouse. II. Cytotoxic cells with specificity for mouse Moloney leukemia cells. Characteristics of the killer cell. *Eur J Immunol*, 5(2):117–121, 1975.

[78] R Kiessling, G Petranyi, K Karre, M Jondal, D Tracey, and H Wigzell. Killer cells: a functional comparison between natural, immune T-cell and antibody-dependent in vitro systems. *J Exp Med*, 143(4):772–780, 1976.

[79] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher a Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–76, March 2012.

[80] Astrid Krmpotić, Dirk H Busch, Ivan Bubić, Friedemann Gebhardt, Hartmut Hengel, Milena Hasan, Anthony A Scalzo, Ulrich H Koszinowski, Stipan Jonjić, A Krmpotic, I Bubic, and S Jonjic. MCMV glycoprotein gp40 confers virus resistance to CD8+ T cells and NK cells in vivo. *Nature immunology*, 3(6):529–35, June 2002.

[81] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25, January 2009.

[82] L L Lanier, B C Corliss, J Wu, C Leong, and J H Phillips. Immunoreceptor DAP12 bearing a tyrosine-based activation motif is involved in activating NK cells. *Nature*, 391(6668):703–707, 1998.

[83] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, 00(00):1–3, 2013.

[84] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60, July 2009.

[85] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, August 2009.

[86] Joy Loh, Guoyan Zhao, Christopher a Nelson, Penny Coder, Lindsay Droit, Scott a Handley, L Steven Johnson, Punit Vachharajani, Hilda Guzman, Robert B Tesh, David Wang, Daved H Fremont, and Herbert W Virgin. Identification and sequencing of a novel rodent gammaherpesvirus that establishes acute and latent infection in laboratory mice. *Journal of virology*, 85(6):2642–56, March 2011.

[87] B R Long, J Michaelsson, C P Loo, W M Ballan, B A Vu, F M Hecht, L L Lanier, J M Chapman, and D F Nixon. Elevated frequency of gamma interferon-producing NK cells in healthy adults vaccinated against influenza virus. *Clin Vaccine Immunol*, 15(1):120–130, 2008.

[88] Olga Y Lubman, Marina Cella, Xinxin Wang, Kristen Monte, Deborah J Lenschow, Yina H Huang, and Daved H Fremont. Rodent herpesvirus Peru encodes a secreted chemokine decoy receptor. *Journal of virology*, 88(1):538–546, January 2014.

[89] Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Yong Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jun Jian Wang, and Tak-Wah Lam. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18, January 2012.

[90] D E MacHugh, M D Shriver, and R T Loftus. Microsatellite DNA Variation and the Evolution, Domestication and Phylogeography of Taurine and Zebu Cattle (Bos taurus and Bos indicus). *Genetics*, 146(3):1071–1086, 1997.

[91] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011.

[92] Maureen P Martin, Arman Bashirova, James Traherne, John Trowsdale, and Mary Carrington. Cutting edge: expansion of the KIR locus by unequal crossing over. *Journal of immunology (Baltimore, Md. : 1950)*, 171(5):2192–5, September 2003.

[93] Maureen P Martin, Ying Qi, Xiaojiang Gao, Eriko Yamada, Jeffrey N Martin, Florencia Pereyra, Sara Colombo, Elizabeth E Brown, W Lesley Shupert, John Phair, James J Goedert, Susan Buchbinder, Gregory D Kirk, Amalio Telenti, Mark Connors, Stephen J O'Brien, Bruce D Walker, Peter Parham, Steven G Deeks, Daniel W McVicar, and Mary Carrington. Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1. *Nature genetics*, 39(6):733–40, June 2007.

[94] Karina L McQueen, Brian T Wilhelm, Kristin D Harden, and Dixie L Mager. Evolution of NK receptors: a single Ly49 and multiple KIR genes in the cow. *European journal of immunology*, 32(3):810–7, March 2002.

[95] J S Miller, K A Alley, and P McGlave. Differentiation of natural killer (NK) cells from human primitive marrow progenitors in a stroma-based long-term culture system: identification of a CD34+7+ NK progenitor. *Blood*, 83(9):2594–2601, 1994.

[96] J. S. Miller and V McCullar. Human natural killer cells with polyclonal lectin and immunoglobulinlike receptors develop from single hematopoietic stem cells with preferential expression of NKG2A and KIR2DL2/L3/S2. *Blood*, 98(3):705–713, August 2001.

[97] a. K. Moesta, P. J. Norman, M. Yawata, N. Yawata, M. Gleimer, and P. Parham. Synergistic Polymorphism at Two Positions Distal to the Ligand-Binding Site Makes KIR2DL2 a Stronger Receptor for HLA-C Than KIR2DL3. *The Journal of Immunology*, 180(6):3969–3979, March 2008.

[98] Achim K Moesta, Thorsten Graef, Laurent Abi-Rached, Anastazia M Older Aguilar, Lisbeth a Guethlein, and Peter Parham. Humans differ from other hominids in lacking an activating NK cell receptor that recognizes the C1 epitope of MHC class I. *Journal of immunology (Baltimore, Md. : 1950)*, 185(7):4233–4237, October 2010.

[99] Ashley Moffett and Charlie Loke. Immunology of placentation in eutherian mammals. *Nature reviews. Immunology*, 6(8):584–94, August 2006.

[100] Eugene W Myers, Granger G Sutton, Art L Delcher, Ian M Dew, Dan P Fasulo, Michael J Flanigan, Saul A Kravitz, Clark M Mobarry, Knut H J Reinert, Karin A Remington, Eric L Anson, Randall A Bolanos, Hui-Hsien Chou, Catherine M Jordan, Aaron L Halpern, Stefano Lonardi, Ellen M Beasley, Rhonda C Brandon, Lin Chen, Patrick J Dunn, Zhongwu Lai, Yong Liang, Deborah R Nusskern, Ming Zhan, Qing Zhang, Xiangqun Zheng, Gerald M Rubin, Mark D Adams, and J Craig Venter. A Whole-Genome Assembly of Drosophila. *Science*, 287(5461):2196–2204, 2000.

[101] Zemin Ning, Anthony J Cox, and James C Mullikin. SSAHA: a fast search method for large DNA databases. *Genome Res*, 11(10):1725–1729, 2001.

[102] PJ Norman and L Abi-Rached. Meiotic recombination generates rich diversity in NK cell receptor genes, alleles, and haplotypes. *Genome ...*, pages 757–769, 2009.

[103] H A Noyes, S L Anderson, A L Archibald, K Ashelford, D Bradely, H A Finlayson, S Kay, S J Kemp, Andy Law, Zen Lu, S Smith, R Talbot, M Agaba, and N Hall. SNP discovery In Zebu And African Taurine Cattle by whole genome sequencing of pooled DNA. page abstract P059, 2010.

[104] Anastazia M Older Aguilar, Lisbeth a Guethlein, Erin J Adams, Laurent Abi-Rached, Achim K Moesta, and Peter Parham. Coevolution of killer cell Ig-like receptors with HLA-C to become the major variable regulators of human NK cells. *Journal of immunology (Baltimore, Md. : 1950)*, 185(7):4238–4251, October 2010.

[105] Maureen a O'Leary, Jonathan I Bloch, John J Flynn, Timothy J Gaudin, Andres Giallombardo, Norberto P Giannini, Suzann L Goldberg, Brian P Kraatz, Zhe-Xi Luo, Jin Meng, Xijun Ni, Michael J Novacek, Fernando a Perini, Zachary S Randall, Guillermo W Rougier, Eric J Sargis, Mary T Silcox, Nancy B Simmons, Michelle Spaulding, Paúl M Velazco, Marcelo Weksler, John R Wible, and Andrea L Cirranello. The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science (New York, N.Y.)*, 339(6120):662–667, February 2013.

[106] Lilian J Oliveira, Nadéra Mansourri-Attia, Alan G Fahey, John Browne, Niamh Forde, James F Roche, Patrick Lonergan, and Trudee Fair. Characterization of the Th Profile of the Bovine Endometrium during the Oestrous Cycle and Early Pregnancy. *PloS one*, 8(10):e75571, January 2013.

[107] Mark T Orr, William J Murphy, and Lewis L Lanier. 'Unlicensed' natural killer cells dominate the response to cytomegalovirus infection. *Nature immunology*, 11(4):321–7, April 2010.

[108] Leonor Palmeira, Bénédicte Machiels, Céline Lété, Alain Vanderplasschen, and Laurent Gillet. Sequencing of bovine herpesvirus 4 v.test strain reveals important genome features. *Virology journal*, 8(1):406, January 2011.

[109] P Parham. The genetic and evolutionary balances in human NK cell receptor diversity. *Seminars in Immunology*, 20(6):311–316, 2008.

[110] Peter Parham. MHC Class I molecules and KIRs in human history, health and survival. *Nat Rev ImmunolNat Rev Immunol*, 5(March):201–214, 2005.

[111] Peter Parham and Ashley Moffett. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nature reviews. Immunology*, 13(2):133–44, February 2013.

[112] J R Patel and S Didlick. Epidemiology, disease and control of infections in ruminants by herpesviruses–an overview. *Journal of the South African Veterinary Association*, 79(1):8–14, March 2008.

[113] Virginie Prod'homme, Peter Tomasec, Charles Cunningham, Marius K Lemberg, Richard J Stanton, Brian P McSharry, Eddie C Y Wang, Simone Cuff, Bruno Martoglio, Andrew J Davison, Véronique M Braud, and Gavin W G Wilkinson. Human cytomegalovirus UL40 signal peptide regulates cell surface expression of the NK cell ligands HLA-E and gpUL18. *Journal of immunology (Baltimore, Md. : 1950)*, 188(6):2794–2804, March 2012.

[114] Qiang Qiu, Guojie Zhang, Tao Ma, Wubin Qian, Junyi Wang, Zhiqiang Ye, Changchang Cao, Quanjun Hu, Jaebum Kim, Denis M Larkin, Loretta Auvil, Boris Capitanu, Jian Ma, Harris a Lewin, Xiaoju Qian, Yongshan Lang, Ran Zhou, Lizhong Wang, Kun Wang, Jinquan Xia, Shengguang Liao, Shengkai Pan, Xu Lu, Haolong Hou, Yan Wang, Xuetao Zang, Ye Yin, Hui Ma, Jian Zhang, Zhaofeng Wang, Yingmei Zhang, Dawei Zhang, Takahiro Yonezawa, Masami Hasegawa, Yang Zhong, Wenbin Liu, Yan Zhang, Zhiyong Huang, Shengxiang Zhang, Ruijun Long, Huanming Yang, Jian Wang, Johannes a Lenstra, David N Cooper, Yi Wu, Jun Wang, Peng Shi, and Jianquan Liu. The yak genome and adaptation to life at high altitude. *Nature genetics*, 44(8):946–949, August 2012.

[115] David H Raulet and Russell E Vance. Self-tolerance of natural killer cells. *Nature reviews. Immunology*, 6(7):520–31, July 2006.

[116] M Rokosz. History of the aurochs (Bos taurus primigenius) in Poland. *Animal Genetic Resources Information*, 1995.

[117] K Rutherford, J Parkhill, J Crook, T Horsnell, P Rice, M a Rajandream, and B Barrell. Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10):944–945, October 2000.

[118] Jennifer G Sambrook, Harminder Sehra, Penny Coggill, Sean Humphray, Sophie Palmer, Sarah Sims, Haru-Hisa Takamatsu, Thomas Wileman, Alan L Archibald, and Stephan Beck. Identification of a single killer immunoglobulin-like receptor (KIR) gene in the porcine leukocyte receptor complex on chromosome 6q. *Immunogenetics*, 58(5-6):481–486, June 2006.

[119] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the ...*, 74(12):5463–5467, 1977.

[120] Mark R Schleiss, Alistair McGregor, K Yeon Choi, Shailesh V Date, Xiaohong Cui, and Michael a McVoy. Analysis of the nucleotide sequence of the guinea pig cytomegalovirus (GPCMV) genome. *Virology journal*, 5:139, January 2008.

[121] S. M. Shahjahan Miah, Tracey L. Hughes, Kerry S. Campbell, and S M Shahjahan Miah. KIR2DL4 Differentially Signals Downstream Functions in Human NK Cells through Distinct Structural Modules. *The Journal of Immunology*, 180(5):2922–2932, February 2008.

[122] G E Shook. Major advances in determining appropriate selection goals. *Journal of dairy science*, 89(4):1349–1361, April 2006.

[123] Simona Sivori, Claudia Cantoni, Silvia Parolini, Emanuela Marcenaro, Romana Conte, Lorenzo Moretta, and Alessandro Moretta. IL-21 induces both rapid maturation of human CD34+ cell precursors towards NK cells and acquisition of surface killer Ig-like receptors. *European journal of immunology*, 33(12):3439–47, December 2003.

[124] Simona Sivori, Michela Falco, Emanuela Marcenaro, Silvia Parolini, Roberto Biassoni, Cristina Bottino, Lorenzo Moretta, and Alessandro Moretta. Early expression of triggering receptors and regulatory role of 2B4 in human natural killer cell precursors undergoing in vitro differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, 99(7):4526–31, April 2002.

[125] Hamish R C Smith, Jonathan W Heusel, Indira K Mehta, Sungjin Kim, Brigitte G Dorner, Olga V Naidenko, Koho Iizuka, Hiroshi Furukawa, Diana L Beckman, Jeanette T Pingel, Anthony a Scalzo, Daved H Fremont, and Wayne M Yokoyama. Recognition of a virus-encoded ligand by a natural killer cell activation receptor. *Proceedings of the National Academy of Sciences of the United States of America*, 99(13):8826–31, June 2002.

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

[127] R Staden. The Staden sequence analysis package. *Molecular Biotechnology*, 5(3):233–241, 1996.

[128] MR Stofega and James Herrington. Mutation of the SHP-2 binding site in growth hormone (GH) receptor prolongs GH-promoted tyrosyl phosphorylation of GH receptor, JAK2, and STAT5B. *Molecular Endocrinology*, 14(9):1338–1350, 2000.

[129] Anne K Storset, Imer O Slettedal, John L Williams, Andy Law, and Erik Dissen. Natural killer cell receptors in cattle: a bovine killer cell immunoglobulin-like receptor multigene family contains members with divergent signaling motifs. *European journal of immunology*, 33(4):980–990, April 2003.

[130] Paul Stothard, Jung-Woo Choi, Urmila Basu, Jennifer M Sumner-Thomson, Yan Meng, Xiaoping Liao, and Stephen S Moore. Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. *BMC genomics*, 12(1):559, January 2011.

[131] Tomoko Takahashi, Makoto Yawata, Terje Raudsepp, Teri L Lear, Bhanu P Chowdhary, Douglas F Antczak, and Masanori Kasahara. Natural killer cell receptors in the horse: evidence for the existence of multiple transcribed LY49 genes. *Eur J Immunol*, 34(3):773–784, March 2004.

[132] F Takei, K L McQueen, M Maeda, B T Wilhelm, S Lohwasser, R H Lian, and D L Mager. Ly49 and CD94/NKG2: developmentally regulated expression and evolution. *Immunol Rev*, 181:90–103, 2001.

[133] K Tamura, D Peterson, N Peterson, G Stecher, M Nei, and S Kumar. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol*, 2011.

[134] Koichiro Tamura, Daniel Peterson, Nicholas Peterson, Glen Stecher, Masatoshi Nei, and Sudhir Kumar. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, 28(10):2731–9, October 2011.

[135] MS Tantia, RK Vijh, and V Bhasin. Whole-genome sequence assembly of the water buffalo (Bubalus bubalis). *The Indian Journal of ...*, 81(May):38–46, 2011.

[136] M Thapa, R S Welner, R Pelayo, and D J Carr. CXCL9 and CXCL10 expression are critical for control of genital herpes simplex virus type 2 infection through mobilization of HSV-specific CTL and NK cells to the nervous system. *J Immunol*, 180(2):1098–1106, 2008.

[137] P Tomasec, V M Braud, C Rickards, M B Powell, B P McSharry, S Gadola, V Cerundolo, L K Borysiewicz, A J McMichael, and G W Wilkinson. Surface expression of HLA-E, an inhibitor of natural killer cells, enhanced by human cytomegalovirus gpUL40. *Science*, 287(5455):1031, 2000.

[138] J A Trapani and M J Smyth. Functional significance of the perforin/-granzyme cell death pathway. *Nat Rev Immunol*, 2(10):735–747, 2002.

[139] M F van den Broek, D Kagi, R M Zinkernagel, and H Hengartner. Perforin dependence of natural killer cell-mediated tumor control in vivo. *Eur J Immunol*, 25(12):3514–3516, 1995.

[140] Cornelis Vink, Erik Beuken, and C A Bruggeman. Complete DNA Sequence of the Rat Cytomegalovirus Genome. *Journal of virology*, 74(16), 2000.

[141] Julian P Vivian, Renee C Duncan, Richard Berry, Geraldine M O'Connor, Hugh H Reid, Travis Beddoe, Stephanie Gras, Philippa M Saunders, Maya A Olshina, Jacqueline M L Widjaja, Christopher M Harpur, Jie Lin, Sebastien M Maloveste, David A Price, Bernard A P Lafont, Daniel W McVicar, Craig S Clements, Andrew G Brooks, and Jamie Rossjohn. Killer

cell immunoglobulin-like receptor 3DL1-mediated recognition of human leukocyte antigen B. *Nature*, October 2011.

[142] N Wagtmann, S Rajagopalan, C C Winter, M Peruzzi, and E O Long. Killer cell inhibitory receptors specific for HLA-C and HLA-B identified by direct binding and by functional transfer. *Immunity*, 3(6):801–809, 1995.

[143] Robert P a Wallin, Valentina Screpanti, Jakob Michaëlsson, Alf Grandien, and Hans-Gustaf Ljunggren. Regulation of perforin-independent NK cell-mediated cytotoxicity. *European journal of immunology*, 33(10):2727–35, October 2003.

[144] Eddie C Y Wang, Brian McSharry, Christelle Retiere, Peter Tomasec, Sheila Williams, Leszek K Borysiewicz, Veronique M Braud, and Gavin W G Wilkinson. UL40-mediated NK evasion during productive infection with human cytomegalovirus. *Proceedings of the National Academy of Sciences of the United States of America*, 99(11):7570–5, May 2002.

[145] Makoto Yawata, Nobuyo Yawata, Monia Draghi, Fotini Partheniou, Ann-Margaret Little, and Peter Parham. MHC class I-specific inhibitory receptors and their ligands structure diverse human NK-cell repertoires toward a balance of missing self-response. *Blood*, 112(6):2369–80, September 2008.

[146] W M Yokoyama, J C Ryan, J J Hunter, H R Smith, M Stark, and W E Seaman. cDNA cloning of mouse NKR-P1 and genetic linkage with LY-49. Identification of a natural killer cell gene complex on mouse chromosome 6. *The Journal of Immunology*, 147(9):3229–3236, 1991.

[147] Wayne M Yokoyama and Beatrice F M Plougastel. Immune functions encoded by the natural killer gene complex. *Nat Rev Immunol*, 3(4):304–316, April 2003.

[148] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–829, May 2008.

[149] Guojie Zhang, Christopher Cowled, Zhengli Shi, Zhiyong Huang, Kimberly a Bishop-Lilly, Xiaodong Fang, James W Wynne, Zhiqiang Xiong, Michelle L Baker, Wei Zhao, Mary Tachedjian, Yabing Zhu, Peng Zhou, Xuanting Jiang, Justin Ng, Lan Yang, Lijun Wu, Jin Xiao, Yue Feng, Yuanxin Chen, Xiaoqing Sun, Yong Zhang, Glenn a Marsh, Gary Crameri, Christopher C Broder, Kenneth G Frey, Lin-Fa Wang, and Jun Wang. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. *Science (New York, N.Y.)*, 339(6118):456–60, January 2013.

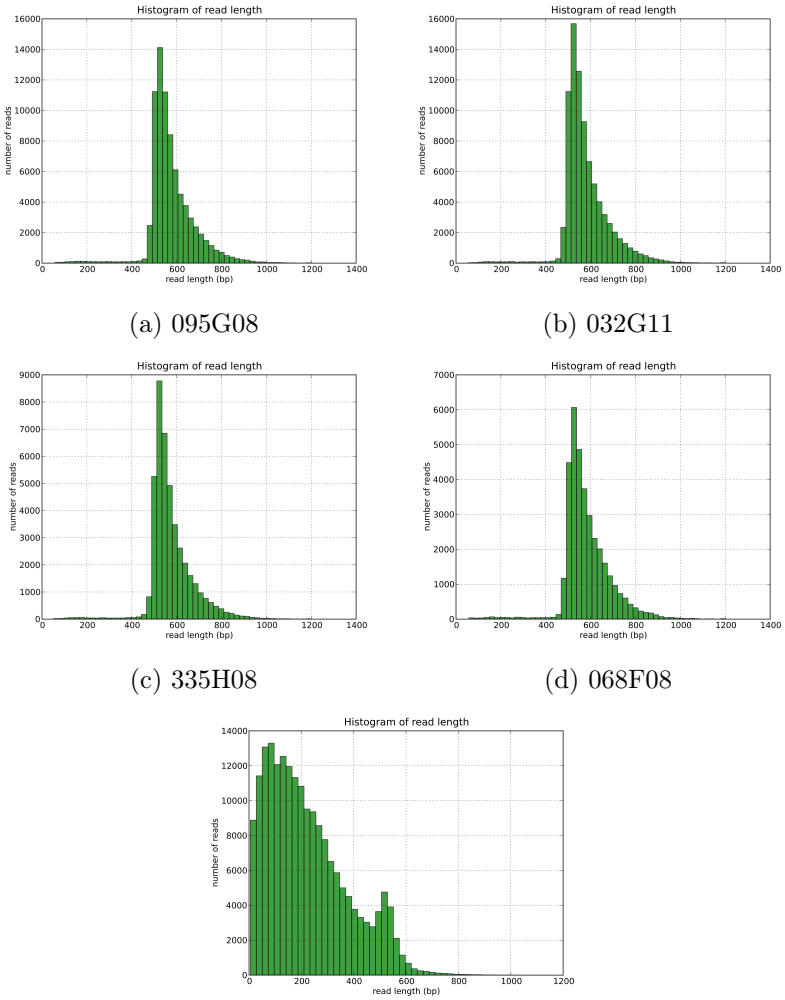
[150] Huajun Zhang, Shawn Todd, Mary Tachedjian, Jennifer a Barr, Minhua Luo, Meng Yu, Glenn a Marsh, Gary Crameri, and Lin-Fa Wang. A novel bat herpesvirus encodes homologues of major histocompatibility complex

classes I and II, C-type lectin, and a unique family of immune-related genes. *Journal of virology*, 86(15):8014–8030, August 2012.

[151] Aleksey V Zimin, Arthur L Delcher, Liliana Florea, David R Kelley, Michael C Schatz, Daniela Puiu, Finnian Hanrahan, Geo Pertea, Curtis P Van Tassell, Tad S Sonstegard, Guillaume Marçais, Michael Roberts, Poorani Subramanian, James a Yorke, and Steven L Salzberg. A whole-genome assembly of the domestic cow, Bos taurus. *Genome biology*, 10(4):R42, January 2009.

9 Appendix

9.1 Chapter 2 Appendix



(a) 095G08

(b) 032G11

(c) 335H08

(d) 068F08

(e) 303D02

Figure S1: Histogram of read lengths for BAC clone 454 sequences

Template	Primer 1 name	Primer 1 sequence	Primer 2 name	Primer 2 sequence
303D2	303D2_9KR_+	ACATGGGCTCAGTTTTGCAC	303D2_VX3_-	AGAAGACTCAGGCCCATCAC
303D2	303D2_B04_+	CCTCCTCCAACAGCCATTTTG	303D2_9KR_-	AAGGACTGATGCTGAGGCTG
303D2	303D2_LRG_S	ACTGCTGGGCATACACACTG	303D2_20Kb_AS	TCTGGGTTCACAATGCAGGG
303D2	303D2_WV8_+	GCCAAGGCTTTACATCCAATG	303D2_B04_-	AGAACCCATTCTCAAGGCCG
303D2	303D3_UXS_CJ_S	TCCTAAGTATTTTATTCTTTCCGTTGC	303D4_UXS_CJ_AS	TGTTAAGGTGGACACAGCCC
4222	303D2_UXS_DBJ_S	CTGTTGGTGGGAATGCAAGC	95G8_H8_DBJ_AS	GAAATCCACCTTGCTGTGCG
68F4	303D2_UXS_DBJ_S	CTGTTGGTGGGAATGCAAGC	95G8_H8_DBJ_AS	GAAATCCACCTTGCTGTGCG
303D2	303D2_UXS_DBJ_S	CTGTTGGTGGGAATGCAAGC	95G8_H8_DBJ_AS	GAAATCCACCTTGCTGTGCG
32G11	303D2_UXS_DBJ_S	CTGTTGGTGGGAATGCAAGC	95G8_H8_DBJ_AS	GAAATCCACCTTGCTGTGCG
95G8	MID3RV-1	AGGTATAACACTTTCCCATTCCT	MID34Q+1	GGCCTCATAAAAGATTTCAG
95G8	MID3RV-2	ACAGCTTCGAGAACAAAGG	MID34Q+2	CACTTTCTCTCTCCCTTATCC
95G8	MID3RV-2	ACAGCTTCGAGAACAAAGG	MID34Q+1	GGCCTCATAAAAGATTTCAG
95G8	MID3RV-1	AGGTATAACACTTTCCCATTCCT	MID34Q+2	CACTTTCTCTCTCCCTTATCC
95G8	Mid3 IB +	GAACCTGATGGTCCAGAG	Mid3 6G -	CTTGGTAATGGTTGCTG
95G8	MID3IntDipS3	CTGGTTTTGCCATACATTAAC	MID3IntDipAS4	CCCAAATTGAAAGAGACAC
95G8	MID3IntDipS3	CTGGTTTTGCCATACATTAAC	MID3IntDipAS3	CCTGTGGTCTCCTCATCTG
95G8	MID3IntDipS4	CTCTGTATAATCGGCTCCAG	MID3IntDipAS3	CCTGTGGTCTCCTCATCTG
95G8	Mid3 p3 -	CCTCCACCAAGTTCCTG	Mid3 7X +	GGGCAGGGAGAGGACTG
95G8 + 335H8	MID37_AP1__pls_1	TTTGTCAGTTGCTTCATTTG	MID37_AP1__mns_1	ATAAATCTGGTCTCCCCTTG
335H8	MID7_X4C_S5	AAGAATCACCAGTCCAAGG	MID7_X4C_AS6	GATCAGCACATCTGGTTG
335H8	MID7_X4C_S6	AACCCTGTATGTGAGACAGC	MID7_X4C_AS5	GATGTTGTTAATTCCTGTTATAC
95G8 + 335H8	MID37_OXP__mns_1	AAGTATTTTATTCTTTTCGTTGC	MID37_AP1__pls_2	AAAGGAAGAAATGAAGAAACC
95G8 + 335H8	MID37_OXP__mns_2	TTCTCCATCCATTAGTGTCC	MID37_AP1__pls_1	TTTGTCAGTTGCTTCATTTG
335H8	MID7_6NY_pls_1	GGCTCAGTGGTAAAGAACCTG	MID7_S13_mns_2	TGGAACCCTATTCTTTTTATG
335H8	MID7_X4C_S6	AACCCTGTATGTGAGACAGC	MID7_X4C_AS6	GATCAGCACATCTGGTTG
335H8	MID7_X4C_S3	ATCATGTAGGAAACGAGTTG	MID7_X4C_AS4	ATAGAACATGGGTTTACCTG
335H8	MID7_X4C_S4	GGATTCATTTTGATGTTTGG	MID7_X4C_AS4	ATAGAACATGGGTTTACCTG
335H8	MID7_X4C_S4	GGATTCATTTTGATGTTTGG	MID7_X4C_AS3	GGGTGTATTTCTGCTGTGTAG
335H8	MID7_X4C_S3	ATCATGTAGGAAACGAGTTG	MID7_X4C_AS3	GGGTGTATTTCTGCTGTGTAG
95G8	Mid3 M3 +	GGGGACAGGGAAAATAAAG	Mid3 7X -	GAGGAGGGTTCGGGATG
95G8 + 335H8	MID37_8T7__S2	TGATCCATGTTGATGTATGG	MID37_8T7__AS2	TGTCAAGGAACAGAGTGATG
95G8 + 335H8	MID37_8T7__S1	GTGCTCCACAACAAGAGAAG	MID37_8T7__AS2	TGTCAAGGAACAGAGTGATG
95G8 + 335H8	MID37_OXP__mns_4	ACAAGCAACTCCTACAGCTC	MID37_AP1__mns_1	ATAAATCTGGTCTCCCCTTG
335H8	Mid7AP1+	CTCTGTGGAGCTTGATTTTC	Mid7XDQ+	TTAACACGTCCTTCTGCAC
335H8	MID7_9CHJ_S1	ACCCAGAGGGATGGTATG	MID7_9CHJ_AS2	TTTCAAGAGAATGGCACATC
335H8	Mid7AP1+	CTCTGTGGAGCTTGATTTTC	Mid7Y19+	TATCAACATGAATCCACCAC
335H8	MID7_X4C_S1	TTATTTGGAGAAGGAAATGG	MID7_X4C_AS2	TTATGGACTCTGTGGGAGAG
95G8 + 335H8	MID37_8T7__S2	TGATCCATGTTGATGTATGG	MID37_8T7__AS1	CGTCTTATCTGAGGTAGATGG
95G8 + 335H8	MID37_8T7__S1	GTGCTCCACAACAAGAGAAG	MID37_8T7__AS1	CGTCTTATCTGAGGTAGATGG
335H8	MID7_X4C_S1	TTATTTGGAGAAGGAAATGG	MID7_X4C_AS1	GCATGATACTGGATGCTTG
95G8 + 335H8	MID37_8T7__pls_2	AAGCCAGAAAGAAAAACACC	MID37_AP1__mns_1	ATAAATCTGGTCTCCCCTTG
95G8 + 335H8	MID37_8T7__pls_2	AAGCCAGAAAGAAAAACACC	MID37_AP1__mns_2	GACAGGAAAGAACCTCAGTAAG
95G8 + 335H8	MID37_8T7__pls_1	CAGCAAAAGAGACACTGATG	MID37_AP1__mns_1	ATAAATCTGGTCTCCCCTTG
95G8 + 335H8	MID37_8T7__pls_1	CAGCAAAAGAGACACTGATG	MID37_AP1__mns_2	GACAGGAAAGAACCTCAGTAAG
335H8	Mid7Y19+	TATCAACATGAATCCACCAC	Mid7XDQ+	TTAACACGTCCTTCTGCAC
335H8	MID7_X4C_S5	AAGAATCACCAGTCCAAGG	MID7_X4C_AS5	GATGTTGTTAATTCCTGTTATAC
95G8 + 335H8	MID37_OXP__mns_2	TTCTCCATCCATTAGTGTCC	MID37_AP1__pls_2	AAAGGAAGAAATGAAGAAACC
335H8	MID7_X4C_mns_2	ATAATTCTTGGCCTCCACTC	MID7_S13_pls_2	AATGGAGTAAGAGTACAGTCACC
335H8	MID7_X4C_mns_2	ATAATTCTTGGCCTCCACTC	MID7_S13_mns_1	ATCCTCACTGTGGTGCTG
335H8	Mid7D9V-	CTTCCTTCTTCTGCATGTTC	Mid7XDQ+	TTAACACGTCCTTCTGCAC
335H8	Mid7AP1-	ACAGCAAATTGATTCAGCTC	Mid7Y19+	TATCAACATGAATCCACCAC
335H8	MID7_6NY_pls_1	GGCTCAGTGGTAAAGAACCTG	MID7_S13_mns_1	ATCCTCACTGTGGTGCTG
335H8	MID7_6NY_pls_1	GGCTCAGTGGTAAAGAACCTG	MID7_9CHJ_mns_1	TCAGTTGTGTCCGACTCTG
335H8	MID7_6NY_pls_2	GTGGGGAGTAATGTTTTCAC	MID7_9CHJ_mns_1	TCAGTTGTGTCCGACTCTG
335H8	MID7_6NY_pls_2	GTGGGGAGTAATGTTTTCAC	MID7_9CHJ_mns_2	TATAGCCCACCAGACTCCTC
95G8 + 335H8	MID37_OXP__mns_3	GAGAAATGCAAATCAAAAGC	MID37_VMA__pls_1	AACTCATTGGAAAAGACTCTG
95G8 + 335H8	MID37_OXP__mns_3	GAGAAATGCAAATCAAAAGC	MID37_VMA__pls_2	GACAACAGAGGATGAGATGG
95G8 + 335H8	MID37_OXP__mns_3	GAGAAATGCAAATCAAAAGC	MID37_VMA__mns_2	AATGGACAGAGGAGTCTGG
95G8 + 335H8	MID37_OXP__mns_4	ACAAGCAACTCCTACAGCTC	MID37_AP1__pls_1	TTTGTCAGTTGCTTCATTTG
95G8 + 335H8	MID37_OXP__mns_4	ACAAGCAACTCCTACAGCTC	MID37_AP1__pls_2	AAAGGAAGAAATGAAGAAACC
335H8	MID7_6NY_S1	ATGTGCTGGGATGTTAATTG	MID7_6NY_AS1	CAGCAAAAGAGACACTGATG
335H8	MID7_6NY_S2	GCATCTGAGTGTATCTGTGG	MID7_6NY_AS1	CAGCAAAAGAGACACTGATG
95G8 + 335H8	MID37_8T7__pls_2	AAGCCAGAAAGAAAAACACC	MID37_VMA__mns_1	AGTGACTAAACCACCACCAC
95G8 + 335H8	MID37_8T7__pls_2	AAGCCAGAAAGAAAAACACC	MID37_VMA__mns_2	AATGGACAGAGGAGTCTGG

Table S1: Table of PCR primers for BAC assembly finishing

9.1.1 Python scripts

Here are details of python scripts written to perform various analyses during this chapter. All scripts have been written by me using guidance from online resources such as biostars, seqanswers, stackoverflow, pythondocs plus the various package documentations such as BioPython, SeqIO and matplotlib. I have largely taught myself python coding so there is inevitably some bad habits and better ways to solve these problems, however the scripts work and have enabled analyses that otherwise would have been too laborious or unrepeatable. All of the source code has been uploaded to my GitHub site under the thesis scripts folder, found here: `https://github.com/nick297/thesis_scripts`.

9.1.2 Sliding window analysis script

This script generates sliding window chart of sequence identity from an aligned fasta sequence file. The scripts requires modules from the BioPython and Matplotlib packages as well a working X server window session. To show annotations this script requires an annotation file, the format for this file was quickly written by me in order to get this script and others working. It takes the common bed format and continues it further, the format is split into tab-delimeted columns with following fields and data types in parenthesis.

Chromosome(string)
gene_start(integer)
gene_stop (integer)
gene_exon_name (string)
Codon_start(0/1/2)
Strand(+/-)
exon_start(integer)
exon_stop (integer)

Other annotation formats could be preferable but I have not coded for them yet. The script can be located at `https://github.com/nick297/thesis_scripts/blob/master/NS_fasta_identy_compare0.1.5.ABC.py`. To use the script run:

Usage:

```
1 python NS_fasta_identy_compare0.1.5.ABC.py alignment.fasta 0
    500 annotation.bed
```

Where alignment.fasta is the aligned sequences, 0 is the sequence within the file to use as the reference (list starts at zero, 0 would be the first sequence, 1 would be the second etc...), 500 is the sliding window size in base pairs, annotation.bed is the annotation file using the format I specified above.

9.1.3 Determining the effects of the Varscan2 SNP caller output

The SNP caller Varscan2 outputs the positions and changes of variable positions from an mpileup output of a sorted SAM file. To determine the effects of these SNPs on the residue sequence of the *KIR* genes I wrote this script. There are other options such as SNPeff [31] which have since been published and may be more robust and have more features. However at the time of conducting this analysis SNPeff was not available and I wanted a quick reliable way to determine the effects of the SNPs I found in the CKHs.

The script requires the same formatted bed file as described above plus a SNP file in Varscan format and the reference sequence. It outputs the original SNP file along with further fields amended to the end of the lines in tab delimited format including the residue changes, S/NS and the exon that the SNP is found in.

Usage:

```
1 python NS_SNP_coder_hap1.3.py reference.fasta bedfile.bed
    snp_file.tab
```

9.1.4 Raw fastq stats and read length histograms

I wrote this python script to quickly assess the read length distributions of sequenced BAC clones using the 454 platform. Since writing this script other programs such as FASTQC have become available that provide a more comprehensive overview of FASTQ details.

This script takes standard input so that files can be piped to it allowing streamed uncompressed data from compressed sources reducing intermediate files.

Usage:

```
zcat file.fastq.gz | python NS_FQ_readlengths.py
```

9.1.5 Structural variation interrogation using paired end read information

This file charts the reads with inferred paired end distances greater than the threshold value. It effectively filters the bam file and creates a new one with the name ".filtered_bam" then generates a read depth coverage chart of that file.

Usage:

```
python NS_chart_bam_filter_inserts.py file.bam threshold(int)
```

9.1.6 P-distance similarity matrix of predicted cDNA sequence

	2DL1*02	2DS1*01N	2DS2*01N	2DS3*01N	2DXP1*01	2DXP2*01	3DXL1*02	3DXL2*01	3DXL3*01	3DXL3*02	3DXL4*02	3DXL5*01	3DXL6*01	3DXL7*01	3DXS1*03	3DXS2*01N	3DXS3*01N	1DP1*01	1DP2*01
2DL1*02																			
2DS1*01N	0.144																		
2DS2*01N	0.222	0.071																	
2DS3*01N	0.227	0.072	0.004																
2DXP1*01	0.458	0.483	0.477	0.479															
2DXP2*01	0.454	0.474	0.483	0.485	0.127														
3DXL1*02	0.382	0.323	0.299	0.302	0.151	0.065													
3DXL2*01	0.380	0.324	0.308	0.312	0.142	0.078	0.080												
3DXL3*01	0.406	0.339	0.312	0.314	0.141	0.107	0.113	0.119											
3DXL3*02	0.405	0.339	0.313	0.314	0.139	0.105	0.113	0.120	0.011										
3DXL4*02	0.383	0.324	0.306	0.310	0.148	0.084	0.077	0.027	0.121	0.122									
3DXL5*01	0.393	0.327	0.304	0.307	0.124	0.097	0.102	0.103	0.049	0.050	0.107								
3DXL6*01	0.395	0.333	0.307	0.310	0.170	0.111	0.085	0.076	0.117	0.119	0.060	0.106							
3DXL7*01	0.393	0.327	0.303	0.306	0.124	0.096	0.101	0.104	0.048	0.050	0.109	0.007	0.107						
3DXS1*03	0.382	0.308	0.278	0.281	0.153	0.057	0.057	0.108	0.126	0.126	0.107	0.113	0.112	0.113					
3DXS2*01N	0.350	0.312	0.315	0.318	0.161	0.097	0.107	0.104	0.133	0.135	0.105	0.124	0.113	0.124	0.085				
3DXS3*01N	0.392	0.312	0.289	0.292	0.161	0.097	0.123	0.122	0.145	0.148	0.122	0.136	0.130	0.137	0.085	0.004			
1DP1*01	0.279	0.085	0.072	0.072	NA	NA	0.133	0.139	0.146	0.146	0.136	0.143	0.133	0.139	0.068	0.067	0.071		
1DP2*01	0.401	0.078	0.122	0.124	NA	NA	0.156	0.163	0.155	0.156	0.158	0.149	0.160	0.148	0.085	0.060	0.095	0.024	

Table S2: Table of cDNA p-distances for each *KIR* aligned sequence

9.2 Chapter 3 Appendix

9.2.1 Python scripts

9.2.2 Aurochs Illumina read extraction

To extract the reads that aligned to the KIR haplotype this script uses a file of a list of names of each read that aligned within the BAM file and extracts it from the bam file. A list of names can be produced using samtools view and awk.

Usage:

```
1 python NS_extract_reads_brokenNames2.py nameoffastqfile.fastq
    nameofnamesfile.txt
```

9.2.3 Aurochs Illumina alignment filtration

This script opens a sam file and prints only the reads that aligned to the group I KIR genes. It will filter out any reads that alternatley align to groups other than group I. This file can be edited to do a similar task with the other KIR gene groups.

Usage:

```
1 python NS_extract_group1_reads.py samfile.sam
```

9.2.4 Genome reference sequence *KIR* removal

This script contains the regions of the cattle chromosome sequences that were cut for containing sequences determined after a blat search. All of the regions cut are within the python script and commented out. To print a chromosome without the KIR regions that line needs to be uncommented.

Usage:

```
1 python NS_print_fasta.py chrom.fasta > KIR_removed/chrom.fasta
```

9.2.5 Simulated datasets

To generate simulated data this file was created which generates fragments from a reference file using kmer and overlap sizes. The fragments can then be aligned and read coverage breadth can be calculated.

Usage:

```
python NS_generate_reads_both_strands.py reference.fasta kmer(
  int) overlap bases (int)
```

Read coverage breadths for each alignment method and each fragment size is plotted using this next script.

Usage:

```
python NS_slope_chart.py stats.txt
```

9.2.6 Coverage depth

Coverage depth of an aligned file is first calculated using the bed tools package and this command.

Command:

```
coverageBed -d -abam file.sorted.bam -b locations.bed >
  coveragedepth.bed
```

The sliding window average of the read depth is calculated and visualised using this next script.

Usage:

```
python NS_chart_KIR_DRX.py coveragedepth.bed
```

9.2.7 High resolution loci defining position analysis

To calculate the level of concordance between the aligned sequences and the reference sequence the next two scripts were written and used within a unix pipe. The csv files contain the different positions between the loci and have been generated using MEGA.

Usage:

```
1 python NS_redvar2.py csv/group3_3DXL33DXL5.csv pileup/buffalo/
    buffalo_seq9.pileup reference(int) | python output/
    NS_redvar_out_cons2.py threshold(float) referencesN(int)
```

For each animal this was repeated with all the different group gene combinations and threshold levels using the commands_buff.sh and commands_do_range_buff.sh bash shell scripts. The output from these files was graphically shown using the next script.

Usage:

```
1 python NS_bar_stacked.py threshold_1.txt threshold_0.75.txt
    threshold_0.5.txt threshold_0.25.txt out.pdf name
```

9.3 Chapter 4 Appendix

9.3.1 PacBio vector screen

To perform the PacBio sequence assembly the smrtportal software needs to be installed. This software needs to be sourced within the users path.

command:

```
source /opt/smrtanalysis/etc/setup.sh
```

Next align reads to vector sequence (pTARBAC1.3.fasta) using the blasr prgram.

command:

```
blasr
    m130723_093514_42149_c100529602550000001823089211101325_s1_p0
    .bas.h5 pTARBAC1.3.fasta -bestn 1 -header > ecoli.align
```

The number of holes needs to be determined.

command:

```
h5dump -y -d /PulseData/BaseCalls/ZMW/HoleNumber
    m130723_093514_42149_c100529602550000001823089211101325_s1_p0
    .bas.h5 | head
```

Output is:

```
DATASPACE  SIMPLE { ( 81741 ) / ( H5S_UNLIMITED ) }
```

Therefore 81741 holes are present, use the script as follows to generate a whitelist of reads for assembly that do no contain vector.

Usage:

```
python whitelist.py ecoli.align 81740 > whitelist.txt
```

This whitelist file needs to be added to the filtering module of the settings.xml file used for assembling the data within the smrtportal pipeline. A settings.xml file can be used from a prevvious assembly attempt without filtering.

xml lines:

```
<param name="whiteList" label="Minimum Subread Length">
<value>/data/sanders/pacbio/vecotor_screen/263M01/A01_1/
    Analysis_Results/whitelist.txt</value>
</param>
```

Then create a file of file names, add it to the input xml and run the smrtpipe.

Commands:

```
ls /home/USERNAME/BACassembly/
    m120729_040044_42134_c100384402550000001523033010171256_s1_p0
    .bas.h5 > input.fofn

fofnToSmrtpipeInput.py input.fofn > input.xml

smrtpipe.py --params=settings.xml xml:input.xml
```

9.4 Chapter 5 Appendix

9.4.1 Read depth coverage of the other animals

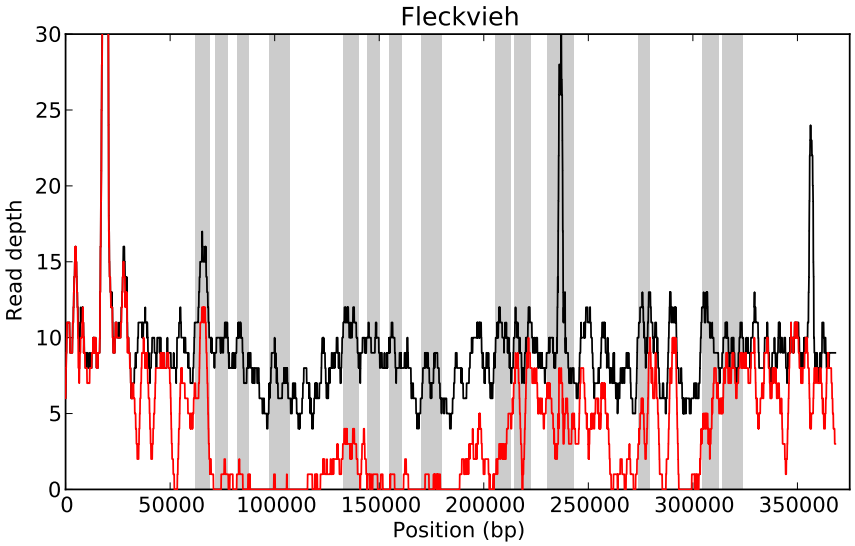


Figure S2: Fleckvieh WGS CKH read coverage. The black line represents the normal unfiltered read coverage and the red line represents reads filtered for unique alignment. The read coverage is a mean of a 300 bp sliding window. For reference the grey columns represent *KIR* loci positions and from left to right are: *3DXL6*, *2DS3*, *3DXS3*, *3DXL7*, *3DXL4*, *2DS2*, *3DXS2*, *3DXL5*, *3DXL2*, *2DS1*, *3DXL3*, *3DXS1*, *3DXL1* and *2DL1*

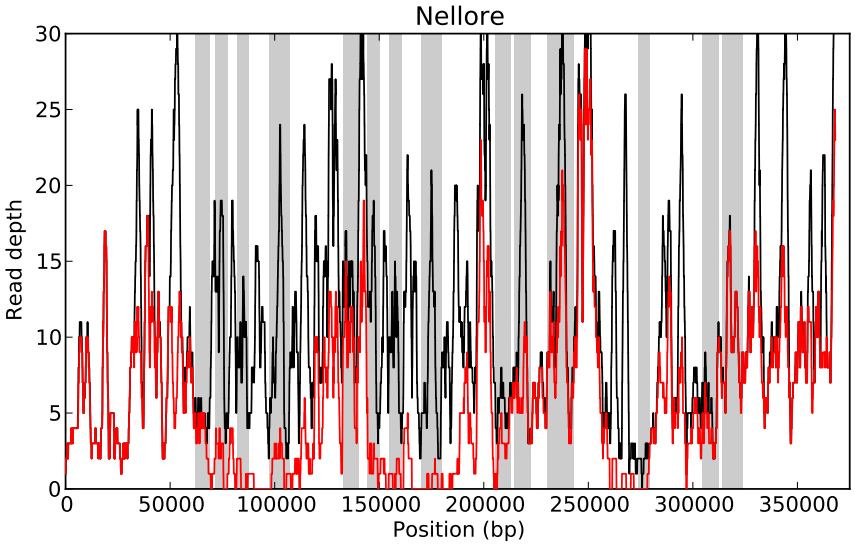


Figure S3: Nellore WGS CKH read coverage. The black line represents the normal unfiltered read coverage and the red line represents reads filtered for unique alignment. The read coverage is a mean of a 300 bp sliding window. For reference the grey columns represent *KIR* loci positions and from left to right are: *3DXL6*, *2DS3*, *3DXS3*, *3DXL7*, *3DXL4*, *2DS2*, *3DXS2*, *3DXL5*, *3DXL2*, *2DS1*, *3DXL3*, *3DXS1*, *3DXL1* and *2DL1*

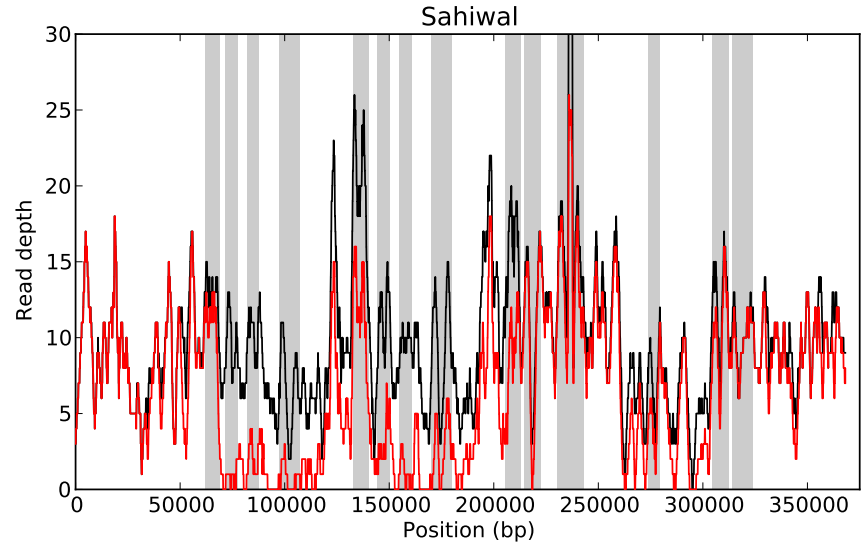


Figure S4: Sahiwal WGS CKH read coverage. The black line represents the normal unfiltered read coverage and the red line represents reads filtered for unique alignment. The read coverage is a mean of a 300 bp sliding window. For reference the grey columns represent *KIR* loci positions and from left to right are: *3DXL6*, *2DS3*, *3DXS3*, *3DXL7*, *3DXL4*, *2DS2*, *3DXS2*, *3DXL5*, *3DXL2*, *2DS1*, *3DXL3*, *3DXS1*, *3DXL1* and *2DL1*

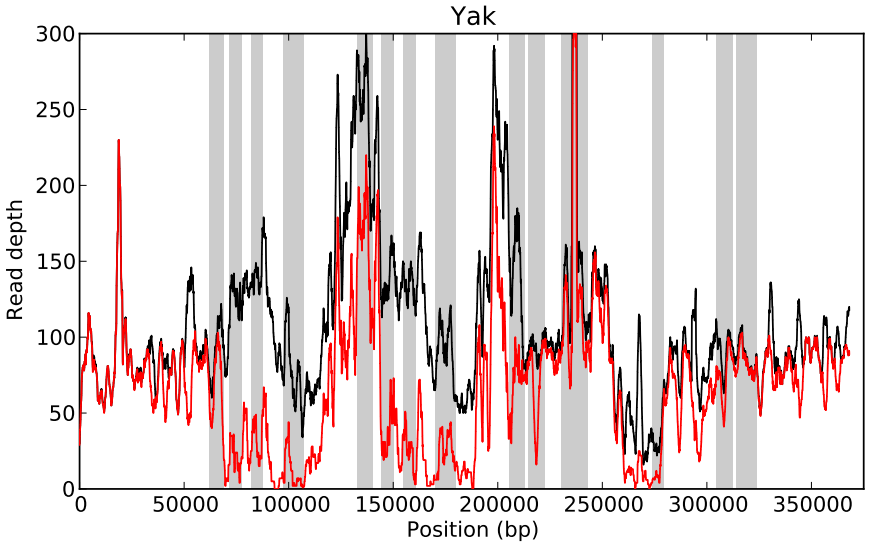


Figure S5: Yak WGS CKH read coverage. The black line represents the normal unfiltered read coverage and the red line represents reads filtered for unique alignment. The read coverage is a mean of a 300 bp sliding window. For reference the grey columns represent *KIR* loci positions and from left to right are: *3DXL6*, *2DS3*, *3DXS3*, *3DXL7*, *3DXL4*, *2DS2*, *3DXS2*, *3DXL5*, *3DXL2*, *2DS1*, *3DXL3*, *3DXS1*, *3DXL1* and *2DL1*

9.5 Chapter 6 Appendix

9.5.1 Filtering Bowtie2 results

To filter the bam file from reads that alternately mapped this python script was used with a unix pipe along with samtools.

Usage:

```
samtools view file.bam | python NS_filter_bowtie.py | samtools view
  -Sbt hap1_genome.fasta.fai - > filtered/file.bam 2> filtered/file.
  bam.log
```

9.5.2 CNV prediction from read depth

To determine the relative increase or decrease in coverage depth which may be indicative of CNV, the next two scripts were written.

Generate base by base relative read changes:

```
python NS_KIR_CNV4_print_ratiolist.py reference_coverage.bed \
coverage_depth.bed exons_detailed.bed > ratios.txt
```

Generate box plot of avbove values for each exon:

```
python NS_box_plot_exons.py exons_detailed.bed ratios.txt
```

9.5.3 Dendrogram of SNP difference numbers

The number SNPs of each animal contains that are different or not seen within in another animal was calculated using MySQL. This was repeated for all the combinations of animals in a pairwise fashion to generate a matrix of SNP differences. The matrix was used to generate a dendrogram and therefore infer phylogenetics using the shared SNP positions between all of the animals. The entire haplotype sequence was used excluding the LILR regions (start at 60 kb into the CKH reference sequence). A bespoke python script was written to generate the dendrogram.

Generate dendrogram of list of files containing SNP difference numbers:

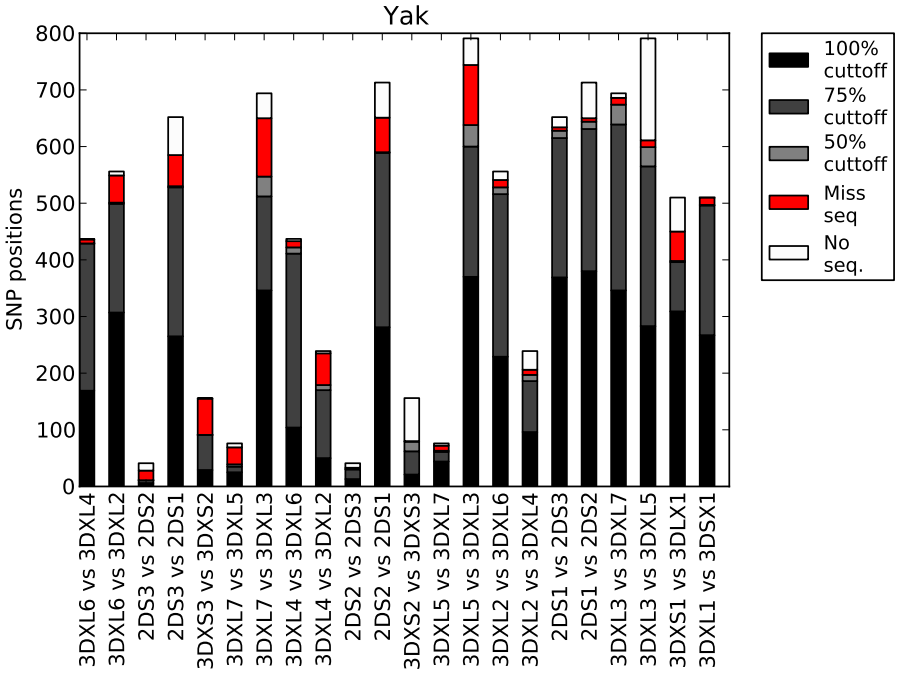


Figure S6: High resolution SNP analysis of yak. Comparison of gene defining SNP positions between gene group loci.

```
1 python NS_dendro.py file.fofn
```

9.5.4 Inferred and actual fragment sizes for each animal

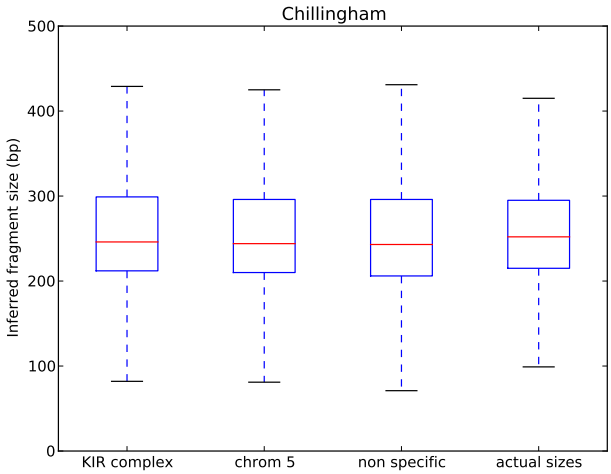


Figure S7

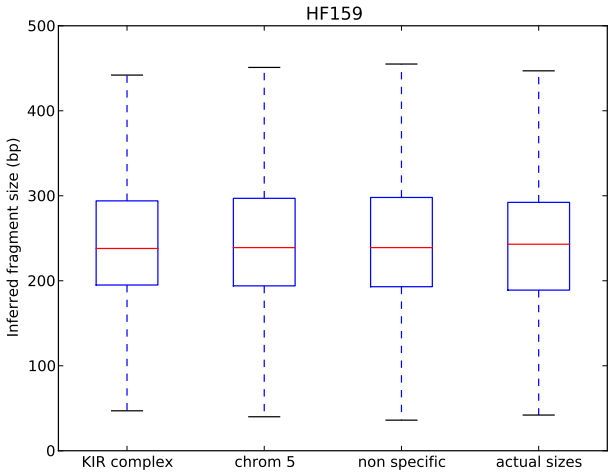


Figure S8

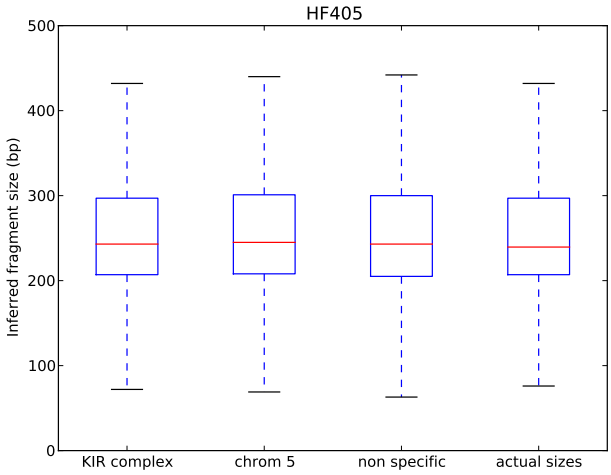


Figure S9

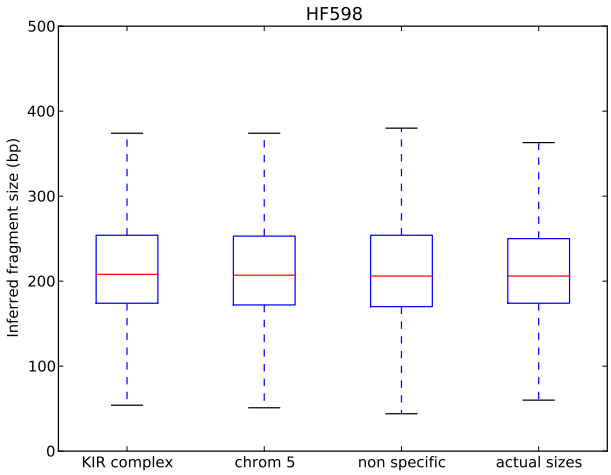


Figure S10

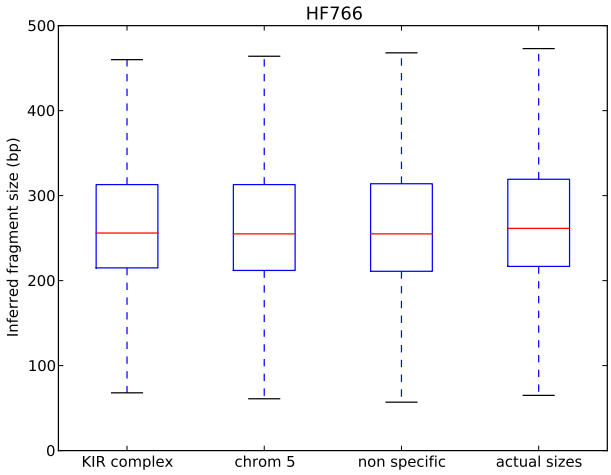


Figure S11

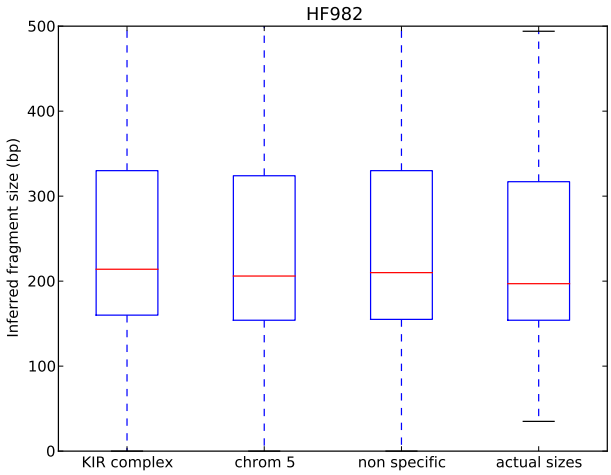


Figure S12

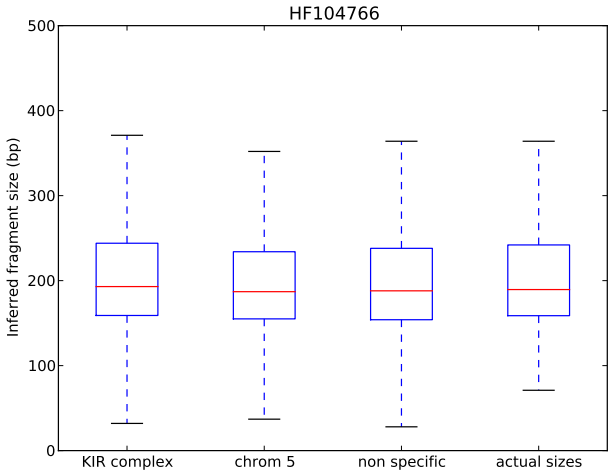


Figure S13

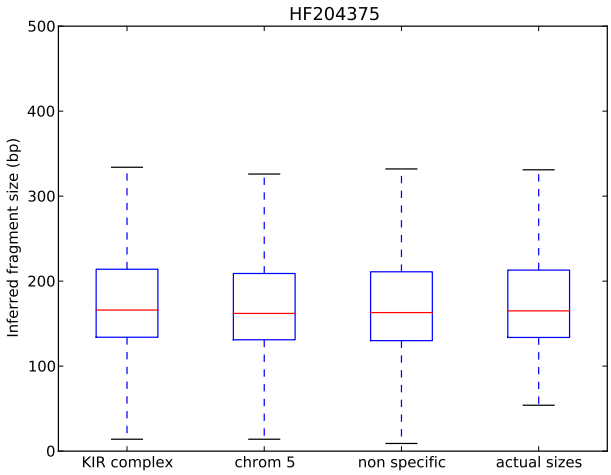


Figure S14

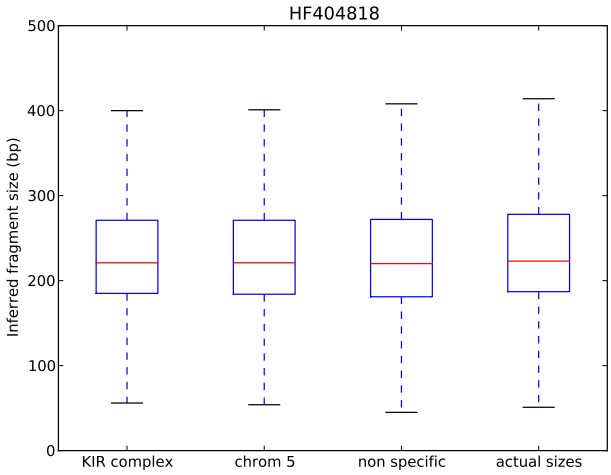


Figure S15

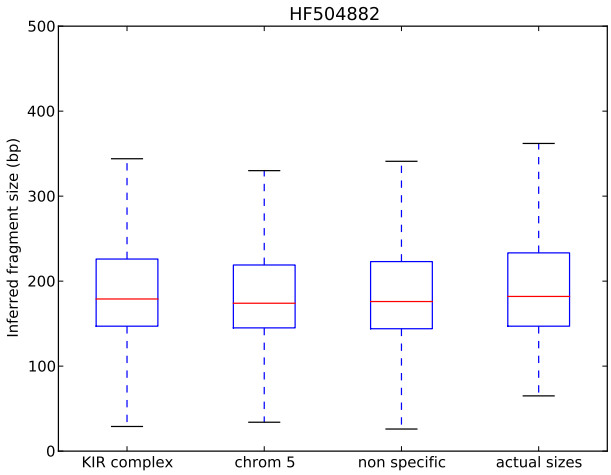


Figure S16

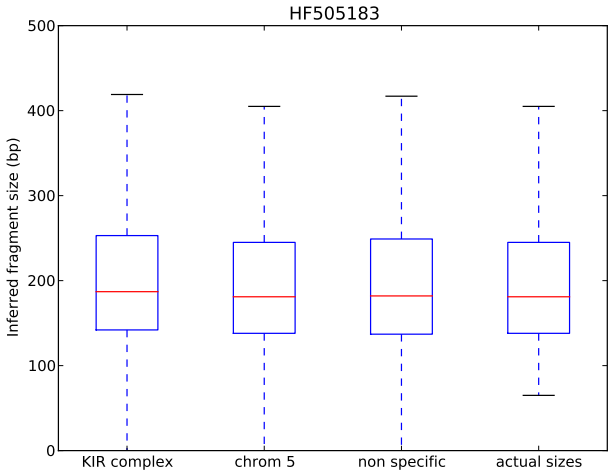


Figure S17

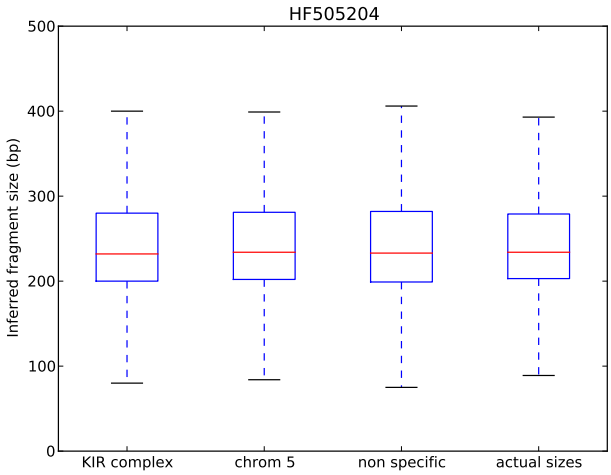


Figure S18

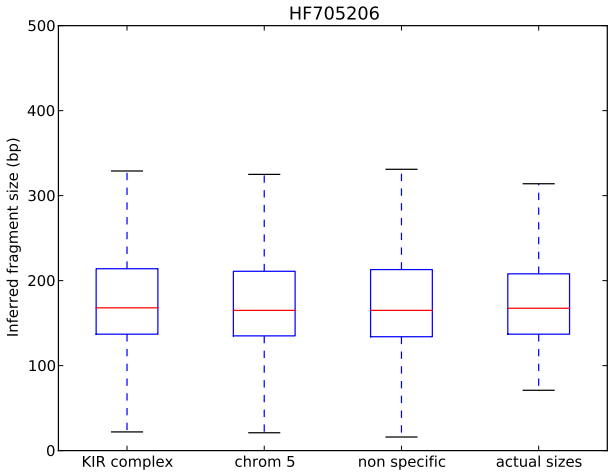


Figure S19

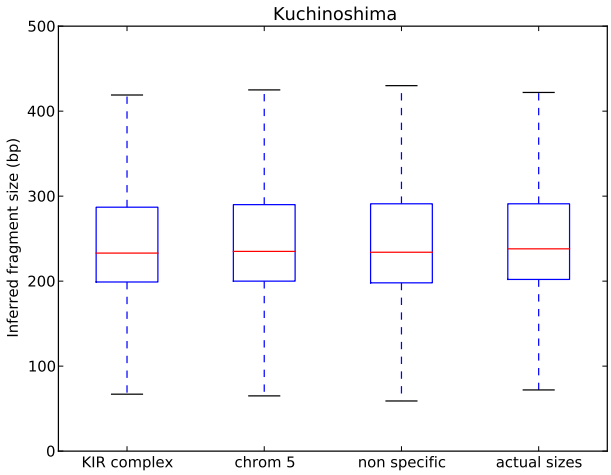


Figure S20

9.5.5 read coverage depth histograms for each animal

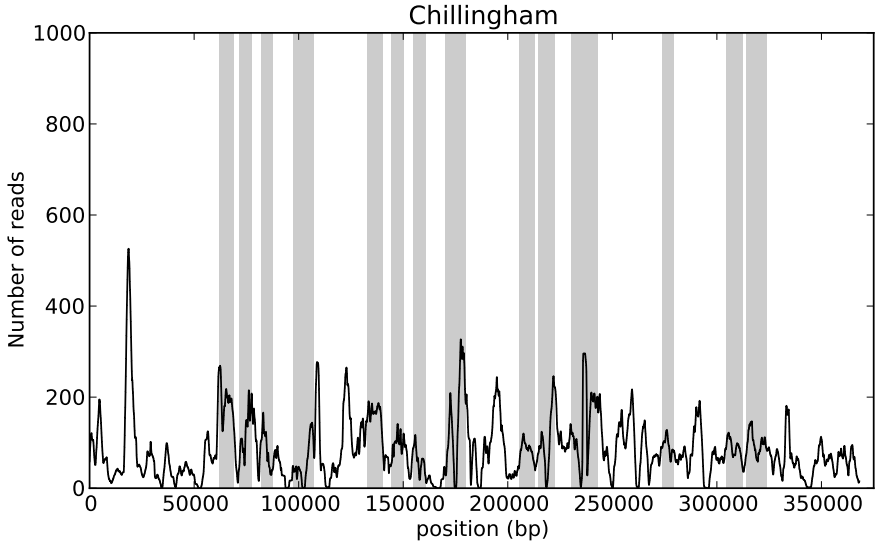


Figure S21

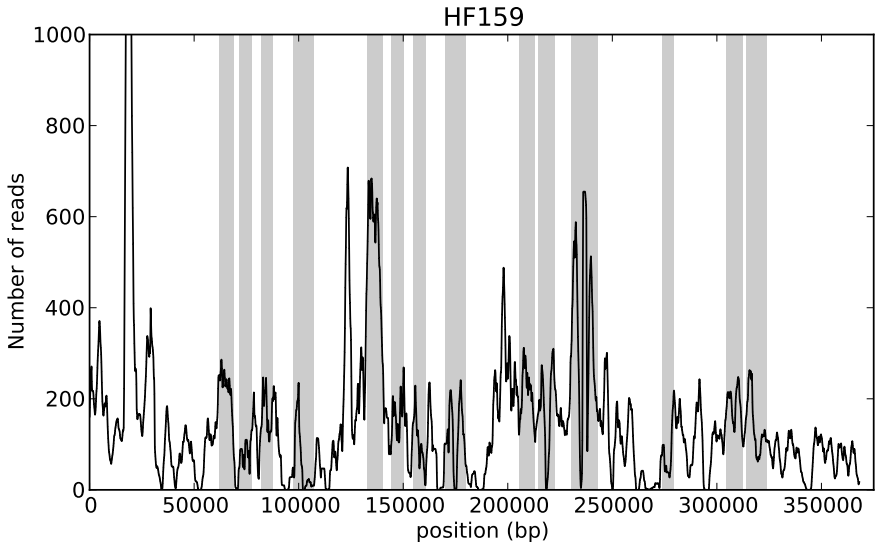


Figure S22

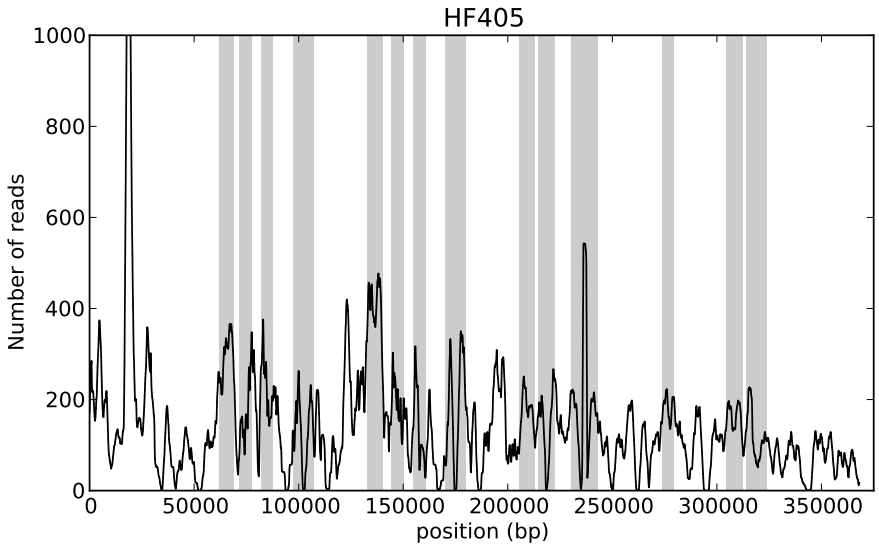


Figure S23

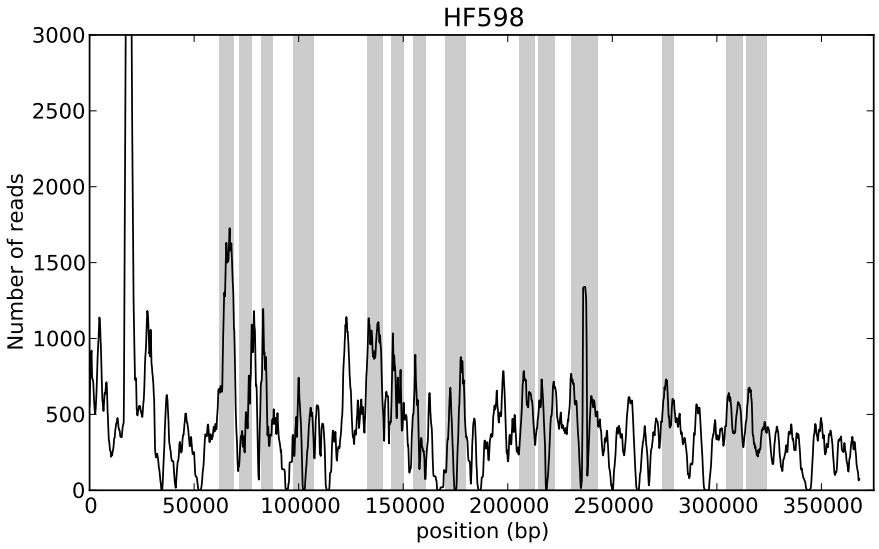


Figure S24

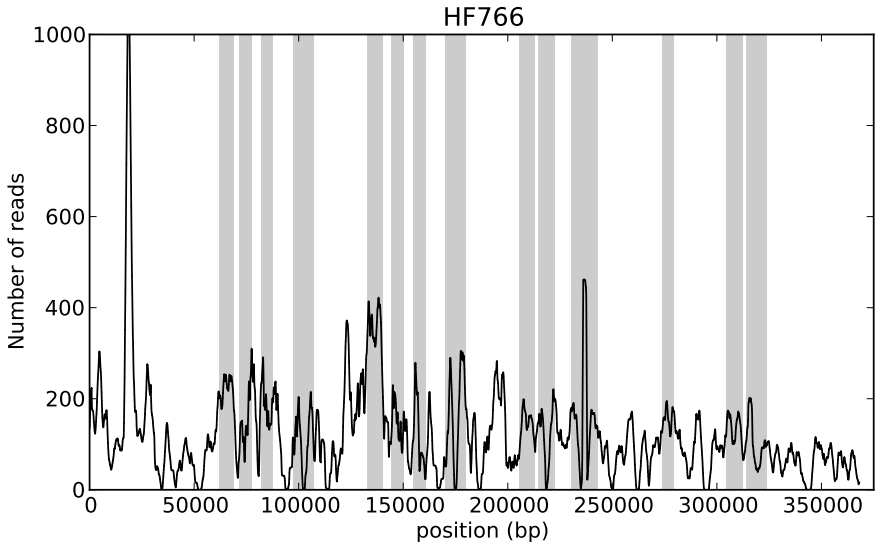


Figure S25

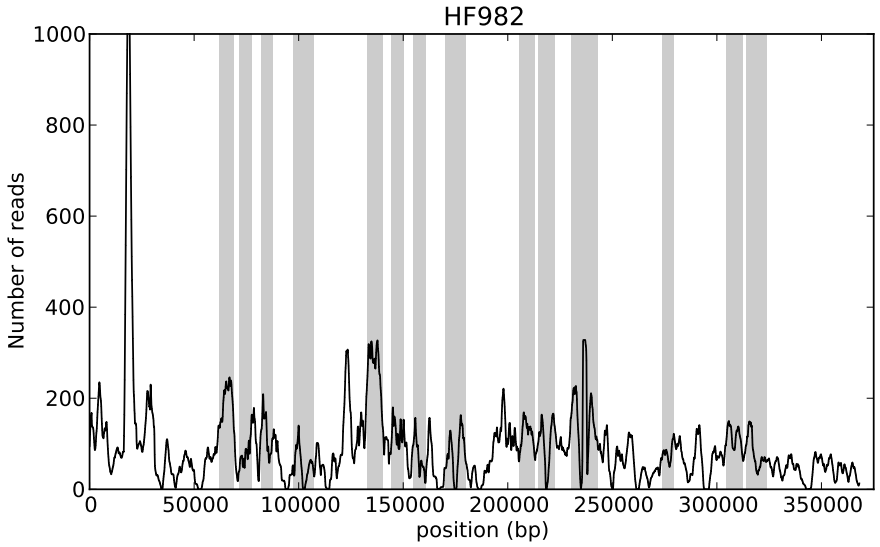


Figure S26

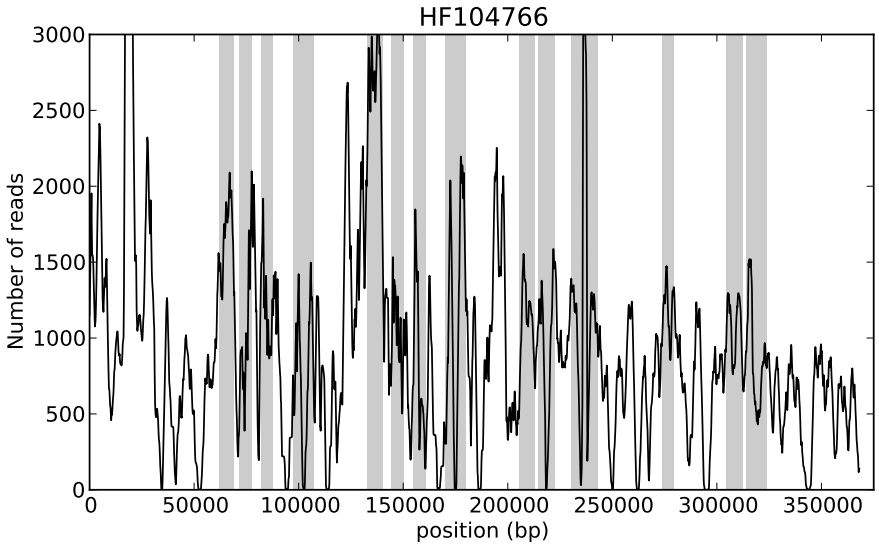


Figure S27

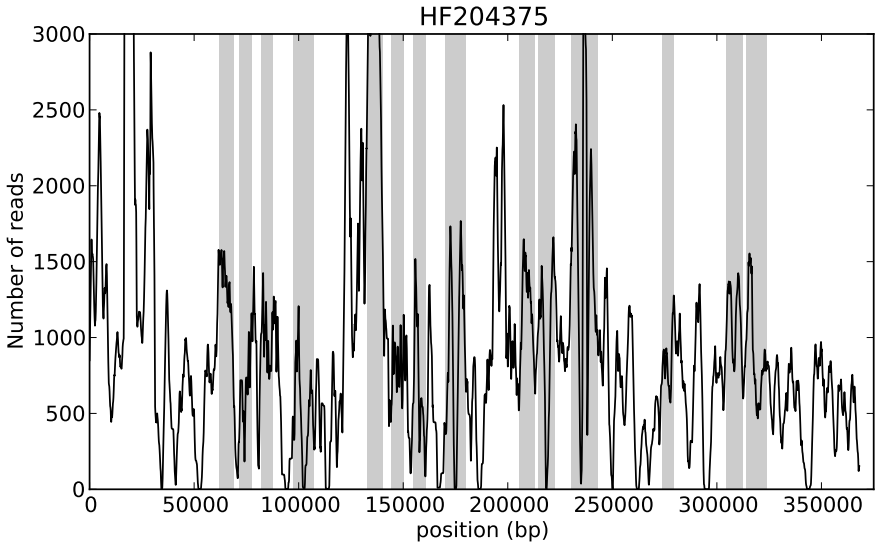


Figure S28

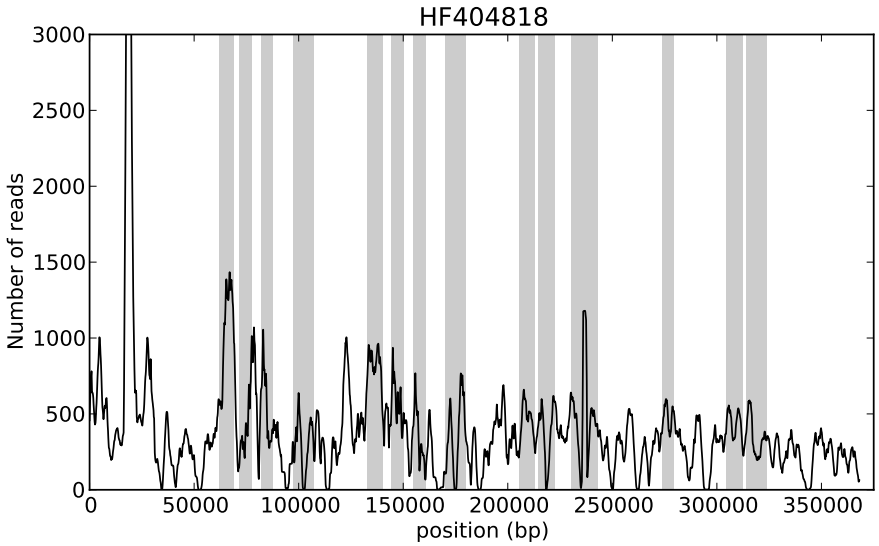


Figure S29

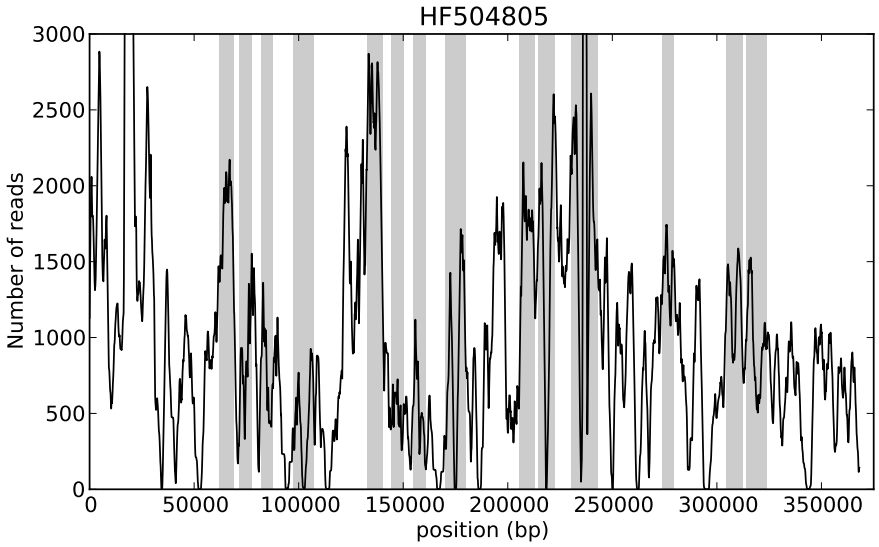


Figure S30

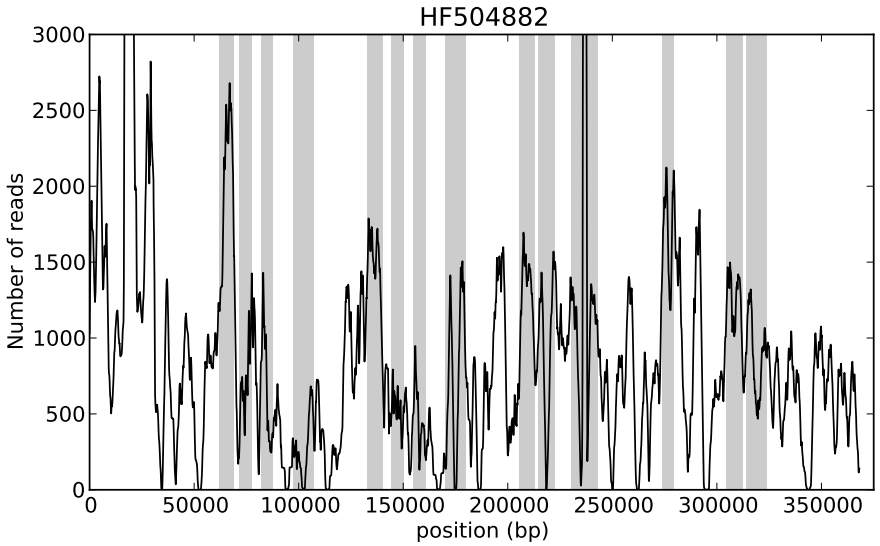


Figure S31

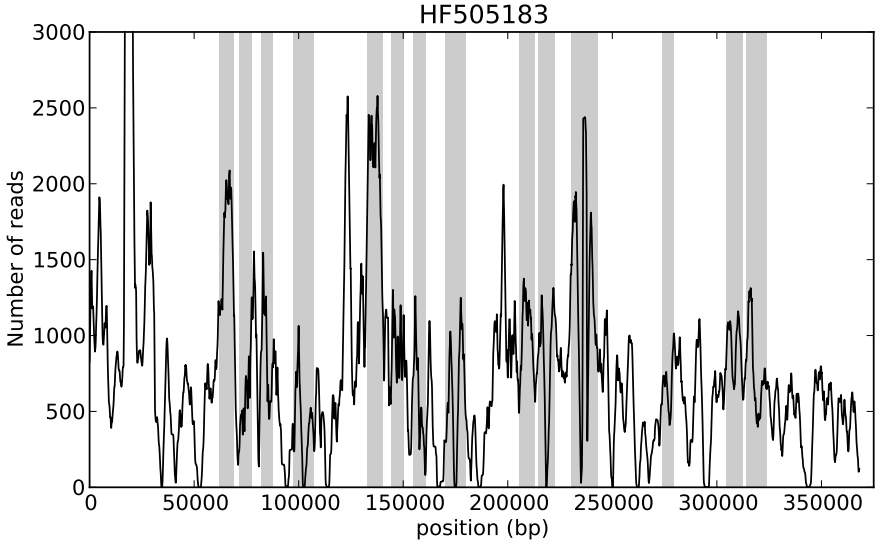


Figure S32

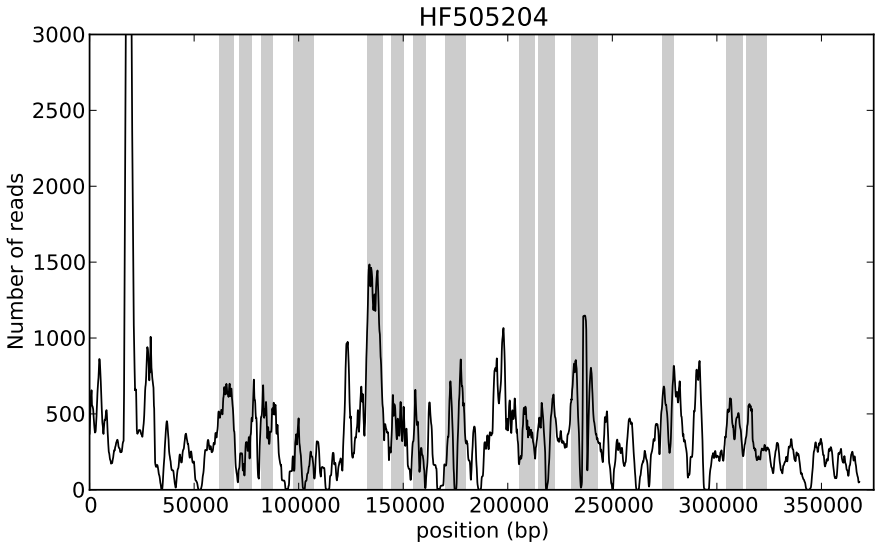


Figure S33

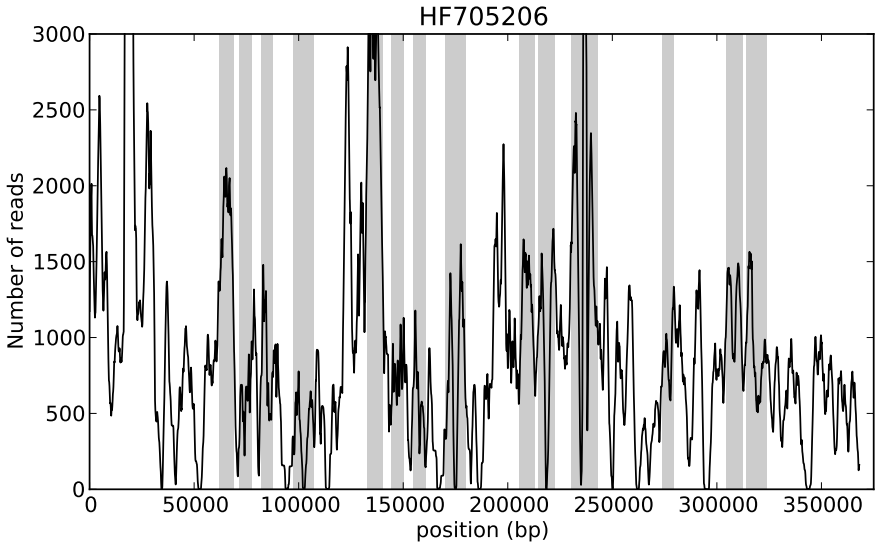


Figure S34

9.5.6 CNV boxplot relative proportions

CNV relative read depth change box plots are shown for each exon of each animal Figures S35 to S58. Each box plot has been produced the same way as Figure 55.

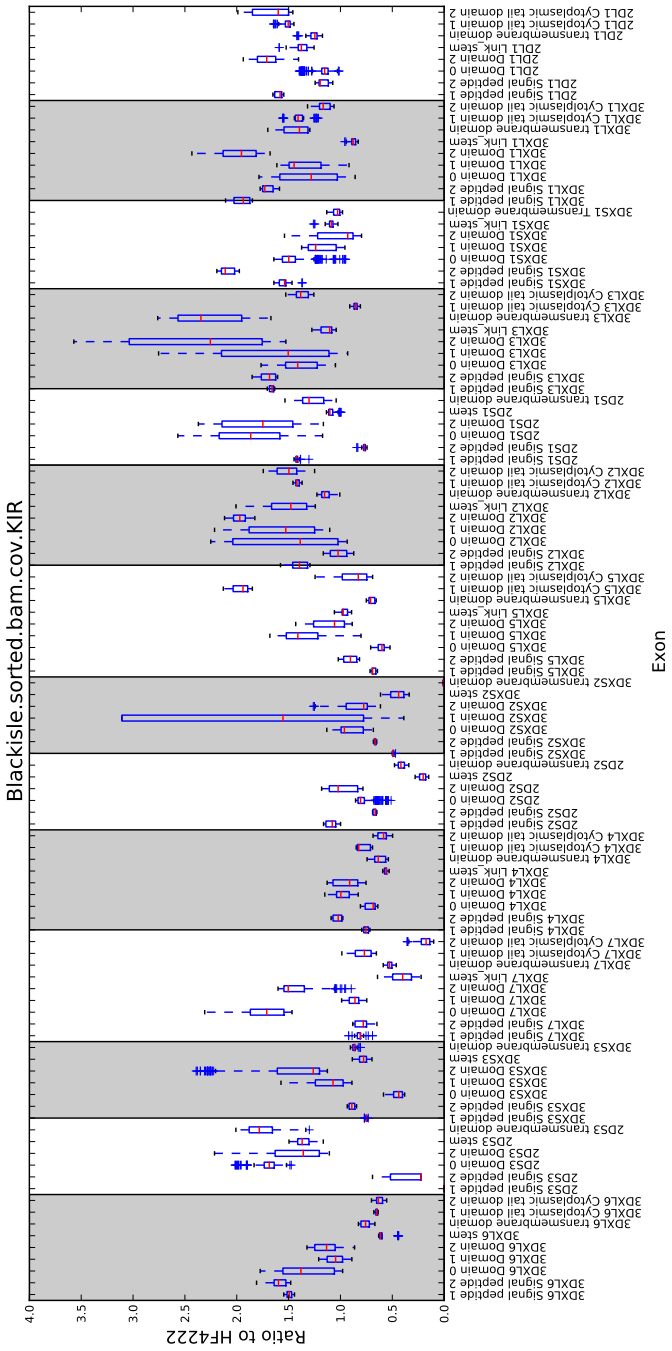


Figure S35: CNV exon prediction of Blackisle. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

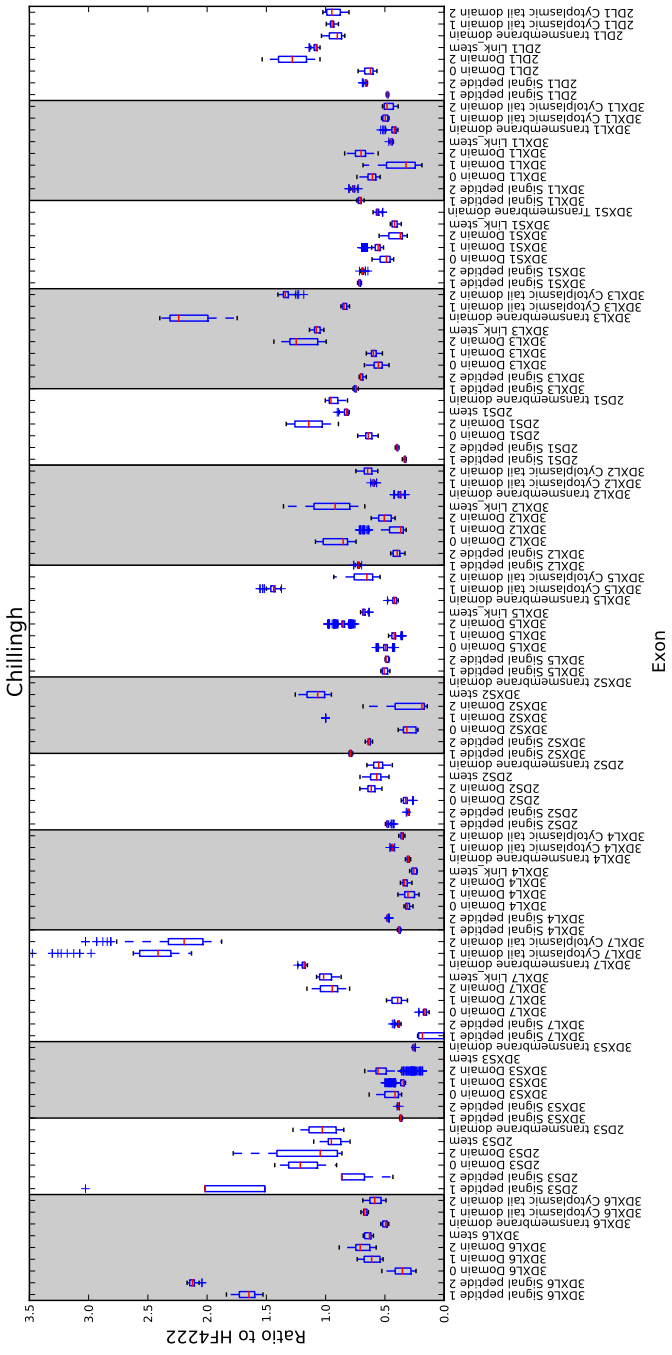


Figure S36: CNV exon prediction of Chillingham250. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

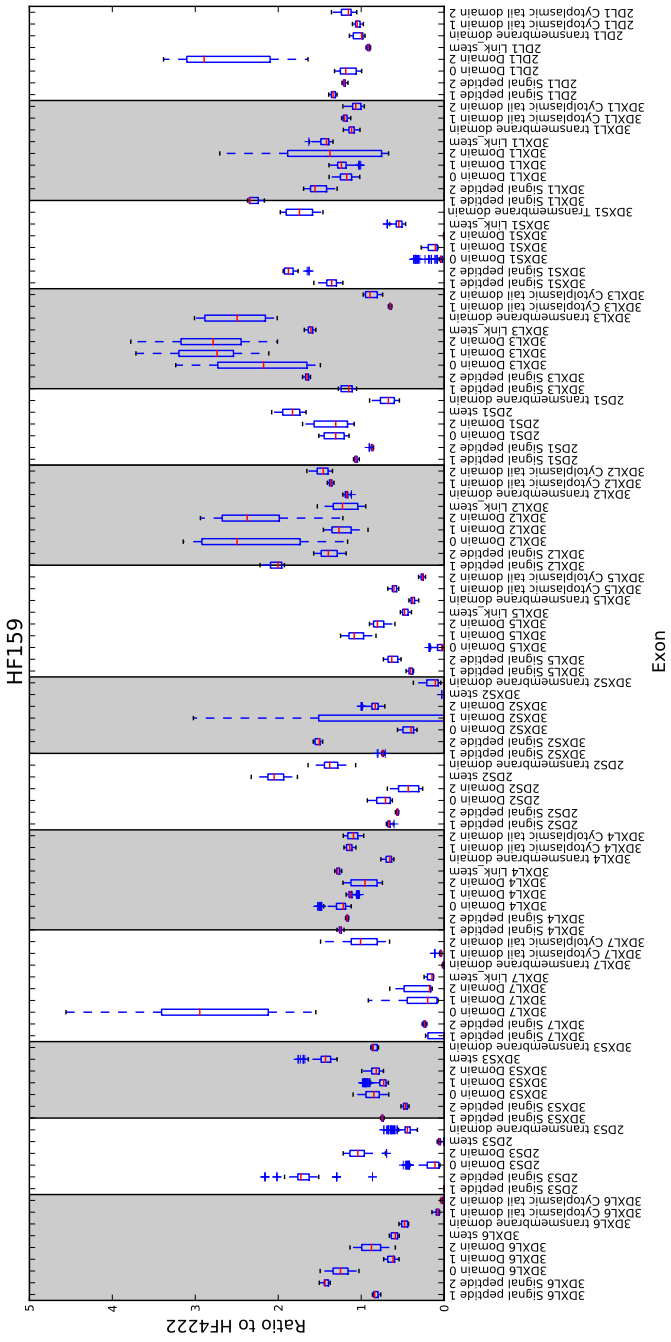


Figure S37: CNV exon prediction of HF159. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

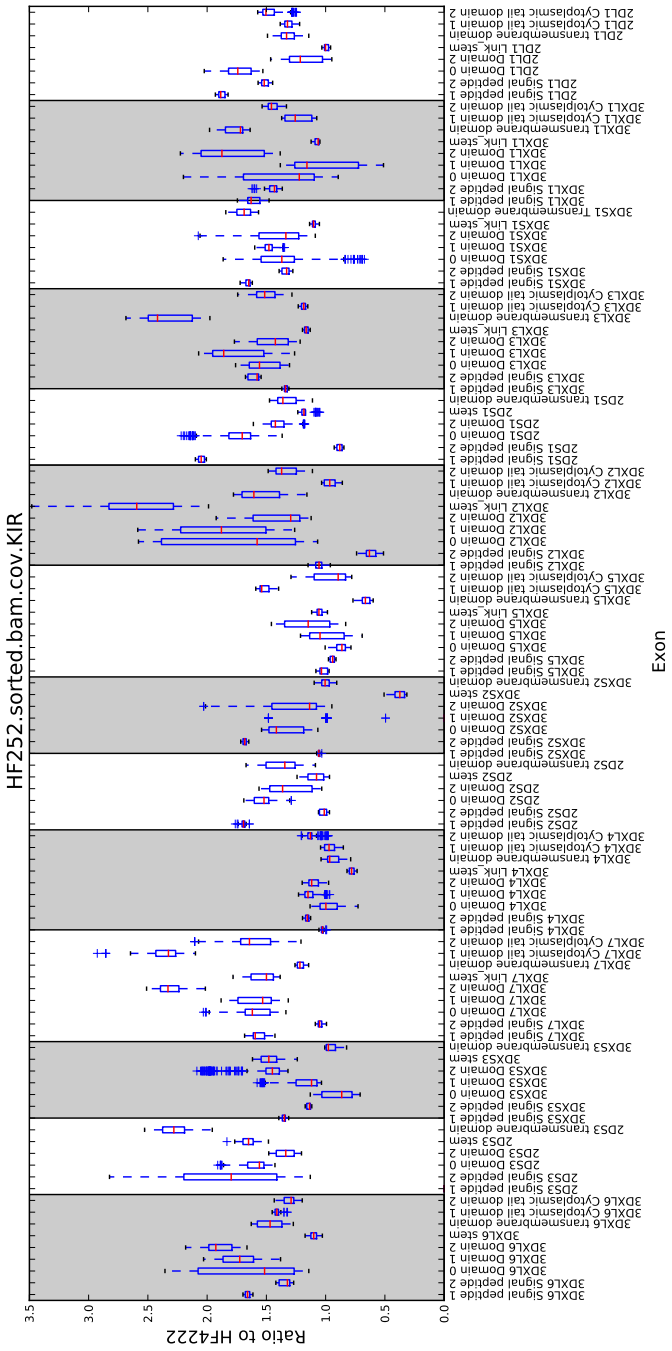


Figure S38: CNV exon prediction of HF252. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

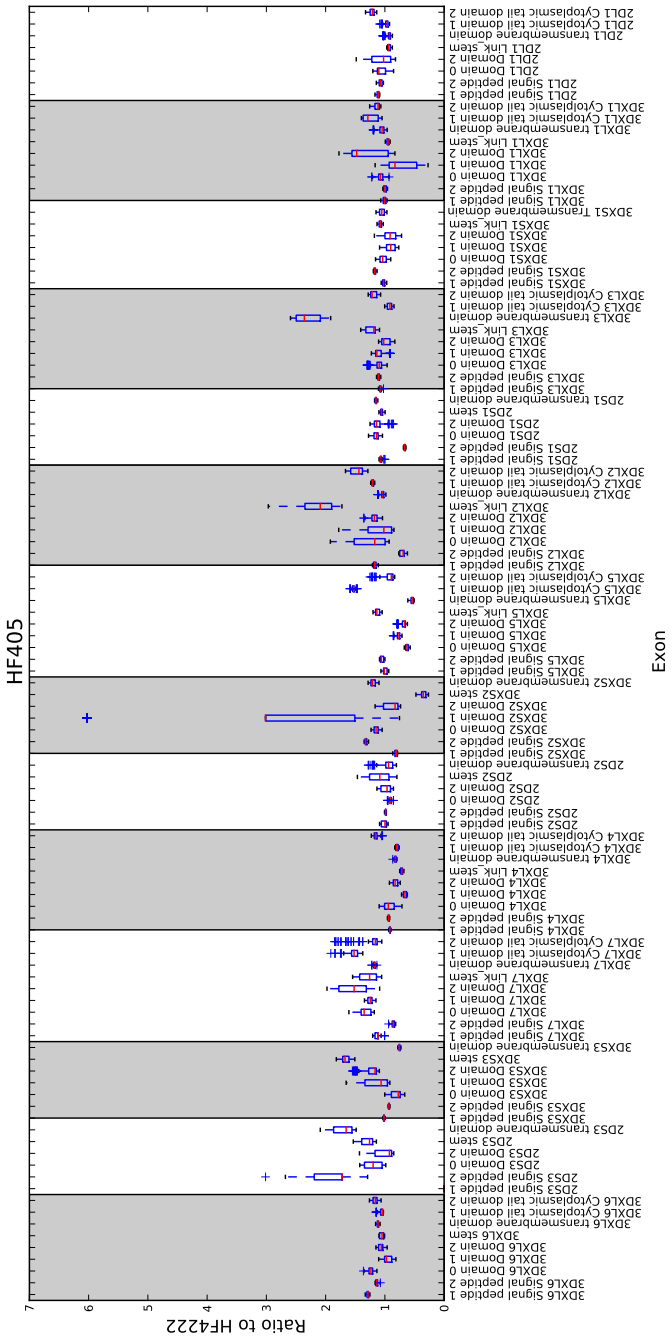


Figure S39: CNV exon prediction of HF405. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

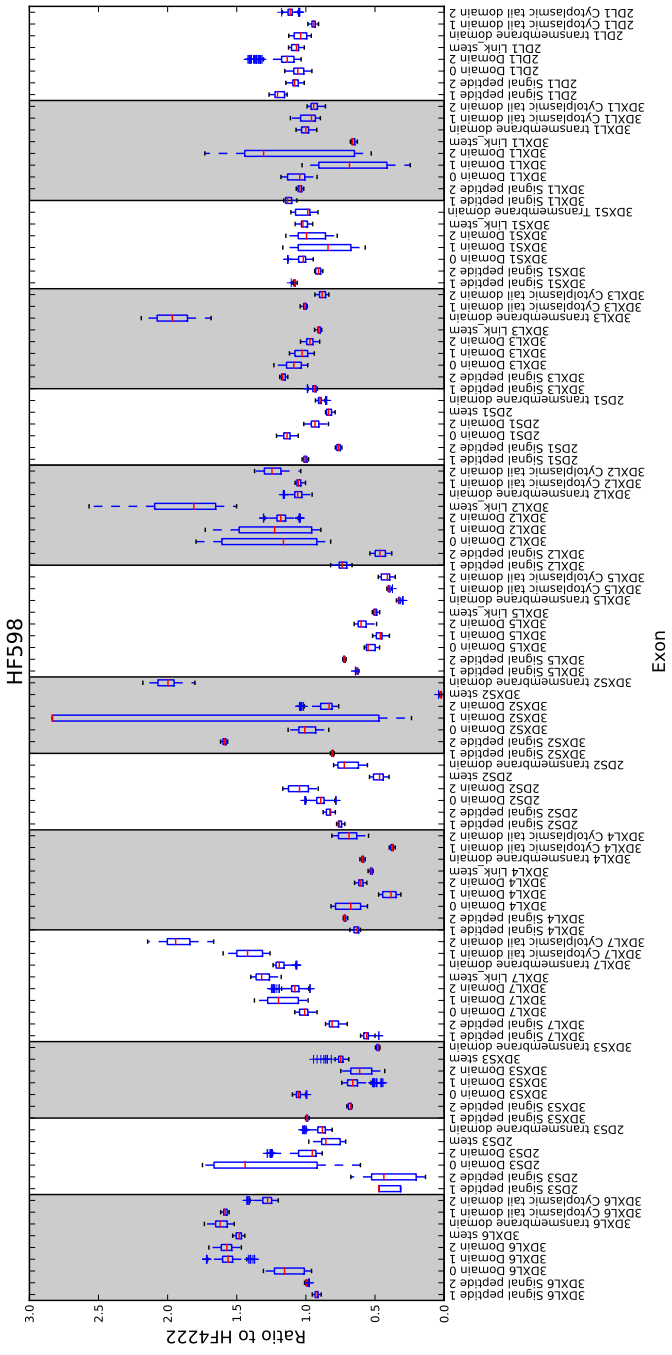


Figure S40: CNV exon prediction of HF598. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

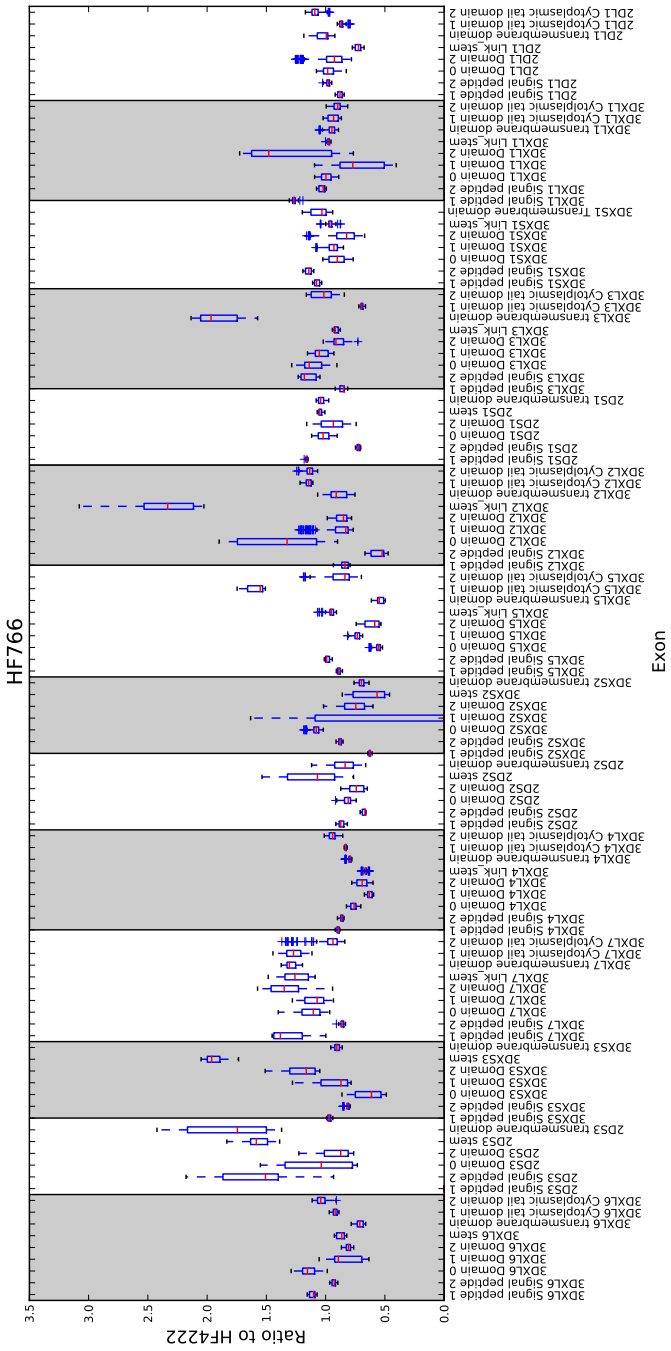


Figure S41: CNV exon prediction of HF766. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

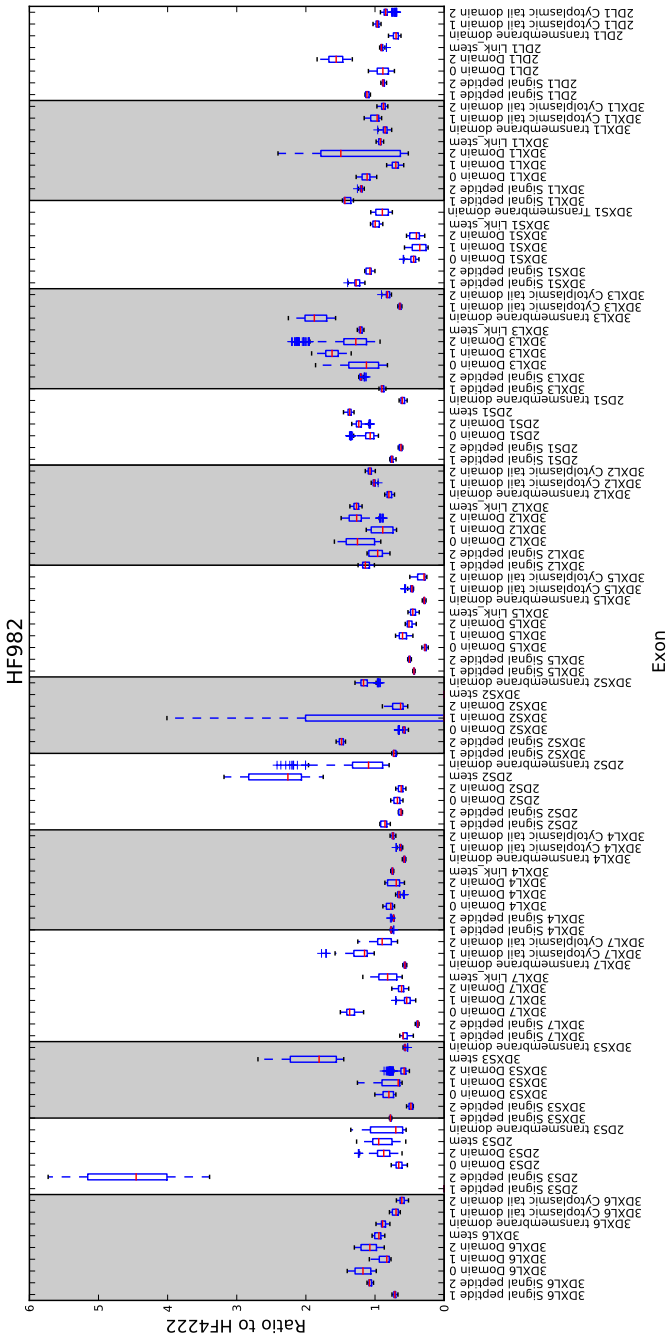


Figure S42: CNV exon prediction of HF982. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

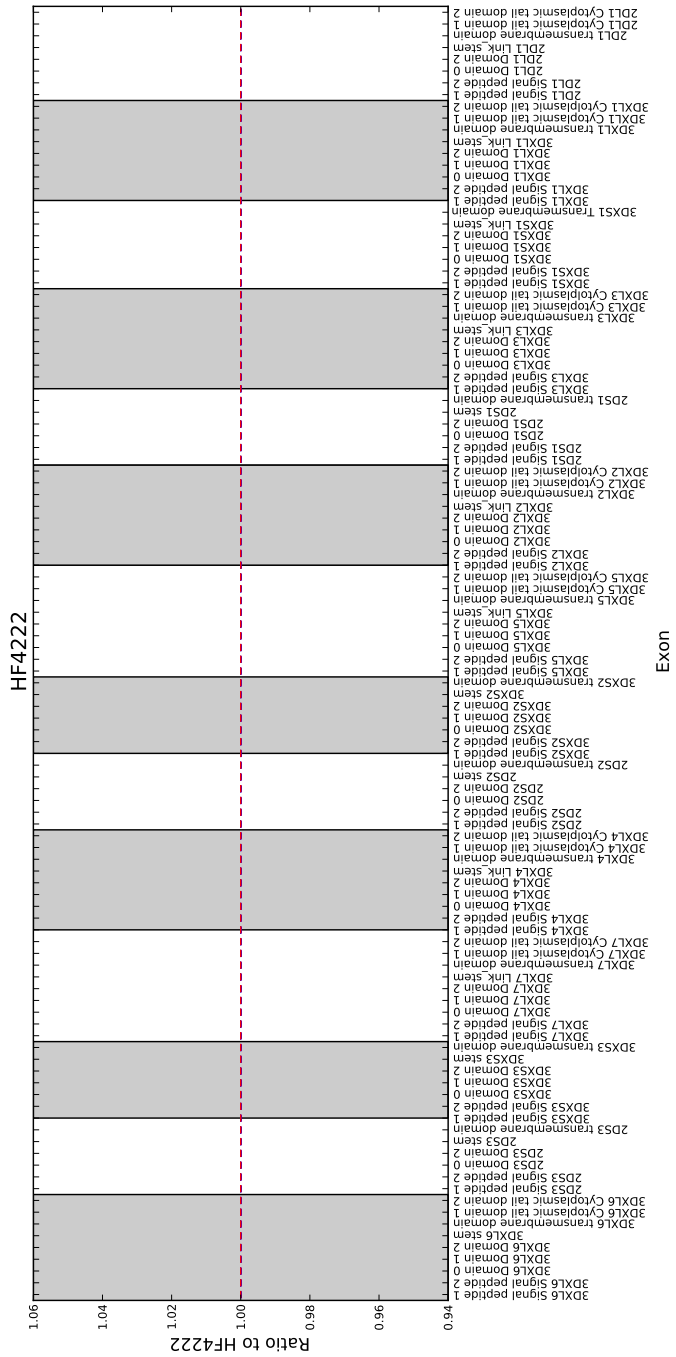


Figure S43: CNV exon prediction of HF4222. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

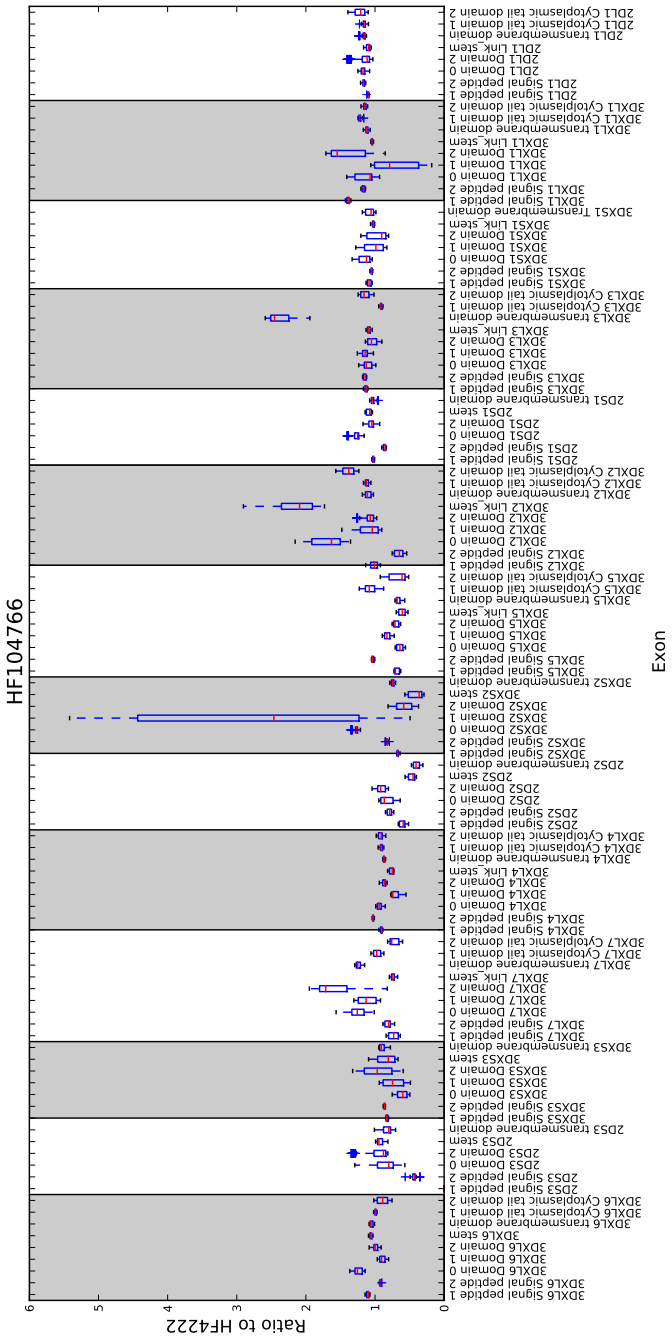


Figure S44: CNV exon prediction of HF104766. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

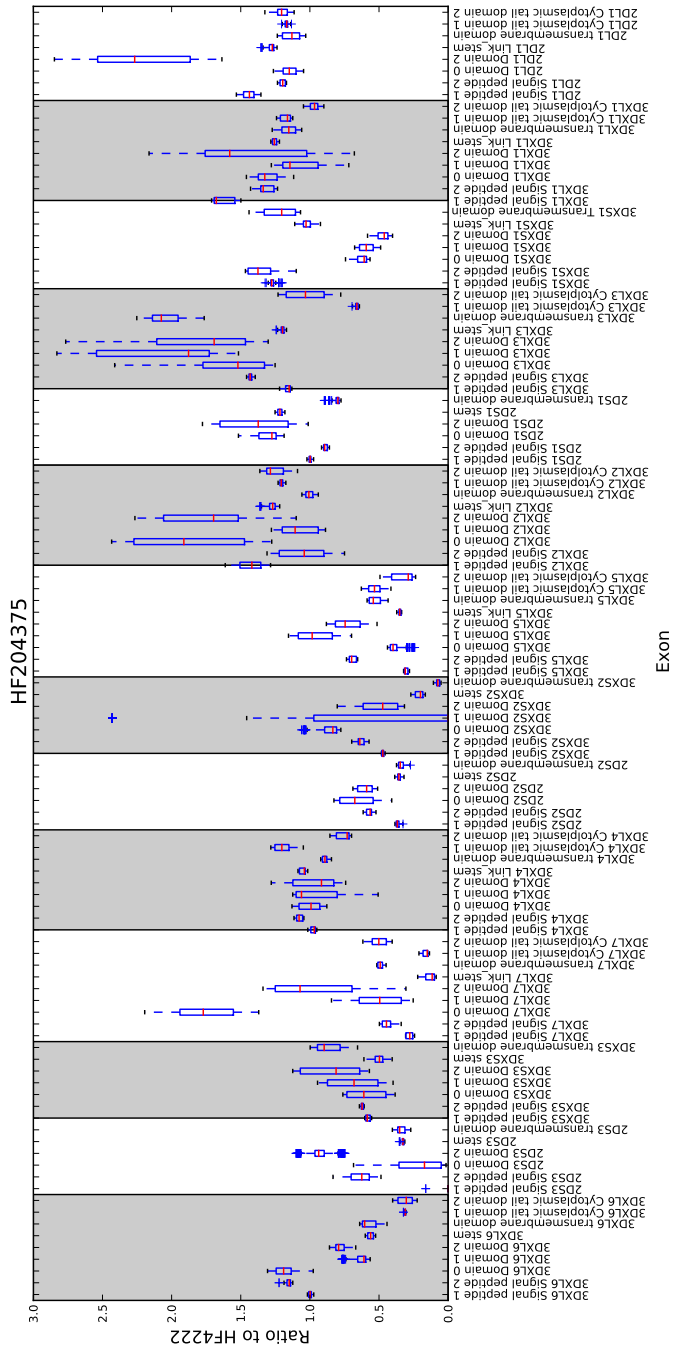


Figure S45: CNV exon prediction of HF204375. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

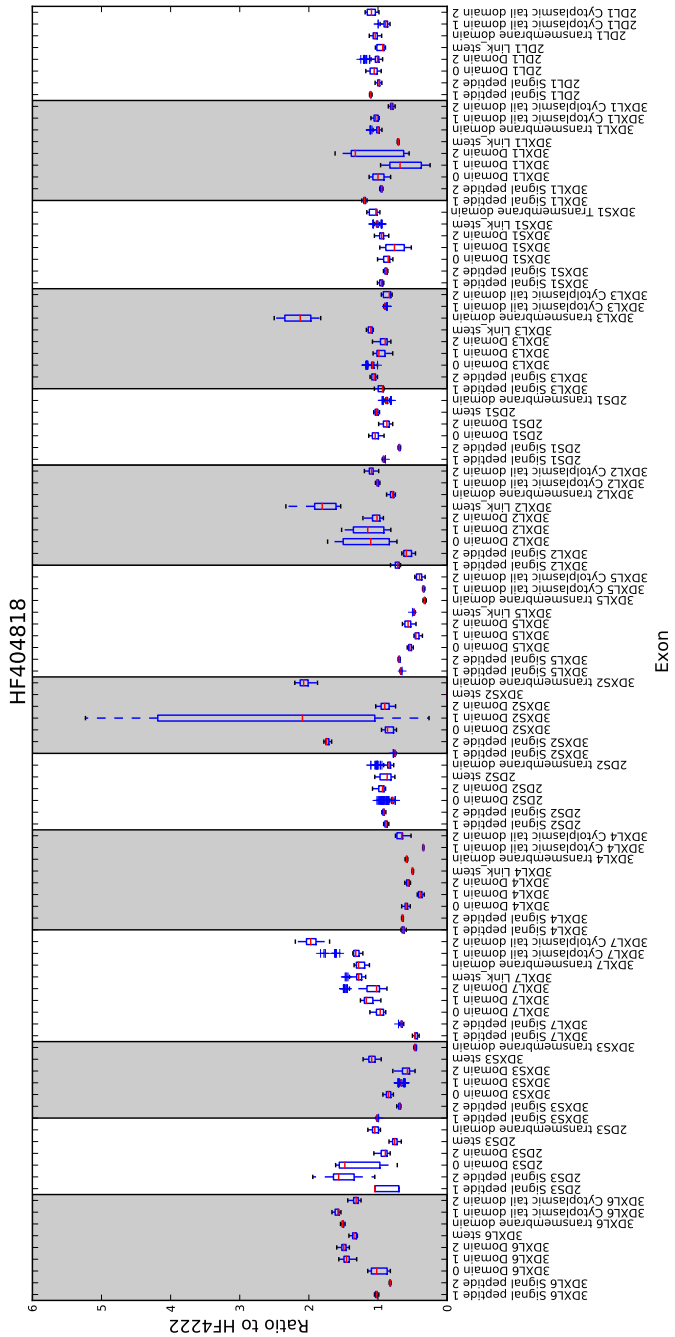


Figure S46: CNV exon prediction of HF404818. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

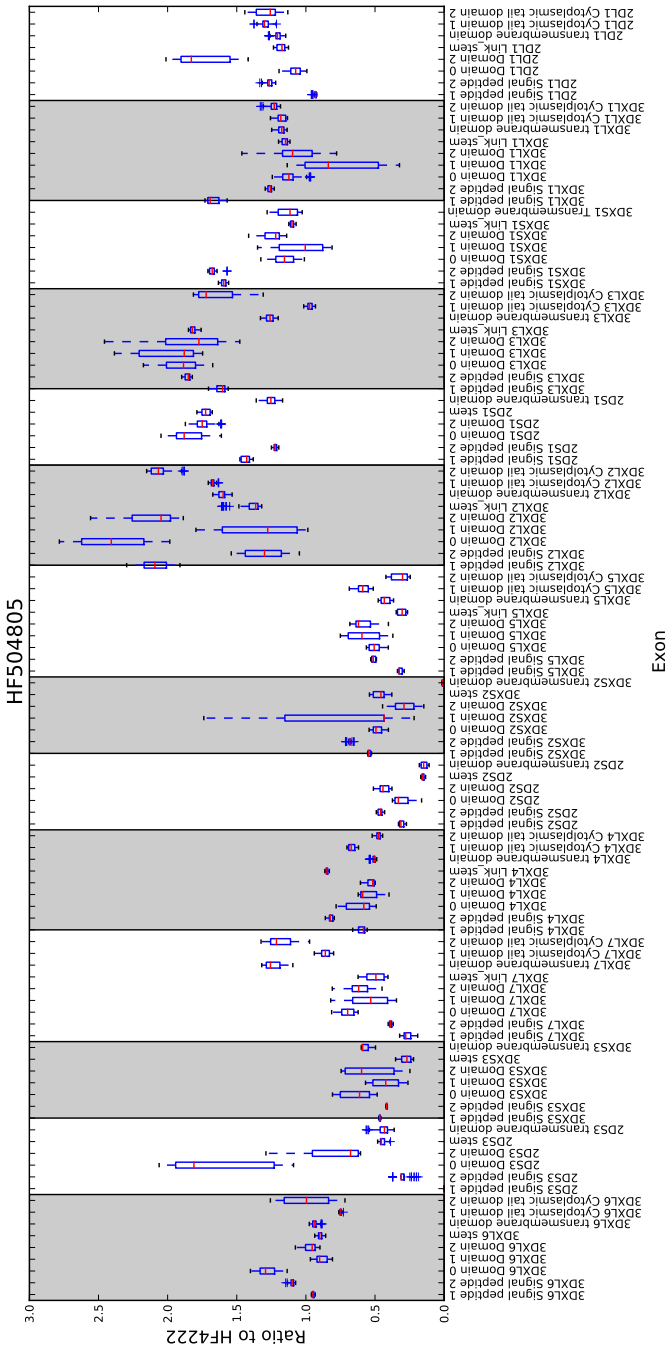


Figure S47: CNV exon prediction of HF504805. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

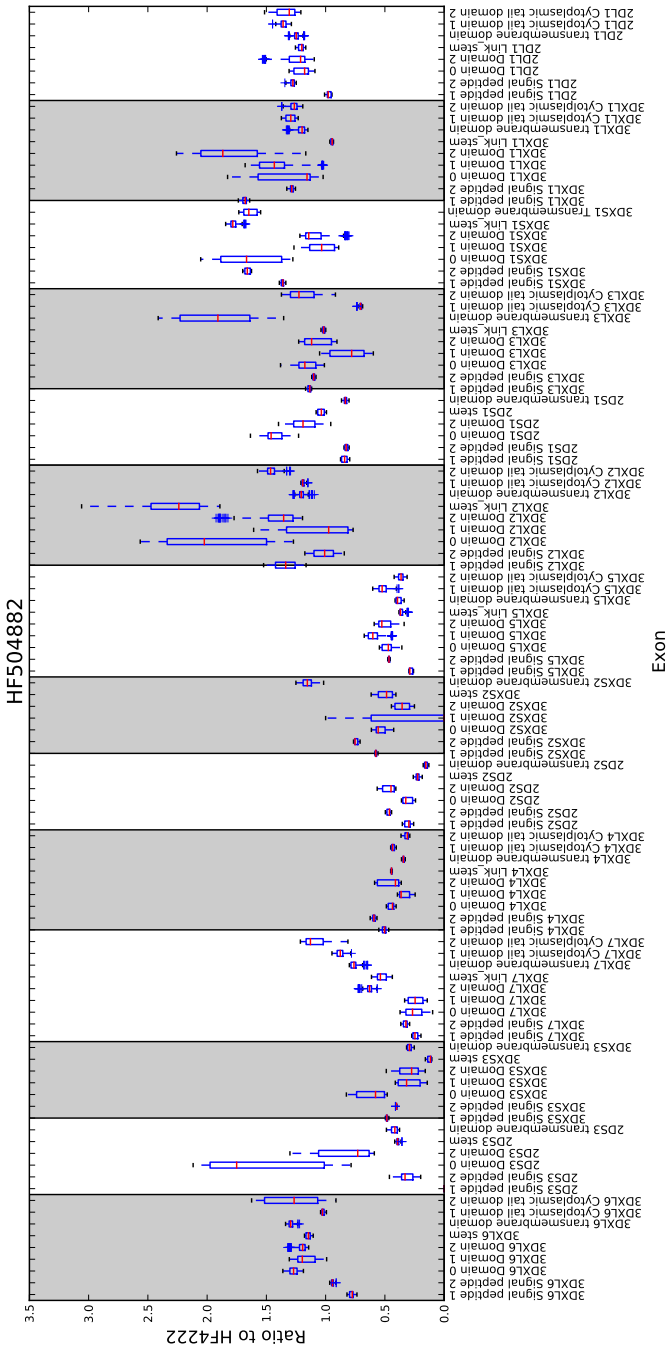


Figure S48: CNV exon prediction of HF504882. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

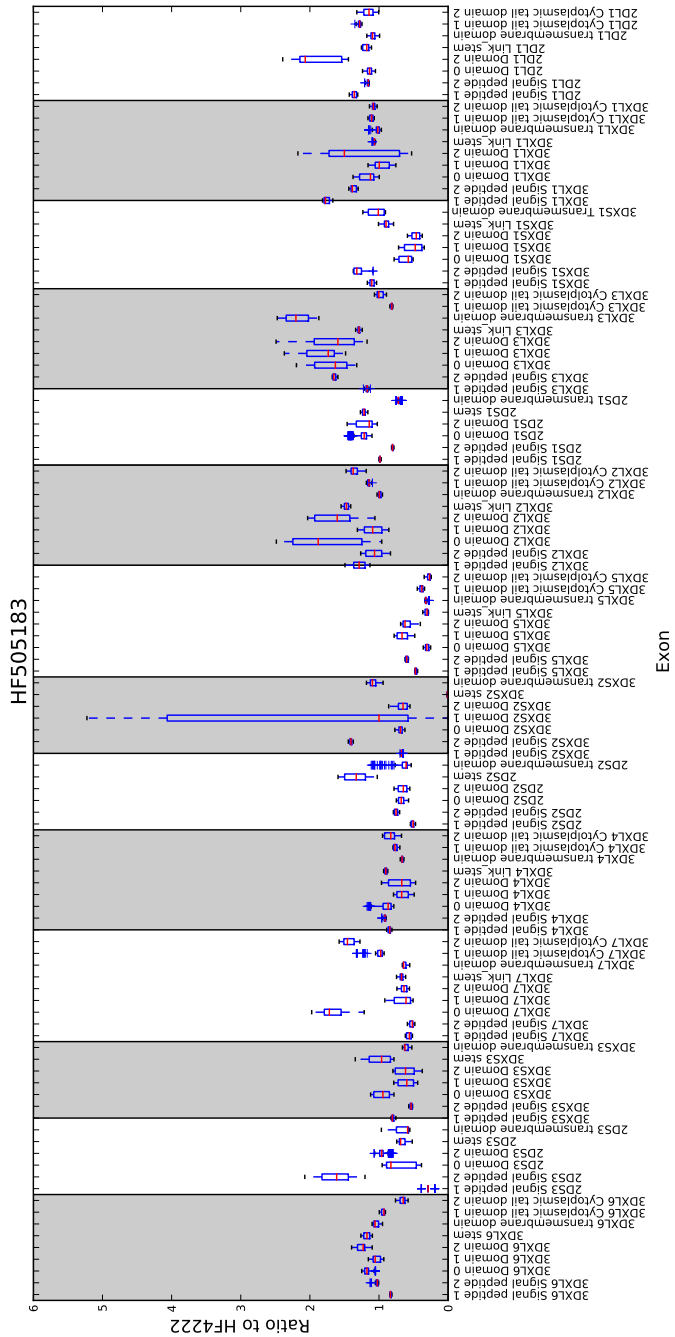


Figure S49: CNV exon prediction of HF505183. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

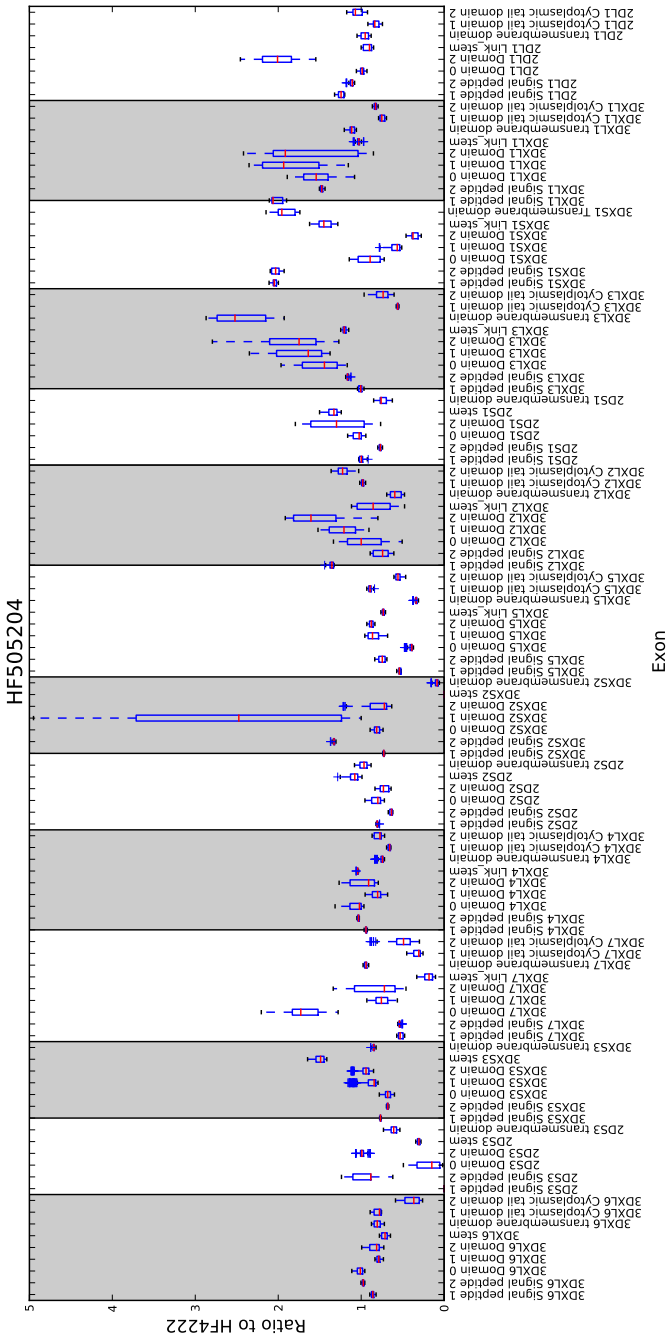


Figure S50: CNV exon prediction of HF505204. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

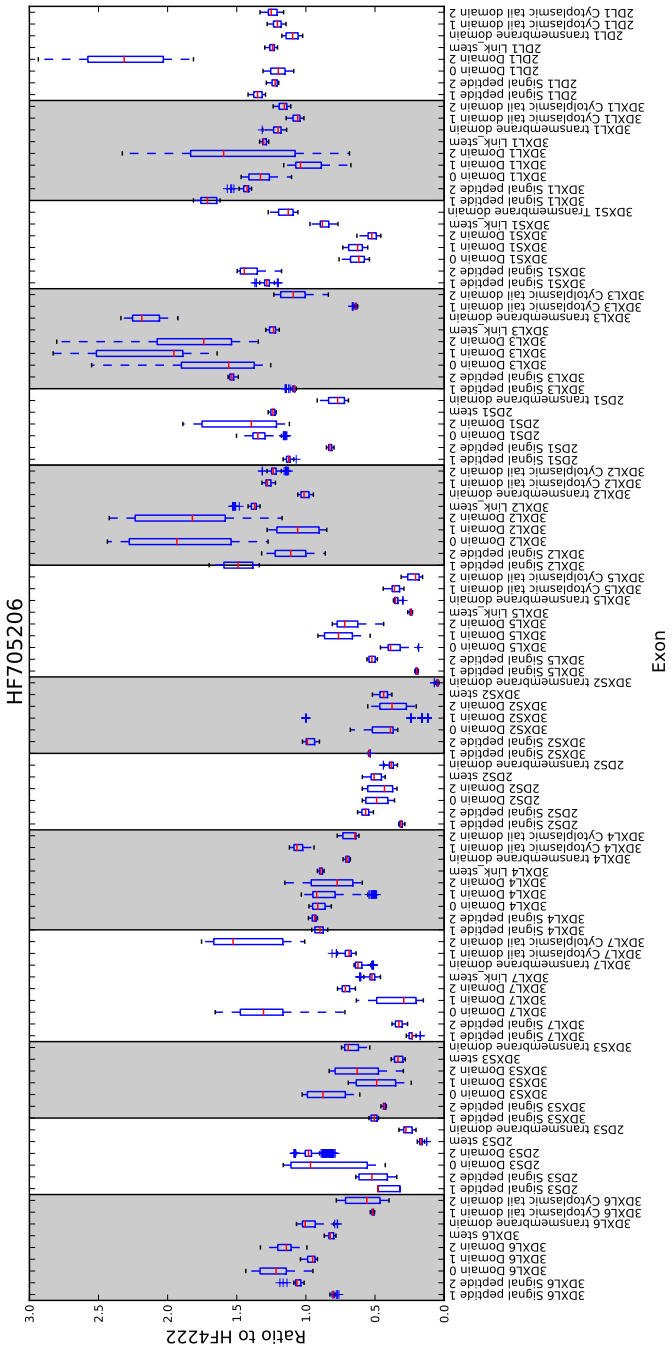


Figure S51: CNV exon prediction of HF705206. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

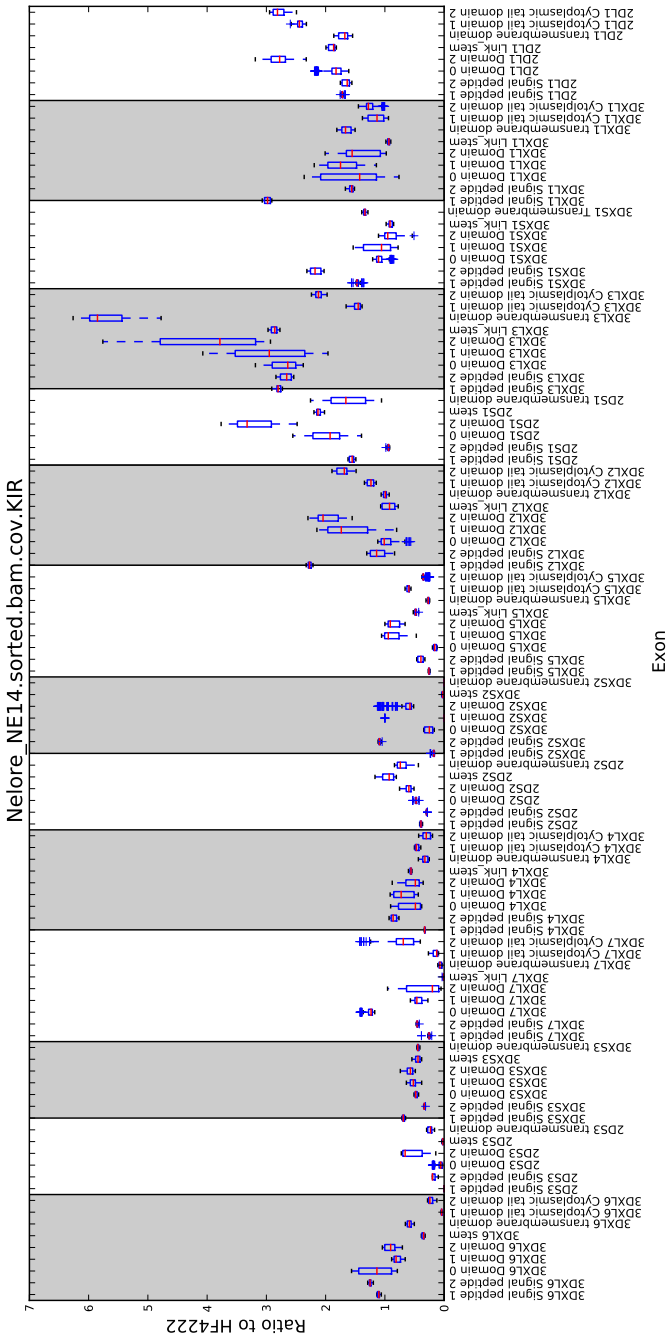


Figure S52: CNV exon prediction of NeloreNE14. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

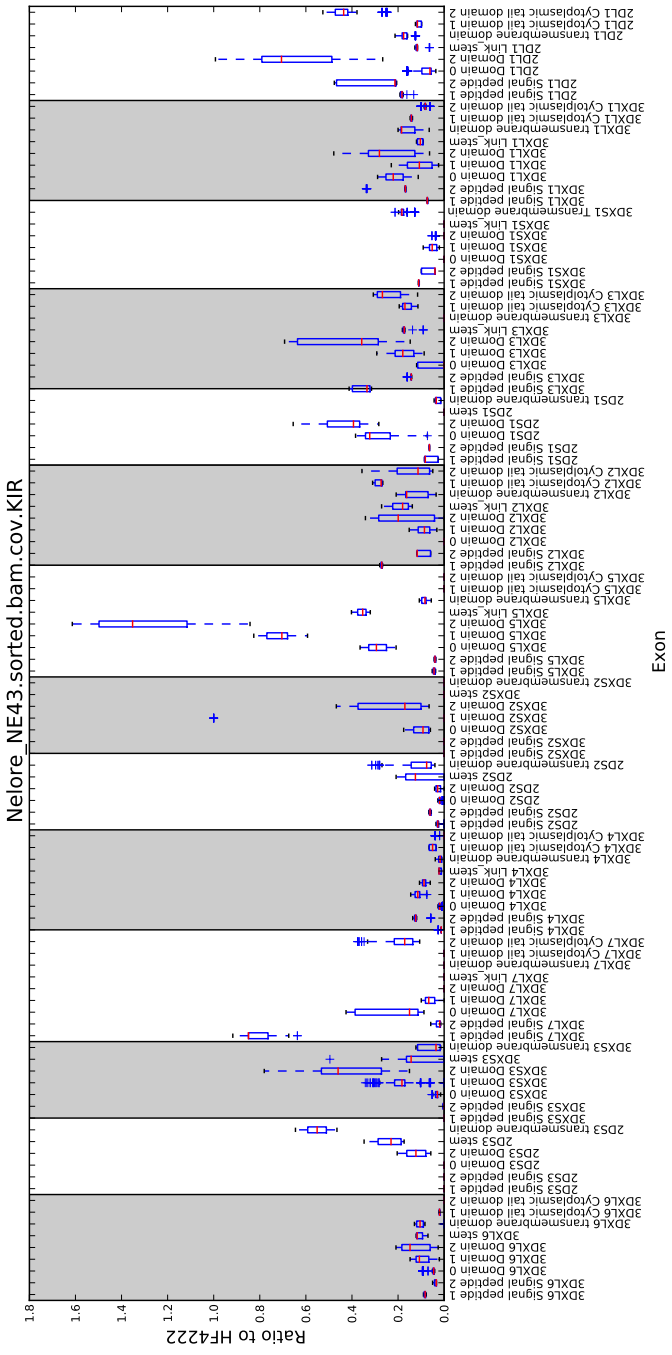


Figure S53: CNV exon prediction of NeloreNE43. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

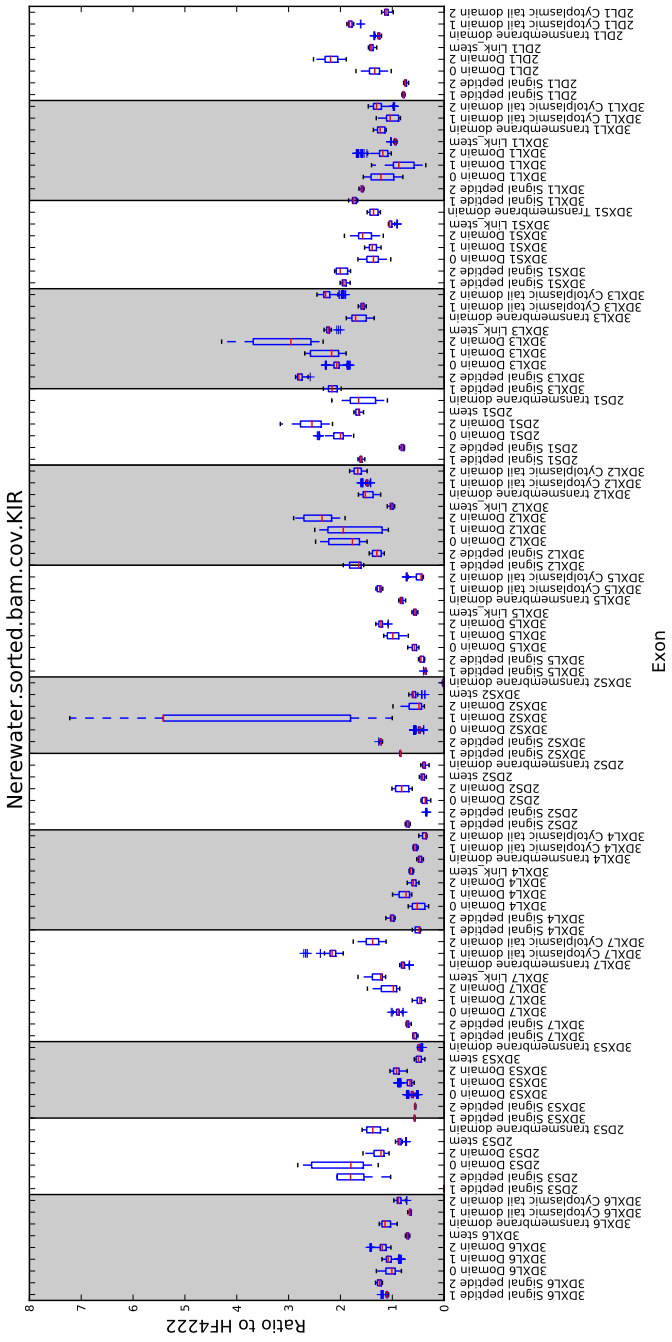


Figure S54: CNV exon prediction of Nerewater. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

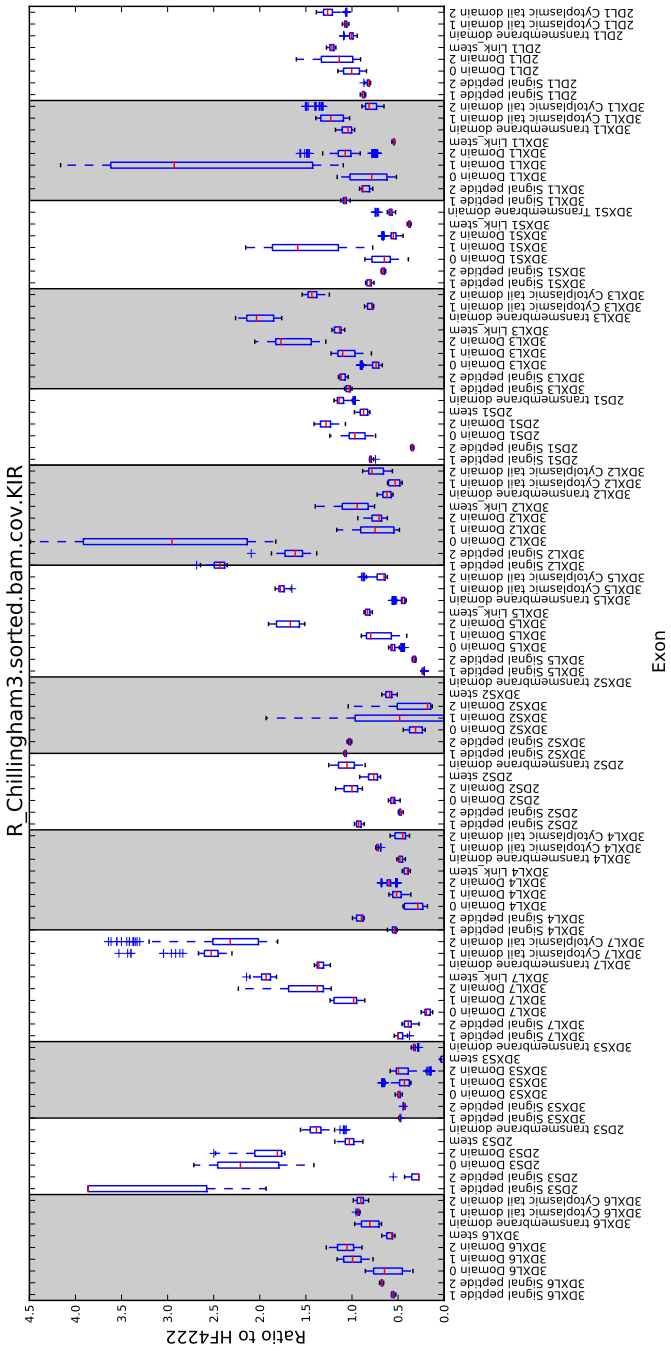


Figure S55: CNV exon prediction of Chillingham3. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

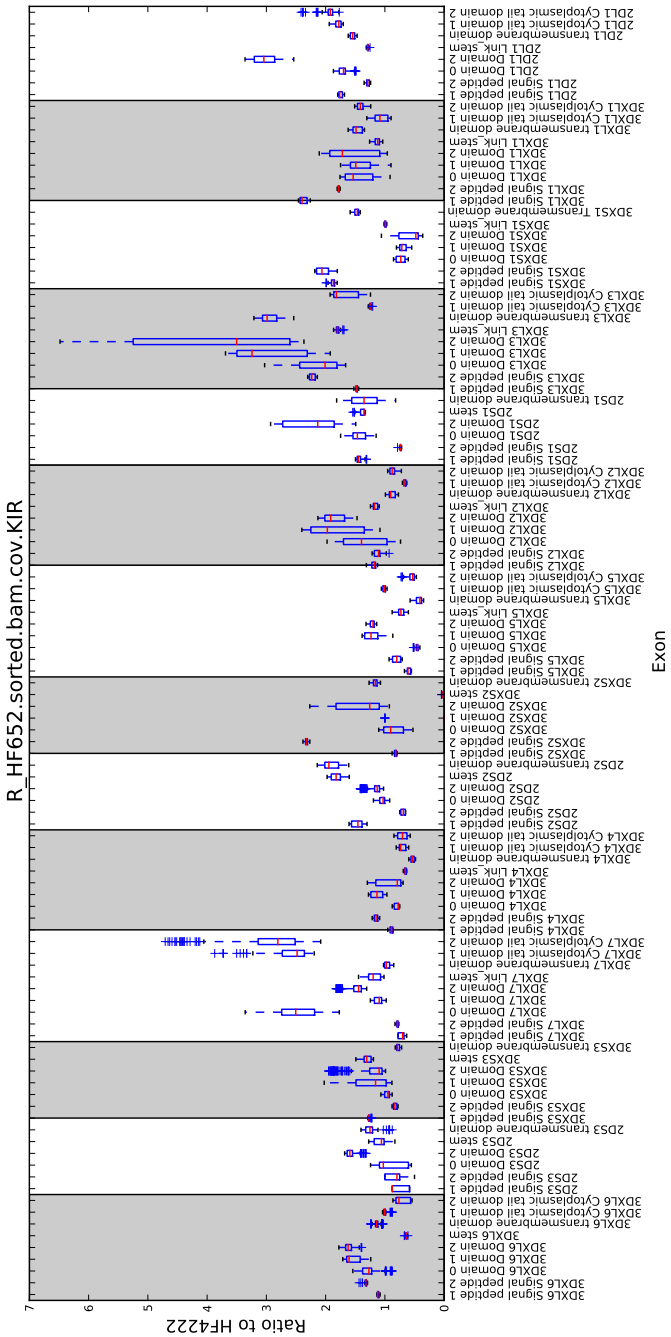


Figure S56: CNV exon prediction of HF652. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

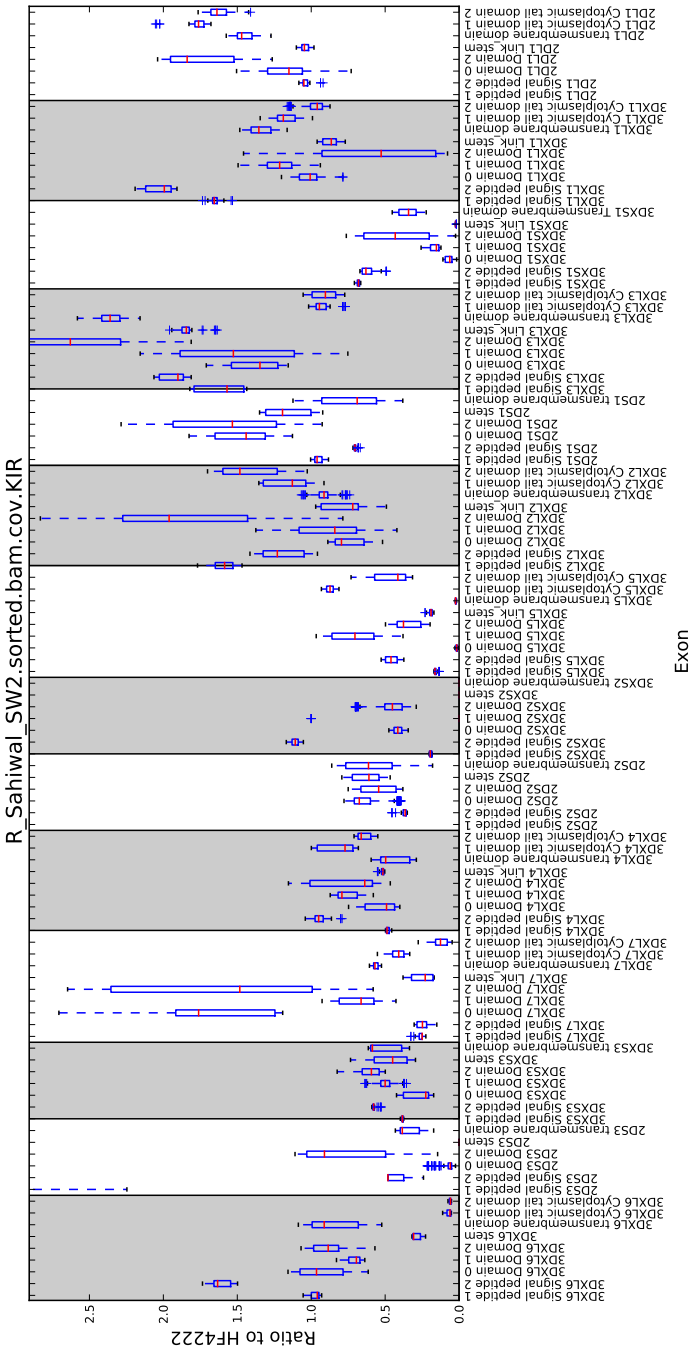


Figure S57: CNV exon prediction of Sahiwal_SW2. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

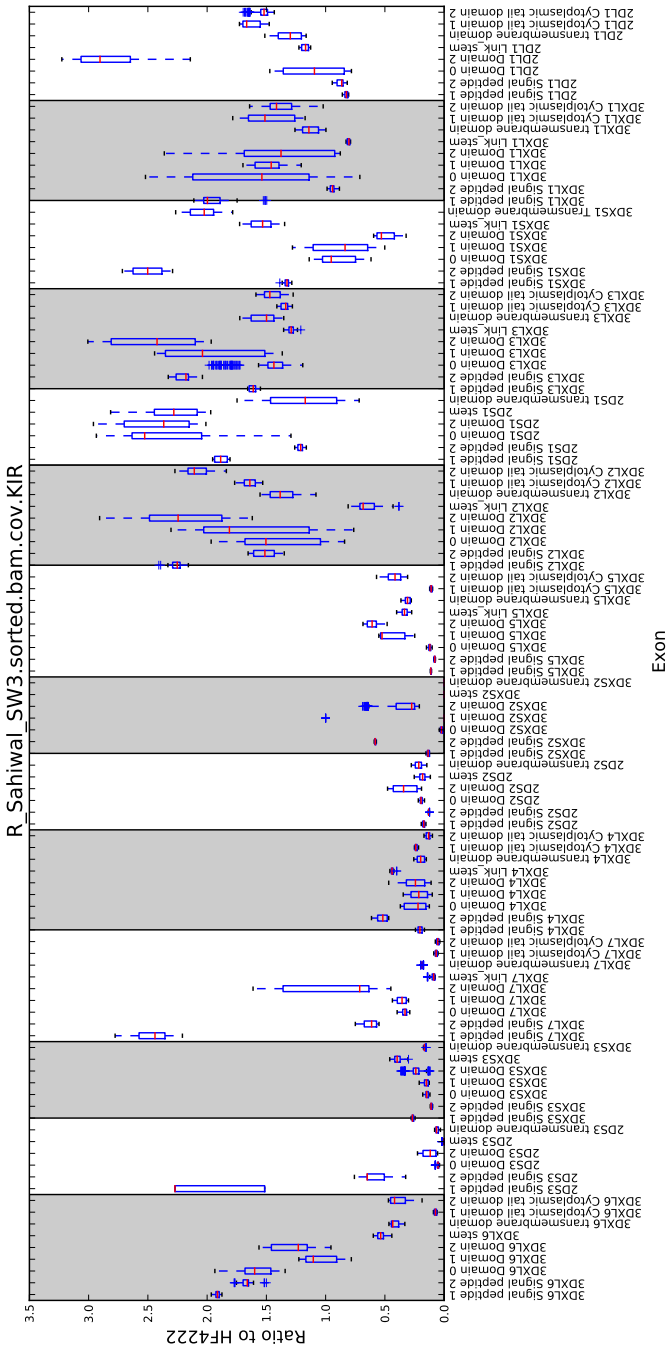


Figure S58: CNV exon prediction of Sahiwal_SW3. Box plots are representative of relative read depth coverage changes compared to animal 4222 after normalisation.

9.5.7 Tables of capture SNPs within KIR exons

These are the tables containing the SNP positions for each *KIR* exon sequence. Each table contains the frequency as a percentage of the number of reads which match the SNP base described in the "var_base" column for each animal. The variable bases, reference bases and resulting residue changes from the reference are described in the left hand columns. For each animal where no value is given the animal has no SNP at that position and contains the same sequence as the reference. SNPs were called using Varscan2 and residue changes were calculated using a bespoke python script, Tables were generated using MySQL.

Haplotype_pos	2DL1 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF503183Freq	HF503183Freq	HF503183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chilliingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	
316363	D 0	G	Q	T	*	97	33																									83	
316364	D 0	C	Q	T	L	98	33																									82	
316380	D 0	G	V	A	V	114	38	44										43													55	89	90
316416	D 0	G	Q	C	H	150	50	45										54								11					59	94	93
316420	D 0	C	P	T	S	154	52	45										46													58		51
316525	D 0	A	K	G	E	259	87	100	100	100	100	100	100	100	100	100	100	53	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
316533	D 0	C	H	T	H	267	89	51										45												56	42	99	97
316536	D 0	G	A	A	A	270	90	50										45												56	42	99	97
316579	D 0	C	L	G	V	313	105	100	100	99	48	52	99	58	53	99	53	54	98		99	98	63		98	99	99	41	57		98	97	98
316607	D 0	A	V	G	G	341	114																						41				
318492	D 2	C	L	G	V	376	126													15				15									
318504	D 2	G	G	T	*	388	130													15													
318505	D 2	G	G	A	E	389	130													16													14
318506	D 2	A	G	G	G	390	130													32							25				17		36
318509	D 2	A	G	C	G	393	131													17													31
318509	D 2	A	P	G	P	393	131																				22						
318510	D 2	C	P	G	A	394	132													15													
318513	D 2	G	V	C	L	397	133													14													
318514	D 2	T	V	A	E	398	133													15													15
318526	D 2	G	G	A	E	410	137													17													17
318527	D 2	G	G	A	G	411	137					27							44					41			32		13		42		28
318528	D 2	G	E	A	K	412	138													22													
318538	D 2	C	T	A	N	422	141													19													15
318554	D 2	C	S	T	S	438	146					12								20													15
318555	D 2	G	E	A	K	439	147	26				29							42								33		19			48	25
318561	D 2	G	A	A	T	445	149					13							12													46	51
318563	D 2	C	A	T	A	447	149																	22									15
318575	D 2	C	F	T	F	459	153	34	98	99	25	26	93	23	97	20			92		94	97	28		71	95	95	23	37		45		24	
318579	D 2	C	L	T	L	463	155	21			32									45								37	15			47		
318585	D 2	A	S	G	G	469	157																						35					
318587	D 2	G	R	A	R	471	157																	25										
318594	D 2	G	V	C	L	478	160																	50									22	
318595	D 2	T	V	A	E	479	160						50						63				47	56			45	21			48	48	
318596	D 2	G	V	C	V	480	160												48								39				45		
318597	D 2	A	N	G	D	481	161												14													25	
318603	D 2	G	G	C	R	487	163					17							25								21						
318604	D 2	G	G	A	E	488	163					14							13													26	
318606	D 2	C	R	T	C	490	164	25				34							50								42	17			45		
318607	D 2	G	R	A	H	491	164					14							13													24	
318610	D 2	C	P	T	L	494	165												13													26	
318611	D 2	G	P	A	P	495	165	28				40							50				38				42	18			46		

Haplotype_pos	2DL1 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq
318611	D 2	G	L	T	L	495	165																									25
318614	D 2	G	L	C	L	498	166	29			35								51			39	52				43	19			48	24
318614	D 2	G	A	T	A	498	166																									26
318621	D 2	G	G	C	R	505	169																28									17
318625	D 2	G	R	A	Q	509	170					41	45		38				49			37	24				36				28
318625	D 2	G	R	C	P	509	170																									
318627	D 2	G	G	A	S	511	171																52									23
318634	D 2	G	R	A	H	518	173												48				53				43	20			48	25
318635	D 2	C	R	T	R	519	173												48								42	20			47	
318636	D 2	G	G	A	R	520	174																53									24
318637	D 2	G	G	A	E	521	174																39									
318640	D 2	C	A	T	V	524	175	25				32							48								43	19			47	
318648	D 2	G	A	T	S	532	178																31									18
318649	D 2	C	A	A	E	533	178												19				23									
318662	D 2	G	L	T	L	546	182					15										13					12					35
318668	D 2	T	P	A	P	552	184					15							19			14					12					
318679	D 2	A	D	C	A	563	188	36				44				53			67			49	28				56	27			31	53	36
318685	D 2	G	S	A	N	569	190																39									
318687	D 2	G	G	A	S	571	191												48								41	20			51	
318688	D 2	G	G	C	A	572	191												48								41	20			51	
318689	D 2	T	G	G	G	573	191	78	100	99	67	68	95	67	99	66	58	99	51	98	96	67	79	72	96	99	58	80	91	76	48	72
318690	D 2	G	V	A	I	574	192												18								14					26
318700	D 2	G	C	T	F	584	195	19			24								46								38				49	
318704	D 2	T	Y	C	Y	588	196	31			27								66				21				50	28		28	49	33
318705	D 2	G	G	A	S	589	197												21				55				11					50
318706	D 2	G	G	A	D	590	197																15									
318708	D 2	T	S	A	T	592	198																15									
318709	D 2	C	S	T	F	593	198												24								19					
318710	D 2	T	S	C	S	594	198	30			28								66				22				50	28			48	34
318715	D 2	C	T	G	S	599	200												44				34									19
318717	D 2	C	R	T	C	601	201	29			22																41				46	
318718	D 2	G	R	A	H	602	201																34									20
318730	D 2	C	S	T	L	614	205				12								22								10					22
318731	D 2	G	S	A	S	615	205	46			37		24		29		50		33			23		15			27	70	60	35		
318738	D 2	G	D	A	N	622	208																16									
318740	D 2	C	D	T	D	624	208	16			19								39								37					
318746	D 2	C	S	T	S	630	210												24								17					
318746	D 2	C	S	A	R	630	210																35									23
318746	D 2	C	S	A	R	630	210																14									23
318757	D 2	T	F	A	Y	641	214	40			39	37	28		36	50			37			24					30	70	32	34	62	31
318761	D 2	G	L	A	L	645	215																13									

Haplotype_pos	2DL1 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq
318764	D 2	T	S	C	S	648	216																32									19
318765	D 2	G	V	A	I	649	217																33									19
319448	Link	C	L	T	L	679	227				49	45				46			100			65					49					
319473	Link	G	S	A	N	704	235				48	47	48		45				100			64					54					
323246	TM	G	A	A	T	736	246																							93		
323279	TM	C	L	T	F	769	257				41	48		42		47			95			31					46	41	89			50
323283	TM	C	T	T	I	773	258	51	100	100	100	100	100	100	99	99		100	100	100	100	100	100	100	96	97	99	99	100	42	98	53
323294	TM	A	I	G	V	784	262				40	49		42		46			94			32					47	98	100		98	98
323306	TM	T	C	G	G	796	266	49	100	100	57	46	97	51	98	51		100		99	98	67	97	98	97	94	51			39		
324538	CT 2	C	D	T	D	873	291				46	51		49		53			96			50					52	99	75		53
324557	CT 2	G	V	T	L	892	298	49	99	100	52	50	99	50	98	47		99		98	98	47		97	99	96	39		22	59		42
324559	CT 2	G	V	T	V	894	298	49	46							46																
324606	CT 2	C	T	G	S	941	314				46	49		49		50			96			50					56	99	62			54
324612	CT 2	C	I	T	I	947	316																									51
324614	CT 2	T	S	C	P	949	317	48																								

Table S3: 2DL1 capture SNPs

Haplotype_pos	3DXL1 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq
305450	SP 1	C	E	G	E	31	11																										24
305602	SP 2	A	Y	G	C	41	14	14	53		22				63					36			20		58			21	54		14	13	25
305617	SP 2	A	N	T	I	56	19	12	53		17				60					30			19		55			17	51		11	13	23
306759	D 0	G	K	A	K	73	25																						38				57
306782	D 0	C	S	T	S	96	32	52										54											53		54		
306783	D 0	G	I	T	F	97	33																					30					15
306845	D 0	C	S	T	S	159	53								32																		63
306848	D 0	A	M	G	M	162	54																										24
306852	D 0	A	Q	G	E	166	56																					30			94		29
306874	D 0	G	R	C	T	188	63		41						28									10	46				41				67
306882	D 0	A	K	G	E	196	66								28										40								
306888	D 0	C	N	T	Y	202	68																						50	64			
306889	D 0	A	H	G	R	203	68	47										52					14						67	100	52	98	40
306900	D 0	C	G	T	W	214	72																						19				
306909	D 0	G	D	C	H	223	75		45						63								22	41					47				95
306912	D 0	C	H	A	N	226	76																74										
306944	D 0	C	F	T	F	258	86																										20
306978	D 0	G	T	A	T	292	98																						28				
307008	D 0	C	Q	G	E	322	108	37										42											32	60	37		
307902	D 1	G	I	A	I	340	114																										13
307908	D 1	C	Q	A	K	346	116	64		86	87	100	98	32	97	99			100	100	98	100	100	100	29	99	95	99	96		99		98
307909	D 1	A	T	C	T	347	116																						96		57		11
307941	D 1	C	V	A	I	379	127																						75				
307951	D 1	A	W	G	W	389	130																						76				
307956	D 1	G	D	C	H	394	132							86		86			100			96	28		20	96	88	22	17		92		
307958	D 1	C	D	G	E	396	132		100	89				89	42	89			100			96	27		99	98	97	24	98		98		
307972	D 1	G	R	A	H	410	137	30																					23		38		
307984	D 1	T	S	C	T	422	141																						63				
307989	D 1	A	M	C	L	427	143		85						43										73								
307989	D 1	A	L	G	V	427	143																						62				
308046	D 1	G	F	T	F	484	162																								46		
308070	D 1	G	I	A	I	508	170																						34				
308092	D 1	G	G	C	A	530	177	42	23	99	25	42	97		96	39			97		99	97	55		45	99	94	40			46		
308104	D 1	T	L	C	S	542	181		23																								
308114	D 1	C	A	T	A	552	184		21																								
308122	D 1	A	Y	G	C	560	187		19																								
308132	D 1	C	Y	T	Y	570	190		17																								32
308144	D 1	A	K	C	N	582	194																						42				
308145	D 1	A	K	C	Q	583	195																						43				
308146	D 1	G	K	A	K	584	195																						42				
308147	D 1	A	K	C	N	585	195																						43				

Haplotype_pos	3DXL1 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF7052O6_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	
308150	D 1	C	S	T	S	588	196		12																			10					
308153	D 1	C	I	T	I	591	197																										58
308156	D 1	T	K	C	N	594	198																										29
308180	D 1	C	P	T	P	618	206	45																						32			30
309552	D 2	A	K	G	K	648	216																24										41
309573	D 2	G	H	A	Q	669	223																					56					31
309574	D 2	G	G	A	R	670	224	14																				35		18			
309580	D 2	T	S	C	P	676	226	13	13															14				96		29		41	
309585	D 2	G	M	T	I	681	227																								79		
309586	D 2	G	F	A	I	682	228																					37					
309590	D 2	G	S	T	F	686	229																										31
309603	D 2	C	N	T	N	699	233	10																				41		15			
309660	D 2	G	R	A	R	756	252																	12									
309703	D 2	T	S	C	P	799	267																					29				76	97
309740	D 2	G	G	A	D	836	279																									65	
309782	D 2	T	F	A	Y	878	293		15																			58	88				
309787	D 2	T	C	C	R	883	295	100	100	100	100	100	99	100	99	99	62	100	100	99	99	98	100	99	100	100	100	100	100	100	100	100	
309801	D 2	G	S	T	S	897	299																91		23								
309808	D 2	A	D	G	D	904	302																							24			
309816	D 2	C	S	T	S	912	304	29																30									
310819	Link	G	T	A	T	948	316										31						22					74		30			
310831	Link	A	P	G	P	960	320	27									32											75		29		46	
310838	Link	A	M	T	L	967	323	27																									
310841	Link	G	D	T	Y	970	324				31																						46
311603	TM	C	V	G	V	994	332																										
311604	TM	A	V	T	V	995	332																					50	83				45
311604	TM	A	V	G	G	995	332																										
311607	TM	T	S	C	T	998	333																					47					
311609	TM	T	P	C	P	1000	334																					51	83				47
311616	TM	C	S	A	N	1007	336																					46					
311616	TM	C	S	G	S	1007	336																										46
311617	TM	A	S	C	S	1008	336																										46
311623	TM	T	S	C	S	1014	338	51			51	47		55		50	51		98			46	97				49	98	100	56	97	99	
311638	TM	T	L	C	L	1029	343																					45					
311640	TM	T	S	G	S	1031	344																										49
311650	TM	C	L	T	L	1041	347										51											45		50	93	49	
311651	TM	A	I	G	V	1042	348										50											45		50	92	49	
311655	TM	T	Q	C	P	1046	349																					45					
311678	TM	C	L	A	I	1069	357	52			49	50		55		50	54		97			52	97				46	46		50	95	98	
311719	TM	C	N	T	N	1110	370		46					41										48				50					
312493	CT 1	T	G	C	G	1161	387	46			49	49		55		51	47		98			51					54	49		44	96	60	

Table S4: 3DXL1 capture SNPs

Haplotype_pos	3DXL1 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	
312634	CT 2	C	P	T	P	1191	397																									30	
312650	CT 2	A	L	G	V	1207	403																									30	
312661	CT 2	C	L	T	L	1218	406																									28	
312701	CT 2	C	P	A	T	1258	420	44									45						66					48		53	91	35	
312707	CT 2	A	T	G	A	1264	422	44									47						68					50		48	92	37	

Haplotype_pos	3DXL2 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF1047766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF7052O6_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
206269	SP 1	T	M	C	T	2	1			31	27		88		90					44	21	26					38	37			23			
206271	SP 1	C	Y	T	Y	4	2																					23			55		
206282	SP 1	C	L	T	L	15	5		55		31	26		49		30			48			39					26	11		33	15	
206283	SP 1	C	L	T	F	16	6	51	43	98	29	38	93		93	31		89		90	98	30				90	38		73	23		27	
207629	D 0	G	E	A	K	73	25												16				23									14	
207636	D 0	T	V	A	E	80	27	58			14						33		30			20	95				19	98		63	99	95	73
207652	D 0	C	S	T	S	96	32												20														
207652	D 0	C	S	A	S	96	32																29								14		
207667	D 0	C	P	T	P	111	37	40									34		23			17	31				15	31		38	16	78
207698	D 0	C	R	T	W	142	48																26									14	
207700	D 0	G	R	C	R	144	48																26									14	
207705	D 0	A	H	T	L	149	50																26									14	
207711	D 0	G	R	A	H	155	52		13		18								30			23	27				18					14	
207713	D 0	T	T	A	T	157	53																						33				
207714	D 0	C	T	A	K	158	53																						33				
207722	D 0	A	K	G	E	166	56				19								31			22					19						
207724	D 0	G	K	T	N	168	56		34																43								
207732	D 0	T	I	G	R	176	59				44	40		64		39			63			37	11				35	40			30	29	
207733	D 0	A	I	C	I	177	59																26									14	
207744	D 0	C	T	G	R	188	63				21								32			24					20						
207746	D 0	G	D	A	N	190	64				18								15			11											
207753	D 0	G	R	A	K	197	66							33					32			23					20						
207754	D 0	A	R	C	S	198	66												32			23					20						
207756	D 0	G	R	A	K	200	67												32			24					19						
207757	D 0	G	R	A	R	201	67												33			23					20						
207758	D 0	G	G	C	R	202	68							34					32			25	25				20					14	
207759	D 0	G	G	T	V	203	68																24									14	
207764	D 0	C	P	T	S	208	70							35					33			25					20						
207767	D 0	C	Q	G	E	211	71							35					35			26	24				21					14	
207779	D 0	T	Y	C	H	223	75				27			43					42			32					31						
207787	D 0	C	F	A	L	231	77		17										33					11				16			37		
207793	D 0	C	N	G	K	237	79		16		16			23					14			11	10	13			59				87	32	
207808	D 0	T	P	A	P	252	84																16										
207814	D 0	C	T	A	T	258	86				30			52					52			40					38						
207824	D 0	G	A	A	T	268	90				29			52					54			41					39						
207826	D 0	C	A	T	A	270	90				30			51					54			39	15				39						
207835	D 0	C	R	T	S	279	93															23					20			24		12	
207843	D 0	C	S	T	F	287	96				28			52					58			40					39						
207853	D 0	C	Y	T	Y	297	99							53					57			38					37						
207854	D 0	T	W	C	R	298	100	27	10		36	51		94		53	27		98			52	63				60	89		26	82	59	99
207855	D 0	G	W	A	*	299	100	27									25						35					47		27	47	37	100

Haplotype_pos	3DXL2 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF1047 66_Freq	HF5051 83Freq	HF5051 83Freq	HF4048 18_Freq	HF5052 04_Freq	HF598_Freq	HF7052 06_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
207855	D 0	G	W	T	L	299	100	32																						32			100
207858	D 0	C	S	T	F	302	101				26			52					57			39					38						
207863	D 0	C	H	T	Y	307	103							56					58			38					40						
207864	D 0	A	H	C	P	308	103					26		56					37			11					14						99
207864	D 0	A	H	G	R	308	103	25			32				29								54					86		25	75	18	99
207864	D 0	A	H	T	L	308	103																				47					40	99
207889	D 0	C	V	G	V	333	111							59					68			47											
207890	D 0	A	T	T	S	334	112							60					68			47					47						
208812	D 1	C	I	A	I	360	120	31															32							39		18	97
208825	D 1	G	C	C	R	373	125																						18				
208827	D 1	A	C	G	W	375	125																					18					
208833	D 1	C	L	T	L	381	127	32															30							41		36	92
208834	D 1	G	F	A	I	382	128																					17					
208837	D 1	C	*	G	G	385	129																					18					
208852	D 1	G	V	A	M	400	134	34										21					30					20		42		19	97
208934	D 1	G	R	A	H	482	161	45										31					65							40		36	93
208963	D 1	C	I	G	V	511	171																								24		
209003	D 1	C	A	T	V	551	184	30															41							32		24	
209018	D 1	G	C	A	Y	566	189					19	28			25			44			17					20	29				20	
209042	D 1	C	P	G	R	590	197																					22					
209080	D 1	A	I	G	V	628	210	21	16	12	43	38		95	29	37	51		95	21	10	35	48	18		14	55	97		30	86	66	95
210117	D 2	G	R	T	L	641	214							40														43			84		
210154	D 2	C	P	T	P	678	226				27	28		32		32			55			37					33						
210155	D 2	G	V	A	M	679	227	52									48						41					28		53		22	90
210167	D 2	G	G	A	R	691	231				23	26				26			34			22	16					29	21			18	26	
210203	D 2	G	A	A	T	727	243																					25					
210247	D 2	G	Y	A	*	771	257																					21					
210256	D 2	C	L	T	L	780	260																					14			27		
210257	D 2	G	A	T	S	781	261	26										42												24			98
210274	D 2	C	H	G	Q	798	266															11											
210274	D 2	C	H	T	H	798	266			11	13																10					14	
210283	D 2	C	A	A	A	807	269							21																	49		
210284	D 2	C	P	G	A	808	270																							23			
210319	D 2	G	P	A	P	843	281				17	22		21					29			17					25						
210328	D 2	C	S	T	S	852	284		32																								
210359	D 2	G	G	C	R	883	295		43		33	55		98		55			98			49	24	49		55	56			87	22		
210359	D 2	G	G	C	R	883	295		43		33	55		98		55			98			49	24	10		55	56			87	22		
210360	D 2	G	G	A	D	884	295					34		43		32			64			23	22				34	40		43	21		
210372	D 2	C	S	T	L	896	299	47			29	33		43		29	39		61			24	22				34	81		43	45	51	99
210373	D 2	G	S	A	S	897	299																	10									
211409	Link	C	T	G	S	965	322				11								27								17	14					

Haplotype_pos	3DXL2 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
211414	Link	G	V	C	L	970	324																						17			40		
211426	Link	T	S	A	T	982	328	19					22			84				91				92				34	73		33	91	100
212196	TM	T	C	A	S	1015	339					50	41			91	42			96			50	48				45	55			92	45	
212203	TM	G	W	T	L	1022	341					52	43			68	42			96			52	48				45	42			44		
212219	TM	G	G	A	G	1038	346									25													16			90		
212249	TM	C	I	T	I	1068	356	65										46						47					38		67		51	100
212278	TM	T	F	G	C	1097	366									30																		
213199	CT 2	T	D	C	D	1182	394																	10										
213208	CT 2	C	Y	T	Y	1191	397																	11										
213269	CT 2	G	V	A	M	1252	418																	18								95		
213308	CT 2	G	D	A	N	1291	431																	16										
213323	CT 2	C	H	T	Y	1306	436																	16										
213324	CT 2	A	H	G	R	1307	436																	16										

Table S5: 3DXL2 capture SNPs

Haplotype_pos	3DXL3 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
231333	SP 1	C	L	T	F	13	5	31																						19			
231340	SP 1	G	R	A	K	20	7																									34	
231498	SP 2	A	R	G	R	36	12	70	47		64	59		96		62	48		99			49	97	56			55	69		73	20	94	99
232760	D 0	C	N	G	K	84	28															58					51			43			
232762	D 0	T	L	C	P	86	29															58								43			
232763	D 0	C	L	T	L	87	29	25	47								31						49	30						21		21	100
232829	D 0	T	F	C	F	153	51	66	46		52	60		97		61	36		98			67	99	61			61	93		55	98	94	99
232830	D 0	C	R	T	C	154	52																44									36	
232831	D 0	G	V	A	D	155	52																	12			18						
232892	D 0	G	Q	T	H	216	72												15														
232916	D 0	C	T	G	T	240	80												21			11					12				12		
232943	D 0	A	A	C	A	267	89				23								44			29	41				24	19			31	33	
232961	D 0	C	S	T	S	285	95				33			36					54			36		14			29	17		11	33		
232962	D 0	G	Q	A	K	286	96																									31	
232967	D 0	G	L	C	F	291	97	11			35								55			37		15			31	19		11	34		
232980	D 0	C	P	A	T	304	102	12			15			27		28			22			17		13			13	12		12	11		
232980	D 0	C	P	G	A	304	102															23											
233945	D 1	A	I	C	L	382	128					28		26		26			52			29					27						
233946	D 1	T	I	C	T	383	128					29		26		26			52			30					27						
233947	D 1	C	I	T	I	384	128					22		23		21			47			19					22						
233948	D 1	A	S	G	G	385	129					22		24		21			46			20					22						
233960	D 1	A	S	C	R	397	133	64				23		58		22	41		48			21	98	51			23	30		73		66	97
233961	D 1	G	S	A	N	398	133					23		25		22			47			19					22						
233964	D 1	C	A	T	V	401	134					23		24		22			47			19					21						
233965	D 1	A	A	G	A	402	134					23		24		21			47			19					21						
233969	D 1	G	A	A	T	406	136																					28					
234016	D 1	T	I	C	I	453	151	64			34	34		64		30	48		54			37	99	53			29	31	63	69	22	71	94
234019	D 1	T	L	G	L	456	152	25																						31			
234025	D 1	A	Q	G	Q	462	154				32	29		33		34			43			33					43	32			28		
234029	D 1	A	N	C	H	466	156																44										
234040	D 1	A	G	G	G	477	159	26																						28			
234072	D 1	A	H	G	R	509	170	62	14		35	36		66		30	48		53			35	99	51			37	36	71	73	16	74	100
234077	D 1	C	N	T	Y	514	172																									19	
234083	D 1	C	C	T	C	520	174																									21	
234091	D 1	C	T	T	T	528	176																								73		
234094	D 1	C	F	T	F	531	177				27	27		27		23			41			27					30					24	
234095	D 1	G	E	A	K	532	178																										
234107	D 1	A	M	G	V	544	182	51	21					22			44						89	35				23	50	54		48	96
234116	D 1	G	A	T	S	553	185	29	68	37	27		96		95	25		88		40	34	28			86	45	27		19				
234117	D 1	C	A	T	V	554	185							22									40					33			24		
234124	D 1	A	A	G	A	561	187	16			38	38		33		37			52			40		25			34	12		19	33		

Haplotype_pos	3DXL3 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF1047766_Freq	HF5051838Freq	HF5051838Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF7052O6_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
234129	D 1	C	T	G	S	566	189				24	23		23		24			35			26					27					21	
234142	D 1	C	Y	T	Y	579	193	12			18								21			18	15				13	19		12	21	29	
234156	D 1	A	H	G	R	593	198												43			28					30			19			
234157	D 1	C	H	T	H	594	198							44					43			28	16	17			29				18		
234167	D 1	A	K	G	E	604	202												43			27	32	16			30				21		
234168	D 1	A	K	T	M	605	202												43			27	16				29			15	21		
234174	D 1	C	S	G	*	611	204	14			23								25			23					17			23			
234184	D 1	C	S	T	S	621	207	16			25								27			26					19	12	16	23			
234199	D 1	C	I	T	I	636	212	10			27								42			24	18				27			24			
234200	D 1	G	V	A	I	637	213							34																	31		
234203	D 1	A	Q	G	E	640	214																16								16		
234207	D 1	C	T	A	K	644	215	11			26								41			24	18				27			25			
239932	D 2	C	R	T	W	649	217												69			43					53	27			11		
239933	D 2	T	R	G	R	650	217																								11		
239936	D 2	A	Y	C	S	653	218	48	50	55									75			53	61	63		13	59	34		56	44	46	
239940	D 2	G	K	A	K	657	219			19									34			28					26						
239943	D 2	A	*	G	W	660	220																								11		
239944	D 2	C	P	T	S	661	221										43						37					18		19		21	98
239945	D 2	C	P	G	R	662	221										43						37					18		19		21	99
239949	D 2	A	S	T	S	666	222	63	50	52				75			56		74			53	99	65		12	58	51		75	44	69	100
239950	D 2	C	L	G	V	667	223																14										
239965	D 2	G	G	T	C	682	228		47														33	38							18		
239967	D 2	C	G	G	G	684	228															31					36	27					
239970	D 2	C	P	T	P	687	229																13								13		
239974	D 2	G	V	C	L	691	231	24									44						39					21		19	25	99
239984	D 2	A	E	G	G	701	234	19			47								69			43					53	27		26	13		
239991	D 2	T	N	C	N	708	236																21								22		
240013	D 2	A	K	G	E	730	244	54	46								50						90	32				28		48	53	99
240015	D 2	A	K	G	K	732	244																18								16		
240018	D 2	T	S	C	S	735	245			32									58								38						
240033	D 2	C	F	T	F	750	250		45														39	38							20		
240045	D 2	A	S	G	S	762	254																				42						
240064	D 2	T	C	C	R	781	261																11										
240069	D 2	A	P	G	P	786	262																11										
240072	D 2	T	L	C	L	789	263												37				11	22			30			13	32		
240082	D 2	T	W	C	R	799	267												55				11	30			41			22	39		
240085	D 2	A	S	G	G	802	268												37					22			27			15	27		
240093	D 2	T	H	C	H	810	270																10										
240094	D 2	G	G	A	R	811	271																10										
240098	D 2	C	A	T	V	815	272												36					22			26	11		14	26		
240112	D 2	T	F	C	L	829	277		51														39	34					85			25	

Haplotype_pos	3DXL3 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183FFreq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
240115	D 2	C	R	T	C	832	278	30									48						49					25		31		34	99
240116	D 2	G	R	C	P	833	278				31								59					33			44	22		24	37		
240120	D 2	T	L	G	L	837	279				31								59					33			44	23		24	37		
240128	D 2	G	H	T	L	845	282																						73				
240144	D 2	C	S	T	S	861	287	12			18								28					13			22	14		28	18		
240145	D 2	G	G	A	S	862	288												33				28	21			25	13		14	26		26
240145	D 2	G	S	C	P	862	288																										
240146	D 2	G	G	C	A	863	288												34				27	21			25	13		31	26		
240158	D 2	G	C	T	F	875	292	11			24								35				31	22			26	14		16	29		
240167	D 2	C	S	T	F	884	295				13								17								13						
240171	D 2	C	F	A	L	888	296				18								29				15	14			25	14		17	16		
240175	D 2	C	H	T	Y	892	298	29									48		35				30	24			25	15		39	30	34	100
240176	D 2	A	H	G	R	893	298												34				28	24			25	15		17	27		
240188	D 2	C	S	T	L	905	302							24																			
240195	D 2	A	S	G	S	912	304	12			21								30				18	15			26	15		19	18		
240198	D 2	C	D	T	D	915	305	15			29								38				34	26			28	16		19	29		
240204	D 2	C	S	T	S	921	307				16								19				22				15						
240211	D 2	C	L	G	V	928	310	13			23								29				18	14			26	17		19	20		
241234	Link	A	Q	C	P	968	323																56	16							36		
241239	Link	A	K	G	E	973	325																							38			
241243	Link	T	M	C	T	977	326		40														41	34							25		
242014	TM	T	L	C	P	1004	335		12														43	21							11		
242023	TM	A	Q	T	L	1013	338		13														46	21							12		
242025	TM	A	K	G	E	1015	339																47	21							12		
242049	TM	T	F	G	V	1039	347		21														68	27							19		
242116	TM	T	S	C	S	1106	369																					19				84	
242882	CT 1	G	A	A	T	1126	376																40										
242886	CT 1	T	I	A	N	1130	377			17	21												14				22				22		
242905	CT 1	T	V	G	V	1149	383																				18						
242910	CT 1	A	A	G	G	1154	385																								18		
242912	CT 1	C	L	A	I	1156	386																				19						
242913	CT 1	G	L	A	Q	1157	386																								18		
242920	CT 1	A	M	G	M	1164	388																				37				20		
242924	CT 1	A	G	G	G	1168	390																				39				18		
243050	CT 2	G	A	A	T	1183	395	31									32						48							24	26	99	
243067	CT 2	C	G	T	G	1200	400																41	34				32					
243074	CT 2	C	L	T	F	1207	403																43	21			13						
243083	CT 2	G	G	A	R	1216	406							49									46	21									
243089	CT 2	G	E	C	Q	1222	408																47	21			11						
243113	CT 2	G	A	A	T	1246	416							49									68	27									
243155	CT 2	G	D	T	Y	1288	430		61												27		48	42		26		29	80			34	

Table S6: 3DXL3 capture SNPs

Haplotype_pos	3DXL3 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos		res_pos	HF504805_Freq	HF504882_Freq	HF10476 6_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillinghamβ_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq

Haplotype_pos	3DXL4 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF1047766_Freq	HF5051838Freq	HF5051838Freq	HF4048I8_Freq	HF5052O4_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chilllinghamb_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
133655	SP 1	T	L	C	L	15	5	11																	20			29		20		50	
133656	SP 1	C	A	T	S	16	6																		19								
133829	SP 2	G	S	C	T	47	16			11	13		36	25	36					15	14	17				15	17						66
135003	D 0	G	E	A	K	73	25				12								17			14	14				11	31					
135009	D 0	A	K	G	E	79	27				12								17			14	14		21		11	32		21			
135010	D 0	A	M	T	M	80	27																		22					15			
135026	D 0	C	S	T	S	96	32				11								16			13						28					
135027	D 0	G	P	A	T	97	33																								39		
135041	D 0	C	P	T	P	111	37				10								14			12						27					
135072	D 0	C	R	T	W	142	48																18										
135073	D 0	G	R	A	Q	143	48										33																
135074	D 0	G	R	C	R	144	48																19										
135079	D 0	A	H	T	L	149	50																18										
135085	D 0	G	R	A	H	155	52		24										10				21	13							10		
135096	D 0	A	K	G	E	166	56																21					21					
135106	D 0	T	I	G	R	176	59			15	34			52	47				25	19	17	32	71			29	37	25			32	42	100
135106	D 0	T	H	C	P	176	59																					20					100
135107	D 0	A	I	C	I	177	59																21										
135118	D 0	C	P	G	R	188	63													18								18					
135120	D 0	G	D	A	N	190	64																				11						
135132	D 0	G	G	C	R	202	68																22										
135133	D 0	G	G	T	V	203	68																21										
135141	D 0	C	Q	G	E	211	71																21										
135155	D 0	C	Y	T	Y	225	75	13	13		13								32				19	91	11		15	69		17	36	92	
135161	D 0	C	F	A	L	231	77		13																			10			16		
135167	D 0	C	N	G	K	237	79		14										19			11	63	10		13	54			35	84	
135177	D 0	G	G	C	R	247	83			15	21		47	31	44					20	13	18				24	23					99
135182	D 0	T	P	A	P	252	84																22										
135196	D 0	A	H	G	R	266	89																21										
135200	D 0	T	A	C	A	270	90	25	19		20			24					37			20	69	15		19	75		26	41	59	
135200	D 0	T	H	A	Q	270	90																								29		
135209	D 0	C	M	T	I	279	93																					13				26	
135212	D 0	G	S	A	S	282	94																					18					
135228	D 0	T	W	C	R	298	100	26	22		25			25		24			42			25	75	16		21	78		26	47	88	
135229	D 0	G	W	T	L	299	100																11					32		26	21	32	
135229	D 0	G	W	A	*	299	100	26																									
135232	D 0	C	G	T	V	302	101																					17					
135237	D 0	C	S	T	S	307	103																					17					
135238	D 0	G	R	A	H	308	103																10										
135238	D 0	G	S	C	S	308	103																					17					
135264	D 0	A	T	G	A	334	112				14								24			15	68			16						26	

Haplotype_pos	3DXL4 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF7052O6_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
136184	D 1	C	I	A	I	360	120	15																50							13		19	
136205	D 1	C	L	T	L	381	127																	54							12		43	
136224	D 1	G	V	A	M	400	134																	42									14	
136255	D 1	G	N	T	I	431	144																						10					
136335	D 1	G	A	C	P	511	171													27	20		25	71	24		17	25	48		26	25	81	
136346	D 1	T	V	C	V	522	174																						30					
136358	D 1	C	P	T	P	534	178					10								11			11	75				12	47		11	42	64	96
136390	D 1	G	L	A	*	566	189																										22	
136400	D 1	A	I	T	I	576	192																										18	
136407	D 1	A	D	C	H	583	195																										17	
136414	D 1	C	P	G	R	590	197																						11					
136451	D 1	T	W	C	C	627	209																										15	
137505	D 2	T	L	G	R	641	214					28	23		26		32			42			19					30	71		37	49	
137505	D 2	T	L	A	Q	641	214																	93									41	
137555	D 2	G	G	A	R	691	231					24			24		25			38			17	90				27	43			38	87	
137569	D 2	T	T	C	T	705	235					72	38	44	75	41	51			97	18	11	70	96			17	82	95	71		95	93	99
137622	D 2	A	D	G	G	758	253					67	30	44	76	41	38			96	18	13	64	89			18	79	94	67		95	91	100
137644	D 2	C	C	T	C	780	260																						15				17	
137655	D 2	G	P	A	Q	791	264																										11	
137662	D 2	T	H	C	H	798	266					11								24			14	72				11	56		31	78	
137671	D 2	C	L	A	L	807	269																									20	15	
137692	D 2	G	S	T	S	828	276																										22	
137703	D 2	T	I	C	T	839	280																	12									30	
137707	D 2	G	P	A	P	843	281					13								19				15				18					32	
137716	D 2	C	S	T	S	852	284		11															13										
137734	D 2	T	Y	C	Y	870	290																	15					10				33	
137735	D 2	G	G	A	S	871	291					18			24					32			21	17				27	10				32	
137740	D 2	T	S	C	S	876	292					37	30		26		34			54			39	69				38	41			20	48	
137748	D 2	G	R	A	H	884	295					28	22		26		28			38			24	81				34	54		11	20	80	
137760	D 2	C	S	T	L	896	299																										18	
137777	D 2	G	S	A	S	913	305					12								16			11					21	21					
138800	Link	C	T	G	S	965	322																						19			37		
138805	Link	G	G	C	R	970	324																						13			27		
138817	Link	A	I	T	F	982	328																	55					22				22	
139568	TM	A	Q	C	P	995	332																	62									26	
139582	TM	C	H	G	D	1009	337											47								45			22		34		52	
139593	TM	A	T	G	T	1020	340	29	48								21												25					
139603	TM	G	A	T	S	1030	344																											
139611	TM	G	P	A	P	1038	346																	81									39	
139624	TM	T	F	G	V	1051	351																	90					37			19	88	
139642	TM	C	L	A	I	1069	357																											

Haplotype_pos	3DXL4 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
139670	TM	G	P	T	L	1097	366																									
140427	CT 1	G	E	C	Q	1132	378					21		21					35			23				21	29			53		
140583	CT 2	G	Q	T	*	1177	393																				22				75	
140597	CT 2	C	Y	T	Y	1191	397																36									
140603	CT 2	C	H	A	Q	1197	399																					18			67	
140638	CT 2	A	N	C	T	1232	411																58								32	
140703	CT 2	T	*	C	R	1297	433					33		30		39			68			35					34	47		59	21	

Table S7: 3DXL4 capture SNPs

Haplotype_pos	3DXL5 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
171205	SP 1	C	L	T	L	33	11			34			43							22	27			30		27					72		
171360	SP 2	A	V	C	V	45	15				10								39			23		12			30	32		16	26		
172612	D 0	G	V	A	M	133	45	26	31	16	52		48	68	46	41		42		16	17	44		28	43	30	72						99
172632	D 0	C	F	T	F	153	51			17	43		52		51		31			27	17	46				21	27	89	100	26		86	
172728	D 0	T	P	G	P	249	83	45	48	47	96		98	98	100		58	39		67	53	100		53	37	65	98	94		58		89	99
173759	D 1	T	L	C	L	393	131					30		43		38			60			35		25			53	47		31	98	11	89
173760	D 1	G	V	A	M	394	132					28		38		37			60			34		20			51	43		29	93		
173782	D 1	C	T	T	M	416	139												18								13	23		11	34		
173783	D 1	G	T	T	T	417	139							26		34			49			32		14			40			22	54		
173783	D 1	G	T	C	T	417	139												18					10			13	25		11	34		
173784	D 1	C	L	A	I	418	140												18								13	23		11	34		
173786	D 1	C	L	A	L	420	140							25		32			46			31		14			39	24		22	52		
173787	D 1	C	R	T	C	421	141																71									18	
173795	D 1	C	H	T	H	429	143	27			35	24		26		33			47			33		16			39	25		23	50		
173801	D 1	G	P	A	P	435	145	45			45	31		38		43			67			43		26			53	50		38	79		
173804	D 1	G	L	A	L	438	146	26			35	24		26		33			47			33		15			40	24		22	48		
173807	D 1	G	L	T	F	441	147																14										
173813	D 1	A	K	G	K	447	149	18			11								22			13	20	12			15	23		15	29	16	
173817	D 1	A	I	T	F	451	151	17			11								21			13		12			14	22		14	28		
173843	D 1	G	G	A	G	477	159	23			36	24		26		33			45			32		14			39	23		18	39		
173852	D 1	G	Q	A	Q	486	162																30									14	
173860	D 1	G	G	A	E	494	165	19			15								24			17		11			13	23		18	25		
173868	D 1	C	L	T	F	502	168	23			36	24		26		32			46			33		16			39	25		19	39		
173874	D 1	C	R	G	G	508	170																17									16	
173880	D 1	C	H	T	Y	514	172																17									16	
173886	D 1	C	P	T	S	520	174																17									16	
173898	D 1	G	V	A	I	532	178																17									13	
173898	D 1	G	V	T	F	532	178																29										
173900	D 1	C	V	T	V	534	178																29									13	
173910	D 1	A	M	G	V	544	182			18																		14					
173919	D 1	G	A	T	S	553	185																29									13	
173924	D 1	T	S	A	R	558	186																28									13	
173927	D 1	A	A	G	A	561	187	39			32	21		22		27			43			27		28			37	44		33	62		
173939	D 1	A	R	C	S	573	191																46									19	
173945	D 1	C	Y	T	Y	579	193	20			30			26		27			39			27	21	16			35	30		17	34	16	
173946	D 1	G	G	A	S	580	194																28									13	
173957	D 1	A	R	C	S	591	197																27									12	
173959	D 1	A	H	G	R	593	198												23			15		11			10	18		14	23		
173960	D 1	C	H	T	H	594	198												23			15		10				18		14	22		
173970	D 1	G	E	A	K	604	202	21	11		26			26		25			37			23	23	16			33	29		17	36	15	
173971	D 1	A	E	T	V	605	202	15			15								22			15						18		14	22		

Haplotype_pos	3DXL5 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF1047766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chilliingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
173977	D 1	C	S	G	*	611	204	18			25					22			36			21					32		15	35		
173987	D 1	T	S	C	S	621	207	47	44	28	14	27				36		58	22	24	26	15	44	28	58	24		27		41	24	26	
173990	D 1	T	D	C	D	624	208												22														
174002	D 1	C	I	T	I	636	212	13			12								20			15						16		13	21		
174003	D 1	G	V	A	I	637	213												45												21		
174010	D 1	C	T	A	K	644	215	12			11								19			13						15			19		
177518	D 2	C	G	T	C	649	217																					23					
177519	D 2	T	G	G	G	650	217																					24					
177525	D 2	A	K	G	R	656	219												36												29		
177536	D 2	C	L	A	I	667	223												36												28		
177553	D 2	C	G	G	G	684	228													35		19					18						
177557	D 2	G	V	A	M	688	230												31												23		
177577	D 2	C	N	T	N	708	236	20			35	32		26		33			60			31	31	14			31	62		23	65	21	
177599	D 2	G	E	A	K	730	244	20			36	32		26		33			60			34		12			33	29		22	67		
177604	D 2	C	S	T	S	735	245			14	34		44		51		25		36	28	16	38					19	44	34	98	24	12	43	
177605	D 2	G	A	A	T	736	246												95									32				50		
177619	D 2	C	F	T	F	750	250	21			36	33		27		35			61			33		12			34	30		21	63		
177632	D 2	G	D	A	N	763	255					30		25		32			59			33		11			30	27		16	59		
177634	D 2	T	D	G	E	765	255				31			25		33			59			33		11			30	27		16	58		
177644	D 2	C	L	T	F	775	259	22			35	31		26		35			58			34		11			31	26		18	63		
177648	D 2	A	E	G	G	779	260	22			34	30		26		34			58			33	23	11			30	27		18	64	16	
177669	D 2	A	Q	G	R	800	267																20									15	
177676	D 2	C	P	T	P	807	269	21			34	29		27		33			60			35		11			30	26		16	56		
177692	D 2	G	A	T	S	823	275																15										
177715	D 2	A	G	G	G	846	282	20			32	26		23		29			57			32	13				29	23		14	50		
177720	D 2	C	P	T	L	851	284																11										
177727	D 2	C	H	T	H	858	286																10										
177730	D 2	C	S	T	S	861	287	19			32	26		22		30			56			33					28	23		14	44		
177757	D 2	C	F	A	L	888	296	18			31	24		21		28			55			32					26	23		14	39		
177781	D 2	A	S	G	S	912	304	16			30					26			55			35					25	23		13	38		
177797	D 2	C	L	G	V	928	310	15			28								57			32					25	24		13	39		
178815	Link	C	P	A	Q	968	323		18														63	10							35		
178824	Link	T	M	C	T	977	326		18														65								34		
179609	TM	A	K	G	K	1017	339																72										
179697	TM	T	Y	C	H	1105	369																42								31		
179704	TM	C	S	T	F	1112	371				44								96								40						
180503	CT 1	C	T	T	T	1173	391																	24									
180692	CT 2	C	P	T	P	1251	417			14	32			46		53		20			16	12	26	95		15	36	58		41		94	
180741	CT 2	G	E	T	*	1300	434	37	24	17						33		44		11	19			18	47	20				26			

Table S8: 3DXL5 capture SNPs

Haplotype_pos	3DXL7 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
98657	SP 1	C	L	T	L	33	11			35											45			44		28							
100063	D 0	G	V	A	M	133	45		44																								100
100085	D 0	G	R	A	H	155	52	72			39	39		41		53			56						11		47	77		79	71	12	
100126	D 0	A	T	G	A	196	66	34			25	26		28		34			35			37		19			34	36		30	38		
100128	D 0	C	T	A	T	198	66	37			31	27		25		36			41			26	89	24			29	50		49	29	56	
100137	D 0	C	P	T	P	207	69	34			25	27		30		35			36			37		18			35	36		31	37		95
100138	D 0	G	E	A	K	208	70	28	93																								
100146	D 0	T	M	G	M	216	72																					15					
100150	D 0	A	Q	C	Q	220	74																					15					
100161	D 0	C	N	A	K	231	77	32																16						27			
100170	D 0	C	T	G	T	240	80	28			26	23		23		34			38			23		21			26	28		38	27		
100185	D 0	C	T	A	T	255	85	32			25	27		29		33			38			30		14			31	38		25	37		
100197	D 0	C	A	A	A	267	89																89									50	
100198	D 0	G	G	T	W	268	90	56			25	27		30		33			38			31		30			32	38		57	35		
100209	D 0	A	T	G	T	279	93	36			24	24		28		32			39			31		14			31	38		28	36		
100215	D 0	C	S	T	S	285	95	58			48	45		51		67			75			53	93	30			56	62		59	61	57	
100221	D 0	A	L	C	F	291	97	59			48	43		51		67			75			52		29			55	62		59	60		
100221	D 0	A	L	C	F	291	97	59			48	43		51		67			75			52		29			55	28		59	60		
100222	D 0	T	E	C	Q	292	98																					27					
100234	D 0	C	P	G	A	304	102	41			24	23		29					41			31		15			30	38		30	34		
100234	D 0	C	P	A	T	304	102	18			22					34			33			21		15			23	23		27	24		
100234	D 0	C	P	A	T	304	102	18			22					34			33			21		14			23	23		27	24		
100249	D 0	A	S	C	R	319	107	42			25	21		29		30			42			31		15			32	37		32	33		
101189	D 1	C	L	A	L	372	124																					11					
101201	D 1	C	P	T	P	384	128												28														
101202	D 1	A	S	G	G	385	129												28														
101211	D 1	G	V	A	M	394	132	16			26	34		22		41			98			20		12			33	30		22	19		
101215	D 1	G	R	A	H	398	133			59		68		28			24		22	53	59			90		46		79			80		
101218	D 1	C	A	T	V	401	134												20														
101219	D 1	A	A	G	A	402	134												20														
101233	D 1	C	T	T	M	416	139												12														
101234	D 1	G	T	T	T	417	139									31			82			14					24	16		18	16		
101234	D 1	G	T	C	T	417	139												14														
101235	D 1	C	L	T	F	418	140										21																
101235	D 1	C	L	A	I	418	140												12														
101238	D 1	C	R	T	C	421	141																94								93		
101252	D 1	G	P	A	P	435	145	12			20					32			94			16		10			27	22		18	19		
101311	D 1	G	G	A	E	494	165												14														
101324	D 1	A	T	T	T	507	169	98	99	57	27	96	36	96	29	98	58	97	80	59	76	36	97	97	95	58	44	100		99	98	95	100
101361	D 1	A	L	G	V	544	182																					93					
101390	D 1	A	R	C	S	573	191																97										92

Haplotype_pos	3DXL7 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
101422	D 1	A	E	T	V	605	202												25														
101438	D 1	T	S	C	S	621	207												23														
101453	D 1	C	I	T	I	636	212												27														
101461	D 1	C	T	A	K	644	215												26														
104966	D 2	C	W	T	W	649	217																					42					
104967	D 2	T	W	G	W	650	217																					42					
104973	D 2	A	K	G	R	656	219																25										19
104984	D 2	C	L	A	I	667	223																27										19
105004	D 2	C	P	T	P	687	229																36										48
105005	D 2	G	V	A	M	688	230																28										23
105025	D 2	C	N	T	N	708	236					19							89				34					24	39				32
105046	D 2	C	S	T	S	729	243					16							86									22					
105047	D 2	G	E	A	K	730	244																31										43
105049	D 2	A	E	G	E	732	244					16							90				32					21	56				44
105060	D 2	A	D	G	G	743	248				41									30	47			50		34							
105082	D 2	T	D	G	E	765	255					11							87				36					19	62				26
105082	D 2	T	G	A	G	765	255																										15
105089	D 2	A	N	G	D	772	258					11							87				34					18	61				43
105096	D 2	A	E	G	G	779	260																41										
105120	D 2	A	N	G	S	803	268			71		99		97			43			64	82		92	96		64	12	93			93	90	
105128	D 2	G	G	A	R	811	271			27	96		96		98					33	18	89				35	87						
105150	D 2	C	P	G	R	833	278							91																			
105154	D 2	G	L	T	L	837	279							95																			
105163	D 2	A	G	G	G	846	282							96																			
107141	TM	C	H	T	Y	1105	369																48										
107148	TM	T	F	C	S	1112	371			22										15	25			45		19		64		17		

Table S9: 3DXL7 capture SNPs

Haplotype_pos	3DXS1 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF1047 66_Freq	HF5051 83Freq	HF5051 83Freq	HF4048 18_Freq	HF5052 04_Freq	HF598_Freq	HF7052 06_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chilliingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
274910	SP 2	A	L	C	L	51	17	11			26								36			22		12			28	11			42		
274918	SP 2	G	W	T	L	59	20				22	22				21			39			29					19						
276068	D 0	G	K	A	K	84	28																93								30		
276080	D 0	C	S	T	S	96	32												81														
276081	D 0	G	I	T	F	97	33																								11		
276143	D 0	C	S	T	S	159	53		59					93									98	43				76			94		
276144	D 0	C	M	G	V	160	54																					58					
276163	D 0	T	T	G	R	179	60																					14					
276180	D 0	A	K	G	E	196	66							48																	50		
276186	D 0	C	R	T	C	202	68		28															42									
276187	D 0	G	N	A	N	203	68																					13					
276192	D 0	T	S	C	P	208	70		35														94	43				28			47		
276198	D 0	C	G	T	W	214	72																					13					
276207	D 0	C	A	G	A	223	75																					68					
276274	D 0	A	E	G	G	290	97		37															42									
276282	D 0	T	W	C	R	298	100		38															43							21		
276306	D 0	G	E	C	Q	322	108	40	56					96			44						93	47				98		56	85	94	
277200	D 1	G	V	A	M	340	114		28					35									69	43				26				55	
277206	D 1	C	Q	A	K	346	116						47	58	38	48	62	44	91	41	35	38	96	19	76	42	56	59		70	88	71		
277207	D 1	A	Q	C	P	347	116							57			61		88			14	96	46			20	59		54	82	69		
277235	D 1	G	G	A	G	375	125					13							88				25				10				66		
277241	D 1	C	L	T	L	381	127							45									72					28				57	
277248	D 1	C	Q	G	E	388	130												84				22										
277249	D 1	A	Q	T	L	389	130												83				22										
277254	D 1	C	Q	G	E	394	132		13			11		49					84				22				10				30		
277270	D 1	G	R	A	H	410	137	24						43			21			54			75					85		26	65		
277282	D 1	T	L	C	P	422	141													31			13										
277287	D 1	A	M	G	V	427	143																13										
277287	D 1	A	M	C	L	427	143		10										50				12										
277291	D 1	T	F	G	C	431	144																					24					
277298	D 1	G	I	A	I	438	146																					33			15	59	
277313	D 1	T	H	A	Q	453	151							47					30				83			12							
277331	D 1	T	I	A	I	471	157																				10						
277342	D 1	G	R	A	H	482	161	31	21	25	24		28		28	21		29		29	22	17			66	24	27	65		31	24		
277344	D 1	G	L	T	L	484	162																							18			
277368	D 1	G	L	A	M	508	170																					19					
277371	D 1	C	P	T	S	511	171							25									80					19				46	
277383	D 1	T	S	A	T	523	175							22									75					19				44	
277383	D 1	T	S	A	T	523	175							22									75					11				44	
277384	D 1	C	S	T	F	524	175							35									81				14	29		14	51		
277390	D 1	G	G	C	A	530	177			10								16		15					49	18	20				21		

Haplotype_pos	3DXS1 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
277402	D 1	T	L	C	S	542	181		17																									
277412	D 1	C	A	T	A	552	184		19						28																			
277412	D 1	C	A	G	A	552	184		11																									
277420	D 1	A	Y	G	C	560	187		21						29																20			
277430	D 1	C	Y	T	Y	570	190		23						43																			42
277443	D 1	A	R	C	R	583	195																											34
277444	D 1	G	R	A	K	584	195																											24
277444	D 1	G	R	A	K	584	195																											24
277444	D 1	G	K	T	I	584	195																											
277445	D 1	A	R	C	S	585	195																											34
277448	D 1	C	S	T	S	588	196		25						41																			11
277455	D 1	G	E	A	K	595	199																											20
277466	D 1	C	A	T	A	606	202																											16
277472	D 1	C	S	T	S	612	204		29						39																			
277475	D 1	C	D	T	D	615	205																											13
277490	D 1	G	M	C	I	630	210																											
278854	D 2	A	K	G	K	648	216								96																			92
278864	D 2	C	L	A	I	658	220								55																			
278875	D 2	G	Q	A	Q	669	223								96																			92
278876	D 2	A	R	G	G	670	224								64																			12
278882	D 2	C	A	T	S	676	226								90																			
278888	D 2	G	P	A	T	682	228																											
278892	D 2	G	L	T	L	686	229																											33
278905	D 2	C	N	T	N	699	233	28										27																
278919	D 2	G	P	A	Q	713	238																											
278920	D 2	C	P	T	P	714	238																											
278962	D 2	G	E	A	E	756	252		22																33									
278966	D 2	G	R	A	R	760	254																											
279006	D 2	G	R	A	H	800	267	23										27											60			24		
279038	D 2	C	L	T	L	832	278																							64				
279067	D 2	T	Y	C	Y	861	287	13	39						96										24					94		20		92
279084	D 2	T	F	A	Y	878	293		18						55															69				
279089	D 2	T	C	C	R	883	295	56	35						96			54							26					94		61		92
279406	Link	A	T	T	S	943	315																	27										12
279408	Link	T	T	G	T	945	315		63						33									56	39		42			57				
279414	Link	T	S	A	R	951	317																											12
279414	Link	T	S	G	R	951	317																	21										
279421	Link	C	P	T	S	958	320		31															27						14				
279422	Link	C	P	G	R	959	320																				42							
279428	Link	T	I	C	T	965	322																	25						25				12
279429	Link	C	I	T	I	966	322																	80										46
																								45										27

Haplotype_pos	3DXS1 feature	ref_base	ref_residue	var_base	var_res	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF5051833Freq	HF5051833Freq	HF5051833Freq	HF404818_Freq	HF5052O4_Freq	HF598_Freq	HF7052O6_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
279435	Link	T	D	C	D	972	324																	26									13	
279439	Link	C	H	G	D	976	326																	21									15	
279443	Link	C	T	T	I	980	327																	25									13	
280199	TM	C	A	T	V	992	331																	36									13	
280202	TM	G	R	A	Q	995	332																	51									22	
280206	TM	T	L	C	L	999	333																	13										
280224	TM	A	R	C	S	1017	339		64			53	50		99		47				98			44	99	46			52	55	50		86	98
280233	TM	C	H	T	H	1026	342																	27									11	
280243	TM	A	R	C	R	1036	346					45	43		42		39				75			30				38		33			30	
280244	TM	G	Q	A	Q	1037	346																										29	
280248	TM	C	L	T	L	1041	347																	28									12	
280251	TM	C	S	T	S	1044	348		32															45	40				52			86	20	
280252	TM	G	V	A	I	1045	349					51	50		51		49				98			40	29			53		36			43	

Table S10: 3DXS1 capture SNPs

Haplotype_pos	2DS1 feature	ref_base	var_base	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
215378	SP 1	C	T	15	5																						31					
215379	SP 1	G	A	16	6																	26									22	
215383	SP 1	G	A	20	7																	26									21	
216906	D 0	C	G	92	31	32				47	44		93		44	48		96			56	99				46		100	33	97	98	93
216908	D 0	C	A	94	32																	66					32	33	71	94	71	99
216937	D 0	G	A	123	41	32				47	44		92		44	50		96			54	30				44	63	71	33	94	27	99
216969	D 0	T	C	155	52																	35									23	
216971	D 0	T	G	157	53																	33					36	25			48	
216973	D 0	G	A	159	53																	35									22	
216975	D 0	A	G	161	54																	35									23	
216983	D 0	G	A	169	57																						31					
216984	D 0	C	T	170	57																	34									23	
216993	D 0	G	A	179	60																	33									22	
216999	D 0	G	A	185	62																	32									22	
217000	D 0	G	A	186	62																	33									22	
217018	D 0	G	C	204	68																	31									23	
217051	D 0	G	T	237	79	30				47	45		96		55	57		97			53	99				47	99	75	42	97	98	99
217051	D 0	G	A	237	79		44																39									99
217054	D 0	A	G	240	80	30				45	44		94		55	56		95			49	31				45	61	69	41	95	27	100
217095	D 0	G	A	281	94																	28									23	
217096	D 0	G	T	282	94	32				47	57		96		62	60	16	98	12	12	55	99		13		51	99	75	49	98	99	99
217128	D 0	G	A	314	105																	29									22	
217142	D 0	G	A	328	110																	36					29				49	
217161	D 0	A	G	347	116																	28									24	
221267	D 2	G	A	393	131	64	48									56						53	49						69		42	92
221278	D 2	T	C	404	135	65	49			51	50		96		52	56		97			61	100	50			59	98	100	71	99	99	98
221342	D 2	G	A	468	156			35	45			98		96					44	39	39					44	36		29			
221370	D 2	G	A	496	166																										35	
221374	D 2	G	T	500	167													21														
221375	D 2	A	G	501	167													21														
221379	D 2	T	C	505	169					13								23														
221385	D 2	T	C	511	171					15								25			12											
221401	D 2	A	T	527	176					16								23								18						
221435	D 2	C	A	561	187					36	31		62		30			53			41	39				38	86	50		59	55	
221459	D 2	C	T	585	195																						31					
221460	D 2	G	A	586	196																										17	
221473	D 2	G	A	599	200																						32					
221485	D 2	C	T	611	204					10																						
221493	D 2	G	A	619	207																	14										
221497	D 2	T	C	623	208					32								51			30	17				24	15					
221501	D 2	C	T	627	209																	14	13									

Haplotype_pos	2DS1 feature	ref_base	var_base	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillinghamj_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nereswater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
221506	D 2	A	C	632	211				18								52				18				23	15					
221509	D 2	T	G	635	212												36														
221510	D 2	A	G	636	212				18								51				18					14					
222535	Link	A	T	654	218																11										
222547	Link	G	T	666	222				15			30					31			17	22				11						
222550	Link	A	T	669	223				15								32			17	32				12						
222561	Link	C	T	680	227							21					23														
222580	Link	T	A	699	233							25					23			12	25								25		
223334	TM	T	G	706	236													42			22							19			
223335	TM	C	T	707	236																11										
223344	TM	T	C	716	239	46			56	58		98		61	66	33	98	49	46	58	100	38	40	41	63	99		48	94	99	99
223388	TM	G	A	760	254													53													
223403	TM	A	G	775	259	21									20													27			97
223406	TM	A	C	778	260																					30					
223418	TM	A	G	790	264		41										19	58	62		54	49	49	59	26	38		27		67	

Table S11: 2DS1 capture SNPs

Haplotype_pos	2DS2 feature	ref_base	var_base	cDNA_pos	res_pos	HF504805_Fred	HF504882_Fred	HF104766_Fred	HF505183Fred	HF505183Fred	HF404818_Fred	HF505204_Fred	HF598_Fred	HF705206_Fred	HF4222_Fred	Chill_250b_Fred	HF159_Fred	HF405_Fred	HF766_Fred	HF982_Fred	Kuchinoshima_Fred	Blackisle_Fred	Chilliingham3_Fred	HF252_Fred	HF652_Fred	Nelore_NE14_Fred	Nelore_NE43_Fred	Nerewater_Fred	Sahiwal_SW2_Fred	Sahiwal_SW3_Fred	Hap2_Fred
145129	SP 1	A	G	25	9				13								26			14					10	35			95		
146652	D 0	C	A	95	32																					29				79	
146728	D 0	C	T	171	57																33										
146737	D 0	G	A	180	60																44										
146744	D 0	G	A	187	63																48										
146798	D 0	C	A	241	81																59									15	
146800	D 0	C	T	243	81																							11			
146815	D 0	T	C	258	86																61									14	
146826	D 0	T	C	269	90	28	30	30	51	38	44	56	44	52	46	50	61	30	31	53	18	20	44	32	49	30		29	37		63
146830	D 0	G	A	273	91				16	38	45	56	45	52	46	49	60	30	30	53	19	21	44	32	49	30		30	30		
146832	D 0	C	T	275	92						21	22	23					12		14											
146844	D 0	A	G	287	96	29	29	30	50	36	44	54	44	52	44	49	60	30	30	53	17	20	44	33	49	29		30	31		67
146860	D 0	C	T	303	101																69									15	
146871	D 0	T	C	314	105	99	62	80	66	99	45	82	45	99	75	97	100	69	76	75	100	99	99	72	67	99		98	98	97	68
146872	D 0	G	A	315	105																69									18	
146883	D 0	C	A	326	109	26	26	21	13		24		21		38	46		17	23	13		20	44	21	14						
146889	D 0	T	A	332	111																69									20	
146892	D 0	T	C	335	112																72									20	
146908	D 0	C	G	351	117				12								31			15											
146909	D 0	T	A	352	118																75									23	
148816	D 2	C	T	414	138	30		18									16	17				44		15				35			
148834	D 2	C	T	432	144				13							43				16					11						
148935	D 2	A	G	533	178											43															
148942	D 2	T	C	540	180	11	10	19	43		37	76	36	34	37	31	94	22	23	51		10	24	26	50	93		13	92	81	
148986	D 2	G	A	584	195													14		22					21						
148987	D 2	T	C	585	195							42				26		23	14	43			20	18	44	28					
148992	D 2	C	A	590	197												42														
149029	D 2	C	T	627	209			14	29		37	48	30					17	12	22				20	23						
150876	TM	G	A	767	256																					50				100	

Table S12: 2DS2 capture SNPs

Table S13: 2DS3 capture SNPs

Haplotype_pos	2DS3 feature	ref_base	var_base	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF1047766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
76202	D 2	C	T	432	144												25			10											
76303	D 2	A	G	533	178				14								36			17					13						
76354	D 2	G	A	584	195													10													
76355	D 2	C	T	585	195																			10							
76360	D 2	C	A	590	197				15								39			18	29				13						
78210	TM	A	T	726	242				15			27		29			89			16		19			20	85			86		
78227	TM	C	T	743	248			22		26								22	23												
78239	TM	C	T	755	252				11			25		28			95			10					18	87			93		

Haplotype_pos	3DXS2 feature	ref_base	var_base	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF1047766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillingham3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
159801	D 2	C	T	649	217			18	17	29		35		23			29	14		13		15		12							51
159808	D 2	T	C	656	219	64	29	21	15	33				49			32		23			42	91	17	12	29	69	51	35		51
159826	D 2	T	C	674	225	100	24	71	52	98		98		97			97	63										98			96
159875	D 2	G	T	723	241				14																						
159878	D 2	C	T	726	242	25								21																	
159920	D 2	G	C	768	256	31								22																	53
159920	D 2	G	C	768	256	24								22																	53
159920	D 2	G	T	768	256		14																								53
159976	D 2	A	G	824	275	50				23				42																	
160024	D 2	T	C	872	291																	18			22			50			
160031	D 2	C	T	879	293																	18					12				
160045	D 2	C	G	893	298					30		21		26			30					25			15	57		65			
160045	D 2	C	A	893	298	18																16					14				
160058	D 2	A	G	906	302	33	29	19		22					28	38		17	19			21	21	16			33				
160060	D 2	T	C	908	303				18	32		31		28			32			20				14	20	64		77			
160068	D 2	T	G	916	306				13								28			13					10	21		38			
160077	D 2	C	G	925	309																					30				62	
161061	Link	G	A	963	321									48	58								44								59

Table S14: 3DXS2 capture SNPs

Haplotype_pos	3DXS3 feature	ref_base	var_base	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF1047766_Freq	HF5051838Freq	HF5051838Freq	HF404818_Freq	HF5052O4_Freq	HF598_Freq	HF7052O6_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chilllinghamβ_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
83253	SP 2	T	A	44	15																54										
83265	SP 2	T	C	56	19																68										
84440	D 0	G	A	105	35	31																	53					37			
84487	D 0	C	T	152	51	31	31	42		28				20	38	39		28	43				38	35	29			28			
84640	D 0	C	T	305	102	38																	48					29			
85637	D 1	C	T	408	136																		13				71	24			
85639	D 1	A	C	410	137	33	35	27	50	44	34	44	31	60	53	52	65	28	24	49	43	32	50	27	46	70	71	46	44	89	99
85799	D 1	T	C	570	190	41	28	26	55	37	33	41	28	58	47	44	65	27	19	38		32	48	24	46	77		45	60	97	99
85812	D 1	T	C	583	195	43	26	18	22	23				37		48	32	17	12	12		31	46		16	44		45	58		
85819	D 1	G	T	590	197																	14						27			
85837	D 1	G	A	608	203												27								17	37			55		
85851	D 1	C	T	622	208																			17							
87181	D 2	C	T	648	216				18			40													19						
87184	D 2	A	C	651	217	59		52	66			96					98						78		51	60	97		60		93
87188	D 2	T	C	655	219	43	31	17	20							35		12					37						43		
87204	D 2	C	T	671	224				21								34			22					17	26				39	
87253	D 2	G	A	720	240				23								35			24					16	25			40		
87255	D 2	G	T	722	241				20								29								18						
87258	D 2	C	T	725	242	17											38			26					18	25		24	41		
87273	D 2	C	G	740	247				23								32			29					18	24			41		
87300	D 2	G	T	767	256	28	39	21							24	42		14	19					15							
87300	D 2	G	C	767	256	21			23								29			26			21		19	23		30	46		
87300	D 2	G	C	767	256	21			23								29			26			13		19	23		26	46		
87309	D 2	A	G	776	259				23								29			27			13		18	22			41		
87328	D 2	G	A	795	265												27			22					14	22			39		
87329	D 2	G	T	796	266												29			22					15	22			38		
87329	D 2	G	C	796	266	20																	13					24			
87334	D 2	T	C	801	267				19					21			30			23			13		17	23		25	41		
87335	D 2	A	C	802	268																		13					24			
87340	D 2	A	T	807	269				20					21			31			24					16	24			40		
87356	D 2	A	G	823	275	48	36	17	23	27				43	25	39	35	17	19	24		29		16	17	24		56	48		
87395	D 2	T	C	862	288	24			19					21			33								14	19		21	33		
87404	D 2	T	C	871	291	22			17								27								14	16		21	28		
87411	D 2	C	T	878	293	22																						18			
87425	D 2	C	G	892	298												23									15			20		
87425	D 2	C	A	892	298																							15			
87438	D 2	A	G	905	302														17									19			
87440	D 2	T	C	907	303												20									12					
88415	Link	T	C	936	312																88								44		
88418	Link	C	T	939	313																78								38		
88424	Link	G	C	945	315																					49			33		

Table S15: 3DXS3 capture SNPs

Haplotype_pos	3DXS3 feature	ref_base	var_base	cDNA_pos	res_pos	HP504805_Freq	HP504882_Freq	HP1047766_Freq	HP5051838Freq	HP5051838Freq	HP4048I8_Freq	HP505204_Freq	HF598_Freq	HF705206_Freq	HP4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillinghamβ_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
88441	Link	G	A	962	321																90									55
89279	TM	C	T	1052	351	26	49	20							24	37		16	23			16	44	18				28		47	53

Haplotype_pos	3DXL6 feature	ref_base	var_base	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillinghamr3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
62974	SP 1	G	A	5	2	38																25				40		30		20	
62980	SP 1	C	G	11	4																14										
63002	SP 1	C	T	33	11																21										
63164	SP 2	A	G	53	18	11			12								30			18		13			33	23		17	19		
63170	SP 2	G	T	59	20																				11						
63175	SP 2	C	G	64	22												30			18		12			32	23		16	21		
63176	SP 2	T	C	65	22												30			18	52	12			32	23		16	20		30
63178	SP 2	G	A	67	23												98			54	99	13			62	65		21	95	65	
63179	SP 2	C	T	68	23					57		66		56			68			34	61				28	39			74		
63180	SP 2	G	A	69	23					43		43		39							62	60		45						14	
64328	D 0	G	A	73	25												96			52	64				50	40			88	37	
64351	D 0	C	A	96	32				44	45		43		45												44				27	
64352	D 0	G	A	97	33				46	47		44		46			97			53	61				40	85			93	63	
64366	D 0	T	C	111	37			64		98		93			55		91	56	66		62	60		45		20				45	
64387	D 0	C	T	132	44				50	50		48		49			97			51					37	42			91		
64409	D 0	C	T	154	52				49	52		50		48			97			48					37	43			90		42
64426	D 0	A	C	171	57																					41					26
64457	D 0	C	T	202	68				48	53		53		48			98			50	53				50	51			94		
64458	D 0	G	A	203	68																					41					42
64459	D 0	T	A	204	68																					41					42
64479	D 0	A	G	224	75																					41					
64551	D 0	C	A	296	99				44	54		55		51			97			52					57	51			93		
65510	D 1	C	A	358	120																								10		
65525	D 1	G	A	373	125	31																45						32			
65527	D 1	C	T	375	125	34	52	43	48		64	36	71	34	48	48		41	42	52			45	46	43			36			98
65532	D 1	G	A	380	127				25	48		33		33			95			25	40				28	51			94	33	
65538	D 1	C	T	386	129																51									30	
65544	D 1	A	G	392	131				23	47		32		32			96			24	52				28	58			92	30	
65554	D 1	C	T	402	134																50									30	
65556	D 1	C	T	404	135											48								46							
65559	D 1	T	C	407	136				23	45		31		30			95			24	94				26	95			93	94	
65565	D 1	C	G	413	138				23	43		28		29			84			22					22	44			86		
65598	D 1	C	T	446	149																53									33	
65599	D 1	A	G	447	149																					41				28	
65616	D 1	G	A	464	155																53									33	
65632	D 1	G	A	480	160	30																48						25			
65638	D 1	G	A	486	162																52									34	
65672	D 1	T	C	520	174																52									33	
65680	D 1	G	A	528	176			20	23		32	31	28		22			23	16	23				21	27						74
65701	D 1	T	C	549	183												13														
65720	D 1	C	T	568	190																46									29	

Haplotype_pos	3DXL6 feature	ref_base	var_base	cDNA_pos	res_pos	HF504805_Fred	HF504882_Fred	HF104766_Fred	HF505183Fred	HF505183Fred	HF404818_Fred	HF505204_Fred	HF598_Fred	HF705206_Fred	HF4222_Fred	Chill_250b_Fred	HF159_Fred	HF405_Fred	HF766_Fred	HF982_Fred	Kuchinoshima_Fred	Blackisle_Fred	Chillinghamβ_Fred	HF252_Fred	HF652_Fred	Nelore_NE14_Fred	Nelore_NE43_Fred	Nerewater_Fred	Sahiwal_SW2_Fred	Sahiwal_SW3_Fred	Hap2_Fred
65729	D 1	T	C	577	193				21	55		37		38			96			30	98				38	97			90	96	
65732	D 1	C	A	580	194												21								12				18		
65733	D 1	A	C	581	194												21				15				12				18	12	
65733	D 1	A	G	581	194																46									29	
65734	D 1	G	A	582	194																15										
65746	D 1	A	C	594	198																33									25	
65746	D 1	A	T	594	198																								19		
65756	D 1	C	T	604	202												12				22								23		
65765	D 1	T	C	613	205																51									38	
65777	D 1	T	C	625	209	12											22			12					14				16		
65777	D 1	T	A	625	209																42									27	
65777	D 1	T	A	625	209																42									16	
65777	D 1	T	A	625	209																24									27	
65777	D 1	T	A	625	209																24									16	
65778	D 1	G	A	626	209																20										
65785	D 1	C	A	633	211												11												23		
66815	D 2	C	T	638	213																16									14	
66816	D 2	T	G	639	213																16									14	
66826	D 2	A	G	649	217																17									17	
66846	D 2	G	A	669	223																16										
66848	D 2	T	G	671	224																48					15				40	
66850	D 2	C	T	673	225					21	34	25		29			59			23					19						
66853	D 2	T	C	676	226																63									32	
66855	D 2	T	A	678	226																14										
66867	D 2	G	A	690	230																39									26	
66874	D 2	T	C	697	233																50					13				41	
66878	D 2	A	T	701	234																29									22	
66879	D 2	C	A	702	234																11										
66881	D 2	T	G	704	235																30									22	
66895	D 2	C	T	718	240												30			10					13						
66896	D 2	G	C	719	240																35									26	
66896	D 2	G	A	719	240																					15				10	
66901	D 2	T	C	724	242												34			12	17				18	14				18	
66902	D 2	G	A	725	242												29								13	12					
66923	D 2	C	T	746	249																37									26	
66927	D 2	G	T	750	250																38									26	
66931	D 2	T	G	754	252												32				57				17						
66931	D 2	T	A	754	252																36					37				49	
66931	D 2	T	A	754	252																36					37				43	
66931	D 2	T	A	754	252																36					18				49	
66931	D 2	T	A	754	252																36					18				43	

Haplotype_pos	3DXL6 feature	ref_base	var_base	cDNA_pos	res_pos	HF504805_Fred	HF504882_Fred	HF104766_Fred	HF505183Fred	HF505183Fred	HF505183Fred	HF404818_Fred	HF505204_Fred	HF598_Fred	HF705206_Fred	HF4222_Fred	Chill_250b_Fred	HF159_Fred	HF405_Fred	HF766_Fred	HF982_Fred	Kuchinoshima_Fred	Blackisle_Fred	Chillingham3_Fred	HF252_Fred	HF652_Fred	Nelore_NE14_Fred	Nelore_NE43_Fred	Nerewater_Fred	Sahiwal_SW2_Fred	Sahiwal_SW3_Fred	Hap2_Fred
66934	D 2	G	A	757	253																	75									46	
66938	D 2	A	G	761	254													26				48				12	51				61	
66946	D 2	G	T	769	257																	45									30	
66947	D 2	T	C	770	257																	91					36				86	
66948	D 2	G	T	771	257																						36				30	
66952	D 2	A	G	775	259																	79					30				75	
66965	D 2	A	C	788	263																	43									25	
66967	D 2	A	G	790	264													24				46				11	53				63	
66970	D 2	C	T	793	265																	31					29				44	
66974	D 2	A	G	797	266													23				47				11	51				63	
66975	D 2	T	C	798	266													24				92				10	52				89	
66976	D 2	G	A	799	267																	43									24	
66981	D 2	A	G	804	268	32	47	43	40			62	30	62	27	46	52		46	40	48			50	53	42		34				99
66986	D 2	A	T	809	270																			30							27	
66989	D 2	C	A	812	271																	41									23	
67002	D 2	G	T	825	275													22				18					26				20	
67008	D 2	T	A	831	277													22				19					26				20	
67026	D 2	T	C	849	283													24				47									28	
67028	D 2	G	A	851	284																	37									18	
67030	D 2	A	G	853	285																	47									28	
67042	D 2	T	A	865	289																	34									16	
67044	D 2	T	C	867	289													20			10	54					30				39	
67045	D 2	G	A	868	290													14				54					27				36	
67046	D 2	G	A	869	290																	34									16	
67055	D 2	C	T	878	293																	27									18	
67058	D 2	A	G	881	294													12				19					27				22	
67059	D 2	C	G	882	294											22																99
67064	D 2	C	G	887	296	63	48	43	40			62		63	23	48	51		45	39	44	30		55	51	47			65		19	100
67064	D 2	C	A	887	296			35		33			32						29	37			54		25							100
67064	D 2	C	A	887	296			35		33			32						29	37			38		25							100
67069	D 2	T	C	892	298																	12									12	
67071	D 2	C	G	894	298	67	49	78	47	60	62	71	65	41	73	54	54	31	74	75	58	98	98	56	78	56	65		74	11	97	98
67080	D 2	C	A	903	301																	20									14	
67086	D 2	C	T	909	303																	26									17	
67098	D 2	A	C	921	307	69	49	79	77	98	64	98	71	68	69	74	54	98	76	75	78	100	99	55	78	80	98		69	98	98	100
67109	D 2	C	G	932	311				25	37		24			29			70			23					25	32			84		
68877	TM	A	C	992	331																	84									36	
68891	TM	C	G	1006	336																	83									33	
68902	TM	G	A	1017	339																	83									30	
68933	TM	T	G	1048	350																	77									27	
68964	TM	C	T	1079	360				19	41		30		33				90			21	71				26	95			91	90	

Table S16: 3DXL6 capture SNPs

Haplotype_pos	3DXL6 feature	ref_base	var_base	cDNA_pos	res_pos	HF504805_Freq	HF504882_Freq	HF104766_Freq	HF505183Freq	HF505183Freq	HF404818_Freq	HF505204_Freq	HF598_Freq	HF705206_Freq	HF4222_Freq	Chill_250b_Freq	HF159_Freq	HF405_Freq	HF766_Freq	HF982_Freq	Kuchinoshima_Freq	Blackisle_Freq	Chillinghamr3_Freq	HF252_Freq	HF652_Freq	Nelore_NE14_Freq	Nelore_NE43_Freq	Nerewater_Freq	Sahiwal_SW2_Freq	Sahiwal_SW3_Freq	Hap2_Freq
69717	CT 1	T	C	1109	370																79									37	
69736	CT 1	G	C	1128	376																					13					
69892	CT 2	G	T	1173	391																					79				84	
69912	CT 2	C	A	1193	398																					78				88	
70032	CT 2	C	T	1313	438																					89				92	