Investigation of Genetic Variation in the Collagenous Lectins of Livestock with and without Infectious Diseases

by

Russell S. Fraser

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Doctor of Philosophy
in
Pathobiology

Guelph, Ontario, Canada

ABSTRACT

INVESTIGATION OF GENETIC VARIATION IN THE COLLAGENOUS LECTINS OF LIVESTOCK WITH AND WITHOUT INFECTIOUS DISEASES

Russell S. Fraser
University of Guelph, 2018

Advisor:
Brandon N. Lillie

Infectious diseases in livestock are a major source of economic loss, decreased welfare, and antimicrobial usage. Typical interventions rely on treatment of the host and/or environmental modifications to reduce pathogen exposure and disease occurrence and/or severity. Various host genetic factors influence the resistance of animals to infectious diseases. In particular, mutations in genes of the immune system can alter disease susceptibility. Collagenous lectins (CLs) are pattern recognition receptors of the innate immune system that contribute to disease resistance by binding surface glycans of bacteria and other potentially pathogenic organisms. Studies in humans and animals have shown that mutations in certain CL genes are associated with infectious diseases. The main objectives of this thesis were to further investigate genetic variation in CLs in cattle, horses, and pigs, and their relationship to infectious disease resistance. Pooled, targeted next-generation sequencing of the CL genes identified 43 missense mutations in cattle, 11 of which were predicted to impact protein structure. In horses, 1 nonsense and 43 missense mutations were identified, including 14 predicted to be functionally relevant. In particular, one missense mutation in the collagen-like domain of *MBL1* was found that is similar to a triple-helix disrupting mutation in human *MBL2* associated with susceptibility to infectious

disease. Allele frequencies were compared to identify alleles (74 in cattle, 113 in horses) associated with infectious diseases. Additional *in silico* analysis of the equine variants associated with infectious diseases identified 2 variants predicted to impact miRNA binding, 8 variants that impacted transcription factor binding sites, and 1 missense variant. In pigs, expression quantitative trait locus (eQTL) analysis identified 298 eQTLs in innate immune genes, 74 of which were genotyped in 1013 pigs (592 healthy, 421 with infectious disease). Variants that altered expression of these genes were associated with *Mycoplasma*, *E. coli*, swine influenza virus, and porcine reproductive and respiratory syndrome virus infection. These studies identified polymorphisms in CL genes that are associated with infectious diseases of livestock. These alleles represent potential candidates for genetic selection for enhanced resistance to infectious diseases of livestock, and expand our understanding of the roles of collagenous lectins in innate immunity.

Acknowledgements

I would like to start by expressing my sincere and heartfelt gratitude to Dr. Brandon Lillie. As a thesis advisor, as a mentor, as a pathology instructor, and as a friend, Brandon always had an open door and time to talk, whatever the subject. He excelled at providing just the right amount of guidance, largely allowing me to control the direction of the thesis, while at the same time providing gentle nudges in the right direction. I thank him for his positive attitude, his humour, his thoughtfulness, and for being an excellent academic role model.

I also extend my thanks to the members of my advisory committee: Drs. L. Arroyo, T. Hayes, and J. Lumsden, each of whom provided valuable insights and comments as the thesis progressed, and have given me valuable career and life advice over the years. Thank you for your help, your relaxed and friendly attitudes, and for helping mentor me at this stage of my career.

To my fellow graduate students, many of whom have become good friends over the years, my thanks for adding a social element to the degree. Whether it was at fluid rounds or sports, I value and cherish your friendship, and hope to maintain it in the years to come.

My parents and my sisters have been a constant, unwavering source of support over the many years of post-secondary education, and I owe them more than can be put into words. To them I extend my thanks and my love.

Finally, my own budding little family. Sonya and Felix, you bring so much joy into my life every day. You support me with your kindness, your love, your laughter, and your hearts. This endeavour would not have been possible without you, and I cannot wait for the next step in our adventure.

Statement of Work Performed

All of the work described in this thesis was performed by me, with the following exceptions:

Chapter 4: The microarray expression data was obtained from Dr. Hein Snyman's doctoral thesis. Genomic material of normal pigs was isolated by Drs. Snyman, Hammermueller, and Lillie. Genomic material of diseased pigs was isolated by Dr. Lillie. Porcine sequencing library preparation was done by myself and by Dr. Ann Meyer, and it was sequenced by The Center for Applied Genomics (The Hospital for Sick Kids, Toronto, Canada).

Table of Contents

List of Tables

List of Figures

List of Abbreviations

3MC	Carnevale, Mingarelli, Malpuech and Michels syndrome
APP	*Actinobacillus pleuropneumonia*
BRD	bovine respiratory disease
BQSR	base quality score recalibration
BRSV	bovine respiratory syncytial virus
CHiP	chromatin immunoprecipitation
CLD	collagen-like domain
CRD	carbohydrate recognition domain
CTLD	C-type lectin-like domain
DAF	decay accelerating factor
eQTL	expression quantitative trait loci
FBG	fibrinogen-like binding domain
FDR	false discovery rate
GWAS	genome-wide association study
IAD	inflammatory airway disease
IAV	influenza A virus
IL	interleukin
indel	insertion/deletion
LD	linkage disequilibrium
LPS	lipopolysaccharide
MALDI-TOF	Matrix-assisted laser desorption/ionization time of flight
MASP	mannan-binding lectin associated serine protease/mannose-associated serine protease
miRNA	microRNA
MRE	miRNA response element
NGS	next-generation sequencing
NLR	nucleotide-binding domain and leucine-rich repeat containing receptors
NOD	nuclear oligomerization domain
PAMP	pathogen associated molecular pattern
PCV2	porcine circovirus type 2
PI3	parainfluenza virus 3
pQTL	protein quantitative trait locus/loci
PRR	pattern recognition receptor
PRRSV	porcine reproductive and respiratory virus
QTL	quantitative trait loci

QTN	quantitative trait nucleotide
RAO	reactive airway obstruction
RIG-I	retinoic acid-inducible gene I
RLR	RIG-I-like receptor
SCS	somatic cell score
SIV	swine influenza virus
SNP	single nucleotide polymorphism
SNV	short nucleotide variant
TF	transcription factor
TFBS	transcription factor binding site
TLR	toll-like receptor
UHN	University Health Network
UTR	untranslated region
VEP	variant effect predictor
WES	whole exome sequencing
WGS	whole genome sequencing

List of Appendices

Chapter 1: Review of the Literature

1.1 Introduction

Immunity to infection is comprised of two intertwined systems: the innate immune system and the adaptive immune system, each of which involve genetic variation to contend with pathogen variation. Innate immunity is comprised of physical, chemical, and cellular barriers against infection, with evolution and selection of germline variants for functional components. In contrast, adaptive immunity involves lymphocytes that derive functional specificity and immunological memory through somatic mutations (immunoglobulin and T-cell-receptor recombinants) that can be selectively expanded in response to particular pathogens, resulting in enhanced subsequent responses to the same pathogen.

A relatively recent focus of research has been the role of pattern recognition receptors (PRRs) in the function of the innate immune system (Medzhitov and Janeway, 1997). PRRs are proteins that recognize motifs on pathogens called pathogen-associated molecular patterns (PAMPs), as well as host-derived material exposed under abnormal conditions (such as cellular stress or necrosis), termed damage associated molecular patterns (DAMPs) (Kawai and Akira, 2009; Medzhitov and Janeway, 1997). In general, PAMPs are highly conserved across broad classes of pathogens (and non-pathogens), and represent structures important to the survival of the pathogen (Mogensen, 2009). These PAMP structures are generally not expressed by the host, providing a degree of host-pathogen differentiation, or specificity, to the innate immune system.

Pattern recognition receptors are frequently grouped into four major categories, based primarily on their ligand-recognition domains (Iwasaki et al., 2015). The first PRRs to be defined were the Toll-like receptors (TLRs) (Medzhitov et al., 1997). They are found within cellular and endosomal membranes and recognize a variety of PAMPs, including ssRNA, ssDNA, and

lipopolysaccharide (LPS), via a leucine-rich repeat domain (Mogensen, 2009; Takeuchi and Akira, 2008). Two other well characterized groups include the RIG-I like receptors (RLRs), which are cytosolic proteins that primarily recognize viral dsRNA via DExD/H box helicase domains (Kawai and Akira, 2009; Yoneyama and Fujita, 2007), and the nucleotide-binding domain and leucine-rich repeat containing receptors (NLRs) that are found in the cytoplasm of immune cells and recognize bacterial peptidoglycans (Chen et al., 2009; Ting et al., 2008). Although the NLRs contain a leucine-rich repeat ligand-binding domain, they are defined by the presence of a nucleotide-binding domain (NOD) that facilitates oligomerization (Chen et al., 2009). Finally, certain subfamilies of the large C-type lectin-like domain (CTLD) superfamily comprise a fourth broad category of PRR, which includes both membrane-bound and soluble receptors capable of recognizing glycans, lipids, and proteins through a carbohydrate-recognition domain (CRD) (Chiffoleau, 2018). One subfamily of the CTLDs is the collagenous lectins, a well conserved group of plasma and membrane-bound proteins that recognize carbohydrate motifs of bacteria, viruses, protozoa, and fungi (Fujita, 2002; Fujita et al., 2004b). The collagenous lectins will be the primary focus of this literature review.

Altered expression, structure, and function of PRRs are well documented causes of immunodeficiencies and susceptibility to infection. Polymorphisms in TLRs can lead to an increase in susceptibility to malaria, tuberculosis, leprosy, and Legionnaire's disease (Carpenter and O'Neill, 2007). Deficiencies of collagenous lectins in humans are associated with increased risk of HIV infection (Garred et al., 1997), bacterial meningitis (Hibberd et al., 1999), viral hepatitis (Eisen and Minchinton, 2003), and respiratory infections (Super et al., 1989). The situation in animals, however, is less clear. Genetic variants in some collagenous lectins are more common in pigs with common infectious diseases (Lillie et al., 2007), and preliminary data

indicate that certain lectin alleles are associated with bovine mastitis (C. Wang et al., 2011; Wang et al., 2012). However, the exact role that collagenous lectins play in the innate immune response remains insufficiently characterized in many species of veterinary interest, including cattle, horses, and pigs.

This review will provide a summary of the role of collagenous lectins in innate immunity, and will draw on evidence from a variety of veterinary and human studies. It will provide an overview of some common infectious diseases of cattle and horses, and by doing so will establish the basis for the central hypotheses that genetic abnormalities in collagenous lectins are common in animals, and that some of these abnormalities are more common in animals with infectious diseases.

1.2 Overview of the Collagenous Lectins

For the purposes of this review, the term collagenous lectins will be used to include both collectins and ficolins. Collectins are one of seventeen subgroups of the large C-type lectin-like domain (CTLD) superfamily (Zelensky and Gready, 2005). The defining characteristics of members of the CTLD superfamily have been revised several times. Originally, each member of the CTLD superfamily was described as having a C-terminal carbohydrate recognition domain (CRD) that bound carbohydrate residues in a calcium dependent ("C-type") fashion. As new proteins were discovered, however, it was found that not all proteins with a CRD were calcium dependent, nor did they all bind carbohydrates. The broadly inclusive "C-type lectin-like domain" moniker was therefore adopted to encompass these proteins (Zelensky and Gready, 2005). Ficolins are a group of proteins that share a similar collagen-like domain (CLD) with the collectins, but differ at their C-terminus, where they are characterized by a fibrinogen-like binding domain rather than a CRD.

In mammals, thirteen collagenous lectins have been identified, though not all are present in a single species. They include the mannose-binding lectins A and C, surfactant proteins A and D, collectins 10, 11, 12, 43, 46, conglutinin, and ficolins 1, 2, and 3 (Hansen et al., 2003b; Holmskov et al., 2003, 1993; Keshi et al., 2006; Ohtani et al., 2001, 1999) . Some differences in nomenclature and expression exist between species, and the reader is referred to Table 1.1 for an overview of the naming conventions that will be used in this review. The ficolins in particular require disambiguation. For all species discussed in this review and subsequent chapters, the ficolin produced by leukocytes (commonly known as ficolin-M in humans) will be designated FCN-1. The ficolin that is produced by the liver (commonly referred to as ficolin-L in humans) will be designated FCN-2. Hakata antigen, also known as ficolin-H, will be referred to as FCN-3.

In order to avoid confusion while discussing the various collagenous lectins from different species, where necessary, a single letter will be introduced in front of the abbreviated protein name to designate the species of origin. These will be “h” for human, “e” for equine, “b” for bovine, “p” for porcine, and “m” for murine. For example, eSP-D refers to SP-D in horses, while bMBL-A refers to bovine MBL-A.

Table 1.1. Gene symbols, common protein names, and list of abbreviations for known mammalian collagenous lectins.

Gene Symbol	**Protein Name**	**Protein Abbreviation**
*MBL1**	Mannose-binding lectin A	MBL-A
MBL2	Mannose-binding lectin C	MBL-C
SFTPA	Surfactant protein A	SP-A
SFTPD	Surfactant protein D	SP-D
CGN1 (*COLEC8*)	Conglutinin	CGN
CL43 (*COLEC9*)	Collectin 43 (Bovidae only)	CL-43
COLEC10	Collectin 10 (CL-L1)	CL-10
COLEC11	Collectin 11 (CL-K1)	CL-11
COLEC12	Collectin 12	CL-12
CL46	Collectin 46 (Bovidae only)	CL-46
FCN1	Ficolin-1 (ficolin-M [human], FCN-2 [cow], FCN-β [pig], FCN-B [rodent])	FCN-1
FCN2	Ficolin-2 (ficolin-L [human], FCN-α [pig], FCN-A [rodent])	FCN-2
FCN3[§]	Ficolin-3 (ficolin-H [human])	FCN-3

* pseudogene in humans and chimpanzees

[§] pseudogene in rodents

With the exception of FCN-1 (Fujita, 2002), CL-10, and CL-12 (Ohtani et al., 2012), collagenous lectins are found as circulating proteins. The MBLs, CL-10, CL-11, CGN, CL-43, CL-46, and FCN-2 are synthesized predominantly in the liver or gallbladder, the surfactant proteins and FCN-3 are produced primarily in the lung, and FCN-1 is produced mainly by leukocytes (Holmskov et al., 2003; Lillie et al., 2005; van de Wetering et al., 2004). CL-12 was classically considered a transmembrane protein, found primarily on the vascular endothelium, of numerous organs (Ohtani et al., 2001), but recently a soluble, circulating isoform was characterized (Ma et al., 2015). The functions of CL-10, a cytosolic protein produced in the liver and placenta, are as yet unclear (Ohtani et al., 2012, 1999).

1.2.1 Structure

The basic structural unit of all collagenous lectins is a trimer composed of three identical polypeptides (Figure 1.1a). Depending on the lectin, these trimers are then further assembled into a variety of higher order multimers (Figure 1.1b). Each polypeptide is characterized by four distinct domains: an N-terminal domain, a collagen-like domain (CLD), a neck region, and a carbohydrate recognition domain (CRD) or fibrinogen-like binding domain (FBG). The N-terminal domain varies between 7 and 28 amino acids but invariably contains 1 – 3 critical cysteine residues (Brown-Augsburger et al., 1996; Drickamer et al., 1986; McCormack et al., 1999; Ohashi and Erickson, 2004). The cysteine residues on adjacent polypeptides form disulfide bridges that stabilize both the trimers and higher order multimers (Drickamer et al., 1986; Lillie et al., 2005; Ohashi and Erickson, 2004) .

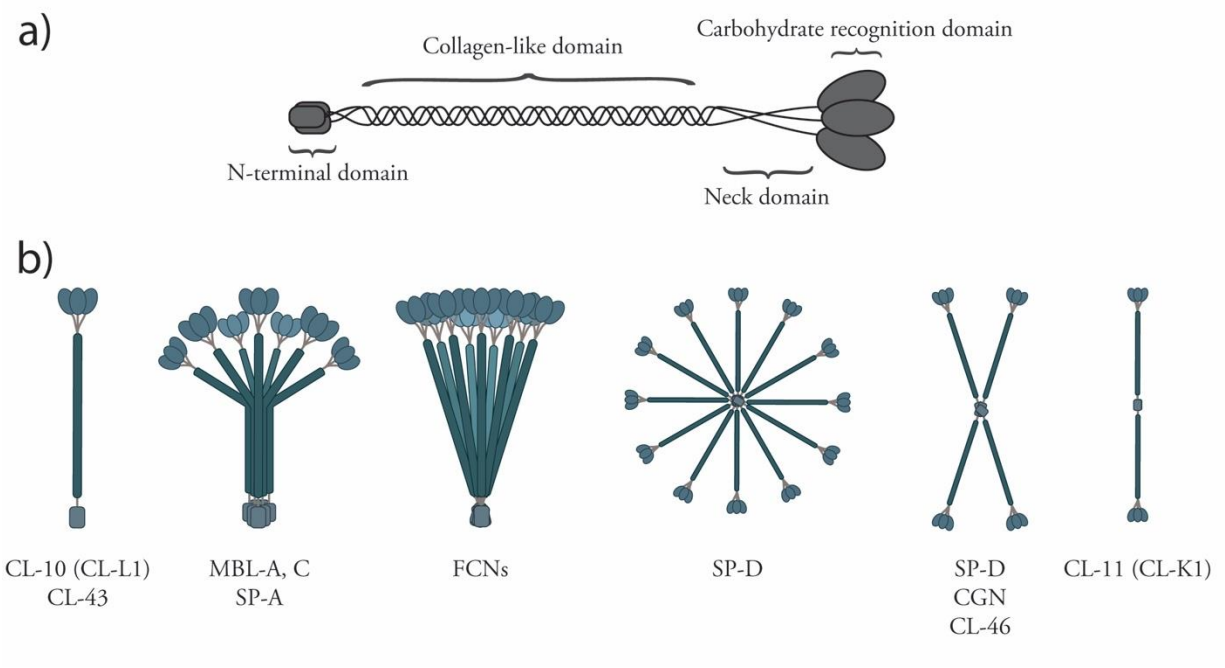


Figure 1.1. Basic trimeric and higher order structures of collagenous lectins.

a) Each monomer of a collagenous lectin trimer is composed a cysteine-rich N-terminal domain, followed by a sequence of repeating Gly-X-Y repeats forming the collagen-like domain. Near the middle of the CLD of MBL and SP-A is a short interruption in the repeating sequence, leading to a bend in the stem (not shown here). The CLD is followed by a short neck region and then the carbohydrate recognition domain, the site of ligand binding. The C-terminal domain of ficolins is structurally different and not depicted here.
b) The multimeric structure of collagenous lectins is variable. CL-43 and CL-10 are secreted as basic trimers, while the MBLs, FCNs, and SP-A form 'bundles of tulips' (sertiform) arrays of varying numbers of trimers. SP-D, CGN, and CL-43 form cruciform structures. CL-12 (CL-P1) is a membrane bound receptor and is not depicted here. Figure adapted from (Lillie et al., 2005; Selman and Hansen, 2012).

The N-terminal domain is followed by the CLD, which is composed of repeating Gly-X-Y amino acid triplets, where X and Y represent any amino acid, though proline, hydroxyproline, and hydroxylysine are overrepresented (Colley and Baenziger, 1987; Crouch et al., 1994; Davis and Lachmann, 1984; Drickamer, 1988). The number of triplets varies substantially between collagenous lectins, ranging from 19 in MBL-C to 59 in SP-D (Holmskov et al., 2003). When in trimeric form, the CLDs of three polypeptides are arranged in a triple helical structure typical of collagens. In addition to its structural features, the CLD of some collagenous lectins contains important functional residues. Near the middle of the CLD of the MBLs and SP-A, there is a short interruption in the Gly-X-Y repeats that creates a bend in the triple helical stem (Benson et al., 1985; Drickamer et al., 1986) that is thought to allow the allow trimers to angle away from the core, facilitating the formation of sertiform ("bundle of tulips") structures (Holmskov et al., 1994; van de Wetering et al., 2004). A relatively well conserved six amino acid sequence, GE(K/R/Q)GEP (residues 37-42 of hMBL-C), also has functional significance. This sequence, found in MBLs, FCNs, and SP-A, is also found in the collagenous domain of C1q, a complement protein. Interactions between this sequence and the C1q receptor, C1qRp, enhance phagocytosis by host monocytes (Arora et al., 2001).

The neck domain is a short sequence of four heptad repeats that link the collagen-like domain to the carbohydrate recognition domain (Zhang et al., 2001). The neck domain is considered essential for intracytoplasmic trimer formation. Initially, it was shown that recombinant proteins consisting of only the neck and carbohydrate recognition domains were able to form trimers, whereas proteins consisting of only the CRD were secreted as monomers (Wang et al., 1995). Subsequently, it was demonstrated that deletion of the amino-terminal heptad of SP-D resulted in complete abrogation of trimer formation (Zhang et al., 2001). It is

thought that the neck region assembles trimers by aligning the collagen-like chains of the CLD, thereby facilitating helix formation (Zhang et al., 2001).

The collectin and ficolin subgroups are primarily differentiated by their ligand binding domain, located at the C-terminus. The collectins have a CRD, a large globular structure responsible for the recognition of and binding to polysaccharide PAMPs. Within the CRD are fourteen invariant and eighteen well conserved amino acid residues (Drickamer, 1988). Four of the conserved residues are cysteines, which aid in the formation of the tertiary structure of the CRD via disulfide bridging. One pair joins the N- and C-terminus of the domain, while the other creates a long loop that contains a carbohydrate and calcium binding groove (van de Wetering et al., 2004; Veldhuizen et al., 2011). Within this groove are several invariant residues, including Glu-Pro-Asn (residues 185 – 187 of hMBL-C) and Trp-Asn-Asp (204 – 206), whose carbonyl groups directly bind Ca^{2+} (Veldhuizen et al., 2011; Weis et al., 1992). The Glu-Pro-Asn motif is the critical determinant of carbohydrate specificity (Drickamer, 1992). The position of the carbonyls of the Glu-Pro-Asn residues favours the binding of monosaccharides containing equatorial hydroxyl groups at the C3 and C4 position, such as mannose or glucose (Veldhuizen et al., 2011). This sequence is present in the collagenous lectins that bind mannose in all species, though at different absolute locations (Veldhuizen et al., 2011).

For the ficolins, the C-terminal globular structure analogous to the CRD of collectins is composed of a fibrinogen-like sequence of 200 – 250 amino acid residues, and is referred to as the fibrinogen-like binding domain (FBG) (Holmskov et al., 2003). Within the FBG are 24 invariant hydrophobic residues, the functional significance of which have not been fully elucidated (Garred et al., 2010). The ultrastructural studies necessary to elucidate the ligand binding sites of ficolins have thus far been restricted to recombinant human proteins (Garlatti et

al., 2010). FCN-1 and FCN-3 share a single conserved ligand binding site found on the external part of the trimer (Garlatti et al., 2007). Significant differences in the amino acid composition at the ligand binding site account for the variation in binding preferences between the two proteins (Garlatti et al., 2010). In contrast, FCN-2 has three external binding sites of significantly different composition, reflecting its broader range of ligand recognition (Garlatti et al., 2010). Ficolins bind to various N-acetylated saccharides by targeting the N-acetyl group itself (Brooks et al., 2003a; Krarup et al., 2004).

The collectins and ficolins differ in their higher order structures. Although the ficolins form a sertiform structure similar to the MBLs and SP-A (Figure 1.1b), the lack of an interruption in the CLD results in a more linear arrangement of the trimeric subunits. Formation of the higher order structures of ficolins is facilitated by N-terminal cysteine residues (Ohashi and Erickson, 2004).

With the exception of CL-43 and CL-10, all collagenous lectins form higher order structures composed of multiple trimers (Figure 1.1b) (Holmskov et al., 1994). Multimers are formed through disulfide bridges linking cysteine residues in the N-terminal domain of adjacent trimers (Drickamer and Taylor, 1993). The formation of multimers reflects a "power in numbers" approach to ligand binding. By itself, the CRD of a single lectin monomer binds weakly to its oligosaccharide ligand. When multiple receptors are engaged, as is the case with multimeric structures, binding avidity exponentially increases (Lee et al., 1992, 1991). Thus, strong binding of collagenous lectins to pathogens requires both high numbers of CRDs as well as numerous saccharide residues on the pathogen.

The structure of CL-12 is unique amongst the collagenous lectins. Unlike most other collagenous lectins, which are found as circulating proteins, CL-12 is primarily found as a

transmembrane protein akin to the type A scavenger receptors (Ohtani et al., 2001). Between the N-terminal domain and the CLD there is a large transmembrane domain and a coiled-coil domain. More recently, however, soluble, circulating CL-12 with the ability to activate the alternative pathway of complement was discovered in human umbilical cord plasma (Ma et al., 2015). Soluble CL-12 was able to recognize and respond to *Aspergillus fumigatus* by activating complement, solidifying its role as a classical collagenous lectin (Ma et al., 2015). The ratio of the soluble and membrane-bound form of CL-12 and their relative contributions have not been reported.

1.2.2 Evolution/Phylogeny

Collagenous lectins are ancient and highly conserved proteins, present in a range of species including sea urchins, frogs, fish, birds, and mammals (Fujita, 2002; Lillie et al., 2005; Ohtani et al., 2012). Though these proteins are shared between invertebrates and vertebrates, the number of collagenous lectin proteins in different species varies. Through phylogenetic analysis of the CRDs of collagenous lectins, it has been proposed that CL-10, CL-11, the MBLs, SP-A, and SP-D evolved from a common ancestral collagenous lectin prior to the divergence of mammals and birds (Hughes, 2007; Selman and Hansen, 2012). The significant variability in the repertoire of collagenous lectins within mammals is the result of species-specific gene duplications and deletions (Hughes, 2007). All mammalian species have two MBL genes (*MBL1* and *MBL2*), though in humans and chimpanzees, a nonsense mutation has resulted *MBL1* becoming a pseudogene (Guo et al., 1998).

The evolution of the ficolins is more complex. Phylogenetic analysis suggests that *FCN3* is the most ancient ficolin gene, and that *FCN1* branched out by gene duplication early in the course of vertebrate evolution, resulting in orthologues in many mammalian species (Garred et

al., 2010). In rodents, *FCN3* has been identified as a pseudogene (Endo et al., 2004). Not all mammalian species possess an *FCN2* gene. Phylogenetic studies suggest that *FCN2* is the result of a duplication of the *FCN1* gene that has occurred independently in multiple different mammalian families (Garred et al., 2010). Furthermore, within each branch of Garred et al.'s (2010) proposed phylogenetic tree, multiple lineage-specific duplication events have occurred, resulting in the expansion of the ficolins to much greater numbers in some lineages. For example, the tropical clawed frog, *Xenopus tropicalis*, possesses over 2 dozen ficolin proteins (Doolittle et al., 2012) possibly a reflection of a greater reliance on a PRR-type immune defence in this species.

Bovidae express three evolutionarily-related collagenous lectins: CL-43, CL-46, and conglutinin (Hansen et al., 2002). Structural similarities and close proximity within a chromosome between these collagenous lectins and SP-D suggests a common ancestry. An early duplication event of the *SFTPD* gene likely led to a common *CGN*/*CL46* gene. After the divergence of *CGN* and *CL46*, *CL43* then likely evolved from *CGN* (Gjerstorff et al., 2004a; Hansen et al., 2002). It is intriguing that these SP-D-like collagenous lectins evolved after the divergence of bovids from other mammals. One suggested explanation for this is that the large load of bacteria and protozoa required for rumination presented an additional need for immunological control (Gjerstorff et al., 2004a; Hansen et al., 2002). In humans, *SFTPD* is expressed and potentially serves a functional role in the stomach and intestine (Gowdy et al., 2012; Lin et al., 2011; Murray et al., 2002), lending weight to the hypothesis that these *SFTPD*-like collagenous lectins may do the same in cattle.

A repeating theme within the evolution of the collagenous lectins is species-dependant gene duplication or gene loss. Examples include the loss of *MBL1* in primates and *FCN3* in

rodents, independent duplication of *FCN1* in multiple species, and the expansion of the *SFTPD* lineage into three descendant genes in Bovidae. The evolution of seemingly redundant genes may be a reflection of the unique infectious challenges posed to each species and the subsequent defensive adaptation. One group has hypothesized that pressure from viruses has selected for gains or losses of various collagenous lectins (Drickamer and Taylor, 2015), but this remains to be proven. Gene redundancies are also frequently thought to provide a fitness advantage to an organism, serving as a 'backup' should an inopportune mutation deactivate one copy (Kafri et al., 2006), and such an advantage would be beneficial in the context of immunity.

1.2.3 Function

Collagenous lectins participate in host defense in various different ways, including agglutination, opsonisation, and complement activation (van de Wetering et al., 2004). They have been shown to bind a wide variety of bacteria, viruses, protozoa, and fungi (Clark et al., 2000; Keirstead et al., 2011; Lillie et al., 2005; Lu et al., 2002). Sertiform and cruciform lectins can effectively bind multiple pathogens, creating large aggregates of microbes and lectins. Aggregation of microbes in turn inhibits their attachment to host cells, preventing invasion, and in the case of some pulmonary pathogens, facilitates mucociliary clearance by the respiratory tract (Ferguson et al., 1999; Hartshorn et al., 1997). The MBLs and SPs have been shown to directly opsonize bacteria, resulting in enhanced phagocytic clearance (Kuhlman et al., 1989; Pikaar et al., 1995; van Iwaarden et al., 1991), and SP-A and SP-D were found to have direct inhibitory effects on bacterial growth (van de Wetering et al., 2004). Some collagenous lectins, particularly the MBLs and FCNs, have important roles in the activation of the complement pathway.

The collagenous lectins have a variety of other immunomodulatory effects (Table 1.2). In rodents and humans, MBL and SP-A have been shown to have anti-inflammatory effects, impacting the response to LPS at several different points in the LPS-CD14-TLR4 pathway (Borron et al., 2000; Nguyen et al., 2012; Sano et al., 2000; Van Rozendaal et al., 1999; M. Wang et al., 2011). The ficolin proteins, as determined through knock-out models, have a pro-inflammatory affect (Luo et al., 2013). The surfactant proteins of rats have the ability to directly increase the cell membrane permeability of *E. coli* (Wu et al., 2003).

Table 1.2: Summary of additional immunomodulatory effects of several collagenous lectins.

Species	**Collagenous lectin**	***In vivo/ vitro***	**Effect**	**Pathogen**	**Reference**
Rat	SP-A, SP-D	*In vitro*	Increases cell membrane permeability	*E. coli* K12	Wu et al., 2003
Rat	SP-D	*In vivo*	Binds (scavenges) LPS, likely CRD dependent	LPS from *Salmonella* Enteriditis	van Rozendaal et al,, 1999
Mouse	SP-A	*In vivo*	Inhibits LPS-induced TNF-alpha and NO_2 production	LPS from *E. coli*	Borron et al,, 2000
Rat	SP-A, SP-D	*In vitro*	Binds to CD14 (co-receptor along with TLR4 for LPS)	LPS from *Salmonella* Minnesota	Sano et al,, 2000
Rat	SP-A	*In vitro*	Dampens PGN-induced inflammatory response in a TLR2-binding dependent manner.	PGN from *S. aureus*	Murakami et al,, 2001
Human, rat	SP-A (human), SP-D (rat)	*In vitro*	Facilitates clearance of apoptotic neutrophils in the lung; non-CRD dependent for SP-A		Schagat et al., 2001
Human	MBL-C	*In vitro*	Suppression of LPS-induced TNF-alpha and IL-12 in a TLR4 dependent manner	LPS from *E. coli*	Wang et al., 2011
Human	SP-A	*In vitro*	Upregulates IRAK-M, a negative regulator of TNF-alpha and IL-6	Source of LPS not specified	Nguyen et al., 2012
Mouse	*MBL1*, *MBL2*	*In vivo*	KO mice have increased TNF-alpha and IL-12	Cecal ligation and puncture	Takahashi et al., 2002
Mouse	*FCN2*	*In vivo*	KO mice have decreased TNF-alpha and IL-17A in response to LPS	LPS from *E. coli*	Luo et al., 2013
Human	SP-D	*In vitro*	Trimers bind to LPS; higher order binds to mannan residues	LPS from *E. coli*	Sorensen et al., 2009

PGN: peptidoglycan; TNF: tumour-necrosis factor; IL: interleukin CRD: carbohydrate recognition domain , KO: knock out, LPS: lipopolysaccharide.

1.2.4 The Lectin Pathway of Complement Activation

A major strategy used by some collagenous lectins in host defense is the activation of complement, an essential component of the immune system. First discovered in the 1890s, the complement system is now known to be comprised of over thirty plasma and membrane bound proteins (Fujita, 2002), which are sequentially activated in cascades of enzymatic activity. Activation of complement can lead to the opsonization of pathogens via the deposition of complement protein C3b, or to the lysis of pathogens by the formation of a membrane attack complex (MAC) (Walport, 2001a). By-products such as C5a are chemoattractants that augment leukocyte responses. Complement can be activated by three different pathways: the classical, the lectin, and the alternative pathways. Each of these activating pathways then leads to a common lytic pathway, the endpoint of complement activation.

Although attempts were made to assign logical names to the proteins of the complement system, the biological inevitability of non-sequential discovery has led to an at times confusing nomenclature. The proteins of the classical pathway are tantalizingly designated C1 to C9, suggesting a potentially intuitive order of activation. Unfortunately, this is not the case. Activation of C1 is followed by C4, C2, C3 and then C5. The activation of C5, the first step of the lytic pathway, signals a return to reason in which proteins are now activated in order from C6 to C9, leading to the formation of the membrane attack complex (Walport, 2001a, 2001b). Activation of the complement proteins is frequently the result of enzymatic cleavage, resulting in two protein fragments and further complications in the nomenclature. In general, the larger fragment is designated ‘b’ (e.g. C3b) and the smaller ‘a’ (e.g. C3a). The requisite exception to the rule is protein C2, whose large and small fragments are named C2a and C2b, respectively (Walport, 2001a).

The classical pathway begins with the recognition of the Fc fragment of antibody bound to antigen by C1q complexed with C1r and C1s (Figure 1.2). The latter two proteins are serine proteases that are activated upon binding of C1q to antibody. C1s first cleaves C4, which binds to the microbial membrane, and then C2, leading to the formation of the C3 convertase, C4b2a. Formation of the C3 convertase is the first step of the lytic pathway, in which C3 is cleaved to form C3b which associates with C4b2a to form the C5 convertase, C4b2a3b. The resulting cleavage of C5 to C5b by the C5 convertase leads to the sequential activation of complement proteins 6 – 9 and the formation of the membrane attack complex (Walport, 2001a, 2001b).

The lectin pathway is similar to the classical pathway (Figure 1.2). Of critical importance to the lectin pathway of activation are four MBL-associated serine proteases (MASPs), MASP-1-3 and sMAP. MASPs are structurally similar to C1r and C1s of the classical pathway (Matsushita et al., 2013; Thiel et al., 1997) and can bind in proenzyme form to the collagen-like domain of certain collagenous lectins. Of the thirteen described mammalian collagenous lectins, only the MBLs, the FCNs, and CL-11 are known to bind MASPs and activate complement, though theoretically the circulating heterocomplex of CL-10 and CL-11 may also play a role (Garred et al., 2016; Gibbs et al., 2009; Hansen et al., 2010; Matsushita et al., 2001; Matsushita and Fujita, 1992). Upon binding of a collagenous lectin to its oligosaccharide ligand, MASP proenzymes are cleaved, thereby activating them as serine proteases. Initially, it was believed that MASP-2 was autoactivated to cleave C4 and C2, resulting in the formation of the C3 convertase C4b2a (Chen and Wallis, 2004; Vorup-Jensen et al., 2000). Recently, however, it was demonstrated that *in vitro* complement activity was abolished in a human patient with a nonsense mutation of MASP-1, but restored upon supplementation with recombinant MASP-1, suggesting an absolute requirement for this protein (Degn et al., 2012). The same study also demonstrated that MASP-1

and MASP-2 associate on the same MBL oligomer, and theorized that MASP-1 was required for the activation of MASP-2, analogous to C1r and C1s of the classical pathway (Degn et al., 2013, 2012). Although it was previously suggested that MASP-1 was capable of cleaving C3 and C2 independently of MASP-2 (Matsushita and Fujita, 1995), it is now thought that MASP-1 is only capable of activating C2 (Degn et al., 2012). The roles of MASP-3 and sMAP are less clear. Both compete with MASP-2 for binding to MBL, and it is thought that they may thereby serve as negative regulatory proteins (Degn et al., 2013; Iwaki et al., 2006; Matsushita et al., 2013).

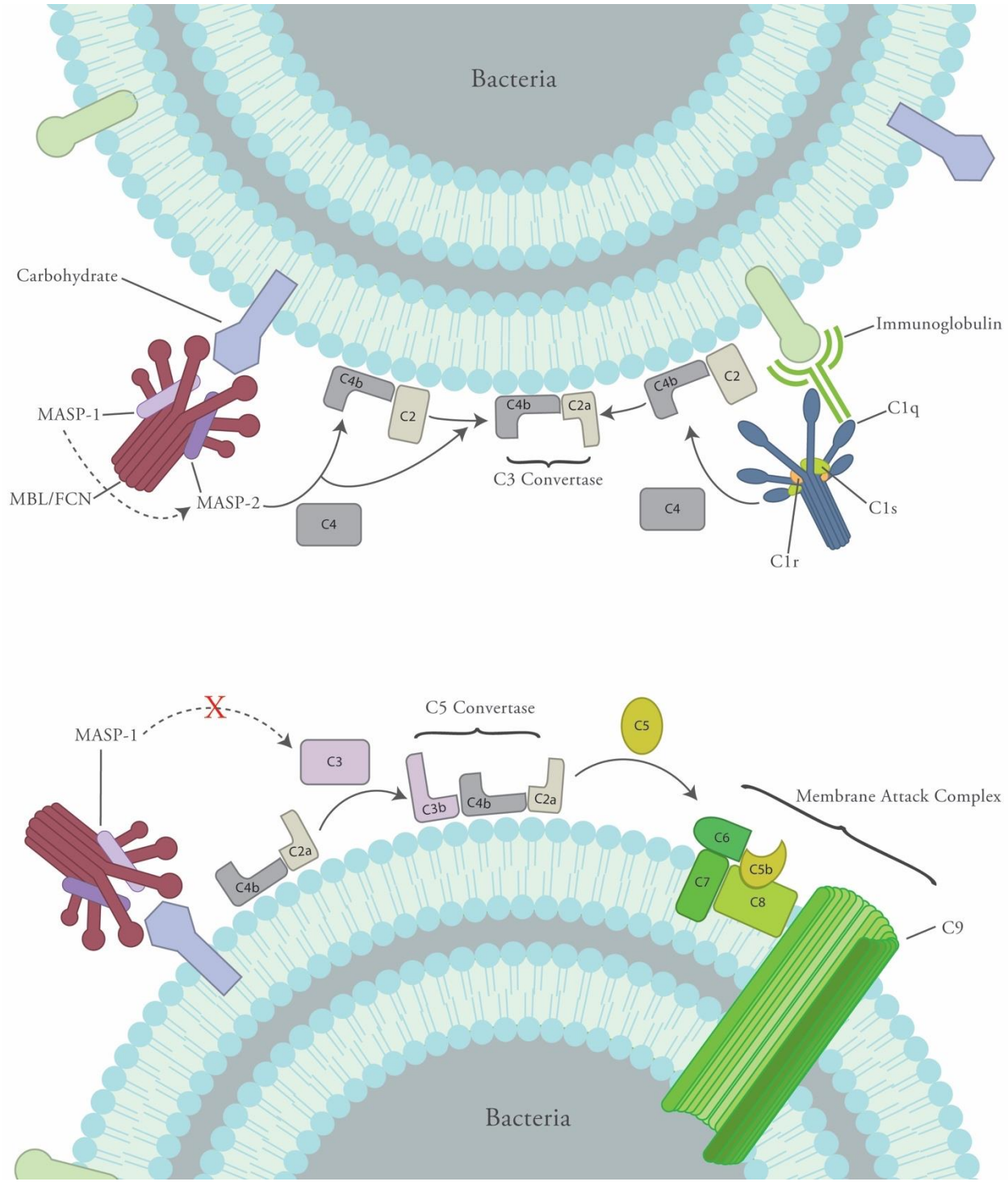


Figure 1.2. The classical and lectin pathways of complement activation.

The classical pathway is characterized by a series of cascading enzymatic protein cleavages, initiated by the binding of C1q to antibody and ending with the formation of the C3 convertase C4b2a. The lectin pathway begins with the binding of a lectin to a sugar residue, leading to the activation of MASP-2 by MASP-1. MASP-2 cleaves C4 and C2, resulting in the formation of the C3 convertase. The C3 convertase initiates the lytic pathway, and results in the formation of the membrane attack complex. Dashed lines indicate hypothesized roles; "x" indicates a contested role. Figure adapted from (Walport, 2001a).

1.2.5 Genetics and the Role of the Collagenous Lectins in Infectious Disease

Genetic polymorphisms affecting the expression, structure and function of collagenous lectins are well documented. The first indication of genetically induced collagenous lectin dysfunction came from children with an opsonic defect and recurring infections (Super et al., 1989). It was demonstrated that these children were heterozygous for a missense mutation (G230A) in the coding region of the collagen-like domain of hMBL-C, resulting in defective trimer assembly (Sumiya et al., 1991). It was later shown that unusual and recurrent infections in adults were associated with homozygous carriage of the same defective *hMBL2* allele (Summerfield et al., 1995). Subsequently, two additional single nucleotide polymorphisms (SNPs) (C223T and G239A) were described in the coding region of *hMBL2*, resulting in four *hMBL2* alleles, designated A, B, C, and D (allele A is the wild-type allele) (Lipscombe et al., 1995; Madsen et al., 1994; Steffensen et al., 2000). The SNPs in alleles B-D result in the destabilization of the basic trimers and inhibit the formation of higher order multimers. In heterozygote carriers, this has a dominant effect, as half the available hMBL-C cannot assemble normally (Sumiya et al., 1991). In addition to the coding region SNPs, two polymorphisms (G550C and G221C) in the promoter region and one polymorphism (C4T) in the 5′-untranslated region of *hMBL2* have been well characterized (Eisen and Minchinton, 2003; Madsen et al., 1998; Minchinton et al., 2002), and are referred to as H/L, X/Y, and P/Q, respectively. These alleles are in complete linkage disequilibrium with the coding region alleles, and only seven of the sixty-four possible haplotypes have been recognized (Munthe-Fog et al., 2008). Genetically determined deficiencies of hMBL-C have been associated with various infectious diseases including HIV, hepatitis B and C, meningococcal disease, and parasitic infections (Eisen and Minchinton, 2003).

Genetic variation of the ficolins has been investigated in the Danish population. Several polymorphisms in *FCN1* were discovered, however, all were sense mutations and none were associated with altered FCN-1 levels. In contrast, numerous polymorphisms were demonstrated in *FCN2*, some of which decreased expression, and one of which affected oligosaccharide binding (Hummelshoj et al., 2005). The relevance of these SNPs in disease susceptibility has not been determined. A frameshift mutation in *FCN3* resulting in decreased FCN-3 expression has also been described (Munthe-Fog et al., 2008), however, it has only been associated with disease in one homozygous patient (Munthe-Fog et al., 2009).

Mutations in *hCOLEC11* and *hMASP1* are the cause of Carnevale, Mingarelli, Malpuech and Michels syndromes, known as 3MC, though the causative mutations appear not to affect the lectin pathway of complement (Rooryck et al., 2011). Several promoter and coding region mutations in *hCOLEC10* and *hCOLEC11* are associated with serum levels of CL-L1 and CL-K1, however, they have not been linked to the increased evidence of infectious disease (Bayarri-Olmos et al., 2015).

Among the surfactant proteins, a missense mutation in *hSFTPD* correlates to lower circulating levels of SP-D and leads to the absence of higher-order oligomerization, resulting in decreased activity against influenza A virus (IAV) and various bacteria (Leth-Larsen et al., 2005). It is also associated with interstitial pneumonia (Ishii et al., 2012). Certain *SFTPA* haplotypes are associated with increased susceptibility to IAV, community-acquired pneumonia and tuberculosis (García-Laorden et al., 2011; Herrera-Ramos et al., 2014; J. Li et al., 2016).

Polymorphisms in the complement activating *MASP* genes are associated with decreased circulating protein levels, and consequently decreased complement activity (Beltrame et al., 2015). Accordingly, haplotypes associated with decreased circulating levels of the various

MASP proteins are more frequent in some infectious diseases, including leprosy (Boldt et al., 2013) and Chagas disease (Boldt et al., 2011).

In animal species of veterinary interest, the collagenous lectins have been most thoroughly investigated in pigs, both in terms of their functional capabilities and genetics. Comparatively little work has been conducted in cattle or horses on the relative role of the collagenous lectins in the innate immune system, or on how the genetics of these proteins might impact susceptibility to disease. The following sections review the existing literature and present a brief overview of some of the common infectious diseases in cattle and horses, and discusses diseases where collagenous lectins have potential roles.

1.2.6 Pigs

Ficolins were first identified as TGF-β-binding proteins in the porcine uterus (Ichijo et al., 1993). FCN-α in porcine plasma was found to bind to *Actinobacillus pleuropneumonia* (APP) (Brooks et al., 2003b) in an N-acetyl-dependent manner (Brooks et al., 2003a), and to several gram-negative and gram-positive organisms, and may bind to lipopolysaccharide (LPS) alone (Nahid and Sugii, 2006). Plasma and recombinant porcine FCN-α (Keirstead et al., 2008) and recombinant SP-A (L. Li et al., 2016) were shown *in vitro* to bind and neutralize porcine reproductive and respiratory syndrome virus (PRRSV), an economically important pathogen of pigs. FCN-β is expressed by porcine neutrophils but does not bind APP, and its innate immune functions remain unclear (Brooks et al., 2003c). Porcine plasma MBL-A binds APP, *Actinobacillus suis*, and *Haemophilus parasuis* (Lillie et al., 2006a), and the porcine pulmonary collagenous lectins, SP-A and SP-D, are, when compared to their homologues in other species, especially capable of inhibiting hemagglutination of IAV, facilitated by a unique asparagine-

linked oligosaccharide in their CRDs (van Eijk et al., 2012, 2003). The binding affinities of the remaining porcine collagenous lectins have not been directly examined.

After showing that porcine collagenous lectins could bind various pathogens, and in combination with observation that some pigs succumbed to infectious disease more readily than others, the role of the genetics of the collagenous lectins was explored. Three SNPs were identified in *MBL1*, one of which (G271T) is predicted to disrupt trimer formation (Lillie et al., 2006b). It was subsequently shown that the presence of this SNP in a population of pigs is associated with decreased serum concentrations of MBL-A (Juul-Madsen et al., 2011). The promoter region of *MBL2* contains several point mutations that are associated with decreased *MBL2* expression, with one mutation (G1081A) having a marked association with expression levels (Lillie et al., 2007). In a large population of normal and diseased pigs subsequently screened for the presence of these alleles, pigs with pneumonia, enteritis, serositis, or septicemia were almost twice as likely to carry the allele as were normal pigs (Lillie et al., 2007). Finally, recent work in Rongchang pigs, an indigenous Chinese breed, showed that a 9 bp deletion in exon 2 of *SFTPA1* was associated with pulmonary edema in a model of acute respiratory distress syndrome (Zhang et al., 2015).

1.2.7 Cattle

Some of the most detailed work on bovine collagenous lectins revolves around their involvement in an infectious disease that is of little concern to the species: influenza A virus (IAV). The bovine specific CL-43, CL-46, and CGN have well characterized anti-viral activity against IAV, a virus which does not appear to successfully infect cattle. Conglutinin aggregates and opsonizes IAV, and inhibits viral infectivity (Hartshorn et al., 1993), while CL-43 strongly enhances neutrophil uptake of IAV (Hartshorn et al., 2002). CL-46 neutralizes virus infectivity,

inhibits viral induced hemagglutination, aggregates IAV, and increases neutrophil uptake of IAV (Hartshorn et al., 2010a, 2010b). In contrast, in an experimental model of an important bovine pathogen, *Mycobacterium bovis*, type II pneumocytes had increased immunolabeling for SP-A as compared to controls (Beytut, 2011), demonstrating a potential response to the pathogen, but the contribution of SP-A to the overall immunopathogenesis is still unknown.

There is limited information regarding the genetics of the bovine collagenous lectins. The plasma concentration of conglutinin is variable and heritable, and low levels are associated with a predisposition to infection (Holmskov et al., 1998). Proteomic analysis of endometrial tissue in cows with and without endometritis identified increased levels of CL-43, however, expression of *CL43* mRNA was not detected, suggesting the increase may have been from circulating CL-43 expressed elsewhere (Choe et al., 2010). Whole transcriptome analysis of cattle with a parasite-resistant phenotype revealed higher levels of expression of *CL46* when compared to parasite-susceptible cattle, but again, the underlying mechanism(s) leading to increased expression are still unknown (Li et al., 2011). Theoretically, in depth analysis may identify regulatory or promoter mutations (similar to those found in *MBL2* of pigs) associated with *CL46* expression. Several studies have implicated the polymorphisms in collagenous lectins with mastitis in Chinese Holstein cattle, and these are reviewed further in the discussion on bovine mastitis.

There are several significant infectious diseases in which the bovine collagenous lectins might play an important immunological role. Bovine respiratory disease (BRD) is a significant cause of morbidity, mortality, and economic loss within the cattle industry (Patrick, 2009; Taylor et al., 2010). In the U.S.A. alone, BRD is estimated to cost approximately $500 million USD per year (Miles, 2009). The current approaches to the control of BRD have not been entirely effective, and have focused primarily on pathogen control rather than host immunity (Miles,

2009). Novel approaches targeting host immunity may present an effective alternative means of controlling BRD. The role of collagenous lectins in pulmonary disease is especially intriguing in cattle, as they uniquely possess three evolutionary descendants of SP-D (conglutinin, CL-43, and CL-46) which have been implicated in the defense against primary pulmonary pathogens such as influenza A virus (Hartshorn et al., 2010b, 2002, 1993).

Bovine respiratory disease frequently involves multiple factors, including immunodeficiency, stress, and infectious agents. The most common agents isolated from cases of BRD are *Mannheimia haemolytica, Pasteurella multocida*, *Histophilus somni*, *Mycobacterium bovis*, *Trueperella pyogenes*, *Mycoplasma bovis*, bovine viral diarrhea virus (BVDV), parainfluenza-3, bovine respiratory syncytial virus (BRSV), and bovine herpesvirus 1 (infectious bovine rhinotracheitis) (Maxie, 2015; Panciera and Confer, 2010; Taylor et al., 2010).

M. haemolytica, *P. multocida*, and *H. somni* are gram-negative bacteria of the family Pasteurellaceae. Although no studies have specifically investigated the interactions between collagenous lectins and these bacteria in cattle, several studies have demonstrated the effects of lectins on gram-negative bacteria of the Pasteurellaceae family in animals (Brooks et al., 2003a, 2003b, Lillie et al., 2006a, 2005). Bovine parainfluenza 3 virus (bPI3) and bovine respiratory syncytial virus (BRSV) are both members of the Paramyxoviridae family. On their own, these pathogens generally cause only mild pneumonia in young calves, however, they are frequently complicated by secondary bacterial infections (Ellis, 2010; Gershwin, 2012). Both viruses are enveloped and encode multiple membrane bound glycoproteins critical to viral attachment and cellular infection (Ellis, 2010; Gershwin, 2012). Investigations focusing on the innate immune response in cattle to these pathogens are sparse, though BRSV has been shown to activate TLR-4 *in vitro* (Lizundia et al., 2008). Supporting evidence from studies on human respiratory syncytial

virus (HRSV), however, have shown that hSP-A and hSP-D bind viral envelope proteins and significantly reduce viral load in a murine model (Ghildyal et al., 1999; Hickling et al., 1999). An experimental model of RSV infection in sheep used BRSV to infect lambs, and noted marked increases in both *SFTPA* and *SFTPD* mRNA expression in airway epithelial cells following challenge (Kawashima et al., 2006).

Bovine herpesvirus 1 (bHV1) is a large, double-stranded DNA α-herpesvirus of the *Herpesviridae* family. It is the causative agent of infectious bovine rhinotracheitis and can also cause immunosuppression leading to secondary bacterial bronchopneumonia (Levings and Roth, 2013). The virus encodes 12 membrane bound proteins, 10 of which are glycosylated (Levings and Roth, 2013). The proteins are variably N- and O- glycosylated, and share a high degree of structural and functional conservation amongst herpesviruses (Turin et al., 1999), but the glycan structures might differ among species. The role of lectins in the defense against bHV1 is unknown, though MBL deficiencies are more common in human patients with herpes simplex virus-associated meningitis (Eisen and Minchinton, 2003). Whether a similar association exists in cattle with bHV1 is as yet unknown.

Bovine viral diarrhea virus (BVDV) is arguably one the most important pathogens in cattle. Although not a cause of primary respiratory disease, the virus infects and functionally impairs leukocytes. The subsequent immunodepression is a frequent cause of secondary bacterial pneumonia (Panciera and Confer, 2010). BVDV is an enveloped virus (family *Flaviviridae*, genus *Pestivirus*) with a positive sense, single-stranded RNA genome encoding up to fourteen proteins (Neill, 2013). Two of these proteins, E1 and E2, are membrane bound glycoproteins accessible to the immune system. Each of these proteins contain N-linked oligosaccharides critical to the replication of the virus (Donis and Dubovi, 1987).

All of the above described bovine viruses contain glycosylated proteins within their envelopes which are exposed to various circulating or cell-associated collagenous lectins. It is thus reasonable to hypothesize that some collagenous lectins might recognize glycans of some of these pathogens. However, further studies are needed to assess the ability of particular lectins to bind and neutralize viruses with envelope glycans generated in bovine cells.

Mastitis is a significant cause of morbidity and economic loss to the dairy cattle industry (Halasa et al., 2007; Heikkilä et al., 2012; Schepers and Dijkhuizen, 1991). Within Canadian dairy herds, the most common causes of mastitis are *Staphylococcus aureus*, *Escherichia coli*, and *Streptococcus uberis* (Olde Riekerink et al., 2008). Pathogens establish infection in the teat cistern, where microbial PAMPS are recognized by PRRs, including TLR-2 and TLR-4 (Goldammer et al., 2004), resulting in the production of IL-1β, TNF-α, and the recruitment of large numbers of neutrophils. The alternative pathway of the complement system is activated in milk and opsonizes bacteria, and recruits and activates neutrophils (Rainard, 2003). Interestingly, the classical complement pathway is virtually non-existent in the bovine mammary gland, where little to no C1q is present. Although there is no specific mention of the lectin pathway of the complement system, it seems unlikely to be active, as an analysis of the proteome of normal and mastitic bovine milk did not find the presence of the requisite molecules, including the MBLs, CL-11, the FCNs, MASP-1 or -2 (lectin pathway) (Reinhardt et al., 2013). However, the same analysis found that CL-12, the collagenous lectin capable of activating the alternate pathway of complement, was present in milk.

A series of studies have highlighted the potential relevance of collagenous lectins in bovine mastitis. A marked decrease in the serum levels of both CGN and MBL-C is associated with mastitis in cattle, and both proteins were shown to recover to levels seen in unaffected

animals upon clearance of infection, suggesting that decreased protein levels were a response to, and not cause of, mastitis (Akiyama et al., 1992). More recent work has described associations between increased somatic cell score (SCS) and two SNPs (c.2534G>A and c.2651G>A) in exon 2 of bovine *MBL1* (*bMBL1*) (C. Wang et al., 2011; Yuan et al., 2012). Another SNP in the promoter region of *bMBL1* (c.1330G>A) was associated with decreased SCS, possibly indicating a marker of mastitis resistance (Liu et al., 2011). The same group described four SNPs in bovine *MBL2*, c.201G>A, c.234C>A, c.1197C>A, and c.1164G>A, which were significantly associated with SCS (Wang et al., 2012; Zhao et al., 2012). The effects of these SNPs on protein expression, structure, and function were not evaluated. A study on *MASP2* similarly found that the missense mutation, c.533G>A, was associated with increased SCS (Wu et al., 2015). However, there is little information directly linking collagenous lectins and host defense against mastitis. Given this information, it is possible that the *bMBL1* and *bMBL2* genes are linked to other genes relevant to the immunological response to mastitis.

1.2.8 Horses

The equine collagenous lectins are perhaps the least characterized of the three species discussed in this thesis. As in cattle and pigs, infectious diseases are a significant concern in horses, and investigation into predisposing factors is a valuable endeavor. The respiratory and gastrointestinal systems of horses are two of the most commonly affected by infectious diseases.

Inflammatory airway disease (IAD) is a significant source of economic loss and morbidity in the thoroughbred racing industry (Wilsher et al., 2006). IAD may be either infectious or non-infectious in origin. Common infectious agents include *Actinobacillus* spp., *Streptococcus equi* var *zooepidemicus*, *Rhodococcus equi.*, equine influenza A virus, and equine herpesvirus 4 (Maxie, 2015). Little research into the role of collagenous lectins and equine

infectious respiratory disease has been performed, though both SP-A and SP-D have been isolated from the bronchoalveolar lavage fluid (BALF) of horses (Hobo et al., 1999a, 1999b). Interestingly, in one study of experimental *S. zooepidemicus* infection, SP-D expression initially rose before tapering off, mirroring the response of the acute phase proteins fibrinogen and serum amyloid A (Hobo et al., 2007), suggesting that SP-D plays a role in the defense against *S. zoopepidemicus,* and potentially a broader role in the innate defense of the airways against infectious agents.

Reactive airway obstruction (RAO), also known as equine asthma (Lavoie, 2015), is a non-infectious cause of respiratory disease in horses. RAO shares a number of similar pathophysiological features with human allergic asthma, making it an attractive target for translational research (Leclere et al., 2011). Studies in rodent models of allergic asthma have demonstrated a potential immunomodulatory role for SP-D (Qaseem et al., 2013). The immunologic hallmark of human allergic asthma is an exuberant T_H2 cell response, characterized by increased levels of IL-13, IL-4, IL-5, IL-6, mast cells and eosinophils (Qaseem et al., 2013). In rodent models of asthma, challenge with an allergen produced an increase in SP-D levels, while administration of exogenous SP-D prior to airway challenge led to an increase in protective type 1 cytokines (IFN-γ, IL-10, IL-12) and decreased airway hyperresponsiveness (Takeda et al., 2003). Eventually, a putative feedback loop was identified between the T_H2 polarizing cytokine IL-13 and SP-D (Haczku et al., 2006). Increased levels of IL-13 lead to an increase in SP-D levels, which in turn negatively feeds back on IL-13, creating a regulatory loop. Decreased IL-13 and increased SP-D shifts T-cell polarization away from an asthmatogenic type 2 response and towards a protective type 1 response (Qaseem et al., 2013). Thus, the rise in SP-D noted after allergic challenge in rodents may be a protective counter-regulatory response.

It is unknown whether SP-D plays a similar role in horses with RAO. In a single study on horses with non-infectious IAD, affected horses had significantly higher SP-D levels than normal controls (Richard et al., 2012). These preliminary results suggest a potential role for SP-D in equine non-infectious respiratory disease. Further work, however, is needed to fully characterize the immunomodulatory effect of SP-D in equine reactive airway obstruction.

Equine enteritis is an important cause of colic, a costly and frequent issue of horses (Egenvall et al., 2008; Feary and Hassel, 2006). Enteritis may be caused by a wide range of pathogens, including various *Salmonella* spp., *Lawsonia intracellularis*, *Clostridium difficile, C. perfringens*, and strongyle parasites (Feary and Hassel, 2006). To date, the role of collagenous lectins in equine enteritis has gone unstudied. Data from studies in other species, however, suggest that collagenous lectins might play a role in mucosal immunity of the gastrointestinal tract. Human gastrointestinal mucosa expresses SP-D, and significantly more SP-D is expressed in gastric pits infected with *Helicobacter pylori* (Murray et al., 2002). An experiment using knockout mice deficient in SP-A showed that SP-A helps protect against gastrointestinal graft-versus-host disease by immunomodulating the T-cell response (Gowdy et al., 2012). At the population level, a variety of SNPs in *SFTPD* were highly associated with the presence of Crohn's disease in humans (Lin et al., 2011). Although these studies provide only weakly associative evidence, they provide a basis of a role for the surfactant proteins in mucosal immunity of the gastrointestinal tract, and justify preliminary investigations on the role of collagenous lectins in equine enteritis.

1.3 Methods of Investigating the Genetics of Infectious Disease

Infectious disease pathophysiology is frequently represented as a triad of interrelated factors: the host, agent, and the environment, each contributing important aspects to the

development of disease. Investigating the genetic basis of infectious disease susceptibility centers on the host, and is dependent on a well-defined host phenotype. The phenotype is composed of both genetic and environmental factors, and the degree to which a phenotype is dominated by the genetic portion is known as heritability. Thus, a good candidate trait for genetic investigation (with respect to selective breeding) should be significantly heritable. In addition, a suitable trait should harbor some amount of economic, welfare, or production importance. Traits can be inherited in a mono- or polygenic fashion, though generally complex phenotypes such as disease resistance are polygenic. Although there is a degree of subjectivity in the evaluation of these parameters, a scoring system based on six categories was devised to help prioritize infectious diseases impacted by genetically-determined variation to resistance (Davies et al., 2009). These criteria included industry concern, economic impact, public concern, food safety and zoonotic threat, animal welfare, and threat to international trade (Davies et al., 2009). In cattle, mastitis was ranked highly, and was followed by a variety of enteric and respiratory infections. In pigs, *E. coli* and porcine respiratory and reproductive syndrome virus (PRRSV) ranked highest, followed by a variety of other bacterial and viral conditions (Davies et al., 2009). Horses were not evaluated. For some of these conditions (e.g. mastitis in cattle, PRRS and various bacterial infections of pigs), there is evidence supporting the involvement of the collagenous lectins, while for others, a role for the collagenous lectins can be reasonably hypothesized based on a) their ability to bind a broad range of pathogens and b) supportive evidence from other species (primarily humans) (discussed in section 1.2.5. Genetics and the Role of the Collagenous Lectins in Infectious Disease).

There are various strategies for the investigation of the genetic basis of infectious disease resistance. Historically, specific genes were targeted, laboriously sequenced in multiple animals

through Sanger sequencing of PCR amplicons, variants were manually identified, and associations with disease phenotype(s) were calculated (Lillie et al., 2007). Modern approaches instead make use of high-throughput technology, often targeting tens to hundreds of thousands of SNPs across a wide range of genes, offering a non-biased approach to the association of variants and disease phenotypes. Genome-wide association studies (GWAS) make use of high density SNP arrays to profile pre-defined variants, often in a population of cases and controls (Goddard and Hayes, 2009). Occasionally, these studies directly identify variants influencing the disease under investigation, but more commonly the results instead implicate entire genes or regions within a chromosome, which are known as quantitative trait loci (QTL). The identification of disease associated QTL is then followed by more detailed examination (fine-mapping) to identify specific, causative mutations.

GWAS has been used to investigate infectious disease resistance in livestock. In pigs, the host genetics of PRRS resistance has been well studied. GWAS identified a locus associated with PRRS in swine on chromosome 4 (ssc4) (Boddicker et al., 2014), and eventually fine-mapping identified an intronic SNP in an innate immune protein, guanylate binding protein 5 (*GBP5*), that led to an alternative transcript with a premature stop codon, and thus decreased levels of GBP5 (Koltes et al., 2015). In cattle, GWAS identified variants in cattle with high and low immune responses (Thompson-Crispi et al., 2014), illustrating the broad influence of genetics on the immune system, and indirectly, on disease resistance. GWAS studies have identified resistance loci to specific pathogens, such as *Mycobacterium bovis* (Bermingham et al., 2014), and have made progress in defining risk loci for BRD (Neibergs et al., 2014; van Eenennaam et al., 2014) and mastitis (Sender et al., 2013). Genomic studies in horses have tended to focus primarily on athletic performance traits, however, an equine 50k SNP array was used to investigate the

pathogenesis of latent equine arteritis virus infection in stallions, and identified several missense mutations in equine *CXCL16* associated with infection (Sarkar et al., 2016). The same array was used to evaluate the susceptibility of horses to bovine papillomavirus, the causative agent of equine sarcoids, and identified QTLs on chromosomes 20 and 22, including one associated with an MHCII related gene (Staiger et al., 2016).

The decreasing cost of next-generation sequencing (NGS) is providing an alternative approach to high density SNP arrays. Whole genome sequencing (WGS) and whole exome sequencing (WES) are two powerful applications of NGS to the discovery of novel or disease-causing mutations. Although WGS bypasses the need for fine-mapping of a gene within a QTL, WGS of the large number of animals required for GWAS is still beyond the economic grasp of many laboratories, and exome capture is not available for all species of livestock (and can also be cost prohibitive). A modified approach is to selectively target and enrich regions of interest in the genome for sequencing. While this provides the advantage of reducing the required sequencing capacity, it also carries the disadvantage of requiring a more laborious and technically demanding sequencing library preparation, along with restricting short nucleotide variant (SNV) detection only to pre-defined DNA sequence. It does, however, sequence a much greater number of SNVs within the targeted region, thus allowing for a more detailed examination of the region of interest. Relatively few studies have used targeted sequencing to investigate the host genetics of infectious disease resistance in animals. A study on cattle native to India identified variants within the exomes of a subset of innate immune genes, however, all animals were healthy (Patel et al., 2015), thus no associations with disease were made. Including animals with infectious diseases can increase the likelihood that variants associated with disease resistance are found.

Variants associated with a trait of interest can then be used to inform breeding programs. In marker-assisted selection, livestock are genotyped for variants of interest, and those with beneficial genotypes are selected for breeding. Markers can be specific, causative alleles, with a known impact on the phenotype, or they may be present in QTL associated with a phenotype and assumed to be in linkage disequilibrium with a causative allele (Dekkers, 2004). A classic example of successful selective breeding approaches include the nearly complete removal of bovine leukocyte adhesion deficiency from Holstein dairy cattle, caused by a mutation to the gene encoding CD18 (*ITGB2*) (Dekkers, 2004; Shuster et al., 1992). Selecting for complex, polygenic traits that are difficult to quantify, such as immunological fitness, is more difficult. Identifying variants associated with such traits is frequently done through a candidate gene approach, in which genes with suspected involvement in the trait are examined in detail.

1.4 Conclusion and Research Proposal

Collagenous lectins, and more broadly, the innate immune system, are critical in the immune response of animals to a variety of pathogens, which represent major causes of morbidity and mortality in cattle, horses, and pigs. The role of collagenous lectins is exemplified by the manifestation of disease in pigs and humans who lack, or who have dysfunctional, proteins, usually as the result of an underlying genetic mutation. Their relative contribution to infectious disease susceptibility in cattle and horses, however, remains poorly defined. Similarly, though the role of the collagenous lectins *MBL1*, *MBL2*, and *FNC1* have been documented in swine, there are gaps in knowledge surrounding the role of the remaining collagenous lectins, along with other innate immune genes, and their contributions to the genetics of infectious disease resistance.

This thesis proposes to investigate the role genetic variation plays in the resistance to common infectious diseases of livestock, specifically horses, cattle, and swine. In horses and cattle, the focus will be primarily on the collagenous lectin and associated MASP genes, while in swine, the investigation will encompass a broader set of genes in the innate immune system.

1.4.1 Objectives

- To identify, via targeted NGS, mutations in the collagenous lectin genes and associated regulatory DNA of horses and cattle, and in a larger subset of innate immune genes in swine.
- To determine if genetic variants in collagenous lectins of cattle and horses are more frequent in animals with evidence of infectious disease.
- To perform *in silico* predictions on the functional impact of any identified single nucleotide variants (horses, cattle) within the coding region, promoter region, and 3′ UTR.
- Identify SNVs associated with altered hepatic expression of innate immune genes in swine (expression quantitative trait loci (eQTL)).
- Determine whether significant eQTL are associated with infectious diseases of swine.

1.4.2 Hypotheses

1) There is substantial genetic variation within the collagenous lectin genes of horses and cattle.
2) Variant density will be lowest in the coding regions and regions with significant functional roles.
3) Some variants will be more or less frequent within populations of animals diagnosed with or without infectious diseases.
4) There are significant local, *cis*-acting eQTL within the porcine innate immune genes.
5) Some eQTL will be significantly more frequent within pigs diagnosed with an infectious disease as compared to normal, healthy control animals.

Chapter 2: Identification of Genetic Variation in Equine Collagenous Lectins Using Targeted Resequencing

This chapter corresponds to the following manuscript currently under review:

Fraser, RS, Arroyo, LG, Meyer, A, Lillie, BN, 2018. Identification of genetic variation in equine collagenous lectins using targeted resequencing. *Vet Immunol and Immunopathol.*

2.1 Abstract

Collagenous lectins are a family of soluble pattern recognition receptors that play an important role in innate immune resistance to infectious disease. Through recognition of carbohydrate motifs on the surface of pathogens, some collagenous lectins can activate the lectin pathway of complement, providing an effective means of host defense. Genetic polymorphisms in collagenous lectins have been shown in several species to predispose animals to a variety of infectious diseases. Infectious diseases are an important cause of morbidity in horses, however little is known regarding the role of equine collagenous lectins. Using a high-throughput, targeted re-sequencing approach, the relationship between genetic variation in equine collagenous lectin genes and susceptibility to disease was investigated. DNA was isolated from tissues obtained from horses submitted for post-mortem examination. Animals were divided into two populations, those with infectious diseases (n = 37) and those without (n = 52), and then subdivided by dominant pathological process for a total of 21 pools, each containing 4-5 horses. DNA was extracted from each horse and pooled in equimolar amounts, and the exons, introns, upstream (approximately 50 kb) and downstream (approximately 3 kb) regulatory regions for 11 equine collagenous lectin genes and related *MASP* genes were targeted for re-sequencing. A custom target capture kit was used to prepare a sequencing library, which was sequenced on an Illumina MiSeq. After implementing quality control filters, 4559 variants were identified. Of these, 92 were present in the coding regions (43 missense, 1 nonsense, and 48 synonymous), 1414 in introns, 3029 in the upstream region, and 240 in the downstream region. *In silico* analysis of the missense SNVs identified 12 mutations with potential to disrupt collagenous lectin protein structure or function, 280 mutations located within predicted transcription factor binding sites, and 95 mutations located within predicted microRNA binding elements. Analysis of allelic association identified 113 mutations that segregated between the infectious and non-

infectious populations. The SNVs discovered in this survey represent potential genetic contributors to disease susceptibility of horses, and might serve as candidates for further population-level genotyping. This study contributes to the growing body of evidence that pooled, high-throughput sequencing is a viable strategy for cost-effective SNV discovery.

Keywords: Collagenous lectins, equine, infectious diseases, innate immunity, next-generation sequencing, pooled sequencing

Acknowledgements: The authors acknowledge the technical help of Jutta Hammermueller and Kristin Wythe.

Funding Sources: This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) [Discovery grant 418356-2012] and an OVC PhD Fellowship to RSF.

2.2 Introduction

Infectious diseases are a significant cause of morbidity and mortality in horses, and the negative impact on their performance can have significant economic consequences for those in the equine industry. Identifying genetic causes of infectious disease susceptibility can help reduce the impact of infectious diseases on performance, and help mitigate the unnecessary use of antimicrobials. Collagenous lectins are a group of evolutionarily conserved pattern recognition proteins with important roles in the innate immune system, and deficiencies in the collagenous lectins are associated with increased susceptibility to infectious diseases in a range of species (Fujita, 2002; Holmskov et al., 2003; Lillie et al., 2005).

The most well-known of the collagenous lectins are the surfactant proteins A and D, while others in the group include mannose (mannan) binding lectin 1 and 2, the ficolins, and a species-dependent number of other collectins. The basic structural unit of all collagenous lectins is a trimer composed of three polypeptides, which then form variable higher order structures. Each polypeptide is composed of an N-terminal domain, collagen-like domain (CLD), a neck region, and a C-type lectin carbohydrate recognition domain (CRD) (collectins) or fibrinogen-like domain (FBG) (ficolins) (Holmskov et al., 2003). The N-terminal domain contains 1 - 3 cysteine residues that stabilize trimer formation (Drickamer et al., 1986; Ohashi and Erickson, 2004). The CLD is composed of a variable number of repeating Gly-X-Y residues that result in a triple-helix formation upon trimer formation (Colley and Baenziger, 1987), and the short neck region composed of heptad repeats is required for trimer formation (Zhang et al., 2001). Finally, the CRD (Weis et al., 1992) or FBG (Garred, 2008) is the site of ligand binding.

After binding of carbohydrate motifs on the surface of pathogens, certain collagenous lectins can activate the lectin pathway of complement (Fujita et al., 2004b; Matsushita, 2010),

agglutinate (Ferguson et al., 1999; Hartshorn et al., 1997) or opsonize pathogens (Kuhlman et al., 1989; Pikaar et al., 1995; van Iwaarden et al., 1991). The activation of the lectin pathway of complement relies on five splice variants derived from the *MASP1* and *MASP2* genes, consisting of the serine proteases MASP-1, MASP-2 and MASP-3 that are analogous with the serine proteases in the classical complement pathway, C1s and C1r , along with two non-enzymatic proteins, sMAP and MAP-1 (Ma et al., 2013).

Genetic variants in the collagenous lectins have been linked to increased susceptibility to both infectious and non-infectious diseases in multiple species. In humans, three coding and three promoter mutations in *MBL2* have been well characterized (Lipscombe et al., 1995; Madsen et al., 1994; Steffensen et al., 2000; Super et al., 1989) and certain haplotypes are associated with a variety of infectious diseases, including childhood respiratory infections, HIV, hepatitis B and C, and meningococcal disease (Minchinton et al., 2002; Super et al., 1989). Several polymorphisms in *FCN2* are associated with lower circulating levels of FCN-2, but it is unclear whether this predisposes to disease (Hummelshoj et al., 2005). Mutations in the *MBL2*, *FCN2,* and *MASP2* genes have been associated with increased risk of sepsis in humans (Bronkhorst et al., 2013). In pigs, a point mutation in the promoter region of *MBL2*, which encodes the MBL-C protein, (rs80823244:G>A) was associated with a significantly decreased level of expression (Lillie et al., 2006b), and this mutation was twice as frequent in animals suffering from pneumonia, enteritis, serositis, or septicemia (Lillie et al., 2007). In cattle, mutations in *MBL1* and *MBL2* are associated with somatic cell score, an indicator of mastitis (C. Wang et al., 2011; Wang et al., 2012).

Relatively little is known about equine collagenous lectins. Three collagenous lectins (MBL-C, SP-A, SP-D) have been biochemically characterized in the horse (Hobo et al., 1999a,

1999b; Podolsky et al., 2006). Low levels of MBL-C were significantly associated with infectious disease in one study, while serum levels of SP-D were increased in an experimental model of *Streptococcus zooepidemicus* infection (Hobo et al., 2007; Podolsky et al., 2006). Most recently, a study found moderately increased levels of circulating SP-D in horses suffering from non-infectious inflammatory airway disease (Richard et al., 2012). To date, no investigations on genetic variation in equine collagenous lectins have been performed. Given that genetic variants in the collagenous lectin genes of other species have a demonstrated impact on susceptibility to infectious disease, we hypothesized that similar differences exist in equine collagenous lectin genes that may contribute to variable resistance to infectious diseases in horses.

To investigate this, we used a pooled, high-throughput, targeted re-sequencing approach to identify genetic variants in the equine collagenous lectins in a population of 89 horses with a mix of infectious and non-infectious disease processes. Pooled sequencing is a cost-effective and accurate method of discovering sequence variants as well as determining population-level allele frequencies (Mullen et al., 2012; Schlötterer et al., 2014). Rather than generating sequence information for individual animals, pooled sequencing combines the genetic material of multiple animals sharing a similar phenotype and identifies variants and generates an accurate estimate of allele frequency (Gautier et al., 2013).

We identified potentially deleterious genetic variants in the genes themselves, and in the upstream and downstream regulatory regions through *in silico* analysis. Additionally, variants associated with infectious disease were identified by comparing estimated allele frequencies in horses diagnosed with infectious diseases to horses lacking any evidence of infectious disease. These variants can serve as candidate quantitative trait nucleotides (QTN) for infectious disease susceptibility in horses.

2.3 Materials and Methods

2.3.1 Animal Selection and Pooling

DNA from horses (n = 89) presenting to the Ontario Veterinary College (OVC) or Animal Health Laboratory (AHL) post-mortem services at the University of Guelph was used for sequencing. Horses underwent a complete post-mortem examination or had a detailed clinical history. Post-mortem examinations were performed under the supervision of a veterinary anatomic pathologist certified by the American College of Veterinary Pathologists. Horses were broadly classified into two populations, Infectious and Non-Infectious, based on evidence of an infectious disease process diagnosed at post-mortem. Where appropriate, ancillary diagnostic testing, such as bacterial culture or pathogen-specific PCR, was performed to confirm the presence of pathogens. Horses were then subdivided into 21 groups of 3-5 animals (as dictated by the composition of cases) (Table 2.1). Grouping of animals was based on the presence or absence of a condition that might be affected by genetic mutations in the collagenous lectins and related genes. The collagenous lectins respond to a wide variety of pathogens, including bacteria, viruses, and fungi (Keirstead et al., 2011; Lillie et al., 2005); inclusion in the study was thus not limited to a single type or species of pathogen. Thirty-four horses had a primary diagnosis of an infectious disease and were divided into 8 groups. This included horses with infections with no predisposing cause; horses with infectious diseases resulting from gastrointestinal accidents, such as colonic rupture, were placed into the Non-Infectious groups, as even immunocompetent animals would be expected to succumb to infection in these cases. An additional subgroup of 3 horses diagnosed with reactive airway obstruction (RAO) was added to the Infectious population (total of 37 animals, 9 groups). Equine RAO is a condition similar to asthma, and has been suggested as an animal model of the human disease (Leclere et al., 2011). The collagenous lectin SP-D has been implicated in asthma in humans (Qaseem et al., 2013), as well as in non-

infectious lower airway inflammatory disease in horses (Richard et al., 2012). Given the similarity between the two conditions and the possible role of a collagenous lectin in the disease pathogenesis, these animals were considered to have a condition potentially impacted by genetic mutations in collagenous lectins, and were included in the Infectious population. Fifty-two horses had no evidence of primary infectious disease (Non-Infectious) and were split into 12 groups (4-5 horses/pool). The study population consisted of 32 Thoroughbreds, 14 Standardbreds, and 43 horses from a variety of other breeds and crosses (Appendix 2.1).

Table 2.1. Categorization of horses by diagnosis obtained at post-mortem.

Status	Group	Diagnosis	Number of animals
Non-Infectious	Group 1	Wobbler (1)	5
	Group 2	Wobbler (2)	5
	Group 3	Fractures (1)	4
	Group 4	Fractures (2)	4
	Group 5	Myocardial disease	4
	Group 6	Idiopathic disease	4
	Group 14	Laminitis	4
	Group 15	Gastrointestinal accident (1)	5
	Group 16	Gastrointestinal accident (2)	5
	Group 17	Gastrointestinal accident (3)	4
	Group 19	Sudden infant death/abortion	4
	Group 24	Miscellaneous	4
Infectious	Group 7	Streptococcal disease	4
	Group 8	Equine Herpes Virus type 1	4
	Group 9	Enteritis	4
	Group 10	Septicemia	5
	Group 11	Miscellaneous infections	4
	Group 12	Bronchopneumonia	4
	Group 13	Colitis	5
	Group 18	Heaves	4
	Group 20	Pneumonia	3

2.3.2 Sample Preparation

Samples of liver, lung, or skeletal muscle were collected from each animal at post-mortem and stored at -20 ˚C. DNA was extracted using a commercial column-based DNA extraction kit (QIAGEN DNeasy Blood and Tissue kit, Mississauga, ON, Canada) and sample concentration was evaluated via fluorometry (Qubit 2.0, Thermo Fisher Scientific, Mississauga, ON, Canada). The DNA from animals in each group was pooled in equimolar amounts to obtain a final concentration of 1 µg of DNA in 50 µl of low-EDTA buffer TE. Each pool of DNA was then acoustically sheared to a target range of 300 bp (peak incident power: 50, duty factor: 20, cycles per burst: 200, time: 175s)(Covaris M220, Woburn, MA, USA).

Following acoustic shearing, each pool of DNA underwent end repair, A-tailing, and adapter ligation (including a unique index) using a KAPA Library Preparation Kit for Illumina Platforms (KAPA Biosystems, Wilmington, MA, USA) as per the manufacturer's instructions, with the following exception: a single cleanup step was performed following adapter ligation. The cleanup was performed by adding 0.55X PEG/NaCL SPRI solution and only the DNA bound to the magnets was retained. The pools were then combined in equimolar amounts to create a single sequencing library.

2.3.3 Target Enrichment and Sequencing

The study was designed to capture exons, introns, and the upstream and downstream regulatory regions of the equine collagenous lectins and two *MASP* genes (Table 2.2). There are seven annotated equine collagenous lectins in the Ensembl database version 90; we identified a further three (*MBL1*, *FCN1*, and a similar gene noted here as *FCN1-like*) through sequence alignment. *FCN1-like* (ENSECAG00000024620) is composed of 4 exons that align to exons 6-9 of *FCN1*. The 633 nt coding sequence of *FCN1-like* completely overlaps the 3′ end of the *FCN1*

coding sequence with 93 % sequence similarity, while the sequence identity between the overlapping portions of FCN1 and FCN1-like amino acid sequences is 88 %. BLAST searches of the first 5 exons of *FCN1* did not identify any stretches of DNA with a similar sequence within the *FCN1-like* region. The predicted protein product of *FCN1-like* does not contain a CLD, instead consisting almost entirely of a FBG. Despite the absence of protein domains consistent with ficolin proteins, it was included in this study based on its current annotation as a gene in the EquCab 2.0 genome (note: the EquCab 3.0 was released following the complete analysis of the data), its high degree of similarity to *FCN1*, and its close genomic proximity to *FCN1*, suggesting it may be the product of a duplication event.

The regulatory regions of the equine collagenous lectin genes are incompletely annotated. To investigate regulatory DNA, regions of up to ~50 kb upstream of the start and up to ~3 kb downstream of the end of gene were included, with upstream and downstream targeted regions adjusted to avoid overlap with unrelated nearby genes, such as occurs in the collectin locus, which in horses includes *SFTPA1*, *MBL1*, and *SFTPD* clustered on chromosome 1. Target coordinates were obtained from the EquCab 2.0 genome (Wade et al., 2009) hosted by University of Santa Cruz, California (Karolchik et al., 2004), and custom target enrichment probes were designed by Nimblegen (Roche, Wisconsin, USA). In total, 658 kb of DNA was targeted for re-sequencing (Table 2.2). Target enrichment was performed using a SeqCap EZ Developer Enrichment Kit as per the SeqCap EZ Library SR User's Guide v.4.2. Following enrichment, the library was sequenced using MiSeq Reagent Kit v3 (600-cycle) sequencing chemistry on an Illumina MiSeq (San Diego, California, USA).

Table 2.2. Genomic coordinates (based on EquCab 2.0) of the regions targeted for resequencing.

Gene name	Ensembl gene ID	Chromosome	Target start (bp)	Target end (bp)	Total sequenced
MBL2	ENSECAG00000018357	1	43,242,494	43,299,035	56,541
SFTPA1[1]	ENSECAG00000018767	1			
MBL1[1]	ENSECAG00000023001	1	88,892,827	89,006,558	113,731
SFTPD[1]	ENSECAG00000006706	1			
FCN3	ENSECAG00000014857	2	28,409,278	28,419,936	10,658
MASP2	ENSECAG00000021756	2	40,412,424	40,446,898	34,474
COLEC12	ENSECAG00000018963	8	40,892,387	40,987,322	94,935
COLEC10	ENSECAG00000018266	9	62,213,522	62,301,779	88,257
COLEC11	ENSECAG00000017488	15	88,547,816	88,602,733	54,917
MASP1	ENSECAG00000008259	19	25,141,928	25,256,816	114,888
FCN1-like[2,*]	ENSECAG00000024620	25	36,785,097	36,875,451	90,354
FCN1[2]	ENSECAG00000000436	25			

[1,2] The genes within these groups are in close proximity on the same chromosome, thus only a single target region was generated for each.

* *FCN1-like* is an unnamed gene predicted by Ensembl with close sequence and protein homology to *FCN1*.

2.3.4 Bioinformatic Analysis of Sequencing Data

Indexed reads were separated into their respective groups and preliminary metrics were generated using FASTQC (Andrews, 2010). Trimmomatic v.0.36 (Bolger et al., 2014) was used to retain reads with a minimum length of 75 and an average Phred score of 20 over a 5 bp sliding window. Reads were aligned to the EquCab 2.0 genome using the BWA-MEM algorithm of the Burrows-Wheeler Aligner (BWA) v.0.7.5 (Li and Durbin, 2009). PCR and optical duplicates were removed using Picard v.1.127 (http://broadinstitute.github.io/picard). The Genome Analysis Toolkit (GATK) Best Practice Guidelines were followed (DePristo et al., 2011; van der Auwera et al., 2013) and GATK v.3.6 was used for in/del realignment and base score recalibration, resulting in a single, analysis ready BAM file for each pool. The ploidy of each group was manually specified where required.

Variants were called using the joint genotyping protocol described by the GATK Best Practice Guidelines, and filtered using separate hard filtering parameters for single nucleotide polymorphisms and insertions/deletions. A minimum of four variant alleles across the entire study population was required to call a variant in order to reduce errors while minimizing false negatives. Mixed and multiallelic variants were excluded and known variants were obtained from dbSNP build 144 (Sherry, 2001). Variant annotation was performed with SnpEff v.4.2 using the EquCab2.82 annotation database (Cingolani et al., 2012).

2.3.5 Variation Rate

To estimate the intergenic and intragenic rates of variation, we first calculated the total number of bases in our target regions that were sequenced to a minimum depth of 5x per animal (445x total). GATK was used to output the depth of coverage for each base in our target coordinates. The output was cross-referenced to an annotation file containing detailed genomic

coordinates obtained from Ensembl (Yates et al., 2016). The untranslated regions (UTRs) of most equine collagenous lectins are either incompletely or totally unannotated; to maintain consistency between genes we thus defined our regions of analysis as the coding sequence (CDS: the start codon to the stop codon), introns, downstream, and upstream DNA. The 50 kb of targeted upstream sequence was split into two regions: the 5 kb proximal to the start codon of the targeted genes was separated from the remaining (up to) 45 kb of upstream DNA. Downstream sequence was defined as the (up to) 3 kb downstream from the stop codon. The same coordinates were used to determine the number of variants in each region. Of note, *MBL1* is 6.87 kb downstream of *SFTPA1*. Variants occurring within, upstream, and downstream of *SFTPA1* were considered to have the potential to affect the expression of *MBL1*. Variants that occurred in overlapping areas were counted once to generate the total number of variants discovered, but were counted separately to generate counts for regional and gene totals for *SFTPA1* and *MBL1*.

2.3.6 *In Silico* Prediction of Functional Effects

All missense mutations were analyzed using both Polyphen2 (Adzhubei et al., 2010) and SIFT (Sim et al., 2012). For Polyphen2, a database comprised of protein sequences corresponding to our genes of interest was obtained from the Ensembl database and the algorithm was run in batch mode at default settings. For SIFT, variants were input into the Variant Effect Predictor of Ensembl and run at default settings.

Promo3 was used to identify potential transcription factor binding sites (TFBSs) in the 5′ promoter region (defined here as within 5 kb of the start codon) (Farré et al., 2003; Messeguer et al., 2002). Only mammalian TFBSs were considered when using Promo3, and a maximum matrix dissimilarity of 0 was selected. Custom Perl scripts were used to identify mutations found in our dataset that intersected with predicted TFBSs.

Equine microRNA sequences were accessed from miRbase (Kozomara and Griffiths-Jones, 2014). All genes analyzed in this study had a single annotated transcript with incompletely defined 3′ UTRs. For the purposes of this analysis, we defined the 3′ UTR as the 3 kb of sequence downstream from the stop codon. The source code for two computational miRNA target prediction algorithms, TargetScan v7.0 (Agarwal et al., 2015) and miRanda v3.3a (Enright et al., 2003), was downloaded and run locally. TargetScan predicts miRNA binding sites through analysis of the miRNA seed sequence (5′ nucleotides 2-8 of the mature miRNA), whereas miRanda uses the entire miRNA sequence and calculates alignment and thermodynamic stability scores. The seed sequences of equine miRNAs were isolated using a custom Python script. TargetScan was run at default parameters using the entire collection of equine miRNA seed sequences against the 3′ UTR regions of the equine collagenous lectins. For miRanda, the alignment score was set at 140 and the free energy score at -20 kcal/mol. The genomic coordinates of the seed region of matching miRNA targets predicted by both algorithms were identified using a custom Python script, and variants falling within these coordinates were identified using GATK's SelectVariant tool.

2.3.7 Allele Frequencies

Allele frequencies were compared by merging individual trimmed, mapped, sorted, and deduplicated Infectious and Non-Infectious pool BAM files into a single BAM file for each population using Samtools v.1.3 (Li et al., 2009). An mpileup file was generated with minimum mapping alignment and base qualities of 20 and then converted to a sync file using Popoolation2 (Kofler et al., 2011). To mitigate the impact of variable depth of sequencing on the output of Popoolation2, the data was first downsampled (as per the software documentation recommendations) to a uniform target coverage of 370x (reflecting approximately 10x coverage

per horse in the Infectious population) across all loci. Fisher's exact tests (calculated by Popoolation2) were used to calculate the significance of allele frequency differences between the Non-Infectious and Infectious pools. The subsampling method of "fraction" was chosen to maintain the exact fraction of the allele frequencies. The Benjamini-Hochberg procedure was used to control for false discovery, with a false discovery rate set to 1×10^{-6} (Benjamini and Hochberg, 1995), using R (R Core Team, 2017).

2.4 Results

Out of a total of 22.1 million reads, approximately 20.9 million reads passed filter, representing approximately 12.6 Gb of total sequence data. The percentage of reads present in each pool ranged from 3.32 - 4.56 % (mean 4.01 %). An average of 56.3 % of reads in each pool covered the regions targeted for enrichment representing a total of 3.4 Gb of DNA. Removal of duplicates and application of quality control filters resulted in an average depth of coverage of 158x per pool (range 128 - 192x) in the targeted regions. Coverage between targeted regions varied significantly (ANOVA, $p < 0.05$), however, all regions achieved a depth of coverage adequate for variant calling (Figure 2.1a). There was no significant variation in coverage between the Non-Infectious and Infectious populations in the targeted regions for any gene (two sample t-test, $p > 0.05$) (Figure 2.1b).

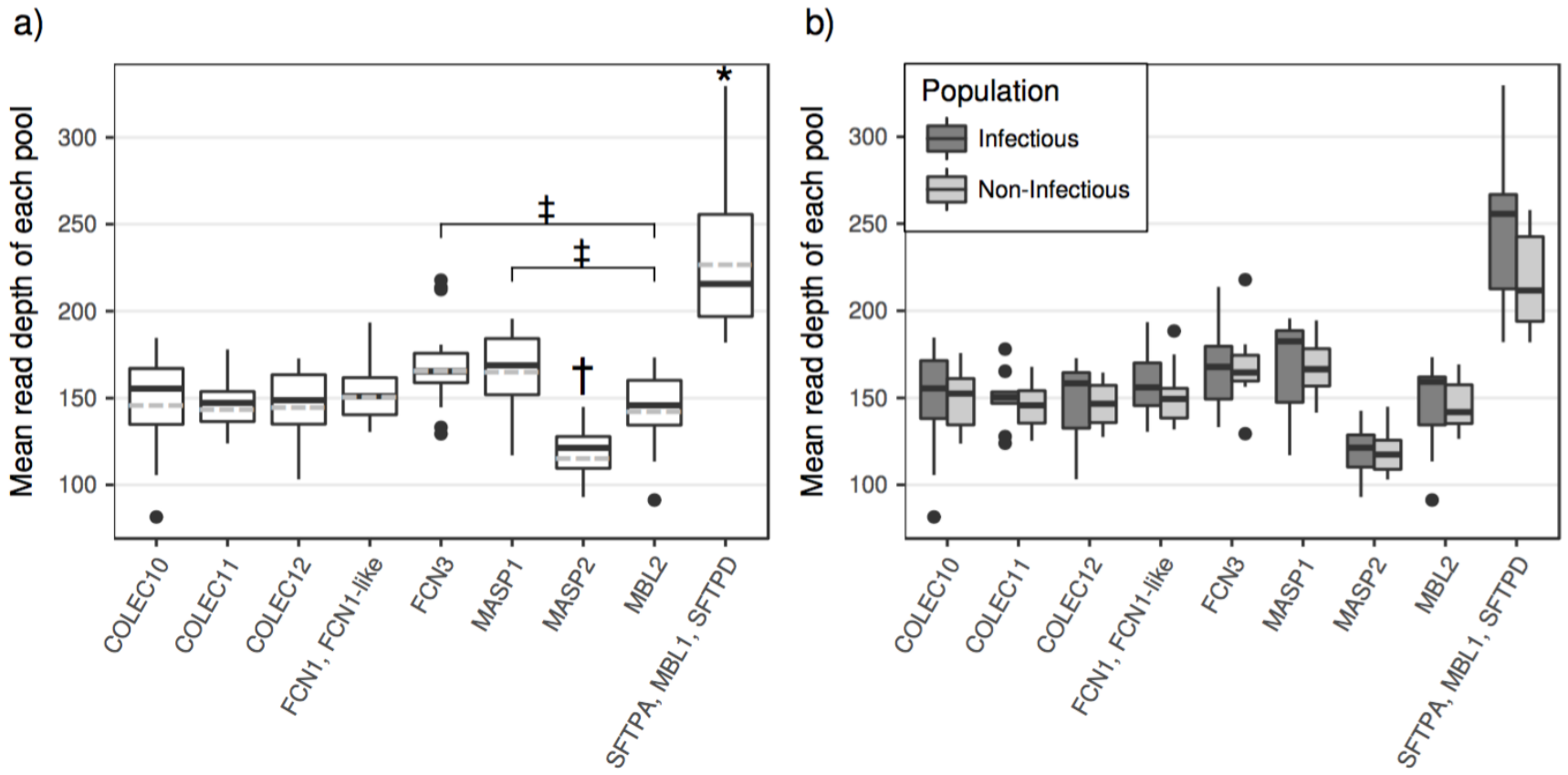


Figure 2.1. Summary of the mean depth of coverage for each pool.

a) The entire study and b) subdivided between the Infectious and Non-Infectious populations of horses. The depth of coverage for the targeted regions differed significantly (a) (ANOVA and Tukey's post-hoc test), however, the mean depth of coverage between the two populations was not significantly different (b) (two sample t-test, $p > 0.05$). Solid lines represent the median, and the dashed line is the mean.
*, †: these target regions differ significantly from all others ($p < 0.05$).
‡ The *MBL2* target region differs significantly from *FCN3*and *MASP1* ($p < 0.05$).

A total of 4559 variants were identified in the targeted regions across all pools, composed of 4174 single nucleotide variants (SNVs) and 385 indels. Our target regions contained 1174 known variants annotated in dbSNP v.144, 962 of which were identified in our dataset. Of the remaining 212 variants present in dbSNP but not our dataset, 66 loci were not variable in our population (despite adequate coverage), and 146 variants were initially identified in our data set but did not pass various filtering parameters. There was a 98.7 % overlap in the variants discovered in the Infectious and Non-Infectious populations. As such, the numbers of variants, and their predicted *in silico* effects, are reported for the entire study, and not subdivided by population, except for the allele frequency comparison between the two groups.

Within the entire study population, there were 92 variants in coding regions, 1414 in introns, 240 in the downstream 3 kb, 513 in the proximal upstream 5 kb, and 2516 in the distal upstream 5 - ~50 kb. The number of variants and variation rate by gene is given in Table 2.3, and a complete breakdown by gene and gene region can be found in Appendix 2.2. The rate of variation within the targeted genes was investigated by comparing the number of variants discovered to the number of bases successfully sequenced. The rate of variation within the *FCN1* and *FCN1-like* target regions was significantly higher compared to the remaining genes (ANOVA and Tukey's post-hoc test, $p < 0.05$) (Figure 2.2a). The rate of variation between different gene regions (exons, introns, upstream, and downstream) was not significantly different (Kruskal-Wallis, $p > 0.05$) (Figure 2.2b). This is likely the result of the elevated variation rate in *FCN1* and *FCN1-like* in all gene regions, thus creating similar variation in the distribution for all gene regions.

Table 2.3. Number of bases targeted for sequencing for each gene, the number of bases sequenced to a minimum of 5x, and the number of variants discovered in each gene region.

Gene	Total targeted (bp)	Sequenced to 5x (bp)	Number of variants	Variation rate (per kb)
COLEC10	88,258	88,258	371	4.2
COLEC11	54,918	54,918	566	10.3
COLEC12	94,479	94,479	521	5.5
FCN1	60,957	60,957	810	13.3
FCN1-like	22,437	22,437	411	18.3
FCN3	10,659	10,054	42	4.2
MASP1	114,890	114,889	639	5.6
MASP2	34,475	34,475	91	2.6
MBL1	45,907	45,025	288	6.4
MBL2	56,542	56,542	468	8.3
SFTPA1	35,227	34,344	217	6.3
SFTPD	61,732	61,732	352	5.7

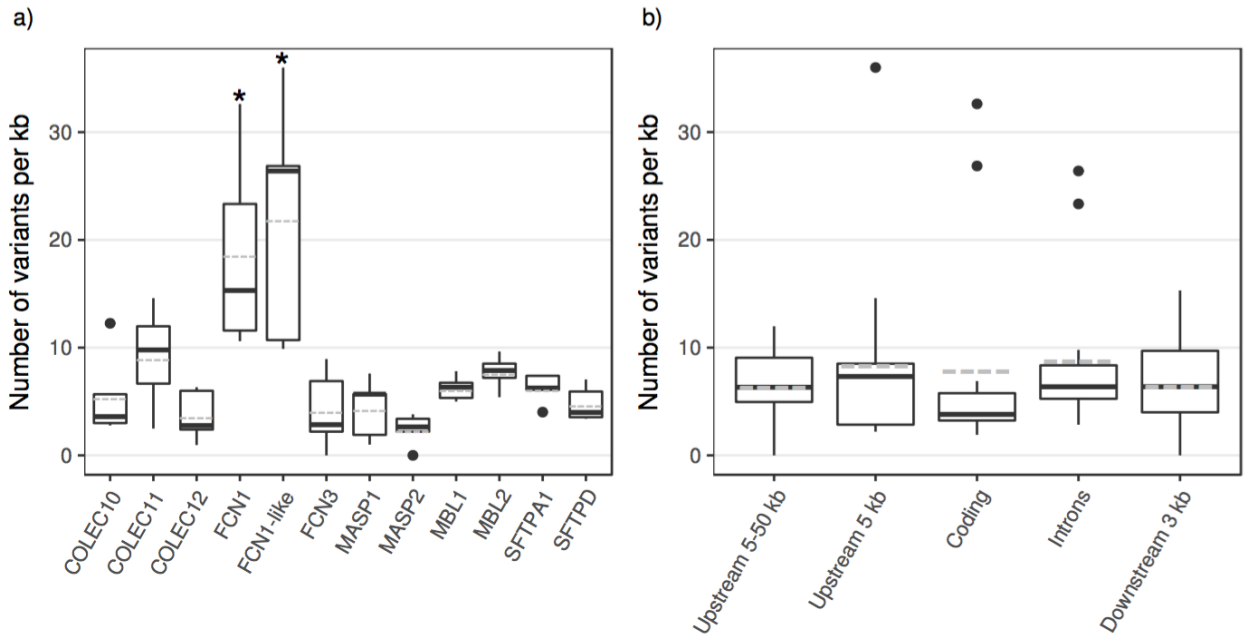


Figure 2.2. Rate of variation (variants / kb) between genes and within gene regions for all horses.

a) Median variation across all gene regions was significantly different between genes (ANOVA and Tukey's post-hoc test). b) The rate of variation within gene regions, taken across all genes, was not significantly different. Solid lines represent the median, and dashed lines are the mean, boxes represent 1st and 3rd quartiles and whiskers represent 1.5 x the interquartile range
*: $p < 0.05$

Of the 92 variants in the coding regions, 44 were non-synonymous. These included 1 nonsense mutation and 43 missense mutations. The nonsense mutation, ENSECAT00000000831.c465A>G, was present in exon 6 of *FCN1*, leading to a TGA stop codon at amino acid position 155 of the protein product. Computational analysis of the missense mutations using Polyphen2 identified 12 mutations with potential deleterious effects to protein structure and function, five of which were assigned a score of 0.99 or higher, predicting a likely damaging effect (Table 2.4). These five variants were also predicted by SIFT to have a deleterious effect. An additional two variants were predicted as deleterious by SIFT, but not by Polyphen2. A missense mutation in *FCN3* resulted in a leucine to serine change for which Polyphen2 was unable to predict an outcome, but SIFT predicted the change to be tolerated.

A total of 513 sequence variants were found in the 5 kb upstream of the targeted genes. TFBS analysis using Promo3 identified 559 putative binding sites in this region for 91 different transcription factors. Two hundred and eighty of the 513 identified variants were located within the predicted TFBS. Although, to our knowledge, there are no studies on the promoter regions of equine collagenous lectins, the promoter region of a variety of collagenous lectins in humans (Grageda et al., 2014; Naito et al., 1999; Tachibana et al., 2013; Wu et al., 2011), chickens (Kjærup et al., 2013), cattle (Hansen et al., 2003a), and mice (Liu et al., 2008) have been characterized, and a series of transcription factors identified or predicted. Eight of the transcription factors (TFs) identified in our dataset that also contained a variant were identified in the aforementioned studies, including c-Ets-1, c-Myc, E47, MZF-1, N-Myc, PEA3, Sp1. These eight TFs accounted for 37 of the predicted TFBSs spread across five genes: *FCN1*, *FCN1-like*, *COLEC11*, *MBL2*, and *MBL1* (Appendix 2.3).

Table 2.4. *In silico* analysis of the missense mutations in the targeted genes using two separate algorithms.

Gene	Chr	Position	Ref allele	Alt allele	Amino acid change	Protein domain	Polyphen2 score	Polyphen2 result	SIFT score	SIFT prediction
FCN1	25	36,824,050	G	T	Pro54Gln	CLD	1	probably damaging	0	deleterious
MBL1	1	88,935,539	G	A	Cys237Tyr	CRD	1	probably damaging	0.04	deleterious
MASP1	19	25,180,346	C	T	Ala137Thr	CUB domain	1	probably damaging	0	deleterious
MBL1	1	88,933,562	G	C	Gly82Ala	CLD	0.999	probably damaging	0	deleterious
FCN3	2	28,415,514	G	T	Gln123His	FBG	0.998	probably damaging	0	deleterious
MASP1	19	25,154,003	A	G	Ile463Thr	Peptidase S1 domain	0.926	possibly damaging	0.05	deleterious
MBL1	1	88,933,006	A	C	Thr36Pro	N-terminal domain	0.86	possibly damaging	0.05	tolerated
FCN1	25	36,824,033	C	T	Val60Met	CLD	0.859	possibly damaging	0.06	tolerated
FCN1	25	36,820,482	C	G	Asp149His	FBG	0.651	possibly damaging	0.18	tolerated
FCN1	25	36,824,026	G	A	Pro62Leu	CLD	0.539	possibly damaging	0.04	deleterious
SFTPD	1	88,950,438	G	A	Pro191Leu	CLD	0.483	possibly damaging	0.23	tolerated
FCN1	25	36,819,956	G	T	Ala174Glu	FBG	0.465	possibly damaging	0.04	deleterious
COLEC10	9	62,298,342	C	T	Thr193Met	CRD	0.284	benign	0.04	deleterious
MASP2	2	40,427,714	G	A	Lys90Arg	CUB domain	0.057	benign	0	deleterious

Abbreviations: Chr: chromosome; Ref: reference; Alt: alternate; SIFT: Sorting Intolerant from Tolerant algorithm; CLD: collagen-like domain; CRD: carbohydrate recognition domain; FBG: fibrinogen like domain; CUB: complement C1r/C1s, Uegf, Bmp1 domain.

microRNA recognition element (MRE) analysis of the downstream 3 kb of the targeted regions was performed using miRanda and TargetScan. The miRanda algorithm identified 1892 potential miRNA recognition elements, while TargetScan identified 7503. Of these, 701 were predicted by both algorithms. Twenty-seven variants were present in the seed region of 33 unique MREs; an additional 68 variants were found in the complete sequence of 70 MREs (Figure 2.3 and Appendix 2.4).

Estimates of allele frequencies were compared between the Infectious and Non-Infectious populations (Figure 2.4). Significant differences, as determined by Fisher's exact test, were found at 113 different loci. The top 15 most significant variants ($p < 7.1 \times 10^{-7}$), along with any variant predicted to have a functional impact, are presented in Table 2.5. The majority (103) occurred in non-coding DNA with no *in silico* predicted function, however, 10 loci intersected with a predicted MRE ($n = 2$), a TFBS ($n = 7$), or a missense mutation ($n = 1$). The missense mutation in *MBL2*, c.521T>A resulting in p.(Gln174Leu), was predicted to be benign by both Polyphen2 and SIFT.

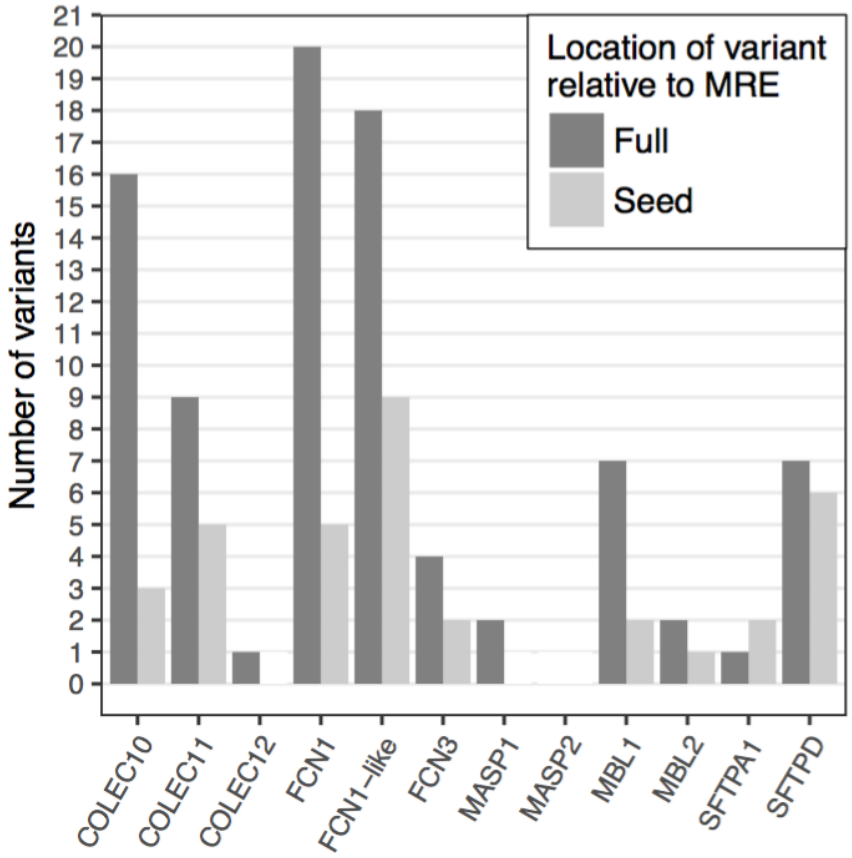


Figure 2.3. Number and location of variants impacting miRNA recognition elements (MREs) in all horses.

Variants are subdivided based on whether they impact the region in which the seed sequence of the miRNA is expected to bind versus the remaining miRNA sequence.

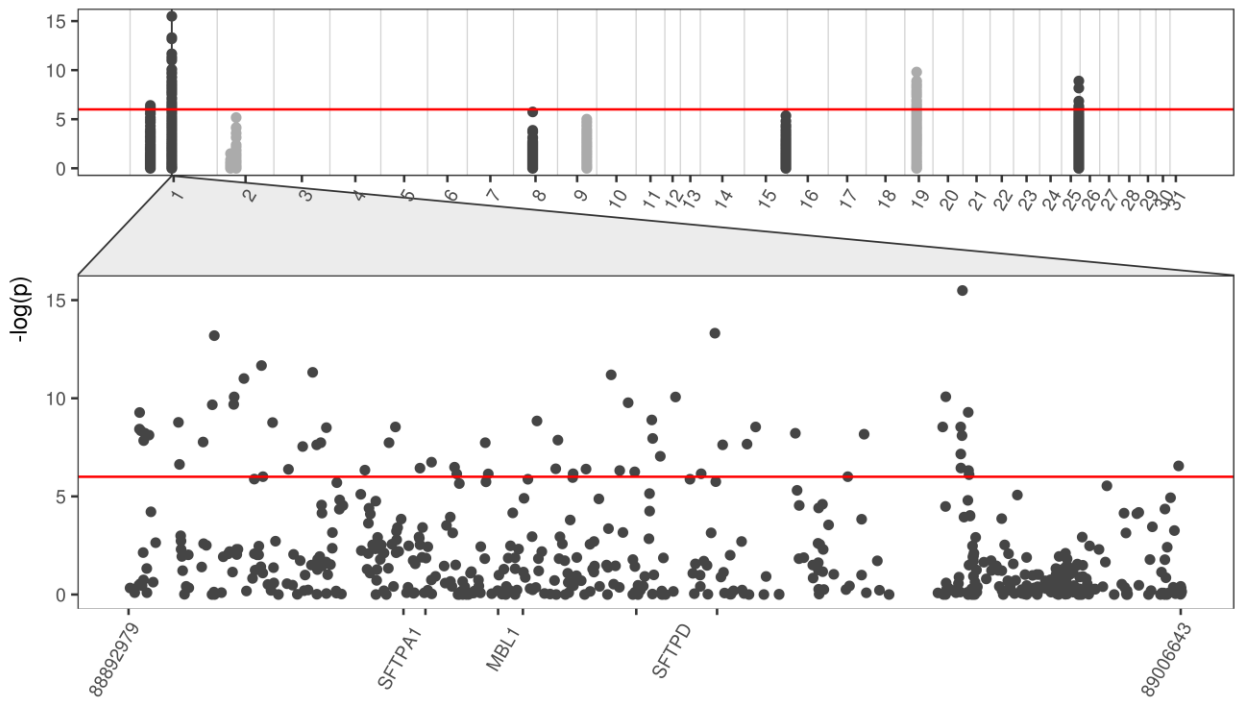


Figure 2.4. Manhattan plot showing p values corrected for multiple testing from allelic association tests performed on variants in horses in the Infectious and Non-Infectious populations.

The horizontal line represents a cut-off value of $p < 1x10^{-6}$ for selecting significant variants as determined by using the Benjamini-Hochberg procedure. Inset: a zoomed in portion of chromosome 1 representing the collectin locus. Sex chromosomes are omitted.

Table 2.5. Frequency of selected variants in Infectious and Non-Infectious horses. *In silico* analysis was performed to identify potential functional consequences for the variants.

Gene	Chr	Position	Ref allele	Alt allele	Freq in Non-Infectious	Freq in Infectious	-log(p)*	Gene region	Predicted impact (*in silico*)
SFTPD	1	88,982,902	T	G	0.338	0.673	15.49	Upstream 5 - 50 kb	-
SFTPD	1	88,956,159	G	C	0.192	0.014	13.32	Intron 2-3	-
SFTPA1	1	88,902,080	G	A	0.208	0.022	13.19	Upstream 5 - 50 kb	-
SFTPA1	1	88,907,177	C	T	0.189	0.019	11.67	Upstream 5 - 50 kb	-
SFTPA1	1	88,912,707	A	G	0.178	0.016	11.32	Upstream 5 - 50 kb	-
SFTPD	1	88,944,953	C	T	0.151	0.008	11.19	Downstream 3 kb	MRE for eca-miR-221
SFTPA1	1	88,905,280	C	T	0.181	0.019	11.01	Upstream 5 - 50 kb	-
SFTPD	1	88,981,103	C	A	0.270	0.538	10.08	Upstream 5 - 50 kb	-
SFTPA1	1	88,904,227	C	G	0.165	0.016	10.07	Upstream 5 - 50 kb	-
SFTPD	1	88,951,906	G	A	0.184	0.024	10.07	Intron 2-3	-
MASP1	19	25,244,848	T	G	0.257	0.065	9.81	Upstream 5 - 50 kb	-
SFTPD	1	88,946,787	G	A	0.151	0.014	9.77	Downstream 3 kb	-
SFTPA1	1	88,904,178	C	T	0.173	0.022	9.69	Upstream 5 - 50 kb	-
SFTPA1	1	88,901,863	G	A	0.159	0.016	9.67	Upstream 5 - 50 kb	-
SFTPD	1	88,983,525	C	G	0.186	0.030	9.28	Upstream 5 - 50 kb	-
MBL1	1	88,936,937	C	T	0.157	0.019	8.84	Downstream 3 kb	MRE for eca-miR-8954
SFTPA1	1	88,921,646	G	A	0.157	0.022	8.54	Upstream 5 kb	TFBS for Msx-1
MBL1	1	88,931,351	G	A	0.149	0.022	7.74	Upstream 5 kb	TFBS for ZF5
SFTPD	1	88,959,630	A	G	0.459	0.689	7.66	Upstream 5 kb	TFBS for Msx-1, Zic1, Zic2, Zic3, and VDR
FCN1-like	25	36,798,301	T	C	0.205	0.062	6.31	Upstream 5 kb	TBFS for En-1
MBL2	1	43,245,721	T	A	0.119	0.016	6.25	ENSECAE00000138507	Gln174Leu
FCN1-like	25	36,796,958	G	A	0.176	0.046	6.15	Upstream 5 kb	TFBS for Zic1, Zic2, Zic3, and VDR
FCN1-like	25	36,796,961	G	A	0.176	0.046	6.15	Upstream 5 kb	TFBS for ZF5
MBL2	1	43,252,523	T	C	0.081	0.003	6.15	Upstream 5 kb	TFBS for Zic2

Abbreviations: Chr: chromosome; Freq: frequency; MRE: miRNA recognition element; TFBS: transcription factor binding site.
* The p value is the result of Fisher's exact test and has been adjusted using the Benjamini-Hochberg procedure.

2.5 Discussion

This study was designed to identify and characterize variants in the equine collagenous lectin genes, as these encode potentially important proteins in the innate immune response against infectious diseases. Horses diagnosed with infectious diseases were included in order to increase the likelihood that disease-associated alleles would be captured in the study population, and allelic association was performed to identify candidate alleles associated with infectious disease resistance. To maximize the number of animals sequenced, we pooled DNA from animals sharing similar phenotypes. Pooled sequencing has been shown to be an accurate and cost-effective method of SNP discovery and allele frequency estimation (Bansal et al., 2011; Bertelsen et al., 2016; Mullen et al., 2012; Rellstab et al., 2013). Fundamental to the accuracy of pooled sequencing experiments is the accurate equimolar pooling of DNA from individual samples. Our two study populations were subdivided into 21 pools of 3 - 5 animals each. Although the percentage of DNA originating from individual horses within a pool cannot be determined, the percentage of reads mapped to each pool ranged from 3.3 - 4.6 % (mean 4.01 %, SD 0.37), indicating a relatively uniform contribution from each pool. However, there were significant differences in average coverage between different targets (Figure 2.1). The collectin locus, a region also found in cattle (Gjerstorff et al., 2004a) and mouse (Akiyama et al., 1999), is comprised of *SFTPA1*, *MBL1*, and *SFTPD*, and had the highest average read depth of 230x, compared to *MASP2*, which had the lowest average read depth of 118x. Although the exact reasons underlying the differences in sequencing depth cannot be explained, the efficacy of genomic capture when using hybridization methods is known to be variable (Bodi et al., 2013). Regardless of capture efficiency, all targets achieved sufficient depth for the purposes of variant calling. Importantly, there was no significant difference in coverage between the Infectious and

Non-Infectious populations at any target region, indicating that target capture was similar across the two populations.

Unlike the design of other pooled-sequencing experiments, in which large numbers of animals are sequenced at relatively shallow depths (Bertelsen et al., 2016; Mullen et al., 2012), we instead used fewer animals and achieved high coverage in each pool. A higher depth of coverage increased the sensitivity of our analysis to low-frequency variants present within our study population; we thus set the minimum required number of variant alleles present in our dataset to four (equivalent to a MAF of 2.3 %). However, the reduction in the total number of animals in the experiment potentially reduces the accuracy of our allele frequency estimates (Kim et al., 2010). Furthermore, some rare variants may not have been present in our study population of 89 horses.

Identifying quantitative trait nucleotides (QTN) in livestock associated with complex traits such as immunity is difficult. Challenges include obtaining adequate sample sizes of animals with a well characterized phenotype in species that have long generation intervals, and for which the production of transgenic animals is frequently not feasible (Ron and Weller, 2007). In this study, animals were stratified into groups based on phenotypes that were the defined either by a detailed clinical history or a comprehensive post-mortem examination. Post-mortem examinations were conducted under the supervision of a veterinary pathologist to ensure the accuracy of the diagnosis. Complete medical histories were not available for all post-mortem cases. It is thus possible that some animals that were assigned into the Non-Infectious group based on post-mortem examination may have had a medical history of infectious disease, but were free from disease or evidence of prior infectious disease at the time of death.

Some collagenous lectins have an established role in the innate immune response in some mammalian species, so variants within these genes are potential QTN for infectious disease susceptibility. Using a pooled and targeted next-generation re-sequencing approach, we identified 4559 sequence variants in the equine collagenous lectin and related *MASP* genes, 113 of which were found to have significantly different allele frequencies in horses with and without infectious diseases. The majority of these variants occurred in non-coding DNA and had no predicted biological relevance. Fifty-nine of the variants were present in the genomic sequence between 5 and 50 kb upstream from the start of our genes of interest. Although no bioinformatic analysis was performed on this location, DNA up to 1 Mb upstream of genes can have a regulatory impact through enhancer, silencer, insulator, or locus control region elements (Maston et al., 2006) that are difficult to predict, and for which the equine genome lacks annotation. Intronic DNA can harbour mutations that affect gene function, as exemplified by a mutation in intron 3 of porcine *IGF2* that impacts muscle growth (van Laere et al., 2003). Of the segregating alleles with predicted function, nine occurred in putative regulatory regions (two downstream, seven upstream). The two downstream regulatory variants were predicted to impact miRNA binding elements, potentially disrupting the typically dampening effect of miRNA binding. The seven upstream regulatory variants bound putative TFBS, potentially impacting TF binding affinity and altering transcription. While it is possible that some variants identified as having significantly different allele frequencies may be in linkage disequilibrium with each other or with other functional mutations in untargeted areas of the genome, the design of this study did not allow for linkage analysis, as genotype information at the individual level is unavailable.

Two studies have investigated quantitative trait loci (QTL) relating to equine infectious diseases. A study on *R. equi* identified a single locus on chromosome 26 associated with

susceptibility in foals (McQueen et al., 2014), while a study on equine arteritis virus identified a locus on chromosome 11 (Go et al., 2011). Our genes of interest were not present on the chromosomes identified in those studies, and thus there was no overlap. None of the variants identified in this study were present within any QTL in the equine QTL database (Hu et al., 2016).

The variation rate amongst the genes targeted in this study varied significantly. The number of variants in the *FCN1* and *FCN1-like* targeted region was higher than in all other genes sequenced. Increased variation can be due to several factors: at the sequence level, high GC content of the DNA and the presence of indels in the region are frequently predictive of variation (Hodgkinson and Eyre-Walker, 2011). The GC content of the *FCN1/FCN1-like* locus was the highest of all the targeted regions at 54.7 % (Appendix 2.5). The same region also contained 85 indels, the most of any region sequenced (Appendix 2.5). At a functional level, genes that interact with the environment, such as genes of the immune response, are often localized to 'hot spots' of variation in the genome, in which they are subject to increased adaptive evolutionary pressure (Chuang and Li, 2004). Of the 15 missense variants found in *FCN1*, seven were found in the FBG, the site of ligand binding, and could reflect evolutionary adaptation to differing pathogens. The elevated mutation rate in the *FCN1-like* gene is more difficult to reconcile, as its status as a functional ficolin remains unclear. The lower rate of variation in the remaining collagenous lectin genes, which also are genes that interact with the environment, may suggest that *FCN1* is under comparatively greater evolutionary pressure, and may play a more significant role in horses.

The *in silico* prediction of the functional effects of missense mutations is imperfect, and even though a mutation may be predicted to be damaging, it is not necessarily pathogenic

(Eilbeck et al., 2017). Evidence from studies in other species on similar proteins can help identify potentially pathogenic missense mutations. In this regard, the p.(Gly82Ala) (Table 2.4) substitution in equine MBL-A is of particular interest. This mutation falls within the collagen-like domain of MBL-A, a region characterized by repeating Gly-X-Y amino acid residues critical to the formation of a collagen-like triple-helix (Holmskov et al., 2003). Comparable mutations to glycine residues in the collagen-like domain have been found in porcine *MBL1* (Lillie et al., 2006b) and in human *MBL2* (Sumiya et al., 1991). In human *MBL2,* the mutation disrupts triple helix formation and acts in a dominant negative fashion, resulting in lower levels of functional, circulating MBL-C (Sumiya et al., 1991). Humans carrying this mutation have been shown to have an increased susceptibility to infectious disease, particularly as children (Sumiya et al., 1991; Super et al., 1990). Similarly, coding mutations in porcine *MBL1* are associated with decreased circulating levels of MBL-A (Juul-Madsen et al., 2011). Whether the equine mutation described here has similar effects on the structure of equine MBL-A is unknown, but is of particular interest as evidence suggests that decreased levels of the related MBL-C (as determined by a complement deposition assay) were significantly associated with infectious disease (Podolsky et al., 2006). The allele frequency of this mutation was not significantly different between Infectious and Non-Infectious populations, though only 14 variant alleles were present in our population. Further elucidation of the relationship of this variant and MBL-A levels and function is thus warranted.

Analysis of the putative upstream and downstream regulatory regions was performed primarily to assign potential function to alleles that segregated in the allele frequency analysis. The results of these predictions should be interpreted with caution, as they have not been functionally validated. Studies of collagenous lectins of other species have identified or predicted

36 potential transcription factors, eight of which were also predicted in our study; however, despite the presence of conserved TFBS sequence identities, the majority of transcription factor binding events appear to be species specific (Schmidt et al., 2010). The accuracy of these predictions awaits confirmation from high quality studies focussed on transcription factor binding.

This is the first study, to our knowledge, that has performed targeted next-generation resequencing on genes of the equine innate immune system. This study has identified genetic variants with predicted impacts on protein structure, function, or expression in genes of importance to the equine innate immune system. Through analysis of allele frequencies, potential quantitative trait nucleotides associated with equine infectious diseases have been described. Future directions include larger scale genotyping studies on individual horses and investigation of the impact of a missense mutation in the CLD of *MBL1* on serum protein levels.

Conflicts of interest: none.

Chapter 3: Identification of Polymorphisms in the Bovine Collagenous Lectins and their Association with Infectious Diseases in Cattle

This chapter is a modified version of the following manuscript currently in press:

Fraser, RS, Lumsden, JS., Lillie, BN, 2018. Identification of polymorphisms in the bovine collagenous lectins and their association with infectious diseases in cattle. *Immunogenetics*. doi:10.1007/s00251-018-1061-7

3.1 Abstract

Infectious diseases are a significant issue in animal production systems, including both the dairy and beef cattle industries. Understanding and defining the genetics of infectious disease susceptibility in cattle is an important step in the mitigation of their impact. Collagenous lectins are soluble pattern recognition receptors that form an important part of the innate immune system, which serves as the first line of host defence against pathogens. Polymorphisms in the collagenous lectin genes have been shown in previous studies to contribute to infectious disease susceptibility, and in cattle, mutations in two collagenous lectin genes (*MBL1* and *MBL2*) are associated with mastitis. To further characterize the contribution of variation in the bovine collagenous lectins to infectious disease susceptibility, we used a pooled NGS approach to identify single nucleotide variants (SNVs) in the collagenous lectins (and regulatory DNA) of cattle with (n = 80) and without (n = 40) infectious disease. Allele frequency analysis identified 74 variants that were significantly ($p < 5 \times 10^{-6}$) associated with infectious disease, the majority of which were clustered in a 29 kb segment upstream of the collectin locus on chromosome 28. *In silico* analysis of the functional effects of all the variants predicted 11 SNVs with a deleterious effect on protein structure and/or function, 148 SNVs that occurred within potential transcription factor binding sites, and 31 SNVs occurring within potential miRNA binding elements. This study provides a detailed look at the genetic variation of the bovine collagenous lectins, and gives and identifies potential genetic markers for infectious disease susceptibility.

KEYWORDS: collagenous lectins, infectious disease, cattle, pooled next-generation sequencing, genetic variants

3.2 Introduction

Infectious diseases are a major source of morbidity, mortality, and economic loss to the cattle industry. Infectious respiratory diseases alone account for close to $0.5-1 billion USD annually in North America (Miles, 2009), and while estimates for other common infectious diseases, such as mastitis and gastrointestinal disease, are difficult to obtain, they undoubtedly add significantly to the economic impact of infectious disease (Halasa et al., 2007; Heikkilä et al., 2012; Schepers and Dijkhuizen, 1991). Infectious diseases also represent a large source of agricultural antimicrobial use, which contributes to the development of antimicrobial resistance (Prescott et al., 2012). The approach to managing the impact of infectious disease has traditionally been to control the pathogen, largely ignoring the potential contributions of an immunologically deficient host (Miles, 2009; Prescott et al., 2012). Recently, however, there has been a broadening in focus to include host factors that contribute to infectious disease susceptibility.

The innate immune system represents the first line of defense against infectious diseases. Pattern recognition receptors (PRRs), a key part of the innate immune system, recognize conserved motifs on pathogens called pattern associated molecular patterns (Janeway, 1989). The collagenous lectins are a subset of membrane bound and/or soluble, circulating C-type lectins that function as PRRs, recognizing carbohydrate residues on the surfaces of bacteria, viruses, and fungi. The collagenous lectin family includes the collectins and ficolins, which share structural and functional similarities. Eleven collagenous lectin genes have been identified in cattle, including the genes encoding mannose-binding lectins 1 and 2 (*MBL1* and *MBL2*), surfactant proteins A and D (*SFTPA1* and *SFTPD*), collectins 10, 11, 12, 43, 46, conglutinin (*CGN1*), and ficolin-1 (*FCN1*). *CL43*, *CL46*, and *CGN1* are found in cattle and a few select

herbivores, and structural similarities between these collagenous lectins and *SFTPD* suggesting suggest that they are evolutionarily related (Gjerstorff et al., 2004a; Hansen and Holmskov, 2002).

Certain collagenous lectins (MBLs, FCNs, and CL-11) can activate the lectin pathway of complement, and can agglutinate or opsonize pathogens (Fujita, 2002; Fujita et al., 2004a). The lectin pathway of complement is activated in part by four MBL-associated serine proteases (MASPs), encoded by two MASP genes, *MASP1* and *MASP2*. The MASP proteins are structurally and functionally similar to C1r and C1s of the classical complement pathway (Matsushita et al., 2013; Thiel et al., 1997) and bind in proenzyme form to the collagen-like domain of the MBLs, FCN-1, and CL-11 (Hansen et al., 2010; Matsushita et al., 2001; Matsushita and Fujita, 1992). Following ligand recognition, the MASP proteins lead to the cleavage of complement components C2 and C4, resulting in the activation of complement.

Single nucleotide variants (SNVs) in the collagenous lectin genes are associated with infectious disease susceptibility in a variety of species. A dominant negative missense mutation in human *MBL2* leads to an opsonic defect and is a cause of recurring infections in children (Sumiya et al., 1991) and adults (Summerfield et al., 1995). Deficiencies of MBL-C resulting from *MBL2* polymorphisms in humans are also associated with HIV, hepatitis B and C, meningococcal disease, and parasitic infections (Eisen and Minchinton, 2003). Polymorphisms in the human ficolin genes are associated with leprosy (Andrade et al., 2017; Boldt et al., 2013), pneumonia (van Kempen et al., 2017), and Chagas disease (Luz et al., 2016), while variation in *SFTPA2* is associated with different outcomes to influenza A virus infection (Herrera-Ramos et al., 2014). In animals, mutations in the promoter region of porcine *MBL2* are associated with decreased expression of MBL-C and are more frequent in animals diagnosed with pneumonia,

enteritis, serositis, or septicemia (Lillie et al., 2007). A missense mutation in porcine *MBL1* is associated with decreased serum concentrations of MBL-A (Juul-Madsen et al., 2011). Relatively little is known about genetic variation in bovine collagenous lectins, and the few investigations that have been done have focused solely on mastitis and the *MBL* genes. Both a missense mutation in exon 2 and a promoter mutation of *MBL1* are associated with altered activity of the classical complement pathway as well as with somatic cell score (SCS), a measure of the inflammatory cell content of milk and an indicator of mastitis (Liu et al., 2011; C. Wang et al., 2011; Yuan et al., 2012). Mutations in the coding region of *MBL2* are also associated with SCS and complement activity (Wang et al., 2012; Zhao et al., 2012).

In order to address this gap in knowledge, we designed a targeted, next-generation sequencing study that captured the bovine collagenous lectin and related *MASP* genes as well as surrounding regulatory DNA, in order to examine and characterize genetic variation of these genes, including *in silico* predictions of functional effects of identified variants. We sequenced our target regions in 120 cattle, 80 of which were diagnosed with infectious disease, and 40 of which lacked any evidence of infectious disease. We also performed association analysis and identified 74 variants significantly associated with infectious disease.

3.3 Materials and Methods

3.3.1 Sample Selection and Library Preparation

Samples of liver or lung were collected from cattle presenting to the post-mortem service at the Ontario Veterinary College or the Animal Health Laboratory at the University of Guelph. Cattle underwent a complete autopsy under the supervision of a veterinary pathologist certified by the American College of Veterinary Pathologists. Ancillary testing (e.g. bacterial culture, viral PCR, etc.) was performed as necessary to confirm the presence of pathogens. Cattle were

broadly divided into two major populations: those with and without evidence of infectious disease (referred to as Infectious and Non-infectious in this article). They were then subdivided into pools of 5 animals each based on the similarity of their diagnosis at autopsy (Table 3.1). The study population consisted predominantly of female Holstein-Friesians (66.7 %), with fewer intact and castrated male Holstein-Friesians (8.3 %). The remainder (25.0 %) was composed of a variety of other breeds and crosses (Table 3.2).

Table 3.1. Pools of cattle grouped together by diagnosis determined at autopsy.

Status	**Group**	**Diagnosis**
Normal	Group 1	Normal (no lesions)
	Group 2	Fractures or trauma
	Group 3	Dental malocclusion, peripheral neuropathy
	Group 4	Intestinal accident or musculoskeletal trauma
	Group 5	Neoplasia
	Group 6	Metabolic disease
	Group 7	Congenital malformations
	Group 8	Organ torsion or rupture
Diseased	Group 9	Endocarditis
	Group 10	Meningitis
	Group 11	Bronchopneumonia 1
	Group 12	Bronchopneumonia 2
	Group 13	Pneumonia (*M. haemolytica*)
	Group 14	Mycoplasma arthritis, osteomyelitis, or pneumonia
	Group 15	Pneumonia (*T. pyogenes*)
	Group 16	Sepsis
	Group 17	Omphalophlebitis
	Group 18	Foot abscess or ulcer
	Group 19	Infectious arthritis
	Group 20	Abortion or perinatal death of infectious cause
	Group 21	Diarrhea
	Group 22	Mastitis
	Group 23	Multifocal abscesses
	Group 24	Metritis or endometritis

Table 3.2. Breakdown of the bovine study population by breed and sex.

	Breed	**F**	**M**	**MN**	**Total**
Non-infectious	Holstein-Friesian	30	1	2	33
	Limousin	1	1		2
	Hereford-Limousin	1			1
	Limousin cross	1			1
	Not available	1			1
	Angus		1		1
	Jersey		1		1
Sub-total		*34*	*4*	*2*	*40*
Infectious	Holstein-Friesian	50	7		57
	Limousin	2	2	1	5
	Shorthorn	1	1		2
	Red Angus	1			1
	Wagyu	1			1
	Black Angus	1	3	2	6
	Simmental cross	1			1
	Unspecified beef breed	2			2
	Angus		1	1	2
	Angus-Simmental		1		1
	Charolais cross			1	1
	Jersey		1		1
Sub-total		*59*	*16*	*5*	*80*
Total		**93**	**20**	**7**	**120**

MN: male neutered (steer).

Tissue samples were stored at -20 ºC until processed. DNA was extracted using a commercial column-based DNA extraction kit (QIAGEN DNeasy Blood and Tissue kit, Mississauga, ON, Canada) and sample concentration was evaluated via fluorometry (Qubit 2.0, Thermo Fisher Scientific, Mississauga, ON, Canada). Equimolar amounts of DNA from cattle in each group were pooled to obtain a final concentration of 1 µg of DNA in 50 µl of low-EDTA buffer TE. Each pool of DNA was acoustically sheared to a target range of 600 bp (peak incident power: 50, duty factor: 20, cycles per burst: 200, time: 29s) (Covaris M220, Woburn, MA, USA).

Following acoustic shearing, each pool of DNA underwent end repair, A-tailing, and adapter ligation (including a unique index) using a KAPA Library Preparation Kit for Illumina Platforms (KAPA Biosystems, Wilmington, MA, USA) as per the manufacturer's instructions, with the following exception: a single cleanup step was performed following adapter ligation. The cleanup was performed by adding 0.6X PEG/NaCl SPRI solution and only the DNA bound to the magnets was retained. The pools were then combined in equimolar amounts to create a single sequencing library.

Target enrichment was performed using a SeqCap EZ Developer Enrichment Kit as per the SeqCap EZ Library SR User's Guide v.4.2. Target regions consisted of the collagenous lectins and related *MASP* genes, and were based on coordinates obtained from the UMD3.1.1 (bosTau8) genome (Zimin et al., 2009) hosted by the University of Santa Cruz, California (Karolchik et al., 2004) (Table 3.3). Up to 50 kb upstream and 3 kb downstream of each gene was targeted for sequencing, in an attempt to capture a portion of regulatory DNA. Due to some inconsistencies in the annotation of the bovine collagenous lectins, annotations from NCBI and Ensembl were compared and reviewed, and the most appropriate annotation for each gene was

chosen. For example, the *MBL1* gene is not annotated in Ensembl or UCSC, while the *SFTPA1* gene, located nearby on the same chromosome, has four annotated transcript variants in Ensembl. The NCBI RefSeq accession for *MBL1*, NM_001010994.3, is identical to the Ensembl *SFTPA1* transcript ENSBTAT00000001165, leading to some uncertainty surrounding the true identity of these transcripts. Alignment of the four bovine *SFTPA1* transcripts to the coding and protein sequences *SFTPA1* and *MBL1* from other species using Clustal Omega (Goujon et al., 2010; Sievers et al., 2011) showed that ENSBTAT00000031298 and ENSBTAT00000001165 had the highest percent identity matrices to *SFTPA1* and *MBL1*, respectively. Thus, in contradiction to the annotation found in Ensembl, ENSBTAT00000001165 was considered to represent the bovine *MBL1* gene, while only the ENSBTAT00000031298 transcript was considered to represent the *SFTPA1* gene. Similarly, we believe the gene annotated as *FCNB* in Ensembl and *FCN1* in NCBI more closely resembles *FCN1*, and is referred to as such in this study. Coordinates for *COLEC10* and *MASP2* were adjusted slightly based on sequence homology to other species in order to ensure they contained start codons. *CL43* was adjusted to correspond to the findings of Hansen et al. (2003). A complete list of the coordinates for gene annotation in this study is provided in Appendix 3.1.

Following enrichment, the library was sequenced using MiSeq Reagent Kit v3 (600-cycle) sequencing chemistry on an Illumina MiSeq (San Diego, California, USA). In order to achieve adequate depth of sequencing, the same library was sequenced twice. All statistical analyses were performed in R unless otherwise specified (R Core Team, 2017).

Table 3.3. The regions targeted for resequencing are given for each gene included in the study. Genes in close proximity to each other were sequenced as a single unit.

Name	Ensembl ID	Chr	Target start	Target end	Total bp
MASP1	ENSBTAG00000012467	1	80,546,924	80,652,367	105,443
COLEC11	ENSBTAG00000016225	8	112,860,707	112,896,491	35,784
FCN1	ENSBTAG00000048155	11	106,773,026	106,834,643	61,617
COLEC10	ENSBTAG00000017343	14	47,260,662	47,359,061	98,399
MASP2	ENSBTAG00000012808	16	43,449,518	43,481,362	31,844
COLEC12	ENSBTAG00000007705	24	35,627,928	35,866,269	238,341
MBL2	ENSBTAG00000007049	26	6,294,785	6,351,912	57,127
CGN1	ENSBTAG00000006536	28			
CL46	ENSBTAG00000048082	28	35,541,900	35,726,104	184,204
CL43	ENSBTAG00000047317	28			
SFTPD	ENSBTAG00000046421	28			
MBL1	ENSBTAT00000001165*	28	35,764,587	35,870,565	105,978
SFTPA1	ENSBTAT00000031298*	28			

* Both of these transcript IDs are from the gene ENSBTAG00000023032 (*SFTPA1*). The transcript given for *MBL1* is identical to the NCBI RefSeq accession for *MBL1* (NM_001010994.3), and percent identity matrices comparing the coding and protein sequences of these accessions to *MBL1* and *SFTPA1* in other species supports their identities as we have determined them based on *in silico* analysis.
Abbreviations: Chr: chromosome

3.3.2 Bioinformatic Analysis of NGS Data

The sequencing data was processed in two stages (Figure 3.1). In the first stage, data from each run was processed separately, while in the second stage, data from the same pool from different runs was merged and then further processed. The data from both sequencing runs was first trimmed using Trimmomatic v. 0.36 (Bolger et al., 2014) based on the following criteria: a) leading and trailing bases with a quality score of less than 20 were trimmed; b) reads were trimmed if quality dropped below an average score of 20 over a 5 bp sliding window; and c) reads were dropped if they were less than 75 bp in length. Reads were then mapped to the bovine genome UMD3.1.1 using BWA-MEM algorithm of BWA v. 0.75 (Li and Durbin, 2009). PCR and optical duplicates were removed with Picard v. 1.127 (http://broadinstitute.github.io/picard/, accessed 2018-01-12). The Genome Analysis Toolkit (GATK) Best Practices Guidelines (DePristo et al., 2011; van der Auwera et al., 2013) were followed for in/del realignment and base quality score recalibration (BQSR) using GATK v. 3.6 (McKenna et al., 2010). At this point, BAM files for each pool generated during the different runs were merged. Each merged BAM file was reprocessed for PCR duplicates and in/del realignment. Variant calling was performed on merged BAM files using the joint genotyping protocol outlined in the GATK Best Practices guidelines. Variants were filtered using separate hard filters for SNVs and in/dels. Multiallelic variants and spanning deletions were excluded, and known variants were obtained from dbSNP v. 150 (Sherry, 2001). Evaluation of target capture between the Non-infectious and Infectious populations was performed by comparing the overall mean of the mean of each pool within each population using a 2-way ANOVA and a least-square means post-hoc test.

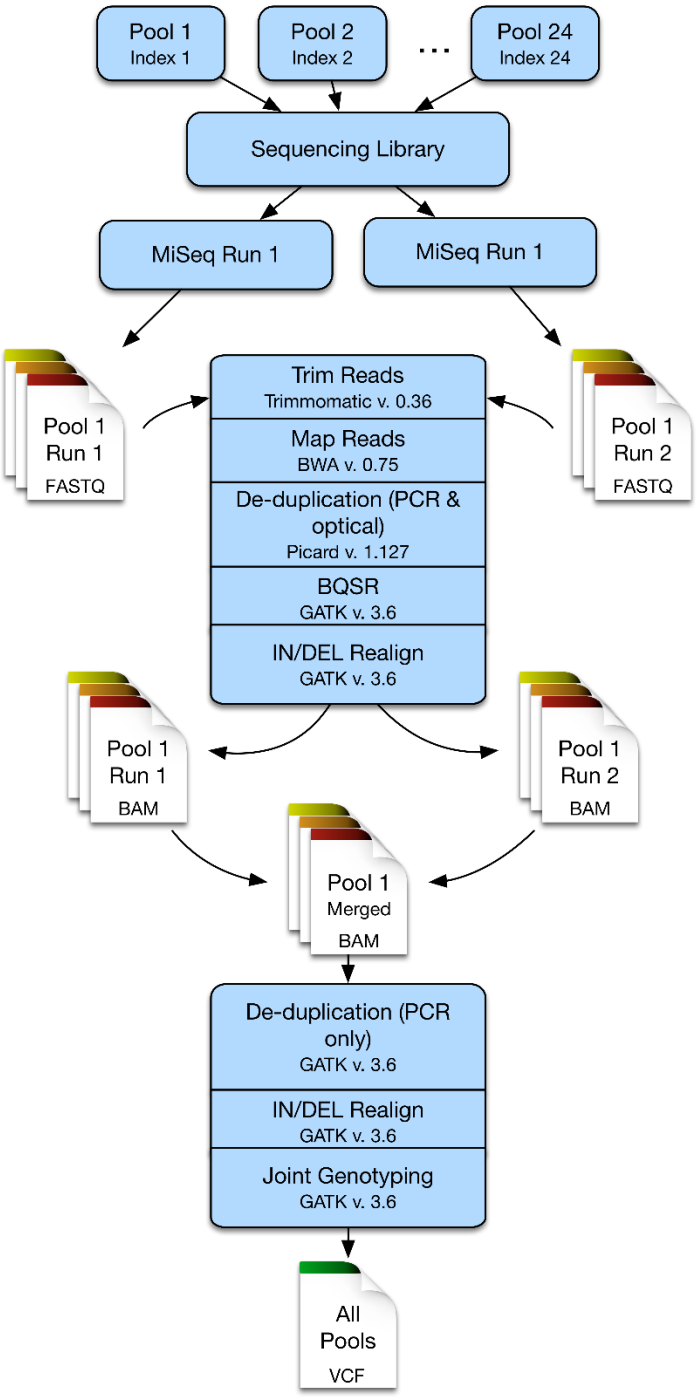


Figure 3.1. Outline of the bioinformatic steps used to call variants following the two sequencing runs.

3.3.3 Variant Analysis

In silico analysis of the variants was performed for coding region variants, downstream 3 kb, and the upstream 5 kb of potential regulatory DNA. Variant density was evaluated both in terms of target regions and functional genomic regions using a 1-way ANOVA and Kruskal-Wallis post-hoc test. Correlations between variant density and GC content as well as indel number were assessed using a Pearson's correlation coefficient. Missense coding variants were analyzed using Polyphen2 (Adzhubei et al., 2010) and the SIFT algorithm (Sim et al., 2012) run by the Variant Effect Predictor (VEP) hosted by Ensembl (McLaren et al., 2010). For Polyphen2, batch submission was used and a FASTA file containing the protein sequences was submitted. For VEP, options were left at their default settings.

The 3 kb downstream from the stop codon of the targeted genes was analyzed for potential miRNA recognition elements (MREs). Although the 3′ UTR was annotated within Ensembl for all genes with the exception of *COLEC10* and *SFTPD*, we opted to analyze the entire 3 kb for each gene to maintain consistency. Multiple MRE discovery algorithms were used to maximize the accuracy of the predictions (Riffo-Campos et al., 2016). miRanda v. 3.3a (Enright et al., 2003) and Targetscan v. 7.0 (Agarwal et al., 2015) were used to identify MREs binding mature miRNA sequences from cattle accessed from miRbase 21 (Kozomara and Griffiths-Jones, 2014). The energy threshold parameter of miRanda was set to -20 kcal/mol, with other parameters for both programs left at their default settings. The intersection of the seed region of predicted MREs from both programs was obtained using BEDtools v. 2.25.0 (Quinlan and Hall, 2010). BEDtools was then used to identify variants in our dataset that intersected with the seed sequence of MREs predicted by both algorithms.

Transcription factor (TF) binding site (TFBS) analysis was performed on the 5 kb upstream to the start codon for each target gene using CIS-BP v. 1.02 (Weirauch et al., 2014). The species was set as *Bos taurus*, and a motif model of "PWM - LogOdds" with a minimum threshold of 8 was selected; only *cis* acting TFBSs were considered. Variants falling within predicted TFBSs were identified using BEDtools. Many of the TFBSs shared identical sequences and bound TFs belonging to the same family, and were thus collapsed into a single result with results reported by TF family. Putative TFBSs were further refined by identifying conserved 50-bp long DNA sequence motifs within the 5 kb upstream from the start codon for each gene across 8 different species, cattle (UMD3.1.1), pig (Sscrofa10.2), horse (EquCab2), rat (Rnor_6.0), mouse (GRCm38), gorilla (gorGor3.1), chicken (Galgal4), and human (GRCh38), using MEME v 4.10.1 (Bailey and Elkan, 1994). The bovine specific genes *CGN*, *CL43*, and *CL46* were aligned to the sequences of *SFTPD* from other species, as they are believed to be evolutionarily related (Gjerstorff et al., 2004b; Hansen and Holmskov, 2002). A minimum E-value of 0.05 was used to consider a motif conserved. The conserved motifs were then examined for the presence of TFBSs predicted by CIS-BP and containing variants.

3.3.4 Allelic Association

The estimated frequency of variant alleles was compared between the Non-infectious and Infectious populations using Popoolation2 (Kofler et al., 2011). Processed BAM files for each pool were combined using Samtools into a single BAM file for each population, and an mpileup file was generated using minimum mapping and base qualities of 20 (Li et al., 2009). Popoolation2 was used to transform the mpileup file into a sync file, which was then down-sampled according to the recommendations of Popoolation2 to 400 reads per population using the method "fraction" to mitigate the impact of variable read depths on statistical testing.

Fisher's exact test was used to determine the significance of allele frequency estimates between the two populations. A minimum of 5 % of the reads (20) across both populations combined was required for the allele to be considered in the allele frequency estimation. The Benjamini-Hochberg procedure was used to correct for multiple testing (Benjamini and Hochberg, 1995) with adjusted p-value cut-offs labeled as described in the BADGE system, which provides recommendations on minimal p-values for significance in allelic association studies (Manly, 2005).

3.4 Results

Evaluation of target capture efficiency showed that the median of the mean target coverage was 172.5, however there were significant differences ($p < 0.05$) both in terms of the depth of coverage for each target region, as well as the total depth of coverage between the Non-infectious and Infectious populations (Figure 3.2). The depth of coverage between the two populations at individual target regions was not statistically different.

Joint genotyping identified a total of 5439 unique variants, all of which were present in both the Non-infectious and Infectious populations. These included 5418 SNVs and 21 in/dels. The majority (5317, 97.7 %) of the variants identified were present in dbSNP 150, while 122 were novel discoveries. A further 32,794 variants were present in dbSNP 150 within our target intervals. Of these, 2922 were present in our population but were excluded due to various filtering parameters, 29,696 loci were not variant in our population, and 356 loci were not successfully sequenced.

A total of 83 coding variants (40 synonymous, 43 missense), 2297 intronic variants, 309 variants within the downstream 3 kb of the stop codon, 414 variants within the 5 kb upstream of the start codon, and 2915 variants in the region between 5-50 kb upstream of the start codon

were identified. The density of variants was examined both by gene and by region (Figure 3.3). Significant differences ($p < 0.05$) in variant density were observed between genes, with the highest density of variants found in the *FCN1* gene and surrounding DNA (Figure 3.3a). Variant density by region was not significantly different, though the coding region tended to have the lowest density (Figure 3.3b).

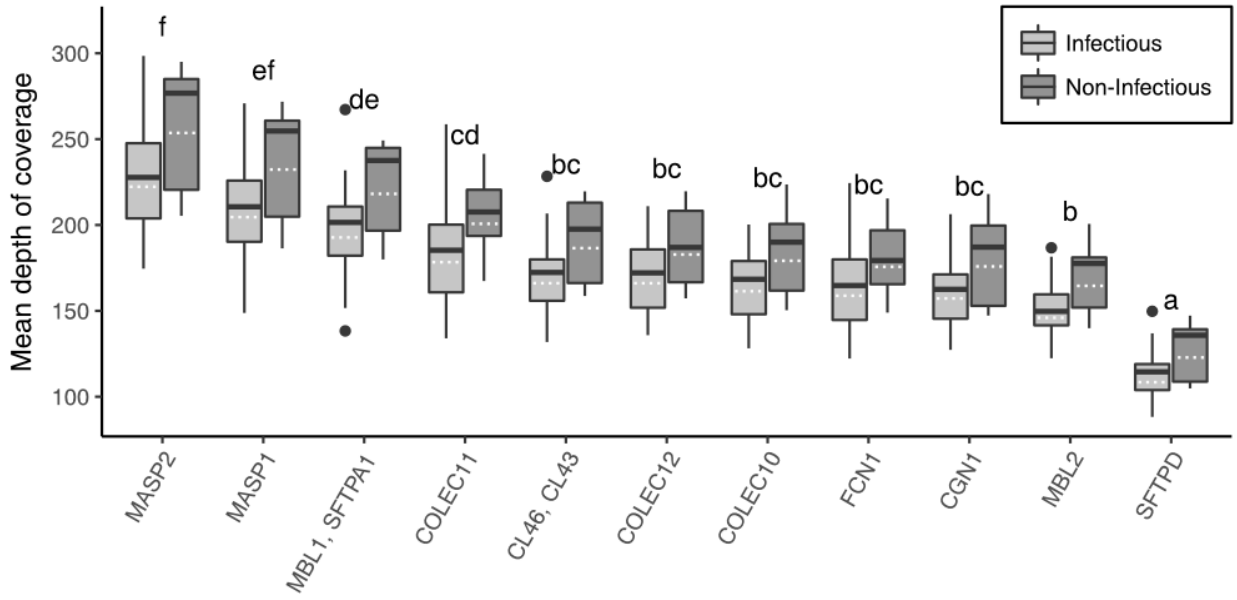


Figure 3.2. The overall mean of the mean depth of coverage for each pool in the Non-infectious and Infectious populations by gene target.

There were no significant differences between populations in a gene target, however there was a significant difference in depth of coverage between different gene targets (2-way ANOVA and least-squares means post-hoc comparison, $p < 0.05$). Genes that share a letter are not significantly different: for example, *MASP2* is not significantly different from *MASP1*, but is significantly different from the remaining gene targets. Solid line represents the median, dashed line is the mean. The hinges represent the 1st and 3rd quantiles, while the whiskers represent 1.5x the interquartile range. Data points beyond this range are illustrated as solid circles.

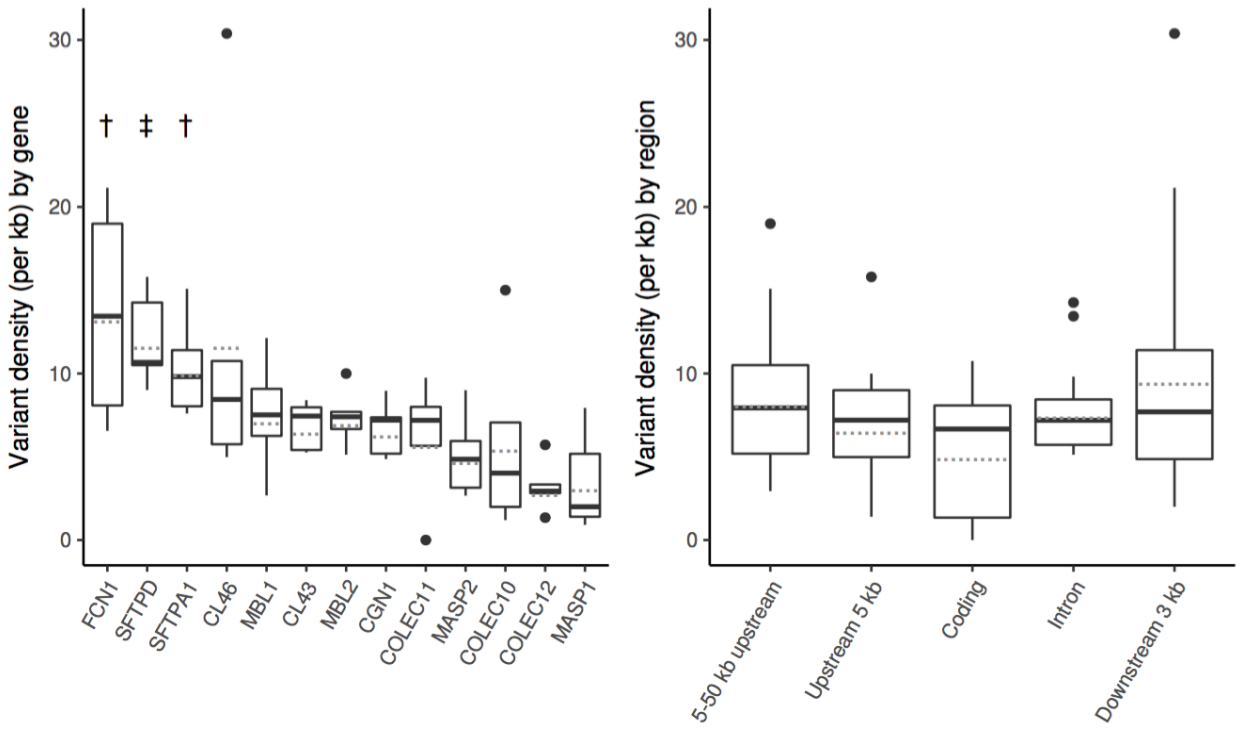


Figure 3.3. Variant density in cattle by gene and gene region.

a Variant density across the entire study population was significantly different amongst the targeted genes. †: Variant density in these genes was significantly different from *COLEC12* and *MASP1*. ‡: Variant density was significantly different from *MASP2*, *COLEC10*, *COLEC12*, and *MASP1* (1-way ANOVA and Kruskal-Wallis post-hoc test, $p < 0.05$). **b** Variant density between functional genomic regions was not significantly different. Note that due to the proximity of some genes (e.g. *MBL1* and *SFTPA1*), some variants occurring downstream, upstream, or within the introns and coding regions of *SFTPA1* were also counted as being upstream from *MBL2*, thus, the total of variants by region is different than the total of unique variants discovered. Solid line represents the median, dashed line is the mean. The hinges represent the 1st and 3rd quantiles, while the whiskers represent 1.5x the interquartile range. Data points beyond this range are illustrated as solid circles.

Amongst the coding variants, 43 missense variants were identified. *In silico* predictions of the effects of the mutations using either algorithm identified a total of 11 SNVs expected to have damaging or possibly damaging effects on protein structure or function (Table 3.4), with three of the variants (within *SFTPD*, *MBL2*, and *CL43*) predicted to be damaging by both algorithms.

A total of 20,147 potential *cis* TFBSs were predicted in the upstream 5 kb of the targeted genes. Of these, 1351 contained a SNV identified in this study. In highly conserved DNA sequence motifs (based on our multispecies comparison of upstream regulatory regions, Appendix 3.2), 148 TFBSs containing a SNV were found across 10 of the targeted genes (Figure 3.4a). These TFBSs were members of 30 TF families (Figure 3.4b).

Analysis of the 3 kb downstream from the stop codon using both algorithms identified 469 potential miRNA recognition elements. Within the seed region of the predicted MREs there were 31 SNVs (Figure 3.5), two of which impacted two separate miRNAs. A total of 28 unique miRNAs were predicted to bind in these regions, 5 of which bound multiple loci. A single MREs intersected with the annotated 3′ UTR of *SFTPA1*.

Evaluation of the frequency of variant alleles identified 25 BADGE class I ($p < 2 \times 10^{-7}$) and 49 class II ($p < 5 \times 10^{-6}$) variants that were significantly associated with either the Non-infectious or Infectious populations (Figure 3.6a). Seventeen associations were found clustered in intron 2 of *MASP1* (Figure 3.6b), and a further 48 associations were present in a ~21 kb region ~29 kb upstream from *CGN1*, the first gene of the bovine collectin locus (Figure 3.6c). Four associations were found distributed upstream, downstream, and within intron 8 of *FCN1* (Appendix 3.3a); 4 associations were found in the introns 4 and 5 and the downstream region of

COLEC11 (Appendix 3.3b); and a single association was found in intron 2 of *COLEC12* (Appendix 3.3c).

Table 3.4. Potentially deleterious missense mutations as determined by in silico analysis with two different prediction algorithms

Gene	Chr	Position	rsID	Ref	Alt	Protein change	Protein domain	Polyphen2 Prediction	Polyphen2 Score	SIFT Prediction	SIFT Score
FCN1	11	106,827,968	rs382216843	C	T	Arg142Cys	FBG	probably damaging	0.986	tolerated	0.06
MBL2	26	6,344,919	rs210611099	C	A	Pro42Gln	CLD	probably damaging	0.974	deleterious	0
SFTPD	28	35,820,078	rs380240341	C	T	Pro132Ser	CLD	probably damaging	0.958	deleterious	0.05
CGN	28	35,598,640	rs208842091	G	A	Arg173His	CLD	possibly damaging	0.72	tolerated	0.16
CL46	28	35,675,371	rs383278255	C	T	Pro185Leu	CLD	possibly damaging	0.672	tolerated	0.32
CL43	28	35,718,034	rs42967143	A	G	Thr117Ala	CLD	possibly damaging	0.659	tolerated	0.9
CL43	28	35,718,807	rs211678602	G	T	Gln185His	neck region	possibly damaging	0.634	deleterious	0.01
MASP2	16	43,463,621	rs207667073	G	A	Gly102Ser	CUB domain	benign	0.191	deleterious	0.01
FCN1	11	106,828,710	rs385211468	C	T	Thr193Met	FBG	benign	0.042	deleterious	0.03
CGN	28	35,602,463	rs466869949	A	C	Glu302Asp	CRD	benign	0.036	deleterious	0.04
SFTPD	28	35,824,136	rs110476851	C	G	Ala288Gly	CRD	benign	0.002	deleterious	0.02

FBG: Chr: chromosome; Ref: reference allele; Alt: alternate allele; SIFT: Sorting Intolerant from Tolerant algorithm; FBG: fibrinogen-like domain; CLD: collagen-like domain; CUB: complement C1r/C1s, Uegf, Bmp1; CRD: carbohydrate recognition domain.

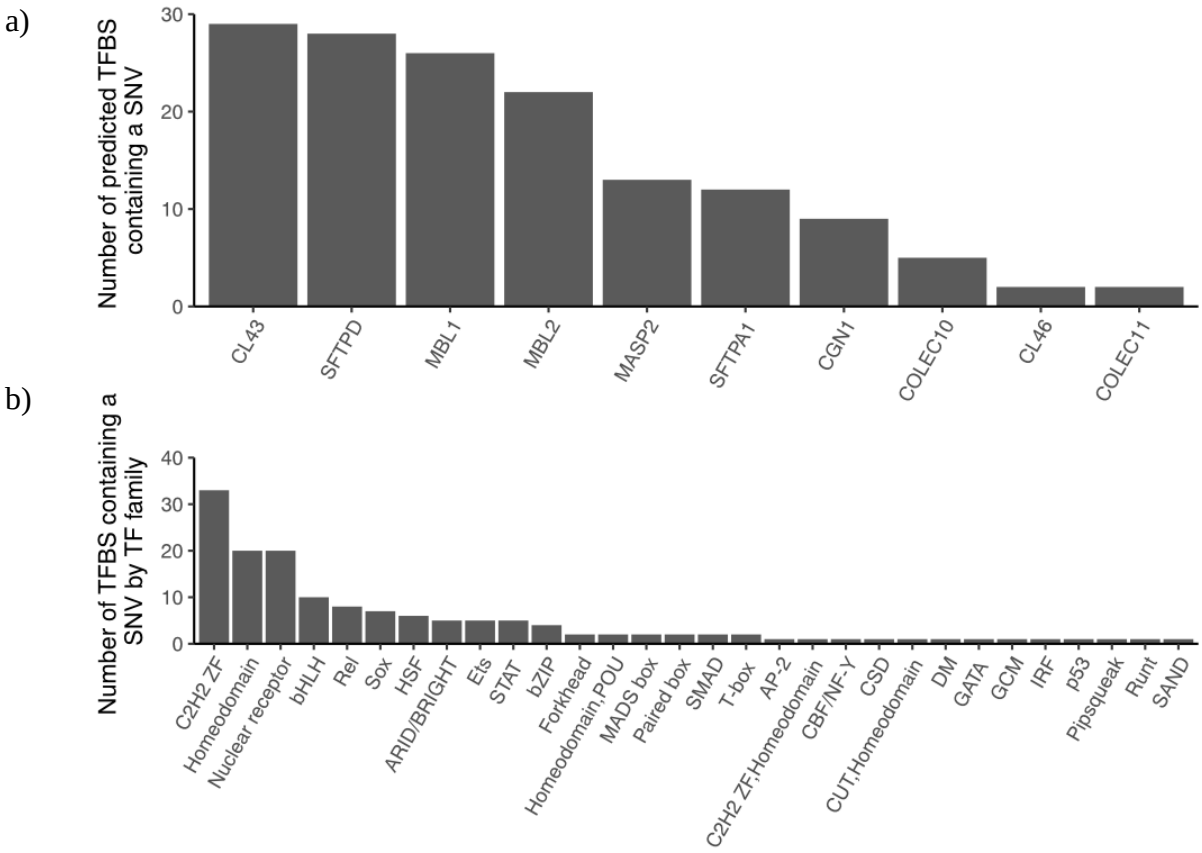


Figure 3.4. TFBS potentially impacted by SNVs.

a) The number of predicted transcription factor binding sites within conserved DNA sequences of the targeted genes that contained a SNV identified in this study. b) Number of predicted transcription factor families containing a SNV identified in this study.

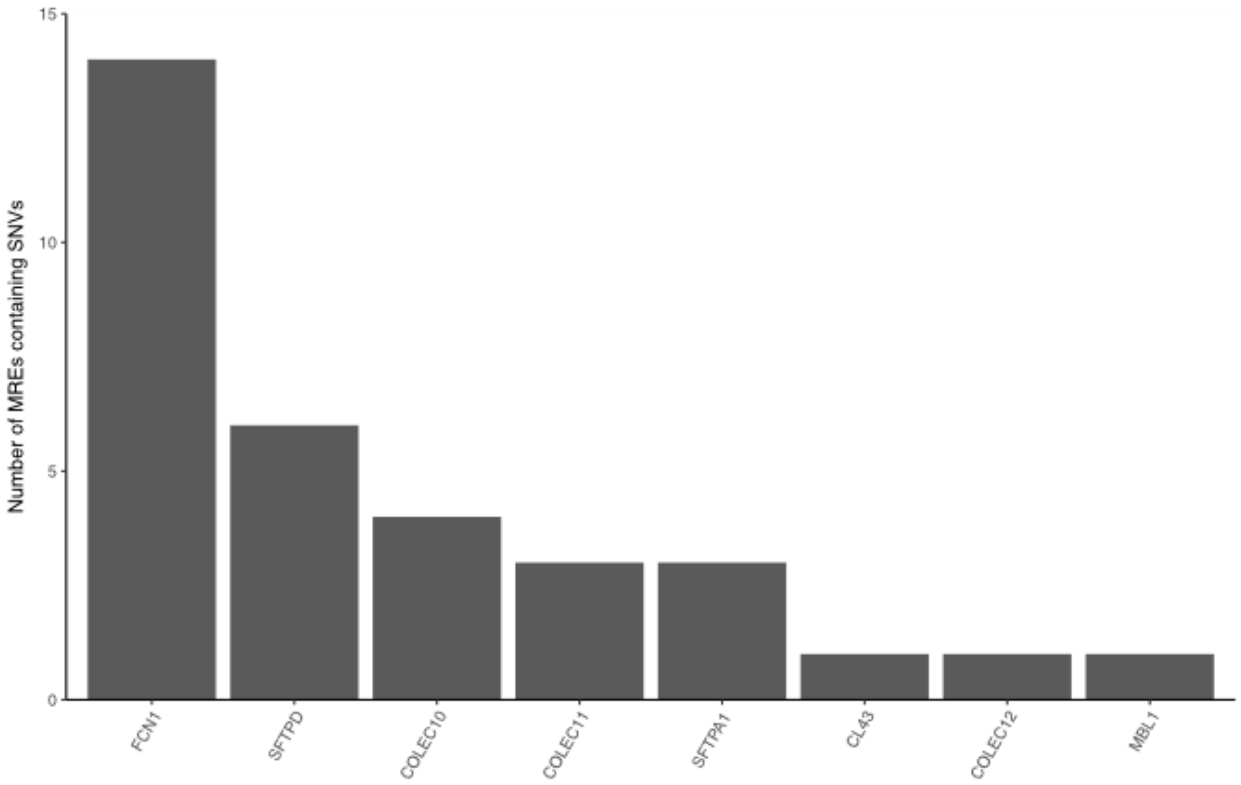


Figure 3.5. The number of in silico predicted miRNA recognition elements found in the targeted genes.

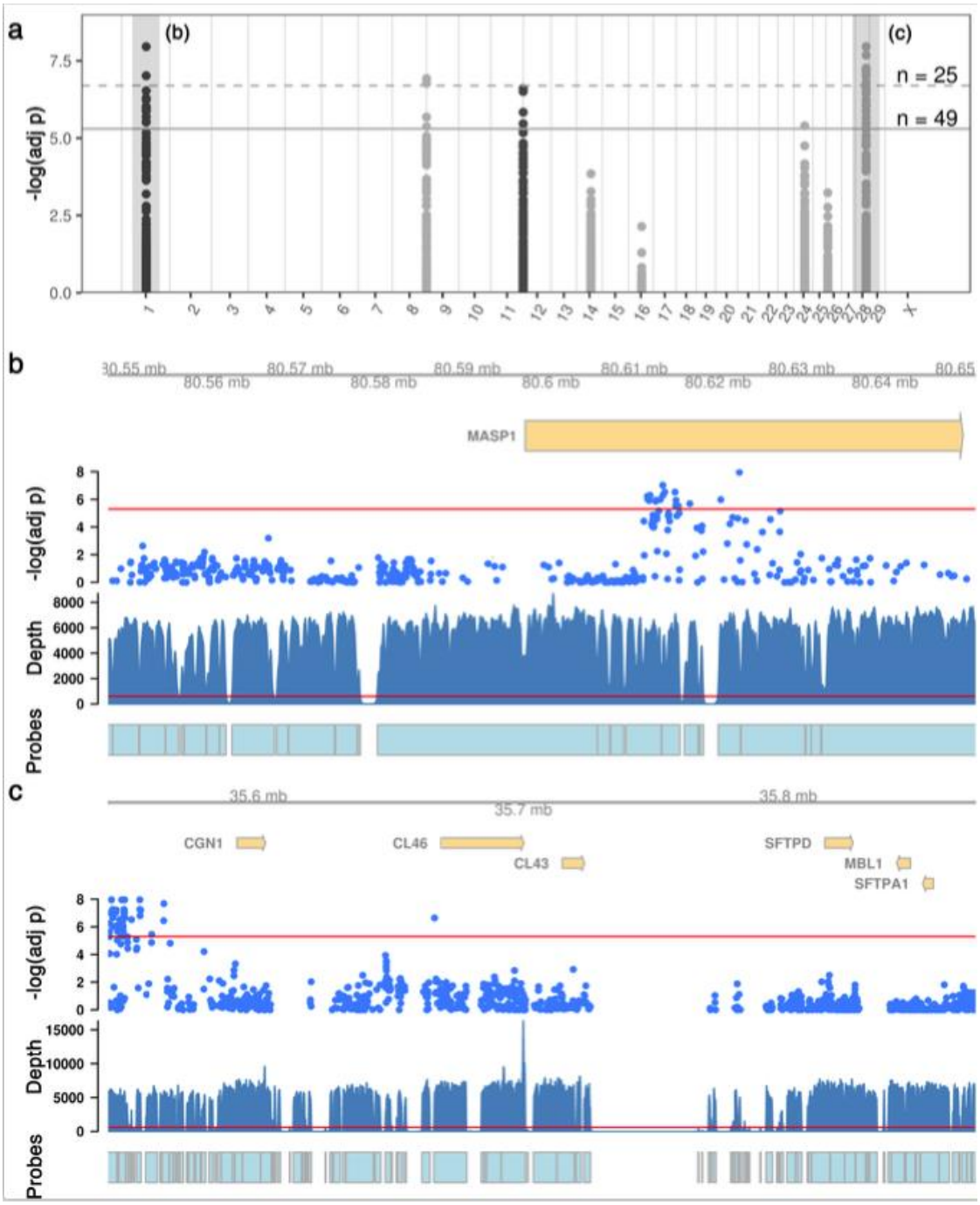


Figure 3.6. Manhattan plot of the variants identified in the infectious and non-infectious pools of cattle.

a Manhattan plot of the allelic association analysis identified 25 variants significant at BADGE class I and 49 variants at class II (Manly, 2005) in 5 different targeted regions. The p value has been adjusted for multiple testing using the Benjamini-Hochberg procedure. **b,c** The areas highlighted in **a** are shown in greater detail. Significant associations were found within the intron of *MASP1* (**b**) and upstream of the collectin locus (**c**). The red line within the Manhattan plot indicates the cutoff for class II significance ($p < 5 \times 10^{-6}$). The depth of sequencing (total from all pools) and the regions with probes designed for target capture are shown to illustrate gaps in sequencing and variant discovery. Red line within the depth indicates the minimum depth required for variant calling.

3.5 Discussion

The overall objective of this study was to characterize the genetic variation in the bovine collagenous lectin genes. Variation in certain collagenous lectins have been shown, both in cattle and other species, to be associated with increased susceptibility to infectious diseases. Associating polymorphisms with infectious disease susceptibility has important implications in agricultural economics, animal breeding, and animal health and welfare. Probing the underlying genetics of complex traits like innate immunity is challenging, and requires large numbers of animals with well-defined phenotypes (Ron and Weller, 2007). The phenotypes of our study samples were determined through complete autopsies supervised by certified veterinary pathologists. This included evaluation of gross tissues, ancillary testing as required, and review of histopathological specimens to define the extent and nature of disease. In order to address the considerable expense required to sequence large numbers of animals, we opted to use a pooled and targeted next-generation sequence approach. Pooled sequencing has been shown to be an accurate and cost-effective method of variant discovery and allele frequency estimation in next-generation sequencing experiments (Bansal et al., 2011; Bertelsen et al., 2016; Mullen et al., 2012; Rellstab et al., 2013), as well as in genome-wide association studies (Keele et al., 2015). While whole genome sequencing of larger groups of disease-specific phenotypes would provide a more complete picture of disease associated variants, the cost remains prohibitive, despite decreasing sequencing costs. Instead, targeted re-sequencing allows specific regions of interest in the genome to be queried, providing much more detail than the available high-density SNP array from Illumina: of the 5439 variants found in our study, only 223 are present in the Illumina BovineHD BeadChip (https://support.illumina.com/array/array_kits/bovinehd_dna_analysis_kit/downloads.html, accessed 2018-01-12).

Allele frequency estimation identified 74 significant variants (BADGE class II or higher), after correction for multiple testing, that were associated with infectious disease. Over half of the significant alleles occurred in a 29 kb segment of DNA upstream from *CGN1*, and these alleles met the highest level of statistical significance proposed by the BADGE system (Manly, 2005). *CGN1* is the first gene in the bovine collectin locus, a 260 kb region on chromosome 28 which includes the *CL46*, *CL43*, *SFTPD*, *MBL1*, and *SFTPA1* genes (Gjerstorff et al., 2004a), and several of these genes have been implicated in infectious disease susceptibility. Plasma concentration of conglutinin is known to be heritable and low levels of plasma conglutinin are associated with increased incidence of respiratory disease in cattle, though whether it is the cause or a consequence (e.g. protein consumption) of respiratory disease is unknown (Holmskov et al., 1998). Whole transcriptome sequencing of the abomasum of parasite-susceptible and resistant calves found that resistant animals expressed higher levels of *CL46* (Li et al., 2011). In cattle breeds native to China, certain *MBL1* haplotypes are associated with somatic cell score, an indirect marker of mastitis (C. Wang et al., 2011). The block of highly significant associations discovered here could theoretically impact any of the genes within the collectin locus, as DNA kilobases to megabases upstream of genes can have a regulatory impact through enhancer, silencer, insulator, or locus control region elements (Maston et al., 2006) that are difficult to predict, and for which the bovine genome lacks annotation. The relative similarity of the allele frequencies in the Non-infectious and Infectious populations was similar for all of these variants, suggesting that they are likely in linkage disequilibrium (Appendix 3.4). The pooled design of this study precludes haplotype analysis so further study of these variants, and the genes of the nearby collectin locus, both in terms of RNA expression and epigenetics, is warranted.

Two class I and 15 class II associations were found in intron 2 of *MASP1*. Human *MASP1* encodes three protein isoforms with distinct functions, MASP-1, MASP-3, and MAP1 (Beltrame et al., 2015). Only a single transcript is annotated for bovine *MASP1*, which best corresponds to the MASP-3 isoform, and an additional 4 transcripts are predicted by NCBI. As with human *MASP1*, the transcript and predicted transcripts that encode the three *MASP1* isoforms share the first 8 exons; thus, the cluster of variants noted in intron 2 is also present in intron 2 for all of these predicted transcripts. Genetic mutations resulting in human *MASP1* deficiency are associated with infectious disease, and several intronic mutations leading to *MASP1* deficiency have been identified, though none are identical to the associations found here (Beltrame et al., 2015; Ingels et al., 2014). Again, the relative similarity of the allele frequencies for the different loci suggests significant linkage disequilibrium. It should be noted that mutations that inhibit the function of the MASPs (and complement-activating collagenous lectins) may also confer benefits: excess activation of complement can contribute to tissue damage or auto-immune disease, and decreased or more moderate levels of complement activity may therefore be beneficial in some scenarios (Beltrame et al., 2015).

A further four associations were found in *COLEC11* and two in *COLEC12*, however no *in silico* consequences were predicted for any of the six. There is little known about the relative importance of CL-K1 and CL-P1, the proteins encoded by these two genes, in the innate immune response to infectious disease, and, to the authors knowledge, these are the first reported associations between mutations in these genes and infectious disease of cattle. Recently, CL-P1 was shown to have a soluble form that can activate the alternative pathway of complement (Ma et al., 2015), and CL-K1 is capable of activating the lectin pathway of complement (Ma et al.,

2013), presenting possible pathways through which genetic mutations could hamper the immune system.

Although *in silico* analysis of the disease-associated alleles did not identify any biological effects, several interesting predictions were made regarding other variants present in our dataset. Two missense mutations found in *MBL2* and one in *MASP2* (Table 3.4) were previously shown to be associated with SCS in Chinese Holstein cattle (Wang et al., 2012). The frequency of these three variants was not significantly different in our populations of Infectious and Non-infectious cattle. Variant rs210611099:c.125C>A was rare, with only 8 alleles predicted across both populations (MAF 3.3 %). Though rare, this is substantially higher than the reported allele frequency of 0.27 % in Chinese Holstein cattle (Wang et al., 2012), and may be the result of different breeds (Chinese Holsteins versus the mixture of breeds common to North America in our study). The mutation occurs in the collagen-like domain (CLD) of MBL-C, and is predicted *in silico* to have a significant impact on protein structure and function. Previously reported mutations in the CLD had an impact on MBL-C driven complement activation (Larsen et al., 2004), as well as on higher-order oligomerization through disruption of the Gly-X-Y collagen-like repeats (Sumiya et al., 1991). Thus, despite its rarity, the role of this variant in infectious disease susceptibility remains of interest. Genotyping of a larger number of animals may be useful in clarifying the discrepancy between studies in allele frequencies, and may provide the statistical power required to determine whether this rare variant plays a role in bovine innate immune defense.

The second *MBL2* missense variant associated with SCS, rs210426415:c.92G>A, leads to the substitution of glutamine for arginine in the N-terminal domain of MBL-C. This domain is believed to utilize conserved cysteine residues to facilitate the functionally critical

oligomerization of MBL-C (Wallis and Drickamer, 1999). This variant was predicted by both *in silico* algorithms to have a low impact on protein structure and function, and was not associated with disease in our study, possibly the result of a smaller cohort of animals diagnosed with mastitis. Thus, although genotyping of larger numbers of North American cattle may reveal an association with infectious disease, this study does not provide evidence for further investigation.

The *MASP2* variant previously shown to be associated with SCS, rs207667073.G>A, results in an amino acid change in the CUB domain that is predicted by the SIFT algorithm, but not Polyphen2, to be deleterious to protein structure or function. This allele showed no significant difference between the Non-infectious and Infectious populations. Again, genotyping of larger number of North American cattle may prove useful in further defining the role of this variant.

Only one variant, rs381773088 in *SFTPA1*, was present within an annotated 3′ UTR and intersected with a predicted MRE for bta-miR-328. To our knowledge, there are no published studies on the role of bta-miR-328 in cattle; however, a study on the human homologue demonstrated that it plays a role *in vitro* in the innate immune defense against *Haemophilus influenza* through negative regulation of phagocytosis (Tay et al., 2015). Gram-negative pulmonary pathogens related to *H. influenza*, notably *M. haemolytica*, *P. multocida*, and *H. somni*, are important causes of respiratory disease in cattle; thus, this predicted MRE affected by a variant may hold relevance for future investigations.

In silico prediction of transcription factor binding sites relies on the observation that the amino acid sequence of the DNA-binding domain of transcription factor proteins largely predicts their DNA-binding specificity, and does so in a highly conserved manner (Jolma et al., 2013; Kasahara et al., 2006; Weirauch et al., 2014). Regulatory DNA is also highly conserved in

animals (Nitta et al., 2015). Thus, to reduce the large number of predicted TFBSs identified in the targeted genes, conserved DNA sequences present in the potential promoter region were identified by comparing sequences from up to 8 different animals including domestic livestock and more distantly related species (human, gorilla, cattle, horse, pig, rat, mouse, and chicken). This conservative approach narrowed the results to 148 potential TFBSs, some of which were similar to TFBSs predicted or shown in previous studies to be involved in the regulation of collagenous lectins. For example, we identified binding sites for SRF (a member of the MADS box family of transcription factors) in *SFTPA1*, and an SRF binding site was also identified in human *SFTPA2* (Grageda et al., 2014). Initial characterization of the promoter of *CL43* included *in silico* prediction of TFBSs (Hansen et al., 2003a), and the same binding sites were predicted in our study for Myb, ARNT, cEBP, Myc, MyoD, Mzf-1, N-Myc, and USF, however, only Mzf-1 was both present within a conserved motif and contained a SNV. This discrepancy may be the result of more stringent requirements in our study (including exclusion of *trans* binding sites and sites outside of conserved sequences), or potentially due to differing prediction algorithms. HNF3alpha (also known as FOXA1) is known to regulate the expression of human and chicken *MBL2* (Kjærup et al., 2013; Naito et al., 1999), and while binding sites were predicted in *CL43* and *COLEC11*, no binding site was predicted for bovine *MBL2*.

The density of variants varied significantly by gene, but not by gene region (Figure 3.3). *FCN1* and the two surfactant protein genes, *SFTPA1* and *SFTPD*, had a significantly higher variant density than the bottom two (for *FCN1*) and four (for *SFTPA1* and *SFTPD*) genes. Both intrinsic and extrinsic factors can affect the degree of sequence variation. Intrinsically, the GC content and the number of indels in a region are predictive of variation (Hodgkinson and Eyre-Walker, 2011). The GC content of the *FCN1* gene was the second highest of the targeted genes

at 55.1%, however, the overall correlation between GC content and the density of variants was not significant (R = 0.36, p = 0.23, Appendix 3.5). Although the *FCN1* target contained the highest number of indels (13/21), there was no correlation between number of indels and variation density (R = 0.36, p = 0.23, Appendix 3.6). Extrinsically, genes that interact with the environment (such as innate immune response genes) can often be found in 'hot spots' in the genome, which show higher levels of variation in response to increased adaptive pressures (Chuang and Li, 2004). Interestingly, a recent study on the equine collagenous lectins found that two equine ficolin genes, *FCN1* and *FCN1-like*, also had the highest variant density amongst the equine collagenous lectins (Fraser, unpublished). Five of the eight coding region mutations found in *FCN1* were located within the fibrinogen-like domain, and three were missense. Although the relative role of the ficolin genes in the innate immune response of these two species is still under investigation, the increased variation observed in both species might suggest increased evolutionary pressure imposed by different pathogens.

Here, we have documented the variation in a subset of innate immune genes, the collagenous lectins, in cattle with and without infectious diseases. Comparison of mutant alleles in the two populations identified 74 alleles associated with infectious disease. These alleles warrant further investigation, both in terms of population level frequencies, and, if confirmed to be significantly associated with disease susceptibility, in terms of their biological mechanisms of action.

3.6 Supporting Information

The variants identified in this study have been submitted to the European Variant Archive (EVA) accession PRJEB24763 and will be released upon publication.

Chapter 4: Identification and Frequency of Local Expression Quantitative Trait Loci in the Innate Immunome of Diseased and Healthy pigs

This corresponds to the following manuscript being prepared for submission:

Fraser, RS, Meyer, A, Snyman, HN, Hayes, MA, Lillie, BN. 2018. Identification and frequency of local expression quantitative trait loci in the innate immunome of diseased and healthy pigs. *Animal Genetics*.

4.1 Abstract

Infectious diseases are a major source of economic loss, antibiotic use, and reduced animal welfare in the swine industry. The innate immune system is an important first line of defence against a variety of different infectious diseases, and some mutations in innate immune genes can impair the ability of pigs to respond to pathogens. Previous work has shown that sequence variants in various genes of the innate immune system are more common in animals suffering from infectious diseases. The objective of this study was to identify sequence variants associated with variability in innate immune gene expression of pigs, and to determine the prevalence of these variants in a population of diseased pigs. Innate immune genes known to have variable hepatic expression were targeted for next-generation sequencing to comprehensively identify mutations in the genes and their surrounding regulatory DNA. Genomic material originated from the liver of 72 healthy market-weight pigs sourced at slaughter, with the same animals used in both the expression microarray and sequencing experiments. After applying quality control filters, 41,331 sequence variants were identified (41,035 single nucleotide variants, 296 indels). Expression quantitative trait loci (eQTL) analysis identified 298 significant local eQTLs associated with 19 different innate immune genes. After linkage disequilibrium analysis, 1013 pigs (592 healthy, 421 with infectious diseases) were genotyped for 74 of the eQTLs. Significant differences in allele frequencies between healthy and diseased pigs were found at 28 loci spread across 8 different genes. These alleles represent significant eQTLs associated with infectious disease.

Keywords: eQTL, porcine, innate immunity, microarray, next-generation sequencing

4.2 Introduction

Genetic variation in the immune system of animals is an important factor in their susceptibility to a variety of infectious diseases. In the swine industry, infectious diseases are a significant source of decreased animal welfare, economic loss, and antimicrobial usage. Mitigating the impact of infectious diseases in pigs through the identification of genetic determinants of infectious disease susceptibility holds potential to reduce these negative effects. Some differences in expression or function of innate immune proteins that interact with pathogens result from genetic variations. We have previously shown that various mutations in a subset of innate immune genes are more prevalent in pigs suffering from infectious diseases (Keirstead et al., 2011). Expression of MBL-C, a collagenous lectin capable of binding to bacterial surfaces and activating the lectin pathway of complement, is reduced by a promoter mutation in *MBL2*, and this mutation was found to be more frequent in pigs with infectious diseases (Lillie et al., 2007). Regulatory variation of this type is proving to be important in the understanding of the genetic basis of disease in humans (Maurano et al., 2012), and it is likely that such mechanisms are similarly important in animals.

Expression quantitative trait loci (eQTL) are genomic regions that carry one or more variants that impact gene expression (Albert and Kruglyak, 2015). eQTLs can be either *cis* or *trans* acting, and can be found close to the genes they influence or distantly. Several eQTLs with disease associations have been described in humans (see review by Albert and Kruglyak (2015)), and the aforementioned porcine *MBL2* promoter mutation is an example of a disease-associated eQTL in pigs. With high throughput technologies, it is now feasible to perform large-scale eQTL studies on the porcine innate immunome. In a previous study, we identified 114 innate immune genes with variable hepatic expression in healthy pigs (Snyman, 2013). Using the genomic material from the same animals, we performed targeted, next-generation re-sequencing on these

genes and their surrounding regulatory DNA and identified sequence variants. We then performed local eQTL analysis and identified variants impacting innate immune gene expression. Finally, we genotyped both healthy pigs and pigs diagnosed with infectious diseases and common swine pathogens and compared the frequency of eQTL alleles in the different populations.

We identified 298 eQTLs that are associated with altered innate immune gene expression, 27 of which are significantly different in healthy pigs and those with infectious disease. This represents the first study of infectious disease associated eQTLs in the porcine innate immunome.

4.3 Materials and Methods

4.3.1 Next-Generation Sequencing

DNA was obtained from samples of liver collected from healthy, market weight, Duroc x Yorkshire x Landrace pigs from a local abattoir as described previously (Snyman, 2013). Briefly, DNA was extracted using a QIAGEN DNeasy tissue kit (QIAGEN Inc, Mississauga, ON, Canada) according to manufacturer's instructions and transferred into low-EDTA buffer TE. The concentration of DNA was fluorometrically determined (Qubit 2.0, Thermo Fisher Scientific, Mississauga, ON, Canada) prior to acoustic shearing to a target range of 300 bp (peak incident power: 50, duty factor: 20, cycles per burst: 200, time: 80s) (Covaris M220, Woburn, MA, USA). A total of 72 pigs representing a range of hepatic innate immune gene expression were chosen for sequencing based on previous results (Snyman, 2013).

A next-generation sequencing library for each pig was prepared with a NEXTFlex Rapid Pre-Capture Combo Kit (BIOO Scientific, Austin, TX, USA) as per the manufacturer's instructions. Innate immune genes previously determined to have variable hepatic expression (n

= 107) (Snyman, 2013) were targeted for resequencing, along with potential regulatory DNA up to approximately 50 kb upstream and downstream of the genes using a custom SeqCap EZ Enrichment kit from Nimblegen (Roche, Madison, WI, USA) (Appendix 4.2). Coordinates were obtained from the Sscrofa 10.2 genome (Groenen et al., 2012) hosted by the University of Santa Cruz, California (Karolchik et al., 2004). Following target capture and enrichment, libraries were sequenced on two lanes of an Illumina HiSeq 2500 (San Diego, CA, USA) using 250 bp PE chemistry at The Center for Applied Genomics (Hospital for Sick Children, Toronto, ON, Canada).

The quality of raw fastq files was assessed using FastQC (Andrews, 2010). Reads were trimmed using Trimmomatic v0.36 (Bolger et al., 2014) based on the following criteria: a) leading or trailing bases below a Phred score of 20 were trimmed; b) reads were trimmed if the average Phred score dropped below 20 over a 5 bp sliding window; and c) reads less than 40 bp in length were dropped. Reads were aligned to the Sscrofa 10.2 genome using the BWA-MEM algorithm of BWA v0.7.5 (Li and Durbin, 2009). PCR and optical duplicates were removed with Picard v1.127. Base quality score recalibration and indel realignment were performed with the Genome Analysis Toolkit (GATK) v3.6 and followed the GATK Best Practices guidelines (DePristo et al., 2011; McKenna et al., 2010; van der Auwera et al., 2013). The GATK joint genotyping protocol for variant calling was followed, using separate hard filters for single nucleotide variants (SNVs) and insertions/deletions. Multi-allelic variants were excluded. Known variants were obtained from the porcine dbSNP build 145 (Sherry, 2001).

4.3.2 Local eQTL Analysis

The NGS results were paired with previously reported gene expression microarray results (Snyman, 2013) for local eQTL analysis. Genetic material originated from the same pigs in both

experiments. The analysis was performed using the Matrix eQTL v2.1.1 (Shabalin, 2012) package for R (R Core Team, 2017). Linear regression assuming an additive model of gene expression was chosen and the maximum distance for local (*cis*) gene-SNP pairs was defined as 50 kb. A minimum of 3 animals of a particular genotype was required for inclusion of that genotype in the analysis. Correction for multiple testing was performed using the Benjamini-Hochberg procedure with a false discovery rate (FDR) of 0.05 (Benjamini and Hochberg, 1995).

4.3.3 Linkage Disequilibrium Analysis

Haploview v4.2 (Barrett et al., 2005) was used to determine whether significant local eQTLs found affecting the same gene were in linkage disequilibrium (LD). Animals missing genotypic information at greater than 50 % of the queried loci were excluded, and only variants that occurred in more than 50 % of the remaining animals were included.

4.3.4 Genotyping of Healthy and Diseased Pigs

Using the results of the LD analysis, animals were genotyped at 86 different loci (31 of which were in LD with 144 additional SNVs). In total, 1013 pigs were genotyped using Sequenom MassARRAY matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (Sequenom, San Diego, CA, USA) at the University Health Network (UHN) Clinical Genomics Centre at the Mount Sinai Hospital, Toronto, Ontario. This included 592 healthy pigs (sourced as described in the section on next-generation sequencing) and 421 pigs diagnosed with infectious disease. Of the 592 healthy pigs, 67 were used in the NGS portion of this study. Genomic material from diseased pigs was obtained from the spleen and liver of pigs presenting to the Animal Health Laboratory, University of Guelph, for post-mortem examination, from the period of 2004 - 2007. Presence of diseases was determined based on the pathological findings from the post-mortem exams and the presence of pathogens was

determined by ancillary diagnostics at the discretion of the pathologist. DNA was extracted from tissue samples as previously described (Lillie et al., 2007). Animals were grouped based on the presence of disease and/or pathogen. If multiple pathogens were present, the animal was assigned to each group; however, co-morbidities were not tracked further.

Following genotyping, local eQTL frequencies were compared between diseased and healthy pigs using PLINK v1.9 (Chang et al., 2015; Purcell et al., 2007). The non-reference allele was always defined as the minor allele using the 'a2-allele' option, regardless of which allele was most frequent within our population. Contingency tables composed of healthy animals versus animals subdivided by disease processes or pathogen types were created, and Fisher's exact test was used to determine significance. Results were corrected for multiple testing using the Benjamini-Hochberg procedure and an FDR cut-off of 0.05.

4.4 Results

After trimming, mapping, and de-duplication of raw data, 120,194,140 million reads remained, approximately 51.2 % of which aligned to the targeted regions. The success of target capture was variable, with a significantly higher average depth of coverage achieved in certain target capture regions (Appendix 4.1). Three pigs were poorly sequenced (< 25,000 total reads) and were excluded from the variant calling results and subsequent analyses. In the 69 pigs that were adequately sequenced, 41,035 single nucleotide variants (SNVs) and 296 small insertions or deletions (indels) were found. Approximately 97.1 % (40,035) of the variants sequenced in this study were present in dbSNP build 145; the remaining 2.9 % (1196) were novel in our population. There were 141,589 variants in the target sequence regions in dbSNP that were not called in our population of pigs. A portion of these variants (19,149) were present in our dataset but did not pass various quality control filters at the time of variant calling. The remainder

(122,440) were not present in our population of pigs. This may be due in part to the origin of the variants in dbSNP, which are derived from a wide variety of pig breeds, whereas our population consisted of breeds common to North American production units (Duroc, Landrace and Yorkshire crosses). Furthermore, some rare variants may not have been present in our population of 69 pigs.

The variants from the 69 pigs were combined with gene expression microarray data of 107 innate immune genes from the same animals generated in a previous study (Snyman, 2013). Analysis using Matrix eQTL identified 298 significant eQTLs in 19 different innate immune genes (Figure 4.1 and Appendix 4.3). We selected the top 86 SNVs with the most significant eQTL results for genotyping (Appendix 4.4), 31 of which were in LD ($R^2 \geq 0.95$) with one or more other SNV(s), such that the genotyping panel represented 230 SNVs (Appendix 4.5 and Appendix 4.6). Following genotyping, ten loci had unresolved clusters and two had poor extension and were excluded from the analysis. The remaining 74 SNVs (28 of which were in LD with 141 other SNVs) had a mean rate of genotyping by SNV of 99.6 %, while the mean rate for genotyping of samples was 98.8 %; all samples and SNVs were included in the subsequent analysis.

We examined the frequency of the local eQTLs in healthy pigs and pigs diagnosed with a variety of different infectious diseases. Animals were grouped into pathogen-specific groups as well as disease-specific groups. Allele frequency analysis using PLINK identified 28 of the 74 genotyped SNVs as having a significantly different allele distribution amongst normal and diseased pigs in at least one of the pathogen or condition subgroups (Appendix 4.7). These 28 SNVs were distributed across 8 different genes: *FCN2*, *SCGB1A1*, *CD55*, *SFTPD*, *CLEC5A*, *NOD1*, *DEFB1*, and *CFB*.

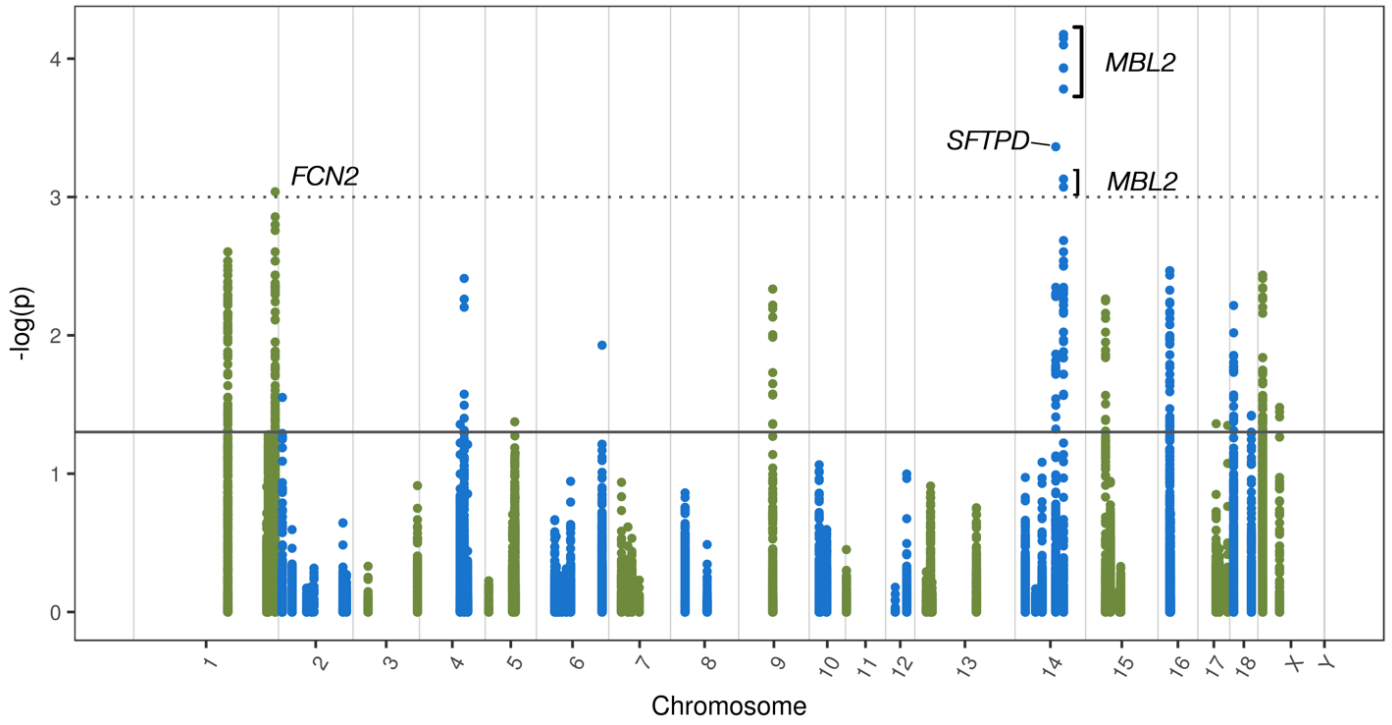


Figure 4.1. Manhattan plot of significant local eQTLs in healthy pigs found within 50 kb of a targeted innate immune gene.

Solid line represents p = 0.05, while the dashed line indicates p = 0.001. P values have been adjusted using Benjamini-Hochberg procedure for multiple testing.

There were 25 alleles with significantly different distributions within pigs diagnosed with a bacterial infection. Only animals diagnosed with *Escherichia coli* or *Mycoplasma* spp. infections had significant differences; animals with *Actinobacillus pleuropneumonia* (APP), *Haemophilus suis*, *Streptococcus suis*, and *Salmonella enterica* Typhimurium were not significantly different from the reference population. Interestingly, animals with *Mycoplasma* spp. infections consistently showed significant differences primarily in *FCN2* alleles (Table 4.1 and Appendix 4.7). The SNVs were re-examined for linkage disequilibrium with the additional genotypes, and nine of the twelve significant alleles were found to be in disequilibrium with an $R^2 \geq 0.75$. Animals diagnosed with *E. coli* k88 infections had eQTLs in four genes, *CD55*, *SFTPD*, *DEFB1*, and *CLEC5A*.

Within animals diagnosed with viral diseases, five eQTLs showed significantly different frequencies (Table 4.2 and Appendix 4.7). Animals with porcine reproductive and respiratory syndrome virus (PRRSV) showed significant differences only within *FCN2* alleles, while those with swine influenza virus (SIV) were more likely to carry mutations in *DEFB1* (a.k.a. *PDB-2*) and *CLEC5A*. Animals with porcine coronavirus 2 (PCV2) carried no eQTLs with any significant differences.

When considering the same animals subdivided by disease, rather than pathogen, animals with pneumonia and septicemia frequently showed significant differences in allele frequencies compared to healthy animals, whereas those with enteritis and serositis showed almost none (3 and 0, respectively) (Table 4.2Table 4.3 and Appendix 4.7).

Table 4.1. Frequency of the alternate allele of local eQTLs in pigs diagnosed with bacterial infections.

Gene	**rsID**	**Expression effect**	**Healthy**	***APP* (n = 25)**	***E. coli* k88 (n = 38)**	***H. parasuis* (n = 21)**	***Mycoplasma* spp. (n = 43)**	***S. typhimurium* (n = 30)**	***S. suis* (n = 98)**
FCN2	rs333222079	↓	8.7 (0)	6.0 (0)	11.8 (0)	17.5 (1)	**19.8 (0)**	15.0 (0)	10.7 (0)
FCN2	rs338072079	↓	18.0 (1)	14.0 (0)	22.4 (0)	23.8 (0)	**33.7 (0)**	31.7 (0)	22.4 (0)
FCN2	rs326158227	↓	76.3 (0)	82.0 (0)	81.6 (0)	85.7 (0)	**91.9 (0)**	83.3 (0)	84.2 (0)
FCN2	rs320342887	↓	76.3 (10)	82.0 (0)	82.4 (1)	85.7 (0)	**91.7 (1)**	81.7 (0)	84.5 (1)
FCN2	rs342195622	↓	76.9 (2)	82.0 (0)	81.6 (0)	85.0 (1)	**93.0 (0)**	80.0 (0)	83.7 (0)
FCN2	rs328892780	↓	73.2 (2)	82.0 (0)	79.0 (0)	73.8 (0)	**88.4 (0)**	71.7 (0)	78.6 (0)
FCN2	rs331664783	↓	76.4 (0)	82.0 (0)	81.6 (0)	83.3 (0)	**91.9 (0)**	75.0 (0)	82.7 (0)
FCN2	rs333340587	↓	76.1 (0)	82.0 (0)	81.6 (0)	81.0 (0)	**91.9 (0)**	75.0 (0)	82.7 (0)
FCN2	rs337034769	↓	76.4 (0)	82.0 (0)	81.6 (0)	85.7 (0)	**91.9 (0)**	83.3 (0)	84.2 (0)
FCN2	rs330321112	↓	76.6 (6)	82.0 (0)	81.6 (0)	85.7 (0)	**91.9 (0)**	80.0 (0)	83.7 (0)
FCN2	rs322664568	↓	75.6 (0)	82.0 (0)	80.3 (0)	83.3 (0)	**91.9 (0)**	78.3 (0)	83.7 (0)
FCN2	rs335102706	↓	53.0 (10)	64.0 (0)	55.3 (0)	70.0 (1)	**72.1 (0)**	60.0 (0)	64.3 (0)
SCGB1A1	rs81227376	↑	42.2 (0)	34.0 (0)	36.8 (0)	45.2 (0)	**58.1 (0)**	40.0 (0)	49.5 (0)
CD55	rs330938856	↑	44.1 (0)	50.0 (0)	**69.7 (0)**	33.3 (0)	55.8 (0)	60.0 (0)	52.0 (0)
CD55	rs322614244	↑	43.9 (8)	50.0 (0)	**67.1 (0)**	33.3 (0)	55.8 (0)	60.0 (0)	52.5 (0)
SFTPD	rs327156991	↑	5.6 (0)	14.0 (0)	**18.4 (0)**	4.8 (0)	8.1 (0)	6.7 (0)	7.1 (0)
DEFB1	rs330573086	↑	14.7 (1)	12.0 (0)	**0.0 (0)**	7.1 (0)	10.5 (0)	6.9 (1)	6.6 (0)
DEFB1	rs319252499	↑	14.9 (0)	12.0 (0)	**0.0 (0)**	7.1 (0)	10.5 (0)	8.3 (0)	6.6 (0)
DEFB1	rs335696039	↑	16.6 (9)	16.0 (0)	**4.1 (1)**	7.1 (0)	19.1 (1)	13.3 (0)	11.3 (1)
CLEC5A	rs343093524	↑	48.2 (2)	60.0 (0)	**13.2 (0)**	50.0 (0)	50.0 (0)	38.3 (0)	39.8 (0)
CLEC5A	rs327744812	↑	57.4 (0)	66.0 (0)	**26.3 (0)**	69.0 (0)	67.4 (0)	51.7 (0)	53.1 (0)
CLEC5A	rs343548293	↑	48.3 (0)	64.0 (0)	**21.1 (0)**	50.0 (0)	50.0 (0)	43.3 (0)	40.3 (0)
CLEC5A	rs346369612	↑	54.1 (0)	72.0 (0)	**25.0 (0)**	57.1 (0)	57.0 (0)	48.3 (0)	46.4 (0)
CLEC5A	rs339063186	↑	16.8 (1)	30.0 (0)	**3.9 (0)**	15.0 (1)	22.1 (0)	10.0 (0)	12.2 (0)
CLEC5A	rs336562067	↑	49.1 (8)	58.0 (0)	**23.7 (0)**	50.0 (0)	48.8 (0)	45.0 (0)	40.2 (1)

Numbers in bold indicate the frequency of the eQTL is significantly different than in healthy animals ($p < 0.05$ after controlling for FDR). Underlined numbers indicate that the eQTL was significantly less frequent as compared to healthy animals.
Arrows indicate the effect of the alternate allele on gene expression.
Numbers in parentheses in the header indicate total number of animals in the group; numbers in brackets in the cells indicate the number of animals with a missed call at that particular locus.
APP: *Actinobacillus pleuropneumonia*

Table 4.2. Frequency of the alternate allele of local eQTLs in pigs diagnosed with viral infections.

Gene	rsID	Expression effect	Healthy	PCV2 (n = 88)	PRRSV (n = 72)	SIV (n = 16)
FCN2	rs339469390	↓	3.5 (0)	8.0 (0)	**9.7 (0)**	6.2 (0)
FCN2	rs333222079	↓	8.7 (0)	14.4 (1)	**17.6 (1)**	21.9 (0)
FCN2	rs338072079	↓	18.0 (1)	21.0 (0)	**29.9 (0)**	25.0 (0)
DEFB1	rs335696039	↑	16.6 (9)	18.0 (2)	14.8 (1)	**43.3 (1)**
CLEC5A	rs343093524	↑	48.2 (2)	51.1 (0)	54.9 (0)	**78.1 (0)**

Numbers in bold indicate the frequency of the eQTL is significantly different than in healthy animals ($p < 0.05$ after controlling for FDR).
Arrows indicate the effect of the alternate allele on gene expression.
Numbers in brackets in the header indicate total number of animals in the group; numbers in brackets in the cells indicate the number of animals with a missed call at that particular locus.
PCV2: porcine circovirus type 2; PRRSV: porcine reproductive and respiratory syndrome virus; SIV: swine influenza virus.

Table 4.3. Frequency of the alternate allele of local eQTLs in pigs classified by disease.

Gene	rsID	Expression effect	Healthy	Enteritis (n = 212)	Pneumonia (n = 267)	Septicemia (n = 92)	Serositis (n = 60)
FCN2	rs326158227	↓	76.3 (0)	82.8 (0)	**83.0 (0)**	**88.0 (0)**	87.5 (0)
FCN2	rs320342887	↓	76.3 (10)	82.9 (2)	**83.0 (2)**	**88.0 (0)**	87.5 (0)
FCN2	rs342195622	↓	76.9 (2)	82.9 (1)	**83.5 (1)**	**88.6 (0)**	89.2 (0)
FCN2	rs328892780	↓	73.2 (2)	78.2 (1)	78.8 (1)	**84.8 (0)**	81.7 (0)
FCN2	rs331664783	↓	76.4 (0)	82.1 (0)	**82.4 (0)**	**88.0 (0)**	86.7 (0)
FCN2	rs333340587	↓	76.1 (0)	80.9 (0)	81.5 (0)	**88.0 (0)**	86.7 (0)
FCN2	rs337034769	↓	76.4 (0)	83.0 (0)	**83.0 (0)**	**88.0 (0)**	86.7 (0)
FCN2	rs330321112	↓	76.6 (6)	82.5 (0)	**82.5 (1)**	**87.5 (0)**	86.4 (1)
FCN2	rs322664568	↓	75.6 (0)	81.6 (0)	**82.0 (0)**	**87.0 (0)**	86.7 (0)
FCN2	rs335102706	↓	53.0 (10)	**62.0 (3)**	**63.4 (3)**	**64.3 (1)**	65.0 (0)
CD55	rs330938856	↑	44.1 (0)	**53.3 (0)**	**51.5 (0)**	52.7 (0)	46.7 (0)
CD55	rs322614244	↑	43.9 (8)	**53.6 (2)**	**52.3 (2)**	53.3 (0)	46.7 (0)
DEFB1	rs330573086	↑	14.7 (1)	11.1 (1)	**9.0 (1)**	**7.6 (0)**	8.3 (0)
DEFB1	rs319252499	↑	14.9 (0)	11.3 (0)	**9.0 (0)**	**7.6 (0)**	8.3 (0)
NOD1	rs343304937	↓	19.9 (0)	24.5 (0)	**27.7 (0)**	23.9 (0)	23.3 (0)
CFB	rs320914345	↓	7.6 (0)	12.0 (0)	**12.2 (0)**	7.1 (0)	11.7 (0)

Numbers in bold indicate the frequency of the eQTL is significantly different than in healthy animals ($p < 0.05$ after controlling for FDR). Underlined numbers indicate that the eQTL was significantly less frequent as compared to healthy animals.
Arrows indicate the effect of the alternate allele on gene expression.
Numbers in parentheses in the header indicate total number of animals in the group; numbers in brackets in the cells indicate the number of animals with a missed call at that particular locus.

4.5 Discussion

Genetic variants that are associated with economically important traits, such as disease susceptibility, can be useful in improving livestock breeding programs. However, identification of candidate eQTLs for complex traits (such as innate immunity) is difficult, and requires adequate sample sizes and well described phenotypes (Ron and Weller, 2007). This study expands upon the current knowledge of infectious disease-associated local eQTLs in pigs. Previous porcine eQTL studies focussed on single pathogen experimental models (Reiner et al., 2014), on production traits such as muscle quality (Cánovas et al., 2012; Ponsuksili et al., 2014) and backfat (Munoz et al., 2013), or on physiological processes (Liaubet et al., 2011). Our study used production animals with and without infectious diseases and/or pathogens diagnosed by necropsy with appropriate laboratory confirmation.

In the first portion of this study, we identified 298 eQTLs spanning 19 different innate immune genes. Linkage disequilibrium analysis of our population identified a subset of 86 eQTLs that, overall, were predictive of 144 of the remaining eQTLs. In the second portion, we genotyped 74 of the 86 eQTLs, and determined that 28 eQTLs in 8 different genes (*FCN2*, *CLEC5A*, *DEFB1*, *SCGB1A1*, *CD55*, *SFTPD*, *NOD1*, and *CFB*) showed significant differences between healthy and diseased pigs. Several of the genes identified in this study have been previously implicated in infectious disease susceptibility. *FCN2* encodes FCN-α, a collagenous lectin with a well-defined role in porcine innate immunity (Brooks et al., 2003b; Keirstead et al., 2008; Nahid and Sugii, 2006). FCN-α was shown to reduce the infectivity of porcine reproductive and respiratory syndrome virus *in vitro* (Keirstead et al., 2008), and we found three *FCN2* eQTLs associated with reduced *FCN2* expression were found to be more frequent in pigs diagnosed with PRRSV (Table 4.2), reinforcing the concept that FCN-α may serve a protective

role against PRRSV infection. Although FCN-α is also reactive to both gram-negative and gram-positive bacteria (Nahid and Sugii, 2006), *FCN2* eQTLs were only found in animals diagnosed with *Mycoplasma* infections.

Of particular interest is the association between an eQTL of *CLEC5A* (rs343093524), and pigs diagnosed with swine influenza virus. This allele was associated with increased expression of *CLEC5A*, and was found at a higher frequency in pigs diagnosed with SIV, suggesting that higher levels of *CLEC5A* are associated with increased incidence of SIV. It was recently shown that the C-type lectin domain containing 5A protein (CLEC5A) interacts with the hemagglutinin protein of influenza A virus and promotes a proinflammatory response in human macrophages, and *CLEC5A*$^{-/-}$ mice had improved survival upon challenge with influenza virus (Teng et al., 2017). It is tempting to extrapolate the beneficial effects of reduced *CLEC5A* expression in mice and hypothesize that a similar mechanism of action may be present in pigs, in which higher levels of CLEC5A result in a more severe inflammatory response upon infection with SIV, leading to a more severe disease presentation. Interestingly, the *CLEC5A* allele present at increased frequency in SIV positive pigs was also present at a significantly lower frequency in pigs diagnosed with *E. coli* k88. CLEC5A has also been identified as a critical receptor in the innate immune response to the gram-negative bacteria *Listeria monocytogenes* (Chen et al., 2017). Mice deficient in CLEC5A showed increased susceptibility to infection, lending a potential biological explanation for increased susceptibility to *E. coli* k88 in pigs with decreased *CLEC5A* expression. Further investigations on the role of *CLEC5A* in the porcine innate immune response to both bacteria and SIV is warranted.

An allele associated with increased expression of *DEFB1* was similarly found to be less frequent in pigs with *E. coli* k88, but more frequent in pigs with SIV. *DEFB1* (also known as

PBD-2) encodes porcine beta defensin 2, an antimicrobial peptide with *in vitro* activity against a variety of gastrointestinal pathogens, including *E. coli* (Veldhuizen et al., 2008). There is no information on the interaction of beta defensin 2 with porcine viruses. An underlying biological explanation for the increased frequency of an eQTL associated with higher expression of *DEFB1* in pigs with SIV is an area to be explored.

An allele associated with increased expression of *SCGB1A1* was found with increased frequency in pigs diagnosed with *Mycoplasma* spp. (rs81227376,Table 4.1). Little information is available on the role of its protein product, secretoglobin (also known as uteroglobin or Club cell secretory protein), in pigs. A recent study on recombinant equine secretoglobin found that the protein reduced neutrophil chemotaxis and reduced neutrophil extracellular traps (Côté et al., 2014). Neutrophils are an important component of arthritic *Mycoplasma* infections in pigs, and increased expression of *SCGB1A1* may contribute to a diminished neutrophilic response.

Alleles associated with increased expression of *CD55* and *SFTPD* were found in pigs diagnosed with *E. coli* k88 (Table 4.1). *SFTPD* is a collagenous lectin that encodes surfactant protein D (SP-D) and recognizes a wide variety of viruses, bacteria, and fungi (Holmskov et al., 2003), while *CD55* encodes the decay accelerating factor (DAF), which inhibits the activation of complement. Increased levels of DAF may inhibit complement activity, thus increasing the susceptibility of pigs to certain pathogens. While it is known that SP-D is expressed in intestinal tissues (Bourbon and Chailley-Heu, 2001), it is unknown how increased levels of SP-D contribute to disease susceptibility, and it is also unknown whether the transcriptional control of *SFTPD* in the porcine intestine is similar to the hepatic expression profile upon which the eQTL was based. It is also uncertain whether *SFTPD* must be expressed at the site of infection in order to be effective. SP-D is present in the serum of humans (Ikeda et al., 2017) and horses (Richard

et al., 2012), and it may be the circulating form of the protein that provides protection against certain pathogens in some tissues. An illustrative example, albeit from a different collectin, comes from an *MBL2* mouse knockout model demonstrating increased susceptibility to influenza A virus (IAV) infection (Chang et al., 2010). Resistance to IAV infection was restored through the systemic administration of MBL-C, suggesting that the site of infection (lung) did not necessarily have to also be the source of MBL-C. A potential mechanism through which altered hepatic expression of *SFTPD* affects gastrointestinal disease can therefore be hypothesized: if hepatic expression contributes to a circulating pool of SP-D relevant to the innate immune response in organs such as the gastrointestinal tract, then altered expression may decrease serum concentrations, thereby contributing to disease pathogenesis.

eQTLs associated with decreased *NOD1* and *CFB* expression were found only within pigs grouped by disease, rather than specific pathogen. This was likely at least partly due to the larger sample sizes within disease groups, resulting in smaller differences achieving statistical significance. *NOD1* encodes the nucleotide-binding oligomerization domain 1 pattern recognition receptor, and is generally associated with the recognition of pathogens in the gastrointestinal tract (Clarke et al., 2010). However, *NOD1* gastrointestinal recognition of bacterial peptidoglycan can enhance systemic innate immunity (Clarke et al., 2010), providing a potential explanation for the increased frequency of *NOD1* eQTLs associated with decreased *NOD1* expression in pigs with pneumonia. A study of polymorphisms in porcine *NOD1* cDNA identified 11 SNPs, 2 of which resulted in decrease ligand binding, however, the eQTL identified here was present with an intron (and was in LD with another intronic SNV, rs337167571) (Shinkai et al., 2015). Little is known about porcine *CFB*, the gene encoding complement factor B, a key component of the alternative pathway of complement that forms the alternative C3

convertase. Hypothetically, decreased levels of CFB could contribute to decreased complement activity, and thus increased susceptibility to pathogens.

A technical issue partially limits the conclusions of this study. There were differences in the depth of sequencing of different targeted regions (Appendix 4.1). Although the exact underlying cause for the differences cannot be explained, the efficacy of genomic capture when using hybridization methods is known to be variable, and is negatively impacted by sequences with highly skewed AT or GC content as well as by pre- and post-hybridization PCRs (Bodi et al., 2013; Mamanova et al., 2010). As a minimum depth of coverage of 10x was required for variant calling, low coverage in certain target regions might have limited variant discovery, and may have subsequently limited the discovery of local eQTLs in these regions. Thus, the absence of eQTLs in some genes should be considered in conjunction with the sequencing results, and not solely as a reflection on the role of those genes in the innate immune response.

This study is the first to examine the porcine innate immunome for genetic variants impacting gene expression and potentially contributing to infectious disease susceptibility. Variants identified here are excellent candidates for future study and potential incorporation into breeding programs, to help reduce the incidence of infectious diseases in pigs.

4.6 Acknowledgements

Funding for the research was provided by Ontario Pork, the Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA), a Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada, with stipend support from the Ontario Veterinary College. The authors thank Jutta Hammermueller for help with sample processing.

4.7 Supporting Information

Data: BAM files aligned to the Sscrofa 10.2 genome (including contigs GL894246_2, GL894265_1, GL896416_1, and JH118996_1) are available in the Sequence Read Archive of NCBI, accession # SRP122501 (to be released upon publication).

Chapter 5: Concluding Discussion

The overall objective of this thesis was to investigate genetic variation in collagenous lectins of horses, cattle, and pigs, and to examine how variants in these and other innate immune genes might be relevant to infectious disease resistance. This work was undertaken as infectious diseases represent significant economic and welfare concern to these species, and are frequently diagnosed within herds and in individual animals. Current attempts at prevention, management, and treatment of disease are only partially successful, and additional strategies to address infectious diseases are required. Additionally, the treatment of bacterial infections with antimicrobials is a contributor to the current state of antimicrobial resistance; reducing the usage of antimicrobials in livestock has been identified as a priority in Canada (Prescott et al., 2012). Targeting host resistance to infectious diseases could therefore be beneficial in a variety of different and important ways.

The studies presented in this thesis used modern techniques, such as pooled, targeted NGS to investigate the innate immune system of livestock. Pooled sequencing is an attractive method for maximizing the number of animals used in an experiment while maintaining reasonable costs. A substantial portion of the sequencing costs is tied to the number of individual indices, or barcodes, used to identify individual samples within the sequencing library. By pooling samples, fewer indices are required, and costs are lowered. Although this comes at the cost of losing individual genotypic data, the estimation of allele frequencies across a pool has been shown to be reliable (Bansal et al., 2011; Mullen et al., 2012; Rellstab et al., 2013). In Chapter 2: and Chapter 3:, cattle and horses were pooled into relatively small groups based on the similarity of disease diagnoses and or pathogen identification. However, given that allele frequency estimation is more robust when pool sizes are larger (Kim et al., 2010), analysis was

performed by merging data from all animals with or without infectious diseases into a single dataset. Thus, in retrospect, all animals could have been placed into two pools for sequencing library preparation for simple allele frequency estimation. This would, however, have increased computing costs for the joint genotyping variant calling protocol.

Somewhat surprisingly, the density of variants followed a similar pattern in both cattle and horses (Figure 5.1). For the most part, homologous genes (where present) exhibited a relatively similar variant density. Of note, *FCN1* (and *FCN1-like* in the horse, which does not exist in cattle) had the highest rate of variation in both species. As discussed in Chapters 2 and 3, variant density is frequently elevated in regions in which there is high GC content and/or elevated incidence of indels (Hodgkinson and Eyre-Walker, 2011). Although in both species, the GC content and indel count were highest in the regions of the *FCN* genes, neither parameter correlated in a statistically significant manner when all genes were accounted for (Appendix 2.5, Appendix 3.5, and Appendix 3.6). Thus, although the elevated variant density may indeed be related to these other structural components, the possibility that the increased variant density might be functionally relevant must be considered. Regions of the genome with high numbers of mutations tend to harbour genes that interact with the environment, such as genes of the immune system (Chuang and Li, 2004). Theoretically, this would give genes in need of adaptation (in this case, those of the innate immune response) the opportunity to evolve in tandem with the evolving threat of pathogens. The elevated number of mutations found in the *FCN* genes may therefore suggest an important evolutionary role for the *FCNs* in innate immunity, one in which frequent adaptation was beneficial to host defense. Supporting this hypothesis is the finding that alleles associated with decreased expression of *FCN2* were more frequent in pigs with infectious disease, as outlined in Chapter 4:. An extension of this observation is that genes with the lowest

variant density, such as *COLEC12*, *MASP1* and *MASP2*, may have a diminished need to adapt in the context of the innate immune response. For the *MASP* genes, this a logical finding: these genes do not respond to evolving pathogens, and instead recognize presumably conserved regions of some collagenous lectins to activate the lectin pathway of complement.

The frequency of mutations can also be related to the number of functionally similar copies of a gene. Gene redundancy can be evolutionarily unstable, as additional functional copies of a gene have the potential to diminish the impact of harmful mutations on selection pressure (Kafri et al., 2006). It is thus arguable that the increased variation found in the *FCN* genes of horses and cattle might be the result of functional redundancy, and that other collagenous lectin proteins (or other innate immune proteins entirely) may compensate for any deleterious *FCN* mutations, allowing them to accumulate. The degree of importance of the ficolins in cattle and horses is therefore debatable, and an opportunity for further research.

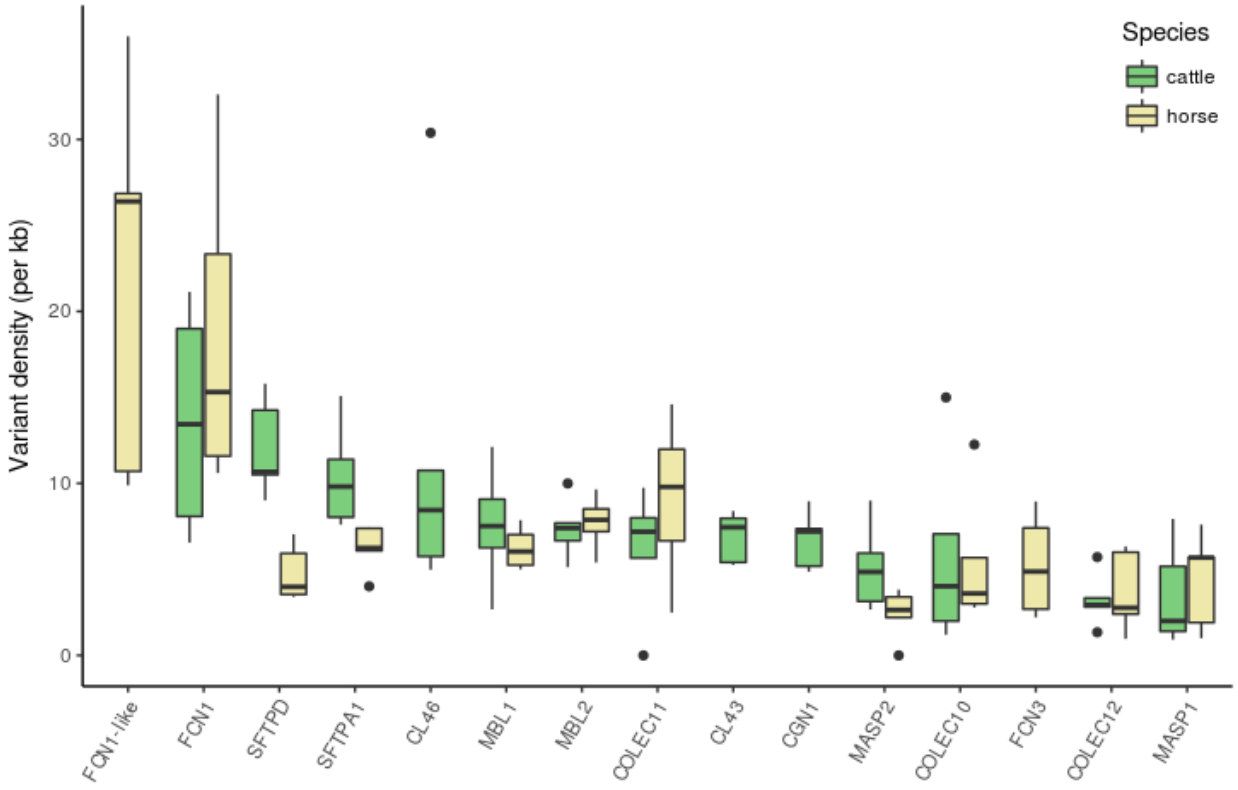


Figure 5.1: Comparison of variant density between different gene regions for horses and cattle.

Solid lines represent the median, hinges represent the 1st and 3rd quartiles, whiskers are 1.5x the interquartile range, and points are outliers. Note that *FCN1-like* and *FCN3* are not present in cattle, while *CGN1*, *CL43*, and *CL46* are not present in horses.

In Chapters 2 and 3, alleles that were significantly associated with animals diagnosed with infectious diseases were identified. In an attempt to discern possible biological relevance of these alleles, *in silico* analysis was used to identify miRNA binding elements and transcription factor binding sides containing sequence variants, along with impactful coding region mutations such as missense mutations. Several algorithms have been developed to assess the effect of missense mutations, most of which align homologous protein sequences, identify conserved residues, and use a scoring system to determine the relative importance of a single amino acid. The algorithms generally differ in the statistical model used to determine significance of conserved residues (Choi et al., 2012). An article reviewing and comparing different algorithms, including the two used in this thesis, Polyphen2 (Adzhubei et al., 2010) and SIFT (Sim et al., 2012), found that they all shared similar sensitivity and specificity when used to identify disease causing variants in a set of control data (Choi et al., 2012). For Polyphen2, the sensitivity and specificity were 88 % and 62 %, while for SIFT, they were 85 % and 69 %. Based on these numbers, these algorithms are likely to accurately reject true negatives, but suffer from potentially identifying false positives. For some of the missense variants identified in this thesis, such as the p.(Gly82Ala) mutation found in equine MBL-A, supporting evidence was obtained from similar mutations in other species known to have adverse effects, and it is possible that the effect of the mutation is similar in the horse. For the remainder, biochemical characterization confirming the predicted effects of the mutations is required, especially if genotyping of larger numbers of animals supports the association with infectious disease.

The biochemical characterization of collagenous lectins in non-reference species is still challenging. Commercial antibodies needed for applications such as enzyme-linked immunosorbent assays or Western blots, useful for the determination of serum collagenous lectin

levels, are not always available, or do not always work (Podolsky et al., 2006). Functional experiments, such as complement deposition assays, could add additional information and help differentiate between mutations that affect function, but not concentration, of the complement-activating collagenous lectins.

It is now well established that immunity is regulated in part by miRNAs (Foster et al., 2013; O'Connell et al., 2012; Taganov et al., 2007; Xiao and Rajewsky, 2009; Zhou et al., 2014). miRNAs can affect various components of the immune system directly, can directly affect some pathogens, and can modulate the magnitude of the immune response. With respect to immunity, studies have typically associated the relative levels of various miRNAs to disease (Iborra et al., 2012). Investigations assessing the effect of mutations, either to the miRNA or to miRNA response elements (MREs), on susceptibility to infectious diseases are sparse. Studies in other fields, however, have found good evidence that mutations in MREs contribute to disease or gene expression. For example, a mutation in a *let-7* binding site in the 3′ untranslated region (UTR) of *KRAS*, a gene known to cause *de novo* endometriosis in mice, was found to be 6 times more frequent in women with endometriosis. The mutation results in decreased *let-7* repression of *KRAS*, and thus increased expression of the gene (Grechukhina et al., 2012). Mutations in the 3′ UTR can also lead to the development of novel miRNA binding sites. In Texel sheep, a mutation in the 3′ UTR of myostatin led to the creation of a novel binding site for two miRNAs, subsequently repressing gene expression and resulting in the increased muscle for which the breed is known (Clop et al., 2006).

In order to assess the potential effects of mutations on miRNA regulation of the genes targeted in this thesis, the 3′ UTR was analyzed for potential MREs. Two algorithms, Targetscan and miRanda, were used to predict MREs. Targetscan identifies MREs based on the

complementarity of the seed sequence to sequences in the 3′ UTR mRNA, and provides information on the type of binding motifs (Agarwal et al., 2015). The highest ranking (strongest) occurs for the 8mer, in which there is perfect complementarity to positions 2-8 and a flanking adenine, followed by 7mer-m8s (position 2-8 match), 7mers (position 2-7 match with a flanking adenine), and offset 6mers (position 3-8 match). On the other hand, miRanda reports scores both for weighted complementarity as well as thermodynamic binding of the entire miRNA (Enright et al., 2003). By using both algorithms independently, and subsequently selecting only MREs predicted by both algorithms, the chances of accurately predicting true MREs is increased (Riffo-Campos et al., 2016).

To determine whether a bioinformatically-predicted MRE has been affected by a short nucleotide variant (SNV), it must first be validated. A common strategy is to clone the 3′ UTR that includes the predicted MRE into a reporter plasmid (typically encoding green fluorescent protein or luciferase) and to transfect cells both with the reporter plasmid and a plasmid encoding an miRNA. Full validation of the interaction between the MRE and miRNA requires demonstrating that the mRNA and miRNA are expressed in the original cell type of interest, and determining the ultimate effect on protein expression (Kuhn et al., 2008). Determining the impact of a SNV once such a reporter system is setup would be (theoretically) as simple as using site-directed mutagenesis to alter the MRE and repeating the above assays.

Predicting transcription factor binding sites (TFBSs) is a difficult task. The binding specificity of a particular transcription factor (TF) is highly predictable based on the amino acid sequence of its DNA-binding domain, and this specificity is largely conserved even amongst divergent species (Jolma et al., 2013; Weirauch et al., 2014). Upstream regulatory DNA is similarly conserved (Nitta et al., 2015). It is disappointing, then, to discover that despite this

promising degree of conservation, the actual binding and use of specific TFBSs is species dependant (Schmidt et al., 2010). Thus, one might predict possible TFBSs, but cannot necessarily infer whether they are used from studies in other species. In species such as horses, where limited experimental data exists to confirm predictions, we are left with simply listing the best guesses derived from prediction algorithms. In cattle, a similar dearth of information exists, though two studies have also attempted to predict TFBSs in the bovine collagenous lectins *SFTPD* and *CL43* (Gjerstorff et al., 2004b; Hansen et al., 2003a). For *CL43*, 8 TFBS were predicted by both this thesis and the original study, though only a single TFBS (Mzf-1) was affected by a SNV. There was no overlap in predicted TFBSs for *SFTPD* between this thesis and Gjerstorff et al. (2004). As noted in Chapter 3, these discrepancies may be due to the restriction of TFBS prediction (in cattle) to conserved motifs only, or perhaps in the different algorithms used. Regardless, the results described in this thesis await validation from experiments capable of identifying DNA sequences bound to TFs, such as chromatin immunoprecipitation (CHiP) sequencing.

Despite the extensive *in silico* investigation of the disease-associated variants, only 10 variants in the horse, and none in cattle, had a predicted putative functional impact (Table 2.4 and Table 3.4). Regardless, the association of alleles with disease is an important finding in and of itself, and can inform further investigations. The first step is confirming the accuracy of the associations, through larger-scale genotyping studies. Obtaining a sufficiently large sample size of animals with the appropriate infectious disease phenotype required for these types of studies is one of the major challenges in genomics research of livestock. This was beyond the scope of the initial survey conducted during the course of this study.

Once associations have been confirmed, or further validated, mechanistic studies on those alleles can then follow. Alleles that fall into coding regions can be investigated using recombinant wildtype and mutated proteins, evaluation of functional parameters, such as complement activity (in those collagenous lectins that have such capabilities) or pathogen binding. As an example, Podolsky et al (2006) used a complement deposition assay to indirectly assess serum levels of MBL-C in healthy and sick horses. Sick horses had significantly lower complement activation when compared to healthy counterparts. Although the cDNA of one sample was synthesized and sequenced, sequencing of the remaining samples was not performed. Hypothetically, decreased MBL-C serum concentration, or, alternatively, normal MBL-C serum concentration but loss of ability to activate complement, may be the result of a SNV. Serum samples from live and affected animals are required to pursue this line of investigation, so that both sequencing and functional assays may be performed.

Alleles predicted to affect the MRE in the 3′ UTR or potential TFBSs in the promoter region generally affect the expression of a gene, and are best investigated using expression studies (e.g. qPCR, microarray, or RNAseq). Effects on expression are not limited to SNVs found within the 3′ UTR or the promoter, however: alleles found in the distant upstream sequence can potentially act as silencers, enhancers, insulators, or locus control regions (Maston et al., 2006), and those within introns can affect gene expression or splicing (van Laere et al., 2003). Any region of the genome that harbours SNV(s) that affect gene expression is known as an expression quantitative trait loci (eQTL), and by performing eQTL analysis, it becomes easier to prioritize variants with no known putative effect.

In part due to the availability of data on the hepatic expression of innate immune genes in pigs, an eQTL approach to variant prioritization was taken in that species as part of this thesis.

Expression levels were combined with NGS-derived genotypes from the same animals, and associations were calculated. With the discovery of significant eQTLs, healthy pigs and pigs diagnosed with infectious diseases were genotyped, attempting to see if alleles associated with altered expression were also associated with infectious disease. Twenty-eight variants affecting the expression of eight genes were associated with infectious diseases in pigs. Further studies are now indicated to disambiguate their potential mechanism(s) of action. Local eQTLs can act in a number of different ways, including affecting TF binding, chromatin accessibility, DNA methylation, splicing, and miRNA binding, amongst others. Determining how and which molecular mechanism is at work would be based on the location of the eQTL relative to the gene. However, despite the relatively large genomic regions targeted for sequencing in this thesis, the experiment lacks the enormous breadth of a whole genome sequencing study. Further, the inherent variability in the success of target capture, resulted in some regions being poorly sequenced. It is thus possible that the significantly associated SNVs identified here are not causative, or biologically relevant, but are instead linked to other mutations. However, careful examination of the Manhattan plots (Figure 3.6 and Appendix 4.3) shows that the significant eQTLs for *MASP1*, *COLEC12*, *FCN2*, and the collectin locus form distinct clusters, suggesting that although there is likely LD present, the causative variant might be present within the cluster. For *FCN2* and the collectin locus, the cluster of variants borders the 5′ edge of the targets, and thus further upstream sequencing and genotyping is needed to further define these two clusters.

Although eQTL studies identify variants associated with gene expression, it is ultimately functional protein levels that determine phenotype. A study investigating the relationship between eQTLs and protein QTL (pQTL) found that although a significant number of eQTLs ultimately affect protein levels, some do not. Furthermore, the effect on protein levels is often

less dramatic than on gene expression (Battle et al., 2015). To further complicate matters, some mutations may affect protein levels without necessarily affecting expression: coding mutations that affect protein folding, post-translational modifications, or higher order structure can result in lower circulating protein levels, as exemplified by the p.Gly54Asp mutation in *hMBL2* (Sumiya et al., 1991). The effect of missense mutations was not examined in this experiment, and thus those that did not impact expression were not genotyped in the diseased and healthy populations of pigs.

Some of the alleles identified in this thesis represent candidate alleles for marker-assisted selection in cattle, horses, and pigs. Although it is unlikely that all of the alleles associated with infectious diseases will have biological relevance, it is possible that a portion may. Validation of associations in larger populations of animals may therefore render a subset of these alleles useful for inclusion in high-throughput genotyping applications, such as high-density SNP chips. Beyond validation of alleles, selection for disease resistance in livestock poses another challenge: some alleles may confer protection to one disease while simultaneously increasing susceptibility to another. The *CLEC5A* mutation described in this thesis is an example of one such allele, in which increased expression was associated swine influenza virus infection, but decreased expression was associated with *E. coli* infection. Quantifying the relative importance of different infectious diseases to different industries, along with quantifying the relative contribution of each allele towards resistance or susceptibility to different diseases represent significant challenges. Furthermore, disease resistance represents only one trait of importance in livestock breeding, and must be balanced with other production traits, such as growth, muscling or milk production, and litter sizes. Ultimately, incorporating these candidates alleles into

genotyping arrays will allow for analysis at the population level, and selection of genotypes with the most beneficial phenotypes.

Apart from use in selective breeding programs, the alleles identified in this thesis have the potential to increase the understanding of the pathophysiology of infectious diseases in livestock. As an example, the mutation to the CLD of equine *MBL1*, similar to a deleterious mutation in human *MBL2*, is worthy of further attention. The human polymorphism has been identified as a cause of recurring pulmonary infections in children. It could therefore by hypothesized that the similar mutation in horses may have a role in equine pneumonia, and studies examining both foals and adult horses are indicated.

Overall, the studies in this thesis have combined to provide a detailed overview of genetic variants that potentially contribute to infectious disease susceptibility of cattle, horses, and pigs. In horses and cattle, variants within or surrounding the collagenous lectins were significantly associated with infectious disease. In pigs, eQTLs associated with infectious diseases were identified. These alleles represent important potential contributions to infectious disease resistance and contribute to the body of knowledge surrounding the genetics of innate immunity of livestock. Further investigation, both in terms of functional and larger-scale genotyping studies, is warranted.

5.1 Summary and Conclusions

1. The DNA within and surrounding the collagenous lectin genes of cattle and horses contains numerous sequence variants.
    a. A substantial amount of genetic variation was discovered within and surrounding the collagenous lectin genes of horses. A total of 4559 SNVs were identified, 74 % of which were previously unknown.
    b. A similar degree of variation was identified within and surrounding the collagenous lectin genes of cattle. A total of 5439 SNVs were identified, however, the better characterization of the genome in this species meant only 122 were novel.
    c. Variation tended to be lowest in coding regions, but this was not significantly different from the distant upstream (5 - 50 kb), upstream, introns, or downstream regions for either species.
    d. The *FCN* genes of both cattle and horses had a significantly higher variant density as compared to the remaining gene targets. Whether this represents increased evolutionary need for variation in the FCN proteins is unknown.
    e. A coding mutation similar to a known deleterious mutation in humans and pigs was found in equine *MBL1*.
2. Variation in the bovine and equine collagenous lectins are associated with infectious diseases, suggesting that these genes play a role in infectious disease resistance in these species.

a. Allele frequency analysis identified 113 alleles associated with infectious disease in horses.
    i. Ten of these alleles had predicted functional impacts, including a single missense mutation, 2 mutations to putative miRNA response elements, and 7 mutations to predicted transcription factor binding sites.
b. In cattle, 74 alleles were associated with infectious diseases. None of the identified alleles had a predicted function. A 21 kb region in the upstream region of the collectin locus accounted for 48 of the significant variants, suggesting a region of interest for further investigation.

3. Expression quantitative trait analysis identified 298 SNVs associated with altered hepatic expression of 19 innate immune genes in pigs. A substantial portion of these SNVs were in strong linkage disequilibrium.
4. By genotyping pigs that were healthy and pigs infected with a variety of pathogens, 28 eQTLs were found to be associated with infectious diseases in pigs. Both increased and decreased hepatic expression of innate immune genes was associated with different infectious diseases. These alleles are potential indicators of infectious disease resistance in pigs, and warrant further investigation.

References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R., 2010. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249. https://doi.org/10.1038/nmeth0410-248

Agarwal, V., Bell, G.W., Nam, J.-W., Bartel, D.P., 2015. Predicting effective microRNA target sites in mammalian mRNAs. Elife 4, 1–38. https://doi.org/10.7554/eLife.05005

Akiyama, J., Volik, S. V., Plajzer-Frick, I., Prince, A., Sago, H., Weier, H.U.G., Vanderbilt, J.N., Hawgood, S., Poulain, F.R., 1999. Characterization of the mouse collectin gene locus. Am. J. Respir. Cell Mol. Biol. 21, 193–199. https://doi.org/10.1165/ajrcmb.21.2.3681

Akiyama, K., Sugii, S., Hirota, Y., 1992. Conglutinin, mannan-binding protein, and serum amyloid P component concentrations in sera from cows: changes associated with mastitis. J. Vet. Med. Sci. 54, 977–81. https://doi.org/10.1248/cpb.37.3229

Albert, F.W., Kruglyak, L., 2015. The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. 16, 197–212. https://doi.org/10.1038/nrg3891

Andrade, F.A., Beltrame, M.H., Bini, V.B., Gonçalves, L.B., Boldt, A.B.W., Messias-Reason, I.J. de, 2017. Association of a new FCN3 haplotype with high ficolin-3 levels in leprosy. PLoS Negl. Trop. Dis. 11, e0005409. https://doi.org/10.1371/journal.pntd.0005409

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. [WWW Document]. URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed 2.28.18).

Arora, M., Munoz, E., Tenner, A.J., 2001. Identification of a site on mannan-binding lectin critical for enhancement of phagocytosis. J. Biol. Chem. 276, 43087–43094. https://doi.org/10.1074/jbc.M105455200

Bailey, T.L., Elkan, C., 1994. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. Proc. Second Int. Conf. Intell. Syst. Mol. Biol. 28–36. https://doi.org/citeulike-article-id:878292

Bansal, V., Tewhey, R., Leproust, E.M., Schork, N.J., 2011. Efficient and cost effective population resequencing by pooling and in-solution hybridization. PLoS One 6, e18353. https://doi.org/10.1371/journal.pone.0018353

Barrett, J.C., Fry, B., Maller, J., Daly, M.J., 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21, 263–265. https://doi.org/10.1093/bioinformatics/bth457

Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., Gilad, Y., 2015. Impact of regulatory variation from RNA to protein. Science (80-.). 347, 664–667. https://doi.org/10.1126/science.1260793

Bayarri-Olmos, R., Hansen, S., Henriksen, M.L., Storm, L., Thiel, S., Garred, P., Munthe-Fog, L., 2015. Genetic variation of COLEC10 and COLEC11 and association with serum levels of collectin liver 1 (CL-L1) and collectin kidney 1 (CL-K1). PLoS One 10, e0114883.

https://doi.org/10.1371/journal.pone.0114883

Beltrame, M.H., Boldt, A.B.W., Catarino, S.J., Mendes, H.C., Boschmann, S.E., Goeldner, I., de Messias-Reason, I.J.T., 2015. MBL-associated serine proteases (MASPs) and infectious diseases. Mol. Immunol. 67, 85–100. https://doi.org/10.1016/j.molimm.2015.03.245

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57, 289–300. https://doi.org/10.2307/2346101

Benson, B., Hawgood, S., Schilling, J., Clements, J., Damm, D., Cordell, B., White, R.T., 1985. Structure of canine pulmonary surfactant apoprotein: cDNA and complete amino acid sequence. Proc. Natl. Acad. Sci. U. S. A. 82, 6379–6383. https://doi.org/10.1073/pnas.82.19.6379

Bermingham, M.L., Bishop, S.C., Woolliams, J.A., Pong-Wong, R., Allen, A.R., McBride, S.H., Ryder, J.J., Wright, D.M., Skuce, R.A., McDowell, S.W., Glass, E.J., 2014. Genome-wide association study identifies novel loci associated with resistance to bovine tuberculosis. Heredity (Edinb). 112, 543–551. https://doi.org/10.1038/hdy.2013.137

Bertelsen, H.P., Gregersen, V.R., Poulsen, N., Nielsen, R.O., Das, A., Madsen, L.B., Buitenhuis, A.J., Holm, L.E., Panitz, F., Larsen, L.B., Bendixen, C., 2016. Detection of genetic variation affecting milk coagulation properties in Danish Holstein dairy cattle by analyses of pooled whole-genome sequences from phenotypically extreme samples (pool-seq). J. Anim. Sci. 94, 1313–1365. https://doi.org/10.2527/jas.2015-9884

Beytut, E., 2011. Immunohistochemical evaluation of surfactant proteins and lymphocyte phenotypes in the lungs of cattle with natural tuberculosis. Res. Vet. Sci. 91, 119–124. https://doi.org/10.1016/j.rvsc.2010.07.023

Boddicker, N.J., Bjorkquist, A., Rowland, R.R.R., Lunney, J.K., Reecy, J.M., Dekkers, J.C.M., 2014. Genome-wide association and genomic prediction for host response to porcine reproductive and respiratory syndrome virus infection. Genet. Sel. Evol. 46, 18. https://doi.org/10.1186/1297-9686-46-18

Bodi, K., Perera, A.G., Adams, P.S., Bintzler, D., Dewar, K., Grove, D.S., Kieleczawa, J., Lyons, R.H., Neubert, T.A., Noll, A.C., Singh, S., Steen, R., Zianni, M., 2013. Comparison of commercially available target enrichment methods for next-generation sequencing. J. Biomol. Tech. 24, 73–86. https://doi.org/10.7171/jbt.13-2402-002

Boldt, A.B.W., Goeldner, I., Stahlke, E.R.S., Thiel, S., Jensenius, J.C., de Messias-Reason, I.J.T., 2013. Leprosy association with low MASP-2 levels generated by MASP2 haplotypes and polymorphisms flanking MAp19 exon 5. PLoS One 8, e69054. https://doi.org/10.1371/journal.pone.0069054

Boldt, A.B.W., Luz, P.R., de Messias-Reason, I.J.T., 2011. MASP2 haplotypes are associated with high risk of cardiomyopathy in chronic Chagas disease. Clin. Immunol. 140, 63–70. https://doi.org/10.1016/j.clim.2011.03.008

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, btu170-2120.

https://doi.org/10.1093/bioinformatics/btu170

Borron, P., Intosh, J.C.M.C., Korfhagen, T.R., Whitsett, J.A., Taylor, J., Wright, J.O.R.A.E., Mcintosh, J.C., Korfha-, T.R., Whitsett, J.A., Taylor, J., Rae, J., Lps-, W.S.A., 2000. Surfactant-associated protein A inhibits LPS-induced cytokine and nitric oxide production in vivo 840–847.

Bourbon, J.R., Chailley-Heu, B., 2001. Surfactant proteins in the digestive tract, mesentery, and other organs: evolutionary significance. Comp. Biochem. Physiol. Part A Mol. Integr. Physiol. 129, 151–161. https://doi.org/10.1016/S1095-6433(01)00312-9

Bronkhorst, M.W.G.A., Lomax, M.A.Z., Vossen, R.H.A.M., Bakker, J., Patka, P., van Lieshout, E.M.M., 2013. Risk of infection and sepsis in severely injured patients related to single nucleotide polymorphisms in the lectin pathway. Br. J. Surg. 100, 1818–1826. https://doi.org/10.1002/bjs.9319

Brooks, A.S., DeLay, J.P., Hayes, M.A., 2003a. Purification and binding properties of porcine plasma ficolin that binds Actinobacillus pleuropneumoniae. Dev. Comp. Immunol. 27, 835–844. https://doi.org/10.1016/S0145-305X(03)00057-0

Brooks, A.S., DeLay, J.P., Hayes, M.A., 2003b. Characterization of porcine plasma ficolins that bind Actinobacillus pleuropneumoniae serotype 5B. Immunobiology 207, 327–337. https://doi.org/10.1078/0171-2985-00245

Brooks, A.S., Hammermueller, J.D., DeLay, J.P., Hayes, M.A., 2003c. Expression and secretion of ficolin β by porcine neutrophils. Biochim. Biophys. Acta 1624, 36–45. https://doi.org/10.1016/j.bbagen.2003.09.004

Brown-Augsburger, P., Hartshorn, K.L., Chang, D., Rust, K., Fliszar, C., Welgus, H.G., Crouch, E.C., 1996. Site-directed mutagenesis of Cys-15 and Cys-20 of pulmonary surfactant protein D: Expression of a trimeric protein with altered anti-viral properties. J. Biol. Chem. 271, 13724–13730. https://doi.org/10.1074/jbc.271.23.13724

Cánovas, A., Pena, R.N., Gallardo, D., Ramírez, O., Amills, M., Quintanilla, R., 2012. Segregation of Regulatory Polymorphisms with Effects on the Gluteus Medius Transcriptome in a Purebred Pig Population. PLoS One 7, e35583. https://doi.org/10.1371/journal.pone.0035583

Carpenter, S., O'Neill, L.A.J., 2007. How important are Toll-like receptors for antimicrobial responses? Cell. Microbiol. 9, 1891–1901. https://doi.org/10.1111/j.1462-5822.2007.00965.x

Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience 4. https://doi.org/10.1186/s13742-015-0047-8

Chang, W.-C., White, M.R., Moyo, P., McClear, S., Thiel, S., Hartshorn, K.L., Takahashi, K., 2010. Lack of the pattern recognition molecule mannose-binding lectin increases susceptibility to influenza A virus infection. BMC Immunol. 11, 64. https://doi.org/10.1186/1471-2172-11-64

Chen, C.-B., Wallis, R., 2004. Two mechanisms for mannose-binding protein modulation of the

activity of its associated serine proteases. J. Biol. Chem. 279, 26058–26065. https://doi.org/10.1074/jbc.M401318200

Chen, G., Shaw, M.H., Kim, Y.-G., Nuñez, G., Nez, G.N., 2009. NOD-like receptors: Role in innate immunity and inflammatory disease. Annu. Rev. Pathol. Mech. Dis. 4, 365–398. https://doi.org/10.1146/annurev.pathol.4.110807.092239

Chen, S.T., Li, F.J., Hsu, T.Y., Liang, S.M., Yeh, Y.C., Liao, W.Y., Chou, T.Y., Chen, N.J., Hsiao, M., Yang, W. Bin, Hsieh, S.L., 2017. CLEC5A is a critical receptor in innate immunity against Listeria infection. Nat. Commun. 8. https://doi.org/10.1038/s41467-017-00356-3

Chiffoleau, E., 2018. C-Type lectin-like receptors as emerging orchestrators of sterile inflammation represent potential therapeutic targets. Front. Immunol. 9, 1–9. https://doi.org/10.3389/fimmu.2018.00227

Choe, C., Park, J.-W., Kim, E.-S., Lee, S.-G., Park, S.-Y., Lee, J.-S., Cho, M.-J., Kang, K.R., Han, J., Kang, D., 2010. Proteomic analysis of differentially expressed proteins in bovine endometrium with endometritis. Korean J. Physiol. Pharmacol. 14, 205–212. https://doi.org/10.4196/kjpp.2010.14.4.205

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., Chan, A.P., 2012. Predicting the functional effect of amino acid substitutions and indels. PLoS One 7, e46688. https://doi.org/10.1371/journal.pone.0046688

Chuang, J.H., Li, H., 2004. Functional Bias and Spatial Organization of Genes in Mutational Hot and Cold Regions in the Human Genome. PLoS Biol. 2, e29. https://doi.org/10.1371/journal.pbio.0020029

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M., 2012. 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 (Austin). 6, 80–92. https://doi.org/10.4161/fly.19695

Clark, H.W., Reid, K.B.M., Sim, R.B., 2000. Collectins and innate immunity in the lung. Microbes Infect. 2, 273–278. https://doi.org/10.1016/S1286-4579(00)00301-4

Clarke, T.B., Davis, K.M., Lysenko, E.S., Zhou, A.Y., Yu, Y., Weiser, J.N., 2010. Recognition of peptidoglycan from the microbiota by Nod1 enhances systemic innate immunity. Nat. Med. 16, 228–231. https://doi.org/10.1038/nm.2087

Clop, A., Marcq, F., Takeda, H., Pirottin, D., Tordoir, X., Bibé, B., Bouix, J., Caiment, F., Elsen, J.M., Eychenne, F., Larzul, C., Laville, E., Meish, F., Milenkovic, D., Tobin, J., Charlier, C., Georges, M., 2006. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nat. Genet. 38, 813–818. https://doi.org/10.1038/ng1810

Colley, K.J., Baenziger, J.U., 1987. Identification of the post-translational modifications of the core-specific lectin. The core-specific lectin contains hydroxyproline, hydroxylysine, and glucosylgalactosylhydroxylysine residues. J. Biol. Chem. 262, 10290–5.

Côté, O., Clark, M.E., Viel, L., Labbé, G., Seah, S.Y.K., Khan, M.A., Douda, D.N., Palaniyar,

N., Bienzle, D., 2014. Secretoglobin 1A1 and 1A1A Differentially Regulate Neutrophil Reactive Oxygen Species Production, Phagocytosis and Extracellular Trap Formation. PLoS One 9, e96217. https://doi.org/10.1371/journal.pone.0096217

Crouch, E.C., Chang, D., Rust, K., Persson, A., Heuser, J., Heusedl, J., 1994. Recombinant pulmonary surfactant protein D. Post-translational modification and molecular assembly. J. Biol. Chem. 269, 15808–15813.

Davies, G., Genini, S., Bishop, S.C., Giuffra, E., 2009. An assessment of opportunities to dissect host genetic variation in resistance to infectious diseases in livestock. Animal 3, 415–436. https://doi.org/10.1017/S1751731108003522

Davis, A.E., Lachmann, P.J., 1984. Bovine conglutinin is a collagen-like protein. Biochemistry 23, 2139–2144. https://doi.org/10.1021/bi00305a006

Degn, S.E., Jensen, L., Hansen, A.G., Duman, D., Tekin, M., Jensenius, J.C., Thiel, S., 2012. Mannan-binding lectin-associated serine protease (MASP)-1 is crucial for lectin pathway activation in human serum, whereas neither MASP-1 nor MASP-3 is required for alternative pathway function. J. Immunol. 189, 3957–3969. https://doi.org/10.4049/jimmunol.1201736

Degn, S.E., Jensen, L., Olszowski, T., Jensenius, J.C., Thiel, S., 2013. Co-complexes of MASP-1 and MASP-2 associated with the soluble pattern-recognition molecules drive lectin pathway activation in a manner inhibitable by MAp44. J. Immunol. 191, 1334–1345. https://doi.org/10.4049/jimmunol.1300780

Dekkers, J.C.M., 2004. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. J. Anim. Sci. 82 E-Suppl, E313-328. https://doi.org/10.2527/2004.8213_supplE313x

DePristo, M.A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498. https://doi.org/10.1038/ng.806

Donis, R.O., Dubovi, E.J., 1987. Glycoproteins of bovine viral diarrhoea-mucosal disease virus in infected bovine cells. J. Gen. Virol. 68, 1607–1616. https://doi.org/10.1099/0022-1317-68-6-1607

Doolittle, R.F., McNamara, K., Lin, K., 2012. Correlating structure and function during the evolution of fibrinogen-related domains. Protein Sci. 21, 1808–1823. https://doi.org/10.1002/pro.2177

Drickamer, K., 1992. Engineering galactose-binding activity into a C-type mannose-binding protein. Nature 360, 183–186. https://doi.org/10.1038/360183a0

Drickamer, K., 1988. Two distinct classes of carbohydrate-recognition domains in animal lectins. J. Biol. Chem. 263, 9557–60.

Drickamer, K., Dordal, M.S., Reynolds, L., 1986. Mannose-binding proteins isolated from rat liver contain carbohydrate-recognition domains linked to collagenous tails. Complete

primary structures and homology with pulmonary surfactant apoprotein. J. Biol. Chem. 261, 6878–87.

Drickamer, K., Taylor, M.E., 2015. Recent insights into structures and functions of C-type lectins in the immune system. Curr. Opin. Struct. Biol. 34, 26–34. https://doi.org/10.1016/j.sbi.2015.06.003

Drickamer, K., Taylor, M.E., 1993. Biology of animal lectins. Annu. Rev. Cell Biol. 9, 237–264. https://doi.org/10.1146/annurev.cb.09.110193.001321

Egenvall, A., Penell, J., Bonnett, B.N., Blix, J., Pringle, J., 2008. Demographics and costs of colic in Swedish horses. J. Vet. Intern. Med. 22, 1029–1037. https://doi.org/10.1111/j.1939-1676.2008.0136.x

Eilbeck, K., Quinlan, A., Yandell, M., 2017. Settling the score: Variant prioritization and Mendelian disease. Nat. Rev. Genet. https://doi.org/10.1038/nrg.2017.52

Eisen, D.P., Minchinton, R.M., 2003. Impact of mannose-binding lectin on susceptibility to infectious diseases. Clin. Infect. Dis. 37, 1496–1505. https://doi.org/10.1086/379324

Ellis, J.A., 2010. Bovine parainfluenza-3 virus. Vet. Clin. North Am. Food Anim. Pract. 26, 575–593. https://doi.org/10.1016/j.cvfa.2010.08.002

Endo, Y., Liu, Y., Kanno, K., Takahashi, M., Matsushita, M., Fujita, T., 2004. Identification of the mouse H-ficolin gene as a pseudogene and orthology between mouse ficolins A/B and human L-/M-ficolins. Genomics 84, 737–744. https://doi.org/10.1016/j.ygeno.2004.07.006

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., Marks, D.S., 2003. MicroRNA targets in Drosophila. Genome Biol. 5, R1. https://doi.org/10.1186/gb-2003-5-1-r1

Farré, D., Roset, R., Huerta, M., Adsuara, J.E., Roselló, L., Albà, M.M., Messeguer, X., 2003. Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN. Nucleic Acids Res. 31, 3651–3653. https://doi.org/10.1093/nar/gkg605

Feary, D.J., Hassel, D.M., 2006. Enteritis and colitis in horses. Vet. Clin. North Am. Equine Pract. 22, 437–479. https://doi.org/10.1016/j.cveq.2006.03.008

Ferguson, J.S., Voelker, D.R., McCormack, F.X., Schlesinger, L.S., 1999. Surfactant protein D binds to Mycobacterium tuberculosis bacilli and lipoarabinomannan via carbohydrate-lectin interactions resulting in reduced phagocytosis of the bacteria by macrophages. J. Immunol. 163, 312–21.

Foster, P.S., Plank, M., Collison, A., Tay, H.L., Kaiko, G.E., Li, J., Johnston, S.L., Hansbro, P.M., Kumar, R.K., Yang, M., Mattes, J., 2013. The emerging role of microRNAs in regulating immune and inflammatory responses in the lung. Immunol. Rev. 253, 198–215. https://doi.org/10.1111/imr.12058

Fujita, T., 2002. Evolution of the lectin–complement pathway and its role in innate immunity. Nat. Rev. Immunol. 2, 346–353. https://doi.org/10.1038/nri800

Fujita, T., Endo, Y., Nonaka, M., 2004a. Primitive complement system?recognition and activation. Mol. Immunol. 41, 103–111. https://doi.org/10.1016/j.molimm.2004.03.026

Fujita, T., Matsushita, M., Endo, Y., 2004b. The lectin-complement pathway--its role in innate immunity and evolution. Immunol. Rev. 198, 185–202. https://doi.org/10.1111/j.0105-2896.2004.0123.x

García-Laorden, M.I., Rodríguez de Castro, F., Solé-Violán, J., Rajas, O., Blanquer, J., Borderías, L., Aspa, J., Briones, M.L., Saavedra, P., Marcos-Ramos, J.A., González-Quevedo, N., Sologuren, I., Herrera-Ramos, E., Ferrer, J.M., Rello, J., Rodríguez-Gallego, C., 2011. Influence of genetic variability at the surfactant proteins A and D in community-acquired pneumonia: a prospective, observational, genetic study. Crit. Care 15, R57. https://doi.org/10.1186/cc10030

Garlatti, V., Belloy, N., Martin, L., Lacroix, M., Matsushita, M., Endo, Y., Fujita, T., Fontecilla-Camps, J.C., Arlaud, G.J., Thielens, N.M., Gaboriaud, C., Arlaud, J., Thielens, N.M., 2007. Structural insights into the innate immune recognition specificities of L- and H-ficolins. EMBO J. 26, 623–633. https://doi.org/10.1038/sj.emboj.7601500

Garlatti, V., Martin, L., Lacroix, M., Gout, E., Arlaud, G.J., Thielens, N.M., Gaboriaud, C., 2010. Structural insights into the recognition properties of human ficolins. J. Innate Immun. 2, 17–23. https://doi.org/10.1159/000233475

Garred, P., 2008. Mannose-binding lectin genetics: from A to Z. Biochem. Soc. Trans. 36, 1461–1466. https://doi.org/10.1042/BST0361461

Garred, P., Genster, N., Pilely, K., Bayarri-Olmos, R., Rosbjerg, A., Ma, Y.J., Skjoedt, M.O., 2016. A journey through the lectin pathway of complement—MBL and beyond. Immunol. Rev. https://doi.org/10.1111/imr.12468

Garred, P., Honoré, C., Ma, Y.J., Rørvig, S., Cowland, J., Borregaard, N., Hummelshøj, T., 2010. The Genetics of ficolins. J. Innate Immun. 2, 3–16. https://doi.org/10.1159/000242419

Garred, P., Madsen, H.O., Balslev, U., Hofmann, B., Pedersen, C., Gerstoft, J., Svejgaard, A., 1997. Susceptibility to HIV infection and progression of AIDS in relation to variant alleles of mannose-binding lectin. Lancet 349, 236–240. https://doi.org/10.1016/S0140-6736(96)08440-1

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C., Estoup, A., 2013. Estimation of population allele frequencies from next-generation sequencing data: Pool-versus individual-based genotyping. Mol. Ecol. 22, 3766–3779. https://doi.org/10.1111/mec.12360

Gershwin, L.J., 2012. Immunology of bovine respiratory syncytial virus infection of cattle. Comp. Immunol. Microbiol. Infect. Dis. 35, 253–257. https://doi.org/10.1016/j.cimid.2012.01.005

Ghildyal, R., Hartley, C., Varrasso, A., Meanger, J., Voelker, D.R., Anders, E.M., Mills, J., 1999. Surfactant protein A binds to the fusion glycoprotein of respiratory syncytial virus and neutralizes virion infectivity. J. Infect. Dis. 180, 2009–2013. https://doi.org/10.1086/315134

Gibbs, R.A., Taylor, J.F., Van Tassell, C.P., Barendse, W., Eversole, K.A., Gill, C.A., Green,

R.D., Hamernik, D.L., Kappes, S.M., Lien, S., Matukumalli, L.K., McEwan, J.C., Nazareth, L. V., Schnabel, R.D., Weinstock, G.M., Wheeler, D.A., Ajmone-Marsan, P., Boettcher, P.J., Caetano, A.R., Garcia, J.F., Hanotte, O., Mariani, P., Skow, L.C., Sonstegard, T.S., Williams, J.L., Diallo, B., Hailemariam, L., Martinez, M.L., Morris, C.A., Silva, L.O.C., Spelman, R.J., Mulatu, W., Zhao, K., Abbey, C.A., Agaba, M., Araujo, F.R., Bunch, R.J., Burton, J., Gorni, C., Olivier, H., Harrison, B.E., Luff, B., Machado, M.A., Mwakaya, J., Plastow, G., Sim, W., Smith, T., Thomas, M.B., Valentini, A., Williams, P., Womack, J., Woolliams, J.A., Liu, Y., Qin, X., Worley, K.C., Gao, C., Jiang, H., Moore, S.S., Ren, Y., Song, X.-Z., Bustamante, C.D., Hernandez, R.D., Muzny, D.M., Patil, S., San Lucas, A., Fu, Q., Kent, M.P., Vega, R., Matukumalli, A., McWilliam, S., Sclep, G., Bryc, K., Choi, J., Gao, H., Grefenstette, J.J., Murdoch, B., Stella, A., Villa-Angulo, R., Wright, M., Aerts, J., Jann, O., Negrini, R., Goddard, M.E., Hayes, B.J., Bradley, D.G., Barbosa da Silva, M., Lau, L.P.L., Liu, G.E., Lynn, D.J., Panzitta, F., Dodds, K.G., 2009. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science (80-.). 324, 528–532. https://doi.org/10.1126/science.1167936

Gjerstorff, M., Hansen, S., Jensen, B., Dueholm, B., Horn, P., Bendixen, C., Holmskov, U., 2004a. The genes encoding bovine SP-A, SP-D, MBL-A, conglutinin, CL-43 and CL-46 form a distinct collectin locus on Bos taurus chromosome 28 (BTA28) at position q.1.8-1.9. Anim. Genet. 35, 333–337. https://doi.org/10.1111/j.1365-2052.2004.01167.x

Gjerstorff, M., Madsen, J., Bendixen, C., Holmskov, U., Hansen, S., 2004b. Genomic and molecular characterization of bovine surfactant protein D (SP-D). Mol. Immunol. 41, 369–376. https://doi.org/10.1016/j.molimm.2004.03.005

Go, Y.Y., Bailey, E., Cook, D.G., Coleman, S.J., MacLeod, J.N., Chen, K.-C., Timoney, P.J., Balasuriya, U.B.R., 2011. Genome-wide association study among four horse breeds identifies a common haplotype associated with in vitro CD3+ T cell susceptibility/resistance to equine arteritis virus Infection. J. Virol. 85, 13174–13184. https://doi.org/10.1128/JVI.06068-11

Goddard, M.E., Hayes, B.J., 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10, 381–91. https://doi.org/10.1038/nrg2575

Goldammer, T., Zerbe, H., Molenaar, A., Schuberth, H.-J., Brunner, R.M., Kata, S.R., Seyfert, H.-M., 2004. Mastitis increases mammary mRNA abundance of beta-defensin 5, toll-like-receptor 2 (TLR2), and TLR4 but not TLR9 in cattle. Clin. Diagn. Lab. Immunol. 11, 174–85. https://doi.org/10.1128/CDLI.11.1.174-185.2004

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., Lopez, R., 2010. A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res. 38, W695-9. https://doi.org/10.1093/nar/gkq313

Gowdy, K.M., Cardona, D.M., Nugent, J.L., Giamberardino, C., Thomas, J.M., Mukherjee, S., Martinu, T., Foster, W.M., Plevy, S.E., Pastva, A.M., Wright, J.R., Palmer, S.M., Palmer, S.M., 2012. Novel role for surfactant protein A in gastrointestinal graft-versus-host disease. J. Immunol. 188, 4897–4905. https://doi.org/10.4049/jimmunol.1103558

Grageda, M., Silveyra, P., Thomas, N.J., DiAngelo, S.L., Floros, J., 2014. DNA methylation

profile and expression of surfactant protein A2 gene in lung cancer. Exp. Lung Res. 41, 93–102. https://doi.org/10.3109/01902148.2014.976298

Grechukhina, O., Petracco, R., Popkhadze, S., Massasa, E., Paranjape, T., Chan, E., Flores, I., Weidhaas, J.B., Taylor, H.S., 2012. A polymorphism in a let-7 microRNA binding site of KRAS in women with endometriosis. EMBO Mol. Med. 4, 206–217. https://doi.org/10.1002/emmm.201100200

Groenen, M.A.M., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J., Li, S., Larkin, D.M., Kim, H., Frantz, L.A.F., Caccamo, M., Ahn, H., Aken, B.L., Anselmo, A., Anthon, C., Auvil, L., Badaoui, B., Beattie, C.W., Bendixen, C., Berman, D., Blecha, F., Blomberg, J., Bolund, L., Bosse, M., Botti, S., Bujie, Z., Bystrom, M., Capitanu, B., Carvalho-Silva, D., Chardon, P., Chen, C., Cheng, R., Choi, S.-H., Chow, W., Clark, R.C., Clee, C., Crooijmans, R.P.M.A., Dawson, H.D., Dehais, P., De Sapio, F., Dibbits, B., Drou, N., Du, Z.-Q., Eversole, K., Fadista, J., Fairley, S., Faraut, T., Faulkner, G.J., Fowler, K.E., Fredholm, M., Fritz, E., Gilbert, J.G.R., Giuffra, E., Gorodkin, J., Griffin, D.K., Harrow, J.L., Hayward, A., Howe, K., Hu, Z.-L., Humphray, S.J., Hunt, T., Hornshøj, H., Jeon, J.-T., Jern, P., Jones, M., Jurka, J., Kanamori, H., Kapetanovic, R., Kim, J.J.-H., Kim, K.-W., Kim, T.-H., Larson, G., Lee, K.K.-T., Leggett, R., Lewin, H.A., Li, Y., Liu, W., Loveland, J.E., Lu, Y., Lunney, J.K., Ma, J., Madsen, O., Mann, K., Matthews, L., McLaren, S., Morozumi, T., Murtaugh, M.P., Narayan, J., Nguyen, D.T., Ni, P., Oh, S.-J., Onteru, S., Panitz, F., Park, E.-W., Park, H.-S., Pascal, G., Paudel, Y., Perez-Enciso, M., Ramirez-Gonzalez, R., Reecy, J.M., Rodriguez-Zas, S., Rohrer, G.A., Rund, L., Sang, Y., Schachtschneider, K., Schraiber, J.G., Schwartz, J., Scobie, L., Scott, C., Searle, S., Servin, B., Southey, B.R., Sperber, G., Stadler, P., Sweedler, J. V, Tafer, H., Thomsen, B., Wali, R., Wang, J.J., White, S., Xu, X., Yerle, M., Zhang, G., Zhang, J.J., Zhao, S., Rogers, J., Churcher, C., Schook, L.B., Pérez-Enciso, M., 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491, 393–398. https://doi.org/10.1038/nature11622

Guo, N., Mogues, T., Weremowicz, S., Morton, C.C., 1998. The human ortholog of rhesus mannose-binding protein-A gene is an expressed pseudogene that localizes to chromosome 10. Mamm. Genome 9, 246–249. https://doi.org/10.1007/s003359900735

Haczku, A., Cao, Y., Vass, G., Kierstein, S., Nath, P., Atochina-Vasserman, E.N., Scanlon, S.T., Li, L., Griswold, D.E., Chung, K.F., Poulain, F.R., Hawgood, S., Beers, M.F., Crouch, E.C., 2006. IL-4 and IL-13 form a negative feedback circuit with surfactant protein-D in the allergic airway response. J. Immunol. 176, 3557–3565. https://doi.org/10.4049/jimmunol.176.6.3557

Halasa, T., Huijps, K., Østerås, O., Hogeveen, H., 2007. Economic effects of bovine mastitis and mastitis management: A review. Vet. Q. 29, 18–31. https://doi.org/10.1080/01652176.2007.9695224

Hansen, S., Holm, D., Moeller, V., Vitved, L., 2003a. Genomic and molecular characterization of CL-43 and its proximal promoter. Biochim. Biophys. Acta (BBA)-General Subj.

Hansen, S., Holm, D., Moeller, V., Vitved, L., Bendixen, C., Reid, K.B.M., Skjoedt, K., Holmskov, U., 2002. CL-46, a novel collectin highly expressed in bovine thymus and liver.

J. Immunol. 169, 5726–5734. https://doi.org/10.4049/jimmunol.169.10.5726

Hansen, S., Holm, D., Moeller, V., Vitved, L., Bendixen, C., Skjoedt, K., Holmskov, U., 2003b. Genomic and molecular characterization of CL-43 and its proximal promoter. Biochim. Biophys. Acta 1625, 1–10. https://doi.org/10.1016/S0167-4781(02)00531-6

Hansen, S., Selman, L., Palaniyar, N., Ziegler, K., Brandt, J., Kliem, A., Jonasson, M., Skjoedt, M.-O., Nielsen, O., Hartshorn, K.L., Jørgensen, T.J.D., Skjødt, K., Holmskov, U., 2010. Collectin 11 (CL-11, CL-K1) Is a MASP-1/3–Associated Plasma Collectin with Microbial-Binding Activity. J. Immunol. 185, 6096–6104. https://doi.org/10.4049/jimmunol.1002185

Hansen, S.S., Holmskov, U., 2002. Lung surfactant protein D (SP-D) and the molecular diverted descendants: conglutinin, CL-43 and CL-46. Immunobiology 205, 498–517. https://doi.org/10.1078/0171-2985-00150

Hartshorn, K.L., Holmskov, U., Hansen, S., Zhang, P., Meschi, J., Mogues, T., White, M.R., Crouch, E.C., 2002. Distinctive anti-influenza properties of recombinant collectin 43. Biochem. J. 366, 87–96. https://doi.org/10.1042/BJ20011868

Hartshorn, K.L., Sastry, K., Brown, D., White, M.R., Okarma, T.B., Lee, Y.M., Tauber, A.I., 1993. Conglutinin acts as an opsonin for influenza A viruses. J. Immunol. 151, 6265–6273.

Hartshorn, K.L., White, M.R., Shepherd, V., Reid, K., Jensenius, J.C., Crouch, E.C., 1997. Mechanisms of anti-influenza activity of surfactant proteins A and D: comparison with serum collectins. Am. J. Physiol. 273, L1156-66.

Hartshorn, K.L., White, M.R., Smith, K., Sorensen, G., Kuroki, Y., Holmskov, U., Head, J., Crouch, E.C., 2010a. Increasing antiviral activity of surfactant protein D trimers by introducing residues from bovine serum collectins: dissociation of mannan-binding and antiviral activity. Scand. J. Immunol. 72, 22–30. https://doi.org/10.1111/j.1365-3083.2010.02409.x

Hartshorn, K.L., White, M.R., Tecle, T., Sorensen, G., Holmskov, U., Crouch, E.C., 2010b. Viral aggregating and opsonizing activity in collectin trimers. Am. J. Physiol. Lung Cell. Mol. Physiol. 298, L79-88. https://doi.org/10.1152/ajplung.00223.2009

Heikkilä, A.-M., Nousiainen, J.I., Pyörälä, S., 2012. Costs of clinical mastitis with special reference to premature culling. J. Dairy Sci. 95, 139–150. https://doi.org/10.3168/jds.2011-4321

Herrera-Ramos, E., López-Rodríguez, M., Ruíz-Hernández, J., Horcajada, J., Borderías, L., Lerma, E., Blanquer, J., Pérez-González, M., García-Laorden, M., Florido, Y., Mas-Bosch, V., Montero, M., Ferrer, J., Sorlí, L., Vilaplana, C., Rajas, O., Briones, M., Aspa, J., López-Granados, E., Solé-Violán, J., de Castro, F., Rodríguez-Gallego, C., 2014. Surfactant protein A genetic variants associate with severe respiratory insufficiency in pandemic influenza A virus infection. Crit. Care 18, R127. https://doi.org/10.1186/cc13934

Hibberd, M.L., Sumiya, M., Summerfield, J.A., Booy, R., Levin, M., 1999. Association of variants of the gene for mannose-binding lectin with susceptibility to meningococcal disease. Lancet 353, 1049–1053. https://doi.org/10.1016/S0140-6736(98)08350-0

Hickling, T.P., Bright, H., Wing, K., Gower, D., Martin, S.L., Sim, R.B., Malhotra, R., 1999. A

recombinant trimeric surfactant protein D carbohydrate recognition domain inhibits respiratory syncytial virus infection in vitro and in vivo. Eur. J. Immunol. 29, 3478–84. https://doi.org/10.1002/(SICI)1521-4141(199911)29:11<3478::AID-IMMU3478>3.0.CO;2-W

Hobo, S., Niwa, H., Anzai, T., 2007. Evaluation of serum amyloid A and surfactant protein D in sera for identification of the clinical condition of horses with bacterial pneumonia. J. Vet. Med. Sci. 69, 827–830. https://doi.org/10.1292/jvms.69.827

Hobo, S., Ogasawara, Y., Kuroki, Y., Akino, T., Yoshihara, T., 1999a. Purification and biochemical characterization of equine pulmonary surfactant protein D. Am. J. Vet. Res. 60, 368–372.

Hobo, S., Ogasawara, Y., Kuroki, Y., Akino, T., Yoshihara, T., 1999b. Purification and biochemical characterization of pulmonary surfactant protein A of horses. Am. J. Vet. Res. 60, 169–173.

Hodgkinson, A., Eyre-Walker, A., 2011. Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. https://doi.org/10.1038/nrg3098

Holmskov, U., Jensenius, J.C., Tornoe, I., Lovendahl, P., Tornøe, I., Løvendahl, P., 1998. The plasma levels of conglutinin are heritable in cattle and low levels predispose to infection. Immunology 93, 431–436. https://doi.org/10.1046/j.1365-2567.1998.00452.x

Holmskov, U., Malhotra, R., Sim, R.B., Jensenius, J.C., 1994. Collectins: collagenous C-type lectins of the innate immune defense system. Immunol. Today 15, 67–74. https://doi.org/10.1016/0167-5699(94)90136-8

Holmskov, U., Teisner, B., Willis, A.C., Reid, K.B.M., Jensenius, J.C., 1993. Purification and characterization of a bovine serum lectin (CL-43) with structural homology to conglutinin and SP-D and carbohydrate specificity similar to mannan-binding protein. J. Biol. Chem. 268, 10120–10125.

Holmskov, U., Thiel, S., Jensenius, J.C., 2003. Collections and ficolins: humoral lectins of the innate immune defense. Annu. Rev. Immunol. 21, 547–78. https://doi.org/10.1146/annurev.immunol.21.120601.140954

Hu, Z.L., Park, C.A., Reecy, J.M., 2016. Developmental progress and current status of the Animal QTLdb. Nucleic Acids Res. 44, D827–D833. https://doi.org/10.1093/nar/gkv1233

Hughes, A.L., 2007. Evolution of the lung surfactant proteins in birds and mammals. Immunogenetics 59, 565–572. https://doi.org/10.1007/s00251-007-0218-6

Hummelshoj, T., Munthe-Fog, L., Madsen, H.O., Fujita, T., Matsushita, M., Garred, P., 2005. Polymorphisms in the FCN2 gene determine serum variation and function of Ficolin-2. Hum. Mol. Genet. 14, 1651–1658. https://doi.org/10.1093/hmg/ddi173

Iborra, M., Bernuzzi, F., Invernizzi, P., Danese, S., 2012. MicroRNAs in autoimmunity and inflammatory bowel disease: Crucial regulators in immune response. Autoimmun. Rev. 11, 305–314. https://doi.org/10.1016/j.autrev.2010.07.002

Ichijo, H., Hellman, U., Wernstedt, C., Gonez, L.J., Claesson-Welsh, L., Heldin, C.H.,

Miyazono, K., 1993. Molecular cloning and characterization of ficolin, a multimeric protein with fibrinogen- and collagen-like domains. J. Biol. Chem. 268, 14505–13.

Ikeda, K., Shiratori, M., Chiba, H., Nishikiori, H., Yokoo, K., Saito, A., Hasegawa, Y., Kuronuma, K., Otsuka, M., Yamada, G., Takahashi, H., 2017. Serum surfactant protein D predicts the outcome of patients with idiopathic pulmonary fibrosis treated with pirfenidone. Respir. Med. 131, 184–191. https://doi.org/10.1016/j.rmed.2017.08.021

Ingels, C., Vanhorebeek, I., Steffensen, R., Derese, I., Jensen, L., Wouters, P.J., Hermans, G., Thiel, S., Van Den Berghe, G., 2014. Lectin pathway of complement activation and relation with clinical complications in critically ill children. Pediatr. Res. 75, 99–108. https://doi.org/10.1038/pr.2013.180

Ishii, T., Hagiwara, K., Ikeda, S., Arai, T., Mieno, M.N., Kumasaka, T., Muramatsu, M., Sawabe, M., Gemma, A., Kida, K., 2012. Association between genetic variations in surfactant protein D and emphysema, interstitial pneumonia, and lung cancer in a Japanese population. COPD J. Chronic Obstr. Pulm. Dis. 9, 409–416. https://doi.org/10.3109/15412555.2012.676110

Iwaki, D., Kanno, K., Takahashi, M., Endo, Y., Lynch, N.J., Schwaeble, W.J., Matsushita, M., Okabe, M., Fujita, T., 2006. Small mannose-binding lectin-associated protein plays a regulatory role in the lectin complement pathway. J. Immunol. 177, 8626–8632. https://doi.org/10.4049/jimmunol.177.12.8626

Iwasaki, A., Medzhitov, R., Akiko, I., Ruslan, M., 2015. Control of adaptive immunity by the innate immune system. Nat. Immunol. 16, 343–353. https://doi.org/10.1038/ni.3123

Janeway, C.A., 1989. Approaching the asymptote? Evolution and revolution in immunology. Cold Spring Harb. Symp. Quant. Biol. 54 Pt 1, 1–13. https://doi.org/10.1101/SQB.1989.054.01.003

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J.M., Vincentelli, R., Luscombe, N.M., Hughes, T.R., Lemaire, P., Ukkonen, E., Kivioja, T., Taipale, J., 2013. DNA-binding specificities of human transcription factors. Cell 152, 327–339. https://doi.org/10.1016/j.cell.2012.12.009

Juul-Madsen, H.R., Kjærup, R.M., Toft, C., Henryon, M., Heegaard, P.M.H., Berg, P., Dalgaard, T.S., 2011. Structural gene variants in the porcine mannose-binding lectin 1 (MBL1) gene are associated with low serum MBL-A concentrations. Immunogenetics 63, 309–317. https://doi.org/10.1007/s00251-011-0512-1

Kafri, R., Levy, M., Pilpel, Y., 2006. The regulatory utilization of genetic redundancy through responsive backup circuits. Proc. Natl. Acad. Sci. 103, 11653–11658. https://doi.org/10.1073/pnas.0604883103

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., Kent, W.J., 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32, D493-6. https://doi.org/10.1093/nar/gkh103

Kasahara, T., Miyazaki, T., Nitta, H., Ono, A., Miyagishima, T., Nagao, T., Urushidani, T.,

2006. Evaluation of methods for duration of preservation of RNA quality in rat liver used for transcriptome analysis. J. Toxicol. Sci. 31, 509–519. https://doi.org/10.2131/jts.31.509

Kawai, T., Akira, S., 2009. The roles of TLRs, RLRs and NLRs in pathogen recognition. Int. Immunol. 21, 317–337. https://doi.org/10.1093/intimm/dxp017

Kawashima, K., Meyerholz, D.K., Gallup, J.M., Grubor, B., Lazic, T., Lehmkuhl, H.D., Ackermann, M.R., 2006. Differential expression of ovine innate immune genes by preterm and neonatal lung epithelia infected with respiratory syncytial virus. Viral Immunol 19, 316–323. https://doi.org/10.1089/vim.2006.19.316

Keele, J.W., Kuehn, L.A., McDaneld, T.G., Tait, R.G., Jones, S.A., Smith, T.P.L., Shackelford, S.D., King, D.A., Wheeler, T.L., Lindholm-Perry, A.K., McNeel, A.K., 2015. Genomewide association study of lung lesions in cattle using sample pooling. J. Anim. Sci. 93, 956–964. https://doi.org/10.2527/jas2014-8492

Keirstead, N.D., Hayes, M.A., Vandervoort, G.E., Brooks, A.S., Squires, E.J., Lillie, B.N., 2011. Single nucleotide polymorphisms in collagenous lectins and other innate immune genes in pigs with common infectious diseases. Vet. Immunol. Immunopathol. 142, 1–13. https://doi.org/10.1016/j.vetimm.2011.02.011

Keirstead, N.D., Lee, C., Yoo, D., Brooks, A.S., Hayes, M.A., 2008. Porcine plasma ficolin binds and reduces infectivity of porcine reproductive and respiratory syndrome virus (PRRSV) in vitro. Antiviral Res. 77, 28–38. https://doi.org/10.1016/j.antiviral.2007.08.001

Keshi, H., Sakamoto, T., Kawai, T., Ohtani, K., Katoh, T., Jang, S.-J., Motomura, W., Yoshizaki, T., Fukuda, M., Koyama, S., Fukuzawa, J., Fukuoh, A., Yoshida, I., Suzuki, Y., Wakamiya, N., 2006. Identification and characterization of a novel human collectin CL-K1. Microbiol. Immunol. 50, 1001–1013. https://doi.org/10.1111/j.1348-0421.2006.tb03868.x

Kim, S.Y., Li, Y., Guo, Y., Li, R., Holmkvist, J., Hansen, T., Pedersen, O., Wang, J., Nielsen, R., 2010. Design of association studies with pooled or un-pooled next-generation sequencing data. Genet. Epidemiol. 34, 479–491. https://doi.org/10.1002/gepi.20501

Kjærup, R.M., Norup, L.R., Skjødt, K., Dalgaard, T.S., Juul-Madsen, H.R., 2013. Chicken mannose-binding lectin (MBL) gene variants with influence on MBL serum concentrations. Immunogenetics 65, 461–471. https://doi.org/10.1007/s00251-013-0689-6

Kofler, R., Pandey, R.V., Schlotterer, C., Schlötterer, C., 2011. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27, 3435–3436. https://doi.org/10.1093/bioinformatics/btr589

Koltes, J.E., Fritz-Waters, E., Eisley, C.J., Choi, I., Bao, H., Kommadath, A., Serão, N.V.L., Boddicker, N.J., Abrams, S.M., Schroyen, M., Loyd, H., Tuggle, C.K., Plastow, G.S., Guan, L., Stothard, P., Lunney, J.K., Liu, P., Carpenter, S., Rowland, R.R.R., Dekkers, J.C.M., Reecy, J.M., 2015. Identification of a putative quantitative trait nucleotide in guanylate binding protein 5 for host response to PRRS virus infection. BMC Genomics 16, 412. https://doi.org/10.1186/s12864-015-1635-9

Kozomara, A., Griffiths-Jones, S., 2014. MiRBase: Annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 42. https://doi.org/10.1093/nar/gkt1181

Krarup, A., Thiel, S., Hansen, A., Fujita, T., Jensenius, J.C., 2004. L-ficolin is a pattern recognition molecule specific for acetyl groups. J. Biol. Chem. 279, 47513–47519. https://doi.org/10.1074/jbc.M407161200

Kuhlman, M., Joiner, K., Ezekowitz, R.A.B., 1989. The human mannose-binding protein functions as an opsonin. J. Exp. Med. 169, 1733–45. https://doi.org/10.1084/jem.169.5.1733

Kuhn, D.E., Martin, M.M., Feldman, D.S., Terry, A. V., Nuovo, G.J., Elton, T.S., 2008. Experimental validation of miRNA targets. Methods 44, 47–54. https://doi.org/10.1016/j.ymeth.2007.09.005

Larsen, F., Madsen, H.O., Sim, R.B., Koch, C., Garred, P., 2004. Disease-associated mutations in human mannose-binding lectin compromise oligomerization and activity of the final protein. J. Biol. Chem. 279, 21302–21311. https://doi.org/10.1074/jbc.M400520200

Lavoie, J.P., 2015. Is the time primed for equine asthma? Equine Vet. Educ. 27, 225–226. https://doi.org/10.1111/eve.12332

Leclere, M., Lavoie-Lamoureux, A., Lavoie, J.-P., 2011. Heaves, an asthma-like disease of horses. Respirology 16, 1027–1046. https://doi.org/10.1111/j.1440-1843.2011.02033.x

Lee, R.T., Ichikawa, Y., Fay, M., Drickamer, K., Shao, M.C., Lee, Y.C., 1991. Ligand-binding characteristics of rat serum-type mannose-binding protein (MBP-A). Homology of binding site architecture with mammalian and chicken hepatic lectins. J. Biol. Chem. 266, 4810–5.

Lee, R.T., Ichikawa, Y., Kawasaki, T., Drickamer, K., Lee, Y.C., 1992. Multivalent ligand binding by serum mannose-binding protein. Arch. Biochem. Biophys. 299, 129–136. https://doi.org/10.1016/0003-9861(92)90254-T

Leth-Larsen, R., Garred, P., Jensenius, H., Meschi, J., Hartshorn, K.L., Madsen, J., Tornøe, I., Madsen, H.O., Sorensen, G., Crouch, E., Holmskov, U., Tornoe, I., Sørensen, G., 2005. A common polymorphism in the SFTPD gene influences assembly, function, and concentration of surfactant protein D. J. Immunol. 174, 1532–1538. https://doi.org/10.4049/jimmunol.174.3.1532

Levings, R.L., Roth, J.A., 2013. Immunity to bovine herpesvirus 1: I. viral lifecycle and innate immunity. Anim. Health Res. Rev. 14, 88–102. https://doi.org/10.1017/S1466252313000042

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T.J., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J., Qi, H., Sun, L., Shen, C., Jiao, W., Xu, F., Xiao, J., Shen, A., 2016. Rs1914663 of SFTPA 1 gene is associated with pediatric tuberculosis in Han Chinese population. Infect. Genet. Evol. 41, 16–20. https://doi.org/10.1016/j.meegid.2016.03.018

Li, L., Zheng, Q., Zhang, Y., Li, P., Fu, Y., Hou, J., Xiao, X., 2016. Antiviral activity of

recombinant porcine surfactant protein A against porcine reproductive and respiratory syndrome virus in vitro. Arch. Virol. 161, 1883–1890. https://doi.org/10.1007/s00705-016-2838-3

Li, R.W., Rinaldi, M., Capuco, A. V., 2011. Characterization of the abomasal transcriptome for mechanisms of resistance to gastrointestinal nematodes in cattle. Vet. Res. 42. https://doi.org/10.1186/1297-9716-42-114

Liaubet, L., Lobjois, V., Faraut, T., Tircazes, A., Benne, F., Iannuccelli, N., Pires, J., Glénisson, J., Robic, A., Le Roy, P., SanCristobal, M., Cherel, P., 2011. Genetic variability of transcript abundance in pig peri-mortem skeletal muscle: eQTL localized genes involved in stress response, cell death, muscle disorders and metabolism. BMC Genomics 12. https://doi.org/10.1186/1471-2164-12-548

Lillie, B.N., Brooks, A.S., Keirstead, N.D., Hayes, M.A., 2005. Comparative genetics and innate immune functions of collagenous lectins in animals. Vet. Immunol. Immunopathol. 108, 97–110. https://doi.org/10.1016/j.vetimm.2005.07.001

Lillie, B.N., Hammermueller, J.D., MacInnes, J.I., Jacques, M., Hayes, M.A., 2006a. Porcine mannan-binding lectin A binds to Actinobacillus suis and Haemophilus parasuis. Dev. Comp. Immunol. 30, 954–65. https://doi.org/10.1016/j.dci.2005.12.008

Lillie, B.N., Keirstead, N.D., Squires, E.J., Hayes, M.A., 2007. Gene polymorphisms associated with reduced hepatic expression of porcine mannan-binding lectin C. Dev. Comp. Immunol. 31, 830–846. https://doi.org/10.1016/j.dci.2006.11.002

Lillie, B.N., Keirstead, N.D., Squires, E.J., Hayes, M.A., 2006b. Single-nucleotide polymorphisms in porcine mannan-binding lectin A. Immunogenetics 58, 983–993. https://doi.org/10.1007/s00251-006-0160-z

Lin, Z., John, G., Hegarty, J.P., Berg, A., Yu, W., Wang, Y., Kelly, A.A., Peterson, B.Z., Poritz, L.S., Floros, J., Koltun, W.A., 2011. Genetic variants and monoallelic expression of surfactant protein-D in inflammatory bowel disease. Ann. Hum. Genet. 75, 559–568. https://doi.org/10.1111/j.1469-1809.2011.00662.x

Lipscombe, R.J., Sumiya, M., Summerfield, J.A., Turner, M.W., 1995. Distinct physicochemical characteristics of human mannose binding protein expressed by individuals of differing genotype. Immunology 85, 660–667.

Liu, D., Yi, M., Smith, M., Mendelson, C.R., 2008. TTF-1 response element is critical for temporal and spatial regulation and necessary for hormonal regulation of human surfactant protein-A2 promoter activity. AJP Lung Cell. Mol. Physiol. 295, L264–L271. https://doi.org/10.1152/ajplung.00069.2008

Liu, J., Ju, Z., Li, Q., Huang, J., Li, R., Li, J., Ma, L., Zhong, J., Wang, C., 2011. Mannose-binding lectin 1 haplotypes influence serum MBL-A concentration, complement activity, and milk production traits in Chinese Holstein cattle. Immunogenetics 63, 727–742. https://doi.org/10.1007/s00251-011-0548-2

Lizundia, R., Sauter, K.S., Taylor, G., Werling, D., 2008. Host species-specific usage of the TLR4-LPS receptor complex. Innate Immun. 14, 223–231.

https://doi.org/10.1177/1753425908095957

Lu, J., Teh, C., Kishore, U., Reid, K.B.M., 2002. Collectins and ficolins: sugar pattern recognition molecules of the mammalian innate immune system. Biochim. Biophys. Acta 1572, 387–400. https://doi.org/10.1016/S0304-4165(02)00320-3

Luz, P.R., Miyazaki, M.I., Chiminacio Neto, N., Padeski, M.C., Barros, A.C.M., Boldt, A.B.W., Messias-Reason, I.J., 2016. Genetically determined MBL deficiency is associated with protection against chronic cardiomyopathy in Chagas disease. PLoS Negl. Trop. Dis. 10, e0004257-16. https://doi.org/10.1371/journal.pntd.0004257

Ma, Y.J., Hein, E., Munthe-Fog, L., Skjoedt, M.-O., Bayarri-Olmos, R., Romani, L., Garred, P., 2015. Soluble collectin-12 (CL-12) is a pattern recognition molecule initiating complement activation via the alternative pathway. J. Immunol. 195, 3365–73. https://doi.org/10.4049/jimmunol.1500493

Ma, Y.J., Skjoedt, M.-O.O., Garred, P., 2013. Collectin-11/MASP complex formation triggers activation of the lectin complement pathway--the fifth lectin pathway initiation complex. J. Innate Immun. 5, 242–250. https://doi.org/10.1159/000345356

Madsen, H.O., Garred, P., Kurtzhals, J.A.L., Lamm, L.U., Ryder, L.P., Thiel, S., Svejgaard, A., 1994. A new frequent allele is the missing link in the structural polymorphism of the human mannan-binding protein. Immunogenetics 40, 37–44. https://doi.org/10.1007/BF00163962

Madsen, H.O., Satz, M.L., Hogh, B., Svejgaard, A., Garred, P., 1998. Different molecular events result in low protein levels of mannan-binding lectin in populations from southeast Africa and South America. J. Immunol. 161, 3169–3175.

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., Turner, D.J., 2010. Target-enrichment strategies for next-generation sequencing. Nat. Methods 7, 111–118. https://doi.org/10.1038/nmeth.1419

Manly, K.F., 2005. Reliability of statistical associations between genes and disease. Immunogenetics 57, 549–58. https://doi.org/10.1007/s00251-005-0025-x

Maston, G.A., Evans, S.K., Green, M.R., 2006. Transcriptional regulatory elements in the human genome. Annu. Rev. Genomics Hum. Genet. 7, 29–59. https://doi.org/10.1146/annurev.genom.7.080505.115623

Matsushita, M., 2010. Ficolins: Complement-activating lectins involved in innate immunity. J. Innate Immun. 2, 24–32. https://doi.org/10.1159/000228160

Matsushita, M., Endo, Y., Fujita, T., 2013. Structural and functional overview of the lectin complement pathway: Its molecular basis and physiological implication. Arch. Immunol. Ther. Exp. (Warsz). 61, 273–283. https://doi.org/10.1007/s00005-013-0229-y

Matsushita, M., Endo, Y., Hamasaki, N., Fujita, T., 2001. Activation of the lectin complement pathway by ficolins. Int. Immunopharmacol. 1, 359–363. https://doi.org/10.1016/S1567-5769(00)00045-X

Matsushita, M., Fujita, T., 1995. Cleavage of the third component of complement (C3) by mannose-binding protein-associated serine protease (MASP) with subsequent complement

activation. Immunobiology 194, 443–448. https://doi.org/10.1016/S0171-2985(11)80110-5

Matsushita, M., Fujita, T., 1992. Activation of the classical complement pathway by mannose-binding protein in association with a novel C1s-like serine protease. J. Exp. Med. 176, 1497–502. https://doi.org/10.1084/jem.176.6.1497

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R., Kaul, R., Stamatoyannopoulos, J.A., 2012. Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190–5. https://doi.org/10.1126/science.1222794

Maxie, M.G., 2015. Jubb, Kennedy, and Palmer's Pathology of Domestic Animals, 6th ed. Elsevier, Edinburgh.

McCormack, F.X., Damodarasamy, M., Elhalwagi, B.M., 1999. Deletion mapping of N-terminal domains of surfactant protein A. The N-terminal segment is required for phospholipid aggregation and specific inhibition of surfactant secretion. J. Biol. Chem. 274, 3173–81. https://doi.org/10.1074/jbc.274.5.3173

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S.B., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–303. https://doi.org/10.1101/gr.107524.110

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., Cunningham, F., 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26, 2069–2070. https://doi.org/10.1093/bioinformatics/btq330

McQueen, C.M., Doan, R., Dindot, S. V., Bourquin, J.R., Zlatev, Z.Z., Chaffin, M.K., Blodgett, G.P., Ivanov, I., Cohen, N.D., 2014. Identification of genomic loci associated with Rhodococcus equi susceptibility in foals. PLoS One 9, e98710. https://doi.org/10.1371/journal.pone.0098710

Medzhitov, R., Janeway, C.A., 1997. Innate immunity: the virtues of a nonclonal system of recognition. Cell 91, 295–8. https://doi.org/10.1016/j.immuni.2007.05.022

Medzhitov, R., Preston-Hurlburt, P., Janeway, C.A., 1997. A human homologue of the Drosophila Toll protein signals activation of adaptive immunity. Nature 388, 394–397. https://doi.org/10.1038/41131

Messeguer, X., Escudero, R., Farré, D., Núñez, O., Martínez, J., Albà, M.M., 2002. PROMO: detection of known transcription regulatory elements using species-tailored searches. Bioinformatics 18, 333–334. https://doi.org/10.1093/bioinformatics/18.2.333

Miles, D.G., 2009. Overview of the North American beef cattle industry and the incidence of bovine respiratory disease (BRD). Anim. Health Res. Rev. 10, 101–103. https://doi.org/10.1017/S1466252309990090

Minchinton, R.M., Dean, M.M., Clark, T.R., Heatley, S., Mullighan, C.G., 2002. Analysis of the

relationship between mannose-binding lectin (MBL) genotype, MBL levels and function in an Australian blood donor population. Scand. J. Immunol. 56, 630–641. https://doi.org/10.1046/j.1365-3083.2002.01167.x

Mogensen, T.H., 2009. Pathogen recognition and inflammatory signaling in innate immune defenses. Clin. Microbiol. Rev. 22, 240–273. https://doi.org/10.1128/CMR.00046-08

Mullen, M.P., Creevey, C.J., Berry, D.P., McCabe, M.S., Magee, D.A., Howard, D.J., Killeen, A.P., Park, S.D., McGettigan, P.A., Lucy, M.C., MacHugh, D.E., Waters, S.M., 2012. Polymorphism discovery and allele frequency estimation using high-throughput DNA sequencing of target-enriched pooled DNA samples. BMC Genomics 13, 16. https://doi.org/10.1186/1471-2164-13-16

Munoz, M., Carmen Rodriguez, M., Alves, E., Maria Folch, J., Ibanez-Escriche, N., Silio, L., Isabel Fernandez, A., 2013. Genome-wide analysis of porcine backfat and intramuscular fat fatty acid composition using high-density genotyping and expression data. BMC Genomics 14. https://doi.org/10.1186/1471-2164-14-845

Munthe-Fog, L., Hummelshøj, T., Honoré, C., Madsen, H.O., Permin, H., Garred, P., 2009. Immunodeficiency associated with FCN3 mutation and ficolin-3 deficiency. N. Engl. J. Med. 360, 2637–2644. https://doi.org/10.1056/NEJMoa0900381

Munthe-Fog, L., Hummelshøj, T., Ma, Y.J., Hansen, B.E., Koch, C., Madsen, H.O., Skjødt, K., Garred, P., 2008. Characterization of a polymorphism in the coding sequence of FCN3 resulting in a Ficolin-3 (Hakata antigen) deficiency state. Mol. Immunol. 45, 2660–2666. https://doi.org/doi:10.1016/j.molimm.2007.12.012

Murray, E., Khamri, W., Walker, M.M.M., Eggleton, P., Moran, A.P., Ferris, J.A., Knapp, S., Karim, Q.N., Worku, M., Strong, P., Reid, K.B.M., Thursz, M.R., 2002. Expression of surfactant protein D in the human gastric mucosa and during Helicobacter pylori infection. Innate Immun. 70, 1481–1487. https://doi.org/10.1128/IAI.70.3.1481

Nahid, A.M., Sugii, S., 2006. Binding of porcine ficolin-α to lipopolysaccharides from Gram-negative bacteria and lipoteichoic acids from Gram-positive bacteria. Dev. Comp. Immunol. 30, 335–343. https://doi.org/10.1016/j.dci.2005.04.002

Naito, H., Ikeda, a, Hasegawa, K., Oka, S., Uemura, K., Kawasaki, N., Kawasaki, T., 1999. Characterization of human serum mannan-binding protein promoter. J. Biochem. 126, 1004–1012. https://doi.org/10.1093/oxfordjournals.jbchem.a022543

Neibergs, H.L., Seabury, C.M., Wojtowicz, A.J., Wang, Z., Scraggs, E., Kiser, J.N., Neupane, M., Womack, J.E., van Eenennaam, A., Hagevoort, G.R., Lehenbauer, T.W., Aly, S., Davis, J., Taylor, J.F., Eenennaam, A., Hagevoort, G.R., Lehenbauer, T.W., Aly, S., Davis, J., Taylor, J.F., 2014. Susceptibility loci revealed for bovine respiratory disease complex in pre-weaned holstein calves. BMC Genomics 15, 1164. https://doi.org/10.1186/1471-2164-15-1164

Neill, J.D., 2013. Molecular biology of bovine viral diarrhea virus. Biologicals 41, 2–7. https://doi.org/10.1016/j.biologicals.2012.07.002

Nguyen, H.A., Rajaram, M.V.S., Meyer, D.A., Schlesinger, L.S., 2012. Pulmonary surfactant

protein A and surfactant lipids upregulate IRAK-M, a negative regulator of TLR-mediated inflammation in human macrophages. AJP Lung Cell. Mol. Physiol. 303, L608–L616. https://doi.org/10.1152/ajplung.00067.2012

Nitta, K.R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E.E.M., Taipale, J., 2015. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. Elife 2015, e04837. https://doi.org/10.7554/eLife.04837

O'Connell, R.M., Rao, D.S., Baltimore, D., 2012. microRNA regulation of inflammatory responses. Annu. Rev. Immunol. 30, 295–312. https://doi.org/10.1146/annurev-immunol-020711-075013

Ohashi, T., Erickson, H.P., 2004. The disulfide bonding pattern in ficolin multimers. J. Biol. Chem. 279, 6534–6539. https://doi.org/10.1074/jbc.M310555200

Ohtani, K., Suzuki, Y., Eda, S., Kawai, T., Kase, T., Keshi, H., Sakai, Y., Fukuoh, A., Sakamoto, T., Itabe, H., Suzutani, T., Ogasawara, M., Yoshida, I., Wakamiya, N., 2001. The membrane-type collectin CL-P1 is a scavenger receptor on vascular endothelial cells. J. Biol. Chem. 276, 44222–44228. https://doi.org/10.1074/jbc.M103942200

Ohtani, K., Suzuki, Y., Eda, S., Kawai, T., Kase, T., Yamazaki, H., Shimada, T., Keshi, H., Sakai, Y., Fukuoh, A., Sakamoto, T., Wakamiya, N., 1999. Molecular cloning of a novel human collectin from liver (CL-L1). J. Biol. Chem. 274, 13681–13689. https://doi.org/10.1074/jbc.274.19.13681

Ohtani, K., Suzuki, Y., Wakamiya, N., 2012. Biological functions of the novel collectins CL-L1, CL-K1, and CL-P1. J. Biomed. Biotechnol. 2012, 1–8. https://doi.org/10.1016/j.molimm.2005.09.015

Olde Riekerink, R.G.M., Barkema, H.W., Kelton, D.F., Scholl, D.T., 2008. Incidence rate of clinical mastitis on Canadian dairy farms. J. Dairy Sci. 91, 1366–1377. https://doi.org/10.3168/jds.2007-0757

Panciera, R.J., Confer, A.W., 2010. Pathogenesis and pathology of bovine pneumonia. Vet. Clin. North Am. Food Anim. Pract. 26, 191–214. https://doi.org/10.1016/j.cvfa.2010.04.001

Patel, S.M., Koringa, P.G., Nathani, N.M., Patel, N. V., Shah, T.M., Joshi, C.G., 2015. Exploring genetic polymorphism in innate immune genes in Indian cattle (Bos indicus) and buffalo (Bubalus bubalis) using next generation sequencing technology. Meta Gene 3, 50–58. https://doi.org/10.1016/j.mgene.2015.01.002

Patrick, R.L., 2009. A dairy producer's view of respiratory disease. Anim. Heal. Res. Rev. 10, 111–112. https://doi.org/10.1017/S1466252309990120

Pikaar, J.C., Voorhout, W.F., Van Golde, L.M.G., Verhoef, J., van Strijp, J.A.G., van Iwaarden, J.F., 1995. Opsonic activities of surfactant porteins A and D in phagocytosis of Gram-negative bacteria by alveolar macrophages. J. Infect. Dis. 172, 481–489. https://doi.org/10.1093/infdis/172.2.481

Podolsky, M.J., Lasker, A., Flaminio, M.J.B.F., Gowda, L.D., Ezekowitz, R.A.B., Takahashi, K., 2006. Characterization of an equine mannose-binding lectin and its roles in disease.

Biochem. Biophys. Res. Commun. 343, 928–936. https://doi.org/doi:10.1016/j.bbrc.2006.03.055

Ponsuksili, S., Murani, E., Trakooljul, N., Schwerin, M., Wimmers, K., 2014. Discovery of candidate genes for muscle traits based on GWAS supported by eQTL-analysis. Int. J. Biol. Sci. 10, 327–337. https://doi.org/10.7150/ijbs.8134

Prescott, J.F., Szkotnicki, J., McClure, J.T., Reid-Smith, R.J., Léger, D.F., 2012. Conference report: antimicrobial stewardship in Canadian agriculture and veterinary medicine. How is Canada doing and what still needs to be done? Can. Vet. J. = La Rev. Vet. Can. 53, 402–7.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575. https://doi.org/10.1086/519795

Qaseem, A.S., Sonar, S., Mahajan, L., Madan, T., Sorensen, G.L., Shamji, M.H., Kishore, U., 2013. Linking surfactant protein SP-D and IL-13: implications in asthma and allergy. Mol. Immunol. 54, 98–107. https://doi.org/10.1016/j.molimm.2012.10.039

Quinlan, A.R., Hall, I.M., 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team, 2017. R: A language and environment for statistical computing.

Rainard, P., 2003. The complement in milk and defense of the bovine mammary gland against infections. Vet. Res. 34, 647–670. https://doi.org/10.1051/vetres:2003025

Reiner, G., Dreher, F., Drungowski, M., Hoeltig, D., Bertsch, N., Selke, M., Willems, H., Gerlach, G.F., Probst, I., Tuemmler, B., Waldmann, K.H., Herwig, R., 2014. Pathway deregulation and expression QTLs in response to Actinobacillus pleuropneumoniae infection in swine. Mamm. Genome 25, 600–617. https://doi.org/10.1007/s00335-014-9536-9

Reinhardt, T.A., Sacco, R.E., Nonnecke, B.J., Lippolis, J.D., 2013. Bovine milk proteome: Quantitative changes in normal milk exosomes, milk fat globule membranes and whey proteomes resulting from Staphylococcus aureus mastitis. J. Proteomics 82, 141–154. https://doi.org/10.1016/j.jprot.2013.02.013

Rellstab, C., Zoller, S., Tedder, A., Gugerli, F., Fischer, M.C., 2013. Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. PLoS One 8, e80422. https://doi.org/10.1371/journal.pone.0080422

Richard, E.A., Pitel, P.H., Christmann, U., Lekeux, P., Fortier, G., Pronost, S., 2012. Serum concentration of surfactant protein D in horses with lower airway inflammation. Equine Vet. J. 44, 277–281. https://doi.org/10.1111/j.2042-3306.2011.00421.x

Riffo-Campos, Á.L., Riquelme, I., Brebi-Mieville, P., 2016. Tools for Sequence-Based miRNA Target Prediction: What to Choose? Int. J. Mol. Sci. 17, 1987. https://doi.org/10.3390/ijms17121987

Ron, M., Weller, J.I., 2007. From QTL to QTN identification in livestock - Winning by points rather than knock-out: A review. Anim. Genet. https://doi.org/10.1111/j.1365-2052.2007.01640.x

Rooryck, C., Diaz-Font, A., Osborn, D.P.S., Chabchoub, E., Hernandez-Hernandez, V., Shamseldin, H., Kenny, J., Waters, A., Jenkins, D., Kaissi, A. Al, Leal, G.F., Dallapiccola, B., Carnevale, F., Bitner-Glindzicz, M., Lees, M., Hennekam, R., Stanier, P., Burns, A.J., Peeters, H., Alkuraya, F.S., Beales, P.L., 2011. Mutations in lectin complement pathway genes COLEC11 and MASP1 cause 3MC syndrome. Nat. Genet. 43, 197–203. https://doi.org/10.1038/ng.757

Sano, H., Chiba, H., Iwaki, D., Sohma, H., Voelker, D.R., Kuroki, Y., 2000. Surfactant proteins A and D bind CD14 by different mechanisms. J. Biol. Chem. 275, 22442–22451. https://doi.org/10.1074/jbc.M001107200

Sarkar, S., Bailey, E., Go, Y.Y., Cook, R.F., Kalbfleisch, T., Eberth, J., Chelvarajan, R.L., Shuck, K.M., Artiushin, S., Timoney, P.J., Balasuriya, U.B.R., 2016. Allelic variation in CXCL16 determines CD3+ T lymphocyte susceptibility to equine arteritis virus infection and establishment of long-term carrier state in the stallion. PLOS Genet. 12, e1006467. https://doi.org/10.1371/journal.pgen.1006467

Schepers, J.A.A., Dijkhuizen, A.A.A., 1991. The economics of mastitis and mastitis control in dairy cattle: a critical analysis of estimates published since 1970. Prev. Vet. Med. 10, 213–224. https://doi.org/10.1016/0167-5877(91)90005-M

Schlötterer, C., Tobler, R., Kofler, R., Nolte, V., 2014. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. Nat. Rev. Genet. 15, 749–763. https://doi.org/10.1038/nrg3803

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., Talianidis, I., Flicek, P., Odom, D.T., 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328, 1036–40. https://doi.org/10.1126/science.1186176

Selman, L., Hansen, S., 2012. Structure and function of collectin liver 1 (CL-L1) and collectin 11 (CL-11, CL-K1). Immunobiology 217, 851–863. https://doi.org/10.1016/j.imbio.2011.12.008

Sender, G., Korwin-Kossakowska, A., Pawlik, A., Hameed, K.G.A., Oprządek, J., 2013. Genetic basis of mastitis resistance in dairy cattle: A review. Ann. Anim. Sci. 13. https://doi.org/10.2478/aoas-2013-0043

Shabalin, A.A., 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28, 1353–1358. https://doi.org/10.1093/bioinformatics/bts163

Sherry, S.T., 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29, 308–311. https://doi.org/10.1093/nar/29.1.308

Shinkai, H., Matsumoto, T., Toki, D., Okumura, N., Terada, K., Uenishi, H., 2015. Porcine NOD1 polymorphisms with impaired ligand recognition and their distribution in pig populations. Mol. Immunol. 63, 305–311. https://doi.org/10.1016/j.molimm.2014.07.020

Shuster, D.E., Kehrli, M.E.J., Ackermann, M.R., Gilbert, R.O., 1992. Identification and prevalence of a genetic defect that causes leukocyte adhesion deficiency in Holstein cattle. Proc. Natl. Acad. Sci. U. S. A. 89, 9225–9229. https://doi.org/10.1073/pnas.89.19.9225

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539. https://doi.org/10.1038/msb.2011.75

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., Ng, P.C., 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 40, W452–W457. https://doi.org/10.1093/nar/gks539

Snyman, H.N., 2013. Constitutive variation in hepatic expression of porcine innate immune genes associated with novel promoter polymorphisms. DVSc Thesis. University of Guelph.

Staiger, E.A., Tseng, C.T., Miller, D., Cassano, J.M., Nasir, L., Garrick, D., Brooks, S.A., Antczak, D.F., 2016. Host genetic influence on papillomavirus-induced tumors in the horse. Int. J. Cancer 139, 784–792. https://doi.org/10.1002/ijc.30120

Steffensen, R., Thiel, S., Varming, K., Jersild, C., Jensenius, J.C., 2000. Detection of structural gene mutations and promoter polymorphisms in the mannan-binding lectin (MBL) gene by polymerase chain reaction with sequence-specific primers. J. Immunol. Methods 241, 33–42. https://doi.org/10.1016/S0022-1759(00)00198-8

Sumiya, M., Tabona, P., Arai, T., Summerfield, J.A.A., Super, M., Levinsky, R.J.J., Turner, M.W.W., Tabona, P., Levinsky, R.J.J., Arai, T., Turner, M.W.W., Summerfield, J.A.A., 1991. Molecular basis of opsonic defect in immunodeficient children. Lancet 337, 1569–1570. https://doi.org/10.1016/0140-6736(91)93263-9

Summerfield, J.A., Ryder, S., Sumiya, M., Thursz, M., Gorchein, A., Monteil, M.A., Turner, M.W., 1995. Mannose binding protein gene mutations associated with unusual and severe infections in adults. Lancet 345, 886–889. https://doi.org/S0140-6736(95)90009-8 [pii]

Super, M., Levinsky, R.J., Turner, M.W., 1990. The level of mannan-binding protein regulates the binding of complement-derived opsonins to mannan and zymosan at low serum concentrations. Clin. Exp. Immunol. 79, 144–50. https://doi.org/10.1111/j.1365-2249.1990.tb05170.x

Super, M., Thiel, S., Lu, J., Levinsky, R.T.J., Turner, M.W., 1989. Association of low levels of mannan-binding protein with a common defect of opsonisation. Lancet 2, 1236–1239. https://doi.org/10.1016/S0140-6736(89)91849-7

Tachibana, K., Takeuchi, K., Inada, H., Sugimoto, K., Ishimoto, K., Yamashita, M., Maegawa, T., Yamasaki, D., Osada, S., Tanaka, T., Rakugi, H., Hamakubo, T., Sakai, J., Kodama, T., Doi, T., 2013. Human mannose-binding lectin 2 is directly regulated by peroxisome proliferator-activated receptors via a peroxisome proliferator responsive element. J. Biochem. 154, 265–273. https://doi.org/10.1093/jb/mvt050

Taganov, K.D., Boldin, M.P., Baltimore, D., 2007. MicroRNAs and immunity: tiny players in a big field. Immunity 26, 133–137. https://doi.org/10.1016/j.immuni.2007.02.005

Takeda, K., Miyahara, N., Rha, Y.-H.H., Taube, C., Yang, E.-S.S., Joetham, A., Kodama, T., Balhorn, A.M., Dakhama, A., Duez, C., Evans, A.J., Voelker, D.R., Gelfand, E.W., 2003. Surfactant protein D regulates airway function and allergic inflammation through modulation of macrophage function. Am. J. Respir. Crit. Care Med. 168, 783–789. https://doi.org/10.1164/rccm.200304-548OC

Takeuchi, O., Akira, S., 2008. MDA5/RIG-I and virus recognition. Curr. Opin. Immunol. 20, 17–22. https://doi.org/10.1016/j.coi.2008.01.002

Tay, H.L., Kaiko, G.E., Plank, M., Li, J., Maltby, S., Essilfie, A.-T., Jarnicki, A., Yang, M., Mattes, J., Hansbro, P.M., Foster, P.S., 2015. Antagonism of miR-328 increases the antimicrobial function of macrophages and neutrophils and rapid clearance of non-typeable Haemophilus influenzae (NTHi) from infected lung. PLoS Pathog. 11, e1004549. https://doi.org/10.1371/journal.ppat.1004549

Taylor, J.D., Fulton, R.W., Lehenbauer, T.W., Step, D.L., Confer, A.W., 2010. The epidemiology of bovine respiratory disease: What is the evidence for predisposing factors? Can. Vet. J. 51, 1095–1102. https://doi.org/10.1128/CMR.16.1.79

Teng, O., Chen, S.-T., Hsu, T.-L., Sia, S.F., Cole, S., Valkenburg, S.A., Hsu, T.-Y., Zheng, J.T., Tu, W., Bruzzone, R., Peiris, J.S.M., Hsieh, S.-L., Yen, H.-L., 2017. CLEC5A-mediated enhancement of the inflammatory response in myeloid cells contributes to influenza virus pathogenicity in vivo. J. Virol. 91, e01813-16. https://doi.org/10.1128/JVI.01813-16

Thiel, S., Vorup-Jensen, T., Stover, C.M., Schwaeble, W., Laursen, S.B., Poulsen, K., Willis, A.C., Eggleton, P., Hansen, S., Holmskov, U., Reid, K.B., Jensenius, J.C., 1997. A second serine protease associated with mannan-binding lectin that activates complement. Nature 386, 506–10. https://doi.org/10.1038/386506a0

Thompson-Crispi, K.A., Sargolzaei, M., Ventura, R., Abo-Ismail, M., Miglior, F., Schenkel, F., Mallard, B.A., 2014. A genome-wide association study of immune response traits in Canadian Holstein cattle. BMC Genomics 15, 559. https://doi.org/10.1186/1471-2164-15-559

Ting, J.P.-Y., Lovering, R.C., Alnemri, E.S., Bertin, J., Boss, J.M., Davis, B.K., Flavell, R.A., Girardin, S.E., Godzik, A., Harton, J.A., Hoffman, H.M., Hugot, J.-P., Inohara, N., Mackenzie, A., Maltais, L.J., Nunez, G., Ogura, Y., Otten, L.A., Philpott, D., Reed, J.C., Reith, W., Schreiber, S., Steimle, V., Ward, P.A., 2008. The NLR gene family: a standard nomenclature. Immunity 28, 285–7. https://doi.org/10.1016/j.immuni.2008.02.005

Turin, L., Russo, S., Poli, G., 1999. BHV-1: new molecular approaches to control a common and widespread infection. Mol. Med. 5, 261–284.

van de Wetering, J.K., Van Golde, L.M.G., Batenburg, J.J., 2004. Collectins: players of the innate immune system. Eur. J. Biochem. 271, 1229–1249. https://doi.org/10.1111/j.1432-1033.2004.04040.x

van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V, Altshuler, D., Gabriel, S.B., DePristo, M.A., 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinforma. 11, 11.10.1-

11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

van Eenennaam, A., Neibergs, H., Seabury, C., Taylor, J., Wang, Z., Scraggs, E., Schnabel, R.D., Decker, J., Wojtowicz, A., Aly, S., Davis, J., Blanchard, P., Crossley, B., Rossitto, P., Lehenbauer, T., Hagevoort, R., Chavez, E., Neibergs, J.S., Womack, J.E., 2014. Results of the BRD CAP project: Progress toward identifying genetic markers associated with BRD susceptibility. Anim. Heal. Res. Rev. 15, 157–160. https://doi.org/10.1017/S1466252314000231

van Eijk, M., Rynkiewicz, M.J., White, M.R., Hartshorn, K.L., Zou, X., Schulten, K., Luo, D., Crouch, E.C., Cafarella, T.R., Head, J.F., Haagsman, H.P., Seaton, B.A., 2012. A unique sugar-binding site mediates the distinct anti-influenza activity of pig surfactant protein D. J. Biol. Chem. 287, 26666–77. https://doi.org/10.1074/jbc.M112.368571

van Eijk, M., White, M.R., Crouch, E.C., Batenburg, J.J., Vaandrager, A.B., Van Golde, L.M.G., Haagsman, H.P., Hartshorn, K.L., 2003. Porcine pulmonary collectins show distinct interactions with influenza A viruses: role of the N-linked oligosaccharides in the carbohydrate recognition domain. J. Immunol. 171, 1431–40. https://doi.org/10.4049/jimmunol.171.3.1431

van Iwaarden, J.F., van Strijp, J.A.G., Ebskamp, M.J., Welmers, A.C., Verhoef, J., Van Golde, L.M.G., 1991. Surfactant protein A is opsonin in phagocytosis of herpes simplex virus type 1 by rat alveolar macrophages. Am. J. Physiol. 261, L204-9.

van Kempen, G., Meijvis, S., Endeman, H., Vlaminckx, B., Meek, B., de Jong, B., Rijkers, G., Bos, W.J., 2017. Mannose-binding lectin and l-ficolin polymorphisms in patients with community-acquired pneumonia caused by intracellular pathogens. Immunology 151, 81–88. https://doi.org/10.1111/imm.12705

van Laere, A.S., Nguyen, M., Braunschweig, M., Nezer, C., Collette, C., Moreau, L., Archibald, A.L., Haley, C.S., Buys, N., Tally, M., Andersson, G., Georges, M., Andersson, L., 2003. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature 425, 832–836. https://doi.org/10.1038/nature02064

Van Rozendaal, B.A.W.M., Van De Lest, C.H.A., Van Eijk, M., Van Golde, L.M.G., Voorhout, W.F., Van Helden, H.P.M., Haagsman, H.P., 1999. Aerosolized endotoxin is immediately bound by pulmonary surfactant protein D in vivo. Biochim. Biophys. Acta - Mol. Basis Dis. 1454, 261–269. https://doi.org/10.1016/S0925-4439(99)00042-3

Veldhuizen, E.J.A., Rijnders, M., Claassen, E.A., van Dijk, A., Haagsman, H.P., 2008. Porcine β-defensin 2 displays broad antimicrobial activity against pathogenic intestinal bacteria. Mol. Immunol. 45, 386–394. https://doi.org/10.1016/j.molimm.2007.06.001

Veldhuizen, E.J.A., van Eijk, M., Haagsman, H.P., 2011. The carbohydrate recognition domain of collectins. FEBS J. 278, 3930–3941. https://doi.org/10.1111/j.1742-4658.2011.08206.x

Vorup-Jensen, T., Petersen, S. V., Hansen, A.G., Poulsen, K., Schwaeble, W., Sim, R.B., Reid, K.B., Davis, S.J., Thiel, S., Jensenius, J.C., 2000. Distinct pathways of mannan-binding lectin (MBL)- and C1-complex autoactivation revealed by reconstitution of MBL with recombinant MBL-associated serine protease-2. J. Immunol. 165, 2093–100. https://doi.org/10.4049/jimmunol.165.4.2093

Wade, C.M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T.L., Adelson, D.L., Bailey, E., Bellone, R.R., Blöcker, H., Distl, O., Edgar, R.C., Garber, M., Leeb, T., Mauceli, E., MacLeod, J.N., Penedo, M.C.T., Raison, J.M., Sharpe, T., Vogel, J., Andersson, L., Antczak, D.F., Biagi, T., Binns, M.M., Chowdhary, B.P., Coleman, S.J., Della Valle, G., Fryc, S., Guérin, G., Hasegawa, T., Hill, E.W., Jurka, J., Kiialainen, A., Lindgren, G., Liu, J., Magnani, E., Mickelson, J.R., Murray, J., Nergadze, S.G., Onofrio, R., Pedroni, S., Piras, M.F., Raudsepp, T., Rocchi, M., Røed, K.H., Ryder, O.A., Searle, S., Skow, L., Swinburne, J.E., Syvänen, A.C., Tozaki, T., Valberg, S.J., Vaudin, M., White, J.R., Zody, M.C., Broad Institute Whole Genome Assembly Team, Broad Institute Genome Sequencing Platform, Lander, E.S., Lindblad-Toh, K., 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326, 865–867. https://doi.org/10.1126/science.1178158

Wallis, R., Drickamer, K., 1999. Molecular determinants of oligomer formation and complement fixation in mannose-binding proteins. J. Biol. Chem. 274, 3580–3589. https://doi.org/10.1074/jbc.274.6.3580

Walport, M.J., 2001a. Complement: First of Two Parts. N. Engl. J. Med. 344, 1058–1066. https://doi.org/10.1056/NEJM200104053441406

Walport, M.J., 2001b. Complement: Second of Two Parts. N. Engl. J. Med. 344, 1140–1144. https://doi.org/10.1056/NEJM200104123441506

Wang, C., Liu, M., Li, Q., Ju, Z., Huang, J., Li, J., Wang, H., Zhong, J., 2011. Three novel single-nucleotide polymorphisms of MBL1 gene in Chinese native cattle and their associations with milk performance traits. Vet. Immunol. Immunopathol. 139, 229–236. https://doi.org/10.1016/j.vetimm.2010.10.023

Wang, J.-Y., Kishore, U., Reid, K.B.M., 1995. A recombinant polypeptide, composed of the α-helical neck region and the carbohydrate recogniton domain of conglutinin, self-associates to give a functionally intact homotrimer. FEBS Lett. 376, 6–10. https://doi.org/10.1016/0014-5793(95)01232-4

Wang, M., Chen, Y., Zhang, Y., Zhang, L., Lu, X., Chen, Z., 2011. Mannan-binding lectin directly interacts with Toll-like receptor 4 and suppresses lipopolysaccharide-induced inflammatory cytokine secretion from THP-1 cells. Cell. Mol. Immunol. 8, 265–275. https://doi.org/10.1038/cmi.2011.1

Wang, X., Ju, Z., Huang, J., Hou, M., Zhou, L., Qi, C., Zhang, Y., Gao, Q., Pan, Q., Li, G., Zhong, J., Wang, C., 2012. The relationship between the variants of the bovine MBL2 gene and milk production traits, mastitis, serum MBL-C levels and complement activity. Vet. Immunol. Immunopathol. 148, 311–319. https://doi.org/10.1016/j.vetimm.2012.06.017

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M.G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J.S., Govindarajan, S., Shaulsky, G., Walhout, A.J.M., Bouget, F.-Y., Ratsch, G., Larrondo, L.F., Ecker, J.R., Hughes, T.R., 2014. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell 158, 1431–1443. https://doi.org/10.1016/j.cell.2014.08.009

Weis, W.I., Drickamer, K., Hendrickson, W.A., 1992. Structure of a C-type mannose-binding protein complexed with an oligosaccharide. Nature 360, 127–134. https://doi.org/10.1038/360127a0

Wilsher, S., Allen, W.R., Wood, J.L.N., 2006. Factors associated with failure of Thoroughbred horses to train and race. Equine Vet. J. 38, 113–118. https://doi.org/10.2746/042516406776563305

Wu, J., Bai, J.Y., Li, L., Huang, S., Li, C.M., Wang, G.L., 2015. Genetic polymorphisms of the BMAP-28 and MASP-2 genes and their correlation with the somatic cell score in Chinese holstein cattle. Genet. Mol. Res. 14, 1–8. https://doi.org/10.4238/2015.January.15.1

Wu, T.-T.T., Chen, T.-L.L., Loon, W.-S.S., Tai, Y.-T.T., Cherng, Y.-G.G., Chen, R.-M.M., 2011. Lipopolysaccharide stimulates syntheses of toll-like receptor 2 and surfactant protein-A in human alveolar epithelial A549 cells through upregulating phosphorylation of MEK1 and ERK1/2 and sequential activation of NF-κB. Cytokine 55, 40–47. https://doi.org/10.1016/j.cyto.2011.03.005

Xiao, C., Rajewsky, K., 2009. MicroRNA control in the immune system: basic principles. Cell 136, 26–36. https://doi.org/10.1016/j.cell.2008.12.027

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Birney, E., Harrow, J.L., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S.J., Cunningham, F., Aken, B.L., Zerbino, D.R., Flicek, P., 2016. Ensembl 2016. Nucleic Acids Res. 44, D710–D716. https://doi.org/10.1093/nar/gkv1157

Yoneyama, M., Fujita, T., 2007. Function of RIG-I-like receptors in antiviral innate immunity. J. Biol. Chem. 282, 15315–8. https://doi.org/10.1074/jbc.R700007200

Yuan, Z., Li, J.J., Li, J.J., Gao, X., Xu, S., 2012. SNPs identification and its correlation analysis with milk somatic cell score in bovine MBL1 gene. Mol. Biol. Rep. 40, 7–12. https://doi.org/10.1007/s11033-012-1934-z

Zelensky, A.N., Gready, J.E., 2005. The C-type lectin-like domain superfamily. FEBS J. 272, 6179–6217. https://doi.org/10.1111/j.1742-4658.2005.05031.x

Zhang, P., McAlinden, A., Li, S., Schumacher, T., Wang, H., Hu, S., Sandell, L., Crouch, E., 2001. The amino-terminal heptad repeats of the coiled-coil neck domain of pulmonary surfactant protein d are necessary for the assembly of trimeric subunits and dodecamers. J. Biol. Chem. 276, 19862–70. https://doi.org/10.1074/jbc.M100597200

Zhang, Y., Zhang, L., Wang, L., Qiao, L., Liang, J., Yan, H., Zhao, K., Liu, X., Wang, L., 2015. Identification and examination of a novel 9-bp insert/deletion polymorphism on porcine SFTPA1 exon 2 associated with acute lung injury using an oleic acid-acute lung injury model. Anim. Sci. J. 86, 573–578. https://doi.org/10.1111/asj.12330

Zhao, Z.L., Wang, C.F., Li, Q.L., Ju, Z.H., Huang, J.M., Li, J.B., Zhong, J.F., Zhang, J.B., 2012.

Novel SNPs of the mannan-binding lectin 2 gene and their association with production traits in Chinese Holsteins. Genet. Mol. Res. 11, 3744–3754. https://doi.org/10.4238/2012.October.15.6

Zhou, A., Li, S., Wu, J., Khan, F.A., Zhang, S., 2014. Interplay between microRNAs and host pathogen recognition receptors (PRRs) signaling pathways in response to viral infection. Virus Res. 184, 1–6. https://doi.org/10.1016/j.virusres.2014.01.019

Zimin, A. V, Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C.P., Sonstegard, T.S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J.A., Salzberg, S.L., 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol. 10, R42. https://doi.org/10.1186/gb-2009-10-4-r42

Appendices

Appendix 1.1. Repository for the scripts for all the bioinformatic manipulations.

All scripts are available at:

https://github.com/russ-dvm/scripts

All scripts were written in Perl, Python, Bash, or R. All scripts were written by me, with the exception of "allele_counts.pl", "make_diploid.pl", and "match_snps.pl", which were written by Ann Meyer.

Appendix 2.1 Breakdown of the breed and sex of horses used in this study.

Population	Breed	F	M	MN	Grand Total
Non-Infectious	Arabian			1	1
	Canadian Sport Horse	1			1
	Clydesdale	1			1
	Dutch Warmblood	1		2	3
	Fjord	1		1	2
	Friesian		1	1	2
	Haflinger	1			1
	Hanoverian	1		1	2
	Hanoverian x Thoroughbred			1	1
	Mixed Breed Horse	1			1
	Unknown	1		1	2
	Oldenburg			1	1
	Paint			2	2
	Pony			1	1
	Quarterhorse	1	1	1	3
	Rocky Mountain Horse			1	1
	Standardbred	3		5	8
	Thoroughbred	6	2	7	15
	Warmblood		2	2	4
	Sub-total:	18	6	28	52
Infectious	Appaloosa			1	1
	Canadian Warmblood	1			1
	Clydesdale cross			1	1
	Fjord	1			1
	Friesian			1	1
	Mixed Breed Horse	1			1
	Unknown	1		2	3
	Paint	1		1	2
	Pony			1	1
	Quarterhorse			1	1
	Standardbred	1	1	4	6
	Thoroughbred	8	4	5	17
	Warmblood			1	1
	Sub-total:	14	5	18	37
	Grand Total	32	11	46	89

Appendix 2.2. Detailed breakdown of variants discovered by gene and by gene region.

Gene	Region	Total targeted (bp)	Total sequenced to a minimum of 445x (5x/animal)	Number of variants	Rate (variant / kb)
COLEC10	Total	88,258	86,356	371	
	Upstream 5-50 kb	45,076	43,174	120	2.78
	Upstream 5 kb	5001	5001	15	3.00
	Introns	34,166	34,166	194	5.68
	Exons	834	834	3	3.60
	Downstream 3 kb	3181	3181	39	12.26
COLEC11	Total	54,918	53,734	566	
	Upstream 5-50 kb	15,340	14,186	170	11.98
	Upstream 5 kb	5001	5001	73	14.60
	Introns	30,770	30,740	301	9.79
	Exons	804	804	2	2.49
	Downstream 3 kb	3003	3003	20	6.66
COLEC12	Total	94,479	91,210	521	
	Upstream 5-50 kb	45,001	44,722	283	6.33
	Upstream 5 kb	5001	5001	12	2.40
	Introns	39,187	36,197	217	5.99
	Exons	2169	2169	6	2.77
	Downstream 3 kb	3121	3121	3	0.96
FCN1	Total	60,957	60,580	810	
	Upstream 5-50 kb	44,999	44,693	518	11.59
	Upstream 5 kb	5001	5001	53	10.60
	Introns	6971	6900	161	23.33
	Exons	981	981	32	32.62
	Downstream 3 kb	3005	3005	46	15.31
FCN1-like	Total	22,437	22,298	411	
	Upstream 5-50 kb	11,377	11,305	121	10.70
	Upstream 5 kb	5001	5001	180	35.99
	Introns	2425	2425	64	26.39
	Exons	633	633	17	26.86
	Downstream 3 kb	3001	2934	29	9.88
FCN3	Total	10,659	9781	42	
	Upstream 5-50 kb	0	0	0	0.00
	Upstream 5 kb	3628	3627	8	2.21
	Introns	3160	3160	9	2.85
	Exons	870	870	6	6.90

	Downstream 3 kb	3001	2124	19	8.95
MASP1	Total	114,890	114,487	639	
	Upstream 5-50 kb	45,240	44,875	255	5.68
	Upstream 5 kb	5001	5001	38	7.60
	Introns	59,548	59,511	339	5.70
	Exons	2100	2100	4	1.90
	Downstream 3 kb	3001	3000	3	1.00
MASP2	Total	34,475	31,165	91	
	Upstream 5-50 kb	10,175	9468	25	2.64
	Upstream 5 kb	5001	2727	6	2.20
	Introns	14,224	13,895	53	3.81
	Exons	2061	2061	7	3.40
	Downstream 3 kb	3014	3014	0	0.00
MBL1	Total	45,907	45,014	288	
	Upstream 5-50 kb	35,226	34,344	217	6.32
	Upstream 5 kb	5001	4990	39	7.82
	Introns	1929	1929	13	6.74
	Exons	750	750	4	5.33
	Downstream 3 kb	3001	3001	15	5.00
MBL2	Total	56,542	55,812	468	
	Upstream 5-50 kb	45,001	44,271	377	8.52
	Upstream 5 kb	5001	5001	36	7.20
	Introns	2792	2792	22	7.88
	Exons	741	741	4	5.40
	Downstream 3 kb	3007	3007	29	9.64
SFTPA1	Total	35,227	33,852	217	
	Upstream 5-50 kb	24,868	23,522	147	6.25
	Upstream 5 kb	5001	4972	37	7.44
	Introns	1489	1489	11	7.39
	Exons	747	747	3	4.02
	Downstream 3 kb	3122	3122	19	6.09
SFTPD	Total	61,732	56,847	352	
	Upstream 5-50 kb	45,001	40,154	283	7.05
	Upstream 5 kb	5001	5001	17	3.40
	Introns	7565	7527	30	3.99
	Exons	1129	1129	4	3.54
	Downstream 3 kb	3036	3036	18	5.93

Appendix 2.3. List of SNVs found within a transcription factor binding site predicted *in silico* in this study, and predicted or identified in studies of collagenous lectins in other species.

Gene	Chr	SNV position	Ref allele	Alt allele	TFBS start	TFBS end	TF	Sequence	RE	Reference
COLEC11	15	88,586,795	G	A	88,586,793	88,586,799	c-Ets-1 [T00112]	CTTCCTG	0.31455	Grageda et al 2015
COLEC11	15	88,585,751	C	G	88,585,749	88,585,755	E47 [T00207]	GCCAGAT	0.31397	Hansen et al 2014
COLEC11	15	88,584,801	C	T	88,584,799	88,584,804	c-Myc [T00140]	CACGTG	1.31228	Hansen et al 2013
COLEC11	15	88,584,801	C	T	88,584,799	88,584,804	N-Myc [T01445]	CACGTG	1.31228	Hansen et al 2016
COLEC11	15	88,585,908	G	T	88,585,906	88,585,911	Sp1 [T00753]	GGGCGG	1.60269	Grageda et al 2015
COLEC11	15	88,585,908	G	T	88,585,905	88,585,913	Sp1 [T00759]	GGGGCGGG	0.03007	Grageda et al 2015
COLEC11	15	88,584,988	G	C	88,584,984	88,584,991	PEA3 [T00684]	TTTCCTGC	0.15133	Grageda et al 2015
FCN1	25	36,826,234	G	A	36,826,230	36,826,235	Sp1 [T00753]	CCGCCC	1.77564	Grageda et al 2015
FCN1-like	25	36,798,021	C	G	36,798,018	36,798,023	Sp1 [T00753]	GGGCGG	2.12888	Grageda et al 2015
FCN1-like	25	36,799,482	C	G	36,799,482	36,799,487	c-Myc [T00140]	CACGTG	1.43378	Hansen et al 2013
FCN1-like	25	36,799,484	C	T	36,799,482	36,799,487	c-Myc [T00140]	CACGTG	1.43378	Hansen et al 2013
FCN1-like	25	36,799,486	T	C	36,799,482	36,799,487	c-Myc [T00140]	CACGTG	1.43378	Hansen et al 2013
FCN1-like	25	36,799,482	C	G	36,799,482	36,799,487	N-Myc [T01445]	CACGTG	1.43378	Hansen et al 2016
FCN1-like	25	36,799,484	C	T	36,799,482	36,799,487	N-Myc [T01445]	CACGTG	1.43378	Hansen et al 2016
FCN1-like	25	36,799,486	T	C	36,799,482	36,799,487	N-Myc [T01445]	CACGTG	1.43378	Hansen et al 2016
FCN1-like	25	36,799,482	C	G	36,799,482	36,799,488	c-Myc [T00143]	CACGTGG	0.39278	Hansen et al 2013
FCN1-like	25	36,799,484	C	T	36,799,482	36,799,488	c-Myc [T00143]	CACGTGG	0.39278	Hansen et al 2013
FCN1-like	25	36,799,486	T	C	36,799,482	36,799,488	c-Myc [T00143]	CACGTGG	0.39278	Hansen et al 2013
FCN1-like	25	36,799,482	C	G	36,799,482	36,799,488	USF-1 [T00875]	CACGTGG	0.71707	Hansen et al 2013
FCN1-like	25	36,799,484	C	T	36,799,482	36,799,488	USF-1 [T00875]	CACGTGG	0.71707	Hansen et al 2013
FCN1-like	25	36,799,486	T	C	36,799,482	36,799,488	USF-1 [T00875]	CACGTGG	0.71707	Hansen et al 2013
FCN1-like	25	36,801,372	A	G	36,801,371	36,801,377	USF-1 [T00875]	CATGTGG	0.71707	Hansen et al 2013
FCN1-like	25	36,801,087	T	C	36,801,084	36,801,090	E47 [T00207]	ATCTGGC	0.32411	Hansen et al 2014
FCN1-like	25	36,801,088	G	A	36,801,084	36,801,090	E47 [T00207]	ATCTGGC	0.32411	Hansen et al 2014
FCN1-like	25	36,796,554	C	T	36,796,548	36,796,554	MZF-1 [T00529]	TCCCCAC	0.39959	Hansen et al 2015
FCN1-like	25	36,801,075	T	C	36,801,075	36,801,081	MZF-1 [T00529]	TCCCCAC	0.39959	Hansen et al 2015
FCN1-like	25	36,801,079	C	G	36,801,075	36,801,081	MZF-1 [T00529]	TCCCCAC	0.39959	Hansen et al 2015

FCN1-like	25	36,800,822	C	G	36,800,822	36,800,828	c-Ets-1 [T00112]	CAGGAAG	0.32852	Grageda et al 2015
MBL1	1	88,931,429	G	A	88,931,426	88,931,431	c-Myc [T00140]	CACGTG	1.17546	Hansen et al 2013
MBL1	1	88,931,429	G	A	88,931,426	88,931,431	N-Myc [T01445]	CACGTG	1.17546	Hansen et al 2016
MBL1	1	88,931,429	G	A	88,931,425	88,931,431	c-Myc [T00143]	CCACGTG	0.28899	Hansen et al 2013
MBL1	1	88,931,429	G	A	88,931,425	88,931,431	USF-1 [T00875]	CCACGTG	0.58899	Hansen et al 2013
MBL1	1	88,928,936	C	T	88,928,936	88,928,942	c-Ets-1 [T00112]	CAGGAAG	0.29968	Grageda et al 2015
MBL1	1	88,929,760	C	T	88,929,757	88,929,764	PEA3 [T00685]	TTACATCC	0.31285	Grageda et al 2015
MBL2	1	43,251,789	C	T	43,251,787	43,251,792	c-Myc [T00140]	CACGTG	0.95652	Hansen et al 2013
MBL2	1	43,251,789	C	T	43,251,787	43,251,792	N-Myc [T01445]	CACGTG	0.95652	Hansen et al 2016
MBL2	1	43,250,717	T	C	43,250,717	43,250,724	PEA3 [T00684]	GGAGGAAA	0.14841	Grageda et al 2015

RE value represents the likelihood of finding the TFBS sequence in a sequence of 1000 random nucleotides.

Appendix 2.4. Variants impacting predicted miRNA recognition elements.

Gene	Chrom	Position	Ref allele	Alt allele	miRNA	MRE location
COLEC10	9	62,298,917	C	T	eca-miR-1244	seed
COLEC10	9	62,299,049	CATCA	C	eca-miR-1301	full
COLEC10	9	62,299,066	G	C	eca-miR-1301	seed
COLEC10	9	62,299,126	C	G	eca-miR-670	full
COLEC10	9	62,298,805	T	A	eca-miR-711	full
COLEC10	9	62,298,823	C	CAA	eca-miR-711	full
COLEC10	9	62,298,805	T	A	eca-miR-8960	full
COLEC10	9	62,299,249	T	C	eca-miR-8970	full
COLEC10	9	62,300,038	G	C	eca-miR-8986a	full
COLEC10	9	62,300,663	A	G	eca-miR-9002	full
COLEC10	9	62,298,661	C	T	eca-miR-9020	full
COLEC10	9	62,298,662	G	A	eca-miR-9020	full
COLEC10	9	62,299,375	G	A	eca-miR-9022	full
COLEC10	9	62,299,249	T	C	eca-miR-9062	full
COLEC10	9	62,301,149	G	A	eca-miR-9134	full
COLEC10	9	62,301,165	A	G	eca-miR-9134	full
COLEC10	9	62,301,173	T	C	eca-miR-9134	seed
COLEC10	9	62,301,176	G	A	eca-miR-9134	full
COLEC10	9	62,300,623	G	A	eca-miR-9147	full
COLEC11	15	88,550,768	C	A	eca-miR-1291a	seed
COLEC11	15	88,550,768	C	A	eca-miR-1291b	seed
COLEC11	15	88,550,707	T	C	eca-miR-330	full
COLEC11	15	88,549,373	T	C	eca-miR-541	seed
COLEC11	15	88,549,379	G	A	eca-miR-541	full
COLEC11	15	88,548,130	G	A	eca-miR-7045	seed

COLEC11	15	88,548,195	A	G	eca-miR-8930	seed
COLEC11	15	88,548,188	G	A	eca-miR-8930	full
COLEC11	15	88,548,845	C	T	eca-miR-8931	full
COLEC11	15	88,548,130	G	A	eca-miR-8934	full
COLEC11	15	88,549,191	G	A	eca-miR-8944	full
COLEC11	15	88,548,188	G	A	eca-miR-8966	full
COLEC11	15	88,548,195	A	G	eca-miR-8966	full
COLEC11	15	88,550,707	T	C	eca-miR-8969	full
COLEC12	8	40,893,161	T	G	eca-miR-8989	full
FCN1	25	36,817,189	A	G	eca-miR-125a-3p	full
FCN1	25	36,816,243	G	A	eca-miR-1379	seed
FCN1	25	36,815,771	C	T	eca-miR-1379	seed
FCN1	25	36,817,105	C	T	eca-miR-138	full
FCN1	25	36,815,529	T	C	eca-miR-221	full
FCN1	25	36,815,536	G	A	eca-miR-221	full
FCN1	25	36,815,641	T	C	eca-miR-326	seed
FCN1	25	36,817,105	C	T	eca-miR-328	full
FCN1	25	36,815,955	C	T	eca-miR-330	full
FCN1	25	36,815,641	T	C	eca-miR-330	seed
FCN1	25	36,817,207	G	A	eca-miR-491-5p	full
FCN1	25	36,814,860	T	C	eca-miR-632	seed
FCN1	25	36,817,105	C	T	eca-miR-8943	full
FCN1	25	36,817,105	C	T	eca-miR-8943	full
FCN1	25	36,817,313	CAT	C	eca-miR-8977	full
FCN1	25	36,817,316	G	C	eca-miR-8977	full
FCN1	25	36,817,105	C	T	eca-miR-8989	full
FCN1	25	36,816,315	C	A	eca-miR-9022	full
FCN1	25	36,817,154	A	T	eca-miR-9048	full
FCN1	25	36,816,703	C	T	eca-miR-9060	full

FCN1	25	36,816,709	C	T	eca-miR-9060	full
FCN1	25	36,816,211	A	G	eca-miR-9098	full
FCN1	25	36,815,221	C	A	eca-miR-9130	full
FCN1	25	36,816,334	G	GGGAAGGCCCTAGAAGCT	eca-miR-9165	full
FCN1	25	36,815,221	C	A	eca-miR-9a	full
FCN1-like	25	36,804,699	C	T	eca-miR-1898	full
FCN1-like	25	36,804,636	T	A	eca-miR-1912	seed
FCN1-like	25	36,805,601	C	T	eca-miR-193a-3p	seed
FCN1-like	25	36,805,407	C	A	eca-miR-24	full
FCN1-like	25	36,805,414	C	T	eca-miR-24	seed
FCN1-like	25	36,804,948	C	T	eca-miR-345-5p	full
FCN1-like	25	36,804,675	G	A	eca-miR-423-5p	full
FCN1-like	25	36,805,953	A	G	eca-miR-483	full
FCN1-like	25	36,805,961	G	C	eca-miR-483	full
FCN1-like	25	36,806,427	G	A	eca-miR-671-3p	seed
FCN1-like	25	36,807,496	T	TGGCTG	eca-miR-8908e	full
FCN1-like	25	36,807,496	T	TGGCTG	eca-miR-8908g	full
FCN1-like	25	36,805,567	G	C	eca-miR-8908i	full
FCN1-like	25	36,804,699	C	T	eca-miR-8941	full
FCN1-like	25	36,804,636	T	A	eca-miR-8975	seed
FCN1-like	25	36,804,636	T	A	eca-miR-9015	seed
FCN1-like	25	36,806,253	T	G	eca-miR-9033	full
FCN1-like	25	36,805,601	C	T	eca-miR-9065	full
FCN1-like	25	36,805,567	G	C	eca-miR-9077	full
FCN1-like	25	36,805,709	G	T	eca-miR-9077	seed
FCN1-like	25	36,805,326	G	A	eca-miR-9083	full
FCN1-like	25	36,805,164	C	T	eca-miR-9122	full
FCN1-like	25	36,805,961	G	C	eca-miR-9134	full
FCN1-like	25	36,805,164	C	T	eca-miR-9146	full

FCN1-like	25	36,805,407	C	A	eca-miR-9151	full
FCN1-like	25	36,805,414	C	T	eca-miR-9151	seed
FCN1-like	25	36,806,460	C	T	eca-miR-9156	seed
FCN3	2	28,419,879	G	A	eca-miR-187	full
FCN3	2	28,417,382	G	A	eca-miR-485-5p	seed
FCN3	2	28,419,517	AG	A	eca-miR-532-3p	full
FCN3	2	28,417,382	G	A	eca-miR-8943	seed
FCN3	2	28,419,879	G	A	eca-miR-9101	full
FCN3	2	28,417,741	G	A	eca-miR-9125	full
MASP1	19	25,144,864	T	C	eca-miR-423-5p	full
MASP1	19	25,144,866	G	T	eca-miR-423-5p	full
MBL1	1	88,935,957	T	A	eca-miR-432	full
MBL1	1	88,936,193	C	G	eca-miR-874	seed
MBL1	1	88,936,937	C	T	eca-miR-8954	seed
MBL1	1	88,936,814	C	A	eca-miR-8975	full
MBL1	1	88,938,572	AGT	A	eca-miR-8990	full
MBL1	1	88,938,468	T	C	eca-miR-9029	full
MBL1	1	88,935,861	CT	C	eca-miR-9108	full
MBL1	1	88,938,468	T	C	eca-miR-9115	full
MBL1	1	88,937,748	G	A	eca-miR-9162	full
MBL2	1	43,245,265	C	T	eca-miR-769a-5p	full
MBL2	1	43,242,585	C	T	eca-miR-8917	seed
MBL2	1	43,242,527	A	G	eca-miR-8978	full
SFTPA1	1	88,927,745	C	G	eca-miR-8919	full
SFTPA1	1	88,925,763	AC	A	eca-miR-9042	seed
SFTPA1	1	88,925,137	C	A	eca-miR-9153	seed
SFTPD	1	88,947,229	C	T	eca-miR-125a-5p	full
SFTPD	1	88,947,229	C	T	eca-miR-125b-5p	full
SFTPD	1	88,947,294	T	A	eca-miR-1902	seed

SFTPD	1	88,947,299	G	A	eca-miR-1902	full
SFTPD	1	88,944,953	C	T	eca-miR-221	seed
SFTPD	1	88,947,229	C	T	eca-miR-331	seed
SFTPD	1	88,947,667	C	T	eca-miR-532-3p	full
SFTPD	1	88,947,519	G	C	eca-miR-671-5p	full
SFTPD	1	88,947,229	C	T	eca-miR-8988	seed
SFTPD	1	88,947,578	G	A	eca-miR-9000	full
SFTPD	1	88,945,223	C	A	eca-miR-9015	seed
SFTPD	1	88,947,294	T	A	eca-miR-9111	seed
SFTPD	1	88,947,299	G	A	eca-miR-9111	full

Appendix 2.5. GC content and number of insertion/deletions by target region.

Target region	**GC content (%)**	**Number of in/dels**
COLEC10	37.7	42
COLEC11	50.1	39
COLEC12	43.4	66
FCN1, FCN1-LIKE	54.7	85
FCN3	52.9	12
MASP1	44.7	49
MASP2	46.1	10
MBL2	39.9	32
SFTPA, MBL1, SFTPD	48.5	50

Appendix 3.1. Coordinates and Ensembl IDs used to annotate the variants identified in cattle.

Gene	Gene Stable ID	Transcript Stable ID	Strand	Feature	Chr	Genomic Start	Genomic End	Length
				45kb	1	80546924	80591982	45059
				5kb	1	80591983	80596983	5001
				5UTR	1	80596924	80596983	60
				exon	1	80596984	80596988	5
				intron	1	80596989	80602802	5814
				exon	1	80602803	80603034	232
				intron	1	80603035	80625335	22301
				exon	1	80625336	80625513	178
				intron	1	80625514	80627138	1625
				exon	1	80627139	80627270	132
				intron	1	80627271	80629378	2108
				exon	1	80629379	80629575	197
				intron	1	80629576	80631780	2205
MASP1	ENSBTAG00000012467	ENSBTAT00000016545	1	exon	1	80631781	80631928	148
				intron	1	80631929	80634052	2124
				exon	1	80634053	80634171	119
				intron	1	80634172	80635102	931
				exon	1	80635103	80635181	79
				intron	1	80635182	80641667	6486
				exon	1	80641668	80641805	138
				intron	1	80641806	80642635	830
				exon	1	80642636	80642710	75
				intron	1	80642711	80646939	4229
				exon	1	80646940	80647823	884
				3UTR	1	80647824	80649363	1540
				3kb_from_stop	1	80647824	80650824	3001
				3kb_from_gene	1	80649363	80652367	3005

				3kb_from_gene	8	112860707	112869670	8963
				3kb_from_stop	8	112867044	112870044	3001
				3UTR	8	112869671	112870044	374
				exon	8	112870045	112870436	392
				intron	8	112870437	112870621	185
				exon	8	112870622	112870717	96
				intron	8	112870718	112872621	1904
				exon	8	112872622	112872675	54
				intron	8	112872676	112875023	2348
COLEC11	ENSBTAG00000016225	ENSBTAT00000021590	-1	exon	8	112875024	112875095	72
				intron	8	112875096	112882013	6918
				exon	8	112882014	112882085	72
				intron	8	112882086	112889628	7543
				exon	8	112889629	112889746	118
				5UTR	8	112889747	112889771	25
				5kb	8	112889747	112894747	5001
				intron	8	112889772	112892870	3099
				5UTR	8	112892871	112892968	98
				45kb	8	112894748	112896491	1743
				45kb	11	106773026	106818024	44999
				5kb	11	106818025	106823025	5001
				exon	11	106823026	106823128	103
				intron	11	106823129	106824397	1269
				exon	11	106824398	106824511	114
				intron	11	106824512	106826831	2320
FCN1	ENSBTAG00000048155	ENSBTAT00000066000	1	exon	11	106826832	106826894	63
				intron	11	106826895	106827305	411
				exon	11	106827306	106827341	36
				intron	11	106827342	106828269	928
				exon	11	106828270	106828302	33
				intron	11	106828303	106828694	392

				exon	11	106828695	106828822	128
				intron	11	106828823	106829287	465
				exon	11	106829288	106829417	130
				intron	11	106829418	106830287	870
FCN1	ENSBTAG00000048155	ENSBTAT00000066000	1	exon	11	106830288	106830422	135
				intron	11	106830423	106831395	973
				exon	11	106831396	106831643	248
				3kb_from_stop	11	106831644	106834644	3001
				3kb_from_gene	11	106831644	106834643	3000
				3kb_from_stop	14	47260661	47263661	3001
				3kb_from_gene	14	47260661	47263661	3001
				exon	14	47263662	47264053	392
				intron	14	47264054	47266479	2426
				exon	14	47266480	47266575	96
				intron	14	47266576	47268309	1734
				exon	14	47268310	47268363	54
COLEC10	ENSBTAG00000017343	ENSBTAT00000023054	-1	intron	14	47268364	47279693	11330
				exon	14	47279694	47279765	72
				intron	14	47279766	47281205	1440
				exon	14	47281206	47281277	72
				intron	14	47281286	47308917	27632
				exon	14	47308914	47309061	148
				5kb	14	47309062	47314062	5001
				45kb	14	47314063	47359061	44999
				45kb	16	43449518	43458114	8597
				5kb	16	43458115	43463115	5001
				5UTR	16	43463070	43463115	46
MASP2	ENSBTAG00000012808	ENSBTAT00000017013	1	exon	16	43463116	43463120	5
				intron	16	43463121	43463203	83
				exon	16	43463204	43463432	229
				intron	16	43463433	43463551	119

				exon	16	43463552	43463729	178
				intron	16	43463730	43464498	769
				exon	16	43464499	43464630	132
				intron	16	43464631	43466756	2126
				exon	16	43466757	43466953	197
				intron	16	43466954	43467026	73
				exon	16	43467027	43467174	148
				intron	16	43467175	43470622	3448
				exon	16	43470623	43470741	119
MASP2	ENSBTAG00000012808	ENSBTAT00000017013	1	intron	16	43470742	43472495	1754
				exon	16	43472496	43472574	79
				intron	16	43472575	43474701	2127
				exon	16	43474702	43474833	132
				intron	16	43474834	43475333	500
				exon	16	43475334	43475408	75
				intron	16	43475409	43476632	1224
				exon	16	43476633	43477399	767
				3UTR	16	43477400	43478362	963
				3kb_from_stop	16	43477400	43480400	3001
				3kb_from_gene	16	43478362	43481362	3001
				3kb_from_gene	24	35627928	35630929	3002
				3kb_from_stop	24	35629346	35632346	3001
				3UTR	24	35630929	35632346	1418
				exon	24	35632347	35632366	20
				intron	24	35632367	35633926	1560
COLEC12	ENSBTAG00000007705	ENSBTAT00000047455	-1	exon	24	35633927	35634072	146
				intron	24	35634073	35642431	8359
				exon	24	35642432	35642541	110
				intron	24	35642542	35643723	1182
				exon	24	35643724	35643860	137
				intron	24	35643861	35644713	853

COLEC12	ENSBTAG00000007705	ENSBTAT00000047455	-1	exon	24	35644714	35645202	489
				intron	24	35645203	35658779	13577
				exon	24	35658780	35659826	1047
				intron	24	35659827	35661030	1204
				exon	24	35661031	35661129	99
				intron	24	35661130	35670671	9542
				exon	24	35670672	35670794	123
				intron	24	35670795	35799347	128553
				exon	24	35799348	35799398	51
				intron	24	35799399	35816124	16726
				exon	24	35816125	35816131	7
				5UTR	24	35816132	35816269	138
				5kb	24	35816132	35821132	5001
				45kb	24	35821133	35866269	45137
MBL2	ENSBTAG00000007049	ENSBTAT00000009270	1	45kb	26	6294785	6339793	45009
				5kb	26	6339794	6344794	5001
				5UTR	26	6344781	6344794	14
				exon	26	6344795	6344984	190
				intron	26	6344985	6345372	388
				exon	26	6345373	6345489	117
				intron	26	6345490	6346824	1335
				exon	26	6346825	6346893	69
				intron	26	6346894	6347903	1010
				exon	26	6347904	6348277	374
				3UTR	26	6348278	6348912	635
				3kb_from_stop	26	6348278	6351278	3001
				3kb_from_gene	26	6348913	6351912	3000
CGN1	ENSBTAG00000006536	ENSBTAT00000018649	1	45kb	28	35541900	35588221	46322
				5kb	28	35588221	35593221	5001
				5UTR	28	35591906	35591977	72
				intron	28	35591978	35592728	751

				5UTR	28	35592729	35592800	72
				intron	28	35592801	35593218	418
				5UTR	28	35593219	35593221	3
				exon	28	35593222	35593420	199
				intron	28	35593421	35597734	4314
				exon	28	35597735	35597851	117
				intron	28	35597852	35598132	281
				exon	28	35598133	35598240	108
				intron	28	35598241	35598546	306
CGN1	ENSBTAG00000006536	ENSBTAT00000018649	1	exon	28	35598547	35598654	108
				intron	28	35598655	35599018	364
				exon	28	35599019	35599135	117
				intron	28	35599136	35599750	615
				exon	28	35599751	35599834	84
				intron	28	35599835	35602290	2456
				exon	28	35602291	35602673	383
				3UTR	28	35602674	35602715	42
				3kb_from_stop	28	35602674	35605674	3001
				3kb_from_gene	28	35602715	35605715	3001
CL46	ENSBTAG00000048082	ENSBTAT00000028716	1	45kb	28	35619502	35664502	45001
				5kb	28	35664503	35669503	5001
CL43	ENSBTAG00000047317	ENSBTAT00000003773	1	45kb	28	35664665	35709665	45001
				5UTR	28	35668986	35669085	100
				intron	28	35669086	35669500	415
				5UTR	28	35669501	35669503	3
				exon	28	35669504	35669702	199
CL46	ENSBTAG00000048082	ENSBTAT00000028716	1	intron	28	35669703	35674001	4299
				exon	28	35674002	35674118	117
				intron	28	35674119	35674453	335
				exon	28	35674454	35674561	108
				intron	28	35674562	35674883	322

				exon	28	35674884	35674991	108
				intron	28	35674992	35675349	358
				exon	28	35675350	35675466	117
				intron	28	35675467	35676081	615
				exon	28	35676082	35676165	84
				intron	28	35676166	35678381	2216
CL46	ENSBTAG00000048082	ENSBTAT00000028716	1	exon	28	35678382	35678435	54
				intron	28	35678436	35699766	21331
				exon	28	35699767	35700095	329
				3UTR	28	35700096	35700161	66
				3kb_from_stop	28	35700096	35703096	3001
				3kb_from_gene	28	35700161	35703161	3001
				5kb	28	35709666	35714666	5001
				5UTR	28	35713329	35713496	168
				5UTR	28	35714664	35714666	3
				exon	28	35714667	35714865	199
				intron	28	35714866	35717562	2696
				exon	28	35717563	35717670	108
				intron	28	35717671	35717992	321
				exon	28	35717993	35718064	72
				intron	28	35718065	35718443	378
CL43	ENSBTAG00000047317	ENSBTAT00000003773	1	exon	28	35718444	35718560	117
				intron	28	35718561	35718748	187
				exon	28	35718749	35718832	84
				intron	28	35718833	35721556	2723
				exon	28	35721557	35721711	155
				intron	28	35721712	35722860	1148
				exon	28	35722861	35722931	71
				3UTR	28	35722932	35723104	173
				3kb_from_stop	28	35722932	35725932	3001
				3kb_from_gene	28	35723105	35726104	3000

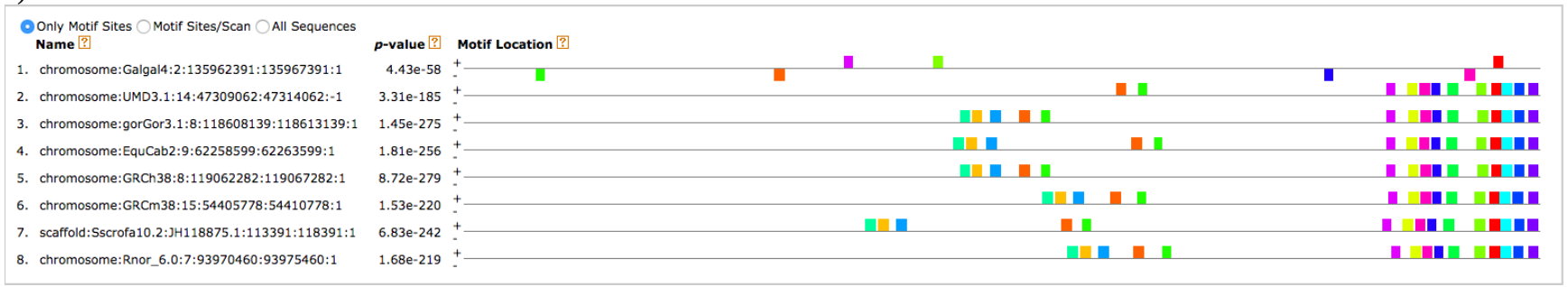
				45kb	28	35764587	35810076	45490
				5kb	28	35810077	35815077	5001
				5UTR	28	35813760	35813798	39
				intron	28	35813799	35814585	787
				5UTR	28	35814586	35814657	72
				intron	28	35814658	35815074	417
				5UTR	28	35815075	35815077	3
				exon	28	35815078	35815276	199
				intron	28	35815277	35819601	4325
				exon	28	35819602	35819718	117
				intron	28	35819719	35820000	282
SFTPD	ENSBTAG00000046421	ENSBTAT00000008579	1	exon	28	35820001	35820108	108
				intron	28	35820109	35820417	309
				exon	28	35820418	35820525	108
				intron	28	35820526	35820887	362
				exon	28	35820888	35821004	117
				intron	28	35821005	35821902	898
				exon	28	35821903	35821986	84
				intron	28	35821987	35824006	2020
				exon	28	35824007	35824383	377
				3kb_from_stop	28	35824384	35827384	3001
				3kb_from_gene	28	35824384	35827384	3001
				3kb_from_gene	28	35837849	35840849	3001
				3kb_from_stop	28	35837918	35840918	3001
				3UTR	28	35840849	35840918	70
				exon	28	35840919	35841292	374
MBL1	ENSBTAG00000023032	ENSBTAT00000001165	-1	intron	28	35841293	35841920	628
				exon	28	35841921	35841995	75
				intron	28	35841996	35842921	926
				exon	28	35842922	35843038	117
				intron	28	35843039	35843425	387

				exon	28	35843426	35843606	181
				5UTR	28	35843607	35843615	9
				5kb	28	35843607	35848607	5001
MBL1	ENSBTAG00000023032	ENSBTAT00000001165	-1	intron	28	35843616	35845921	2306
				5UTR	28	35845922	35846070	149
				45kb	28	35846071	35870565	24495
				3kb_from_gene	28	35847665	35850665	3001
				3kb_from_stop	28	35848716	35851716	3001
				3UTR	28	35850665	35851716	1052
				exon	28	35851717	35852093	377
				intron	28	35852094	35852538	445
				exon	28	35852539	35852616	78
				intron	28	35852617	35853352	736
SFTPA1	ENSBTAG00000023032	ENSBTAT00000031298	-1	exon	28	35853353	35853472	120
				intron	28	35853473	35853722	250
				exon	28	35853723	35853894	172
				5UTR	28	35853895	35853917	23
				5kb	28	35853895	35858895	5001
				intron	28	35853918	35854728	811
				5UTR	28	35854729	35854765	37
				45kb	28	35858896	35870565	11670

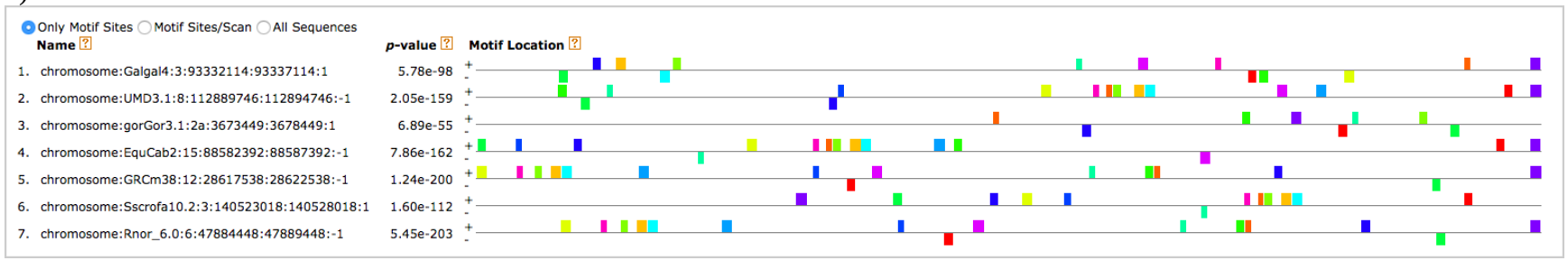
Appendix 3.2. Conserved motifs found within the 5 kb upstream of the start codon of the collagenous lectin and MASP genes.

An individual box represents the location of a conserved motif; similar motifs in different species share a box with the same colour. UMD3.1: cattle genome; gorGor3: gorilla genome; EquCab2: horse genome; GRCh38: human genome; GRCm38: mouse genome; Sscrofa10.2: porcine genome; Rnor_6.0: rat genome; Galgal4: chicken genome.

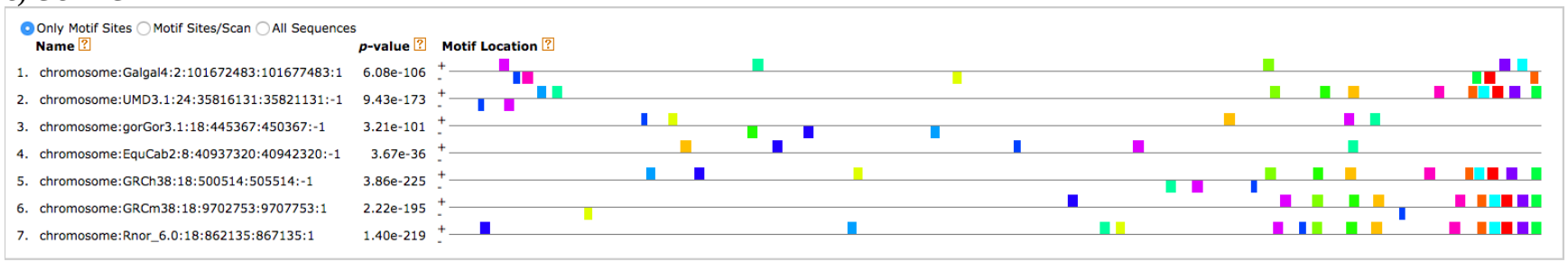
a) *COLEC10*



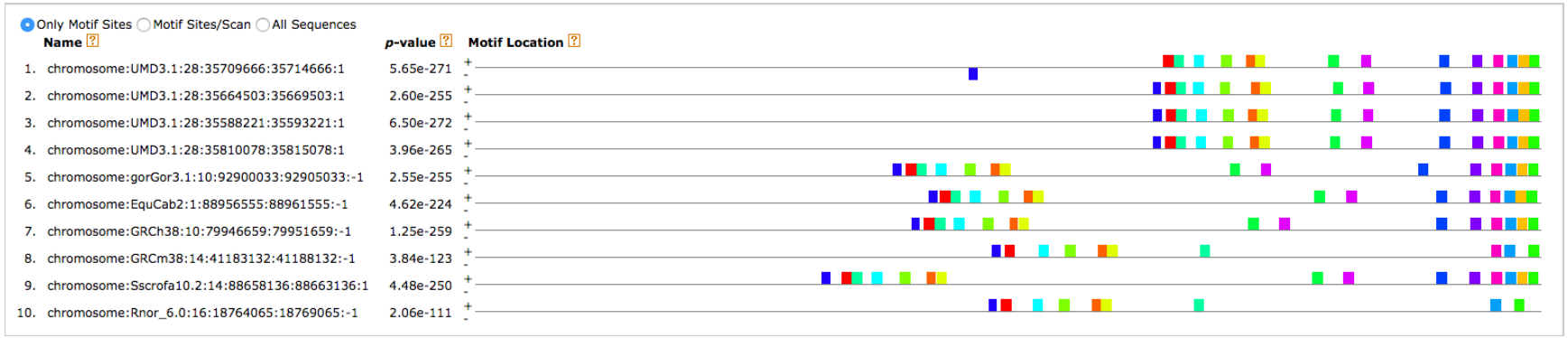
b) *COLEC11*



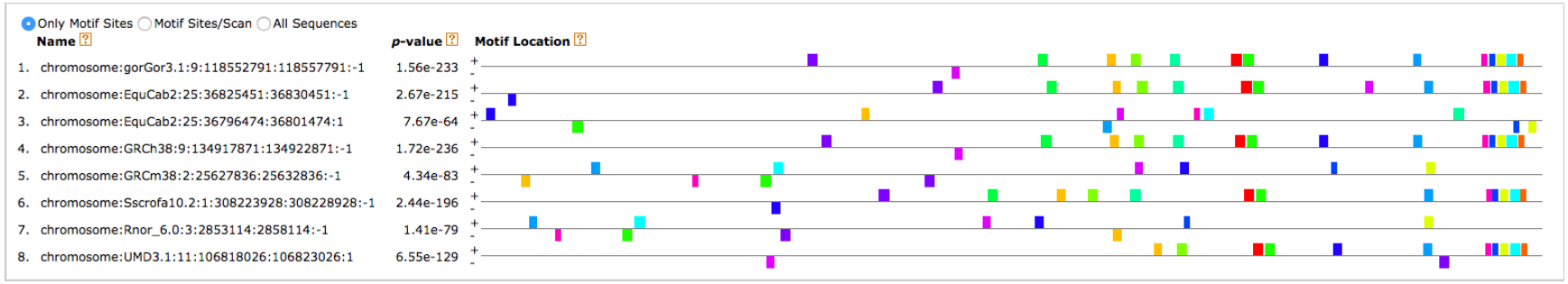
c) *COLEC12*



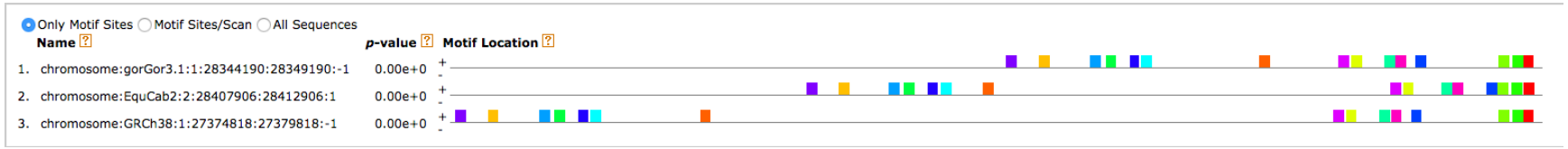
d) *CGN-CL43-CL46*[†]



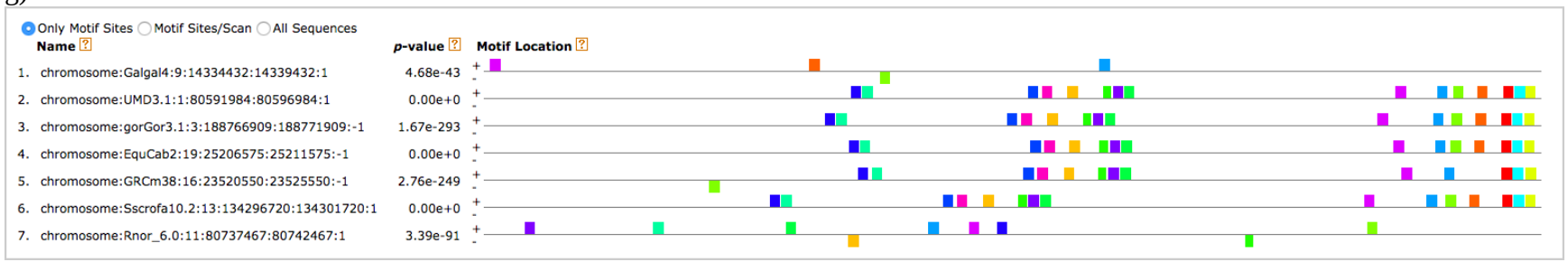
e) *FCN1 (and equine FCN-1 like)*



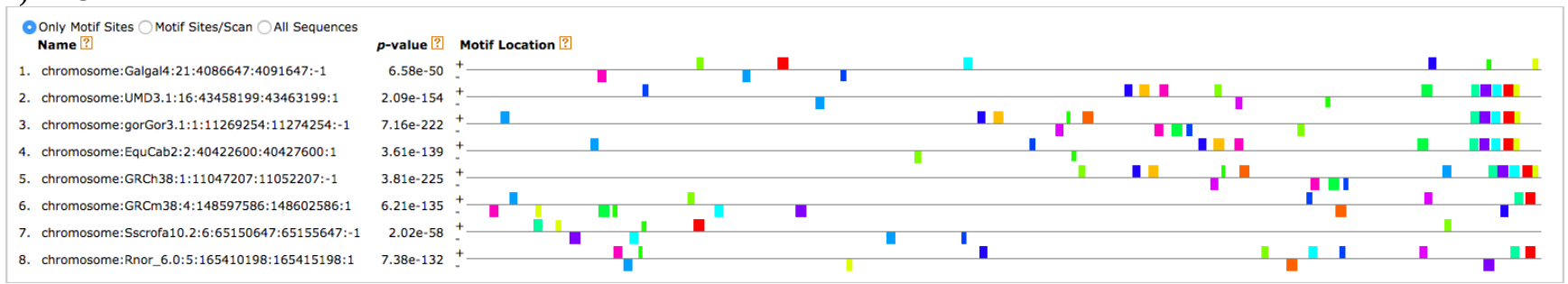
f) *FCN3*



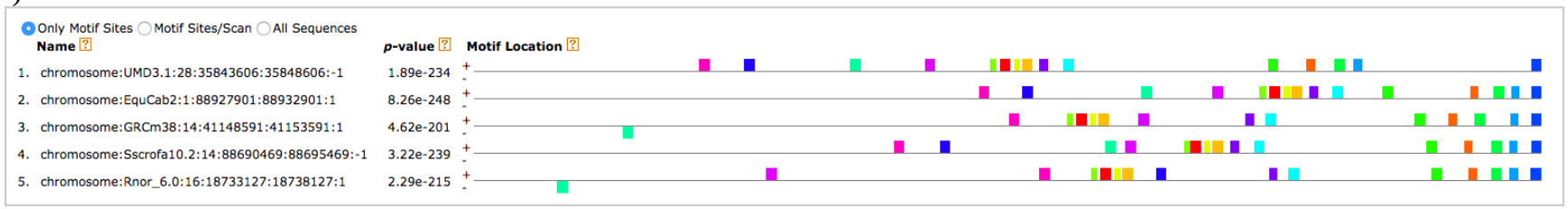
g) *MASP1*



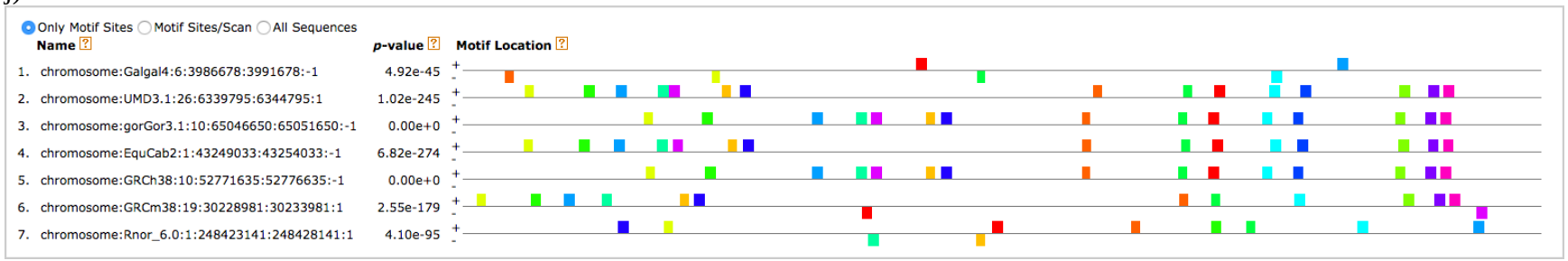
h) *MASP2*



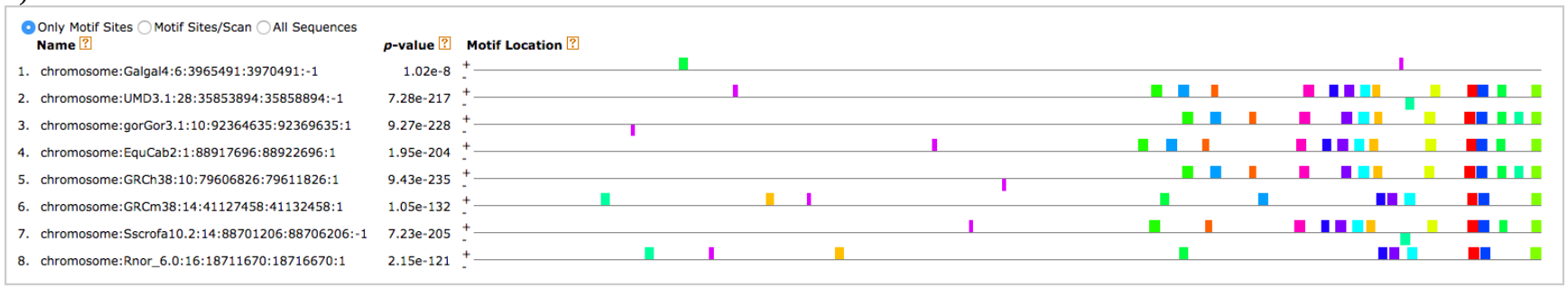
i) *MBL1*[§]



j) *MBL2*



k) *SFTPA*

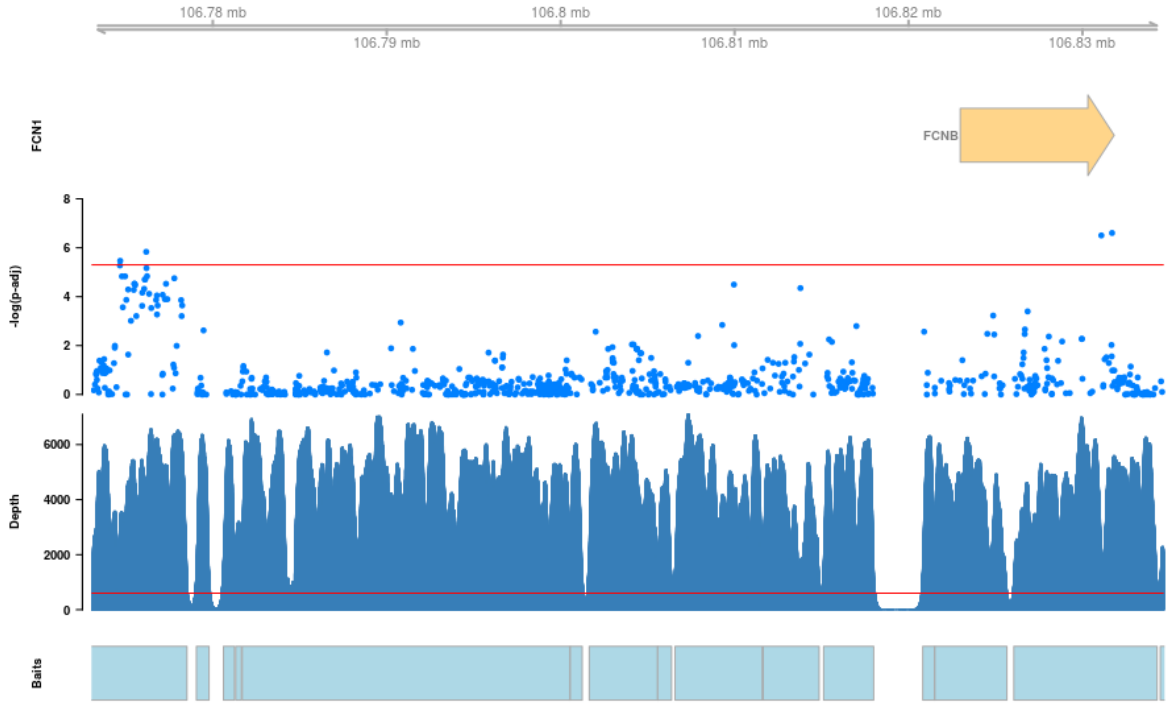


l) *SFTPD*

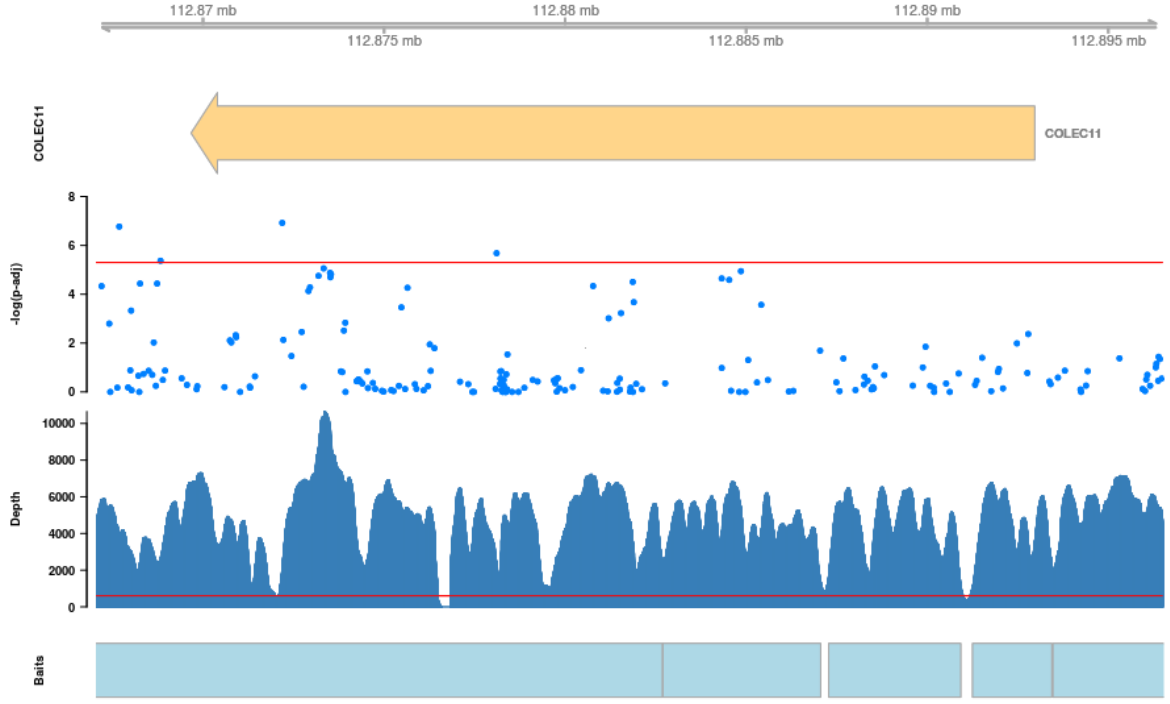
†: Although unique to ruminants, *CGN1*, *CL43*, and *CL46* are likely the result of duplications of the *SFTPD* gene, and were thus compared to *SFTPD*.
§*MBL1* is a pseudogene in primates

Appendix 3.3. Manhattan plots, coverage, and bait information for bovine *FCN1*, *COLEC11*, and *COLEC12*.

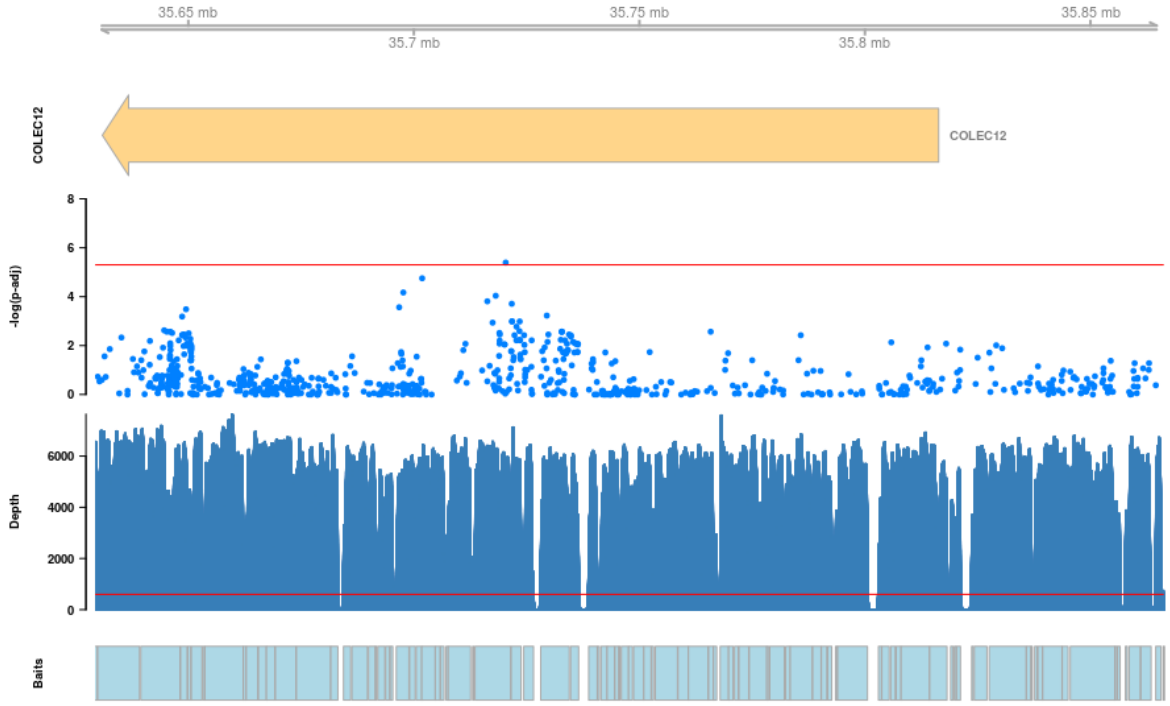
a)



b)



c)



Distribution of significant associations are shown for **a** *COLEC11*, **b** *COLEC12*, and **c** *FCN1*. The red line indicates the cutoff for BADGE class II significance ($p < 5 \times 10^{-6}$). The depth of sequencing (total from all pools) and the probes used for target capture are shown to illustrate gaps in sequencing and variant discovery.

Appendix 3.4. The allele frequency of significantly associated alleles are shown, including estimations of the minor allele frequencies for both the Non-Infectious and Infectious populations.

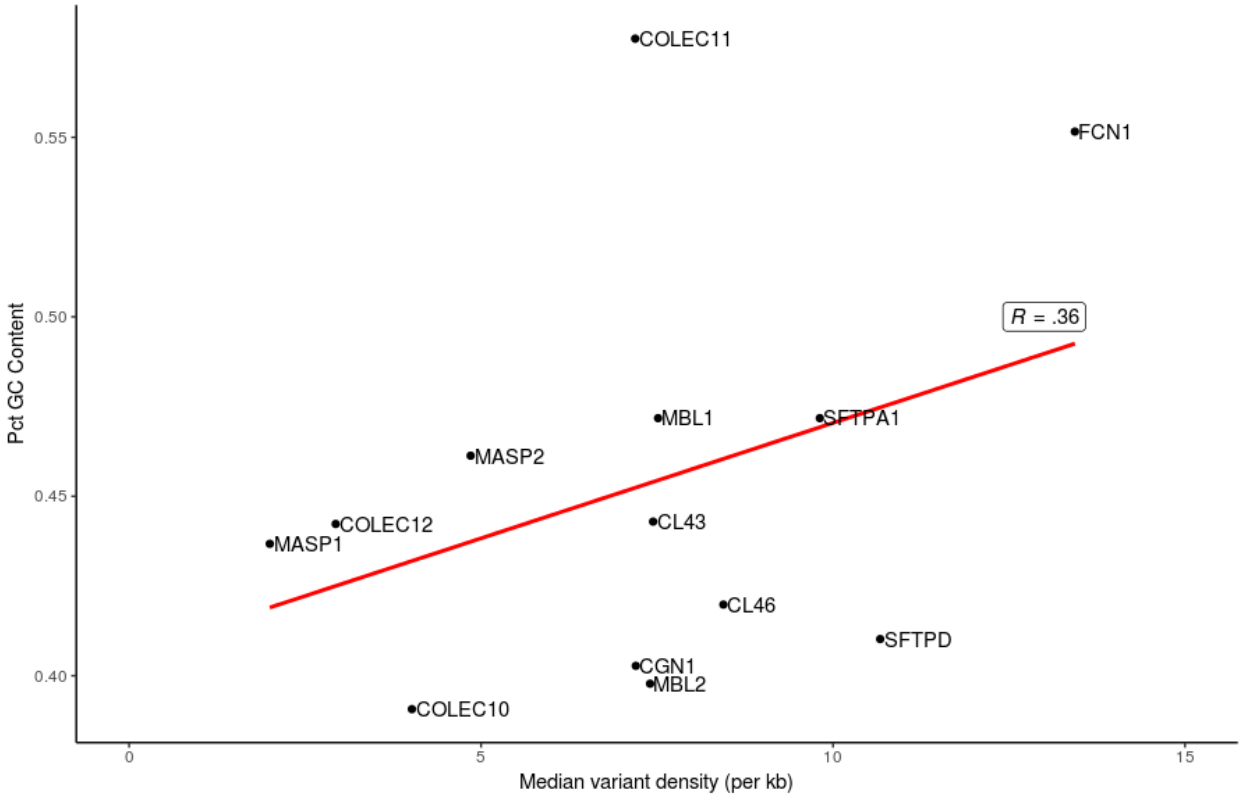
Gene	Chrom	Position	rsID	Ref allele	Alt allele	Freq in Non-Infectious	Freq in Infectious	-log(p-adj)*	Location
MASP1	chr1	80622569	rs43249827	G	A	0.075	0.2475	7.95	intron
CGN1	chr28	35544451	rs209513117	T	A	0.0175	0.14	7.95	upstream 5-50 kb
CGN1	chr28	35544473	rs211258454	C	G	0.0175	0.145	7.95	upstream 5-50 kb
CGN1	chr28	35548333	rs378135096	G	A	0.0325	0.1775	7.95	upstream 5-50 kb
CGN1	chr28	35549433	rs527011314	A	G	0.0175	0.14	7.95	upstream 5-50 kb
CGN1	chr28	35555210	rs208114400	T	C	0.0175	0.14	7.95	upstream 5-50 kb
CGN1	chr28	35564277	rs209926165	G	C	0.02	0.145	7.68	upstream 5-50 kb
CGN1	chr28	35555183	rs211682909	A	G	0.0175	0.135	7.26	upstream 5-50 kb
CGN1	chr28	35544939	rs210425590	A	G	0.02	0.1375	7.26	upstream 5-50 kb
CGN1	chr28	35549175	rs210225284	A	G	0.02	0.1375	7.26	upstream 5-50 kb
CGN1	chr28	35546195	rs211256115	A	T	0.0275	0.155	7.22	upstream 5-50 kb
CGN1	chr28	35546274	rs383948170	G	A	0.0325	0.1625	7.22	upstream 5-50 kb
CGN1	chr28	35549355	rs463399282	A	C	0.0175	0.13	7.21	upstream 5-50 kb
CGN1	chr28	35555503	rs211418990	A	G	0.0175	0.1325	7.21	upstream 5-50 kb
CGN1	chr28	35546132	rs209644363	A	G	0.025	0.1475	7.10	upstream 5-50 kb
MASP1	chr1	80613375	rs43250747	T	A	0.085	0.2475	7.02	intron
CGN1	chr28	35544127	rs211211424	C	T	0.02	0.135	7.02	upstream 5-50 kb
CGN1	chr28	35549164	rs208463943	A	G	0.02	0.135	7.02	upstream 5-50 kb
CGN1	chr28	35549371	rs445770773	G	A	0.02	0.135	7.02	upstream 5-50 kb
CGN1	chr28	35546076	rs208532033	G	A	0.0225	0.14	7.01	upstream 5-50 kb
CGN1	chr28	35546079	rs210766709	A	G	0.0225	0.14	7.01	upstream 5-50 kb
COLEC11	chr8	112872188	rs381186716	C	G	0.1	0.27	6.92	intron
CGN1	chr28	35549260	rs211013887	C	T	0.0225	0.1375	6.80	upstream 5-50 kb
CGN1	chr28	35555289	rs209752875	G	A	0.0225	0.1375	6.80	upstream 5-50 kb
COLEC11	chr8	112867687	rs382787827	G	A	0.1375	0.3175	6.77	downstream 3 kb
CGN1	chr28	35544985	rs211314548	A	G	0.02	0.13	6.64	upstream 5-50 kb
CGN1	chr28	35549055	rs440808750	A	T	0.02	0.13	6.64	upstream 5-50 kb
CGN1	chr28	35549563	rs447690172	A	G	0.02	0.1275	6.64	upstream 5-50 kb
COLEC12	chr28	35666506	rs209858141	T	C	0.02	0.1275	6.64	intron

FCN1	chr11	106831702	rs458265862	G	C	0.14	0.32	6.61	downstream 3 kb
MASP1	chr1	80613612	rs43250749	A	G	0.0925	0.2525	6.53	intron
MASP1	chr1	80614847	rs43250756	G	T	0.0925	0.2525	6.53	intron
CGN1	chr28	35547554	rs207808860	C	T	0.04	0.1675	6.52	upstream 5-50 kb
CGN1	chr28	35552052	rs209143409	C	A	0.0175	0.1225	6.52	upstream 5-50 kb
FCN1	chr11	106831086	rs379035858	C	T	0.1825	0.05	6.50	intron
CGN1	chr28	35548408	rs208236288	G	A	0.0225	0.1325	6.45	upstream 5-50 kb
CGN1	chr28	35564217	rs208961754	A	T	0.025	0.1375	6.45	upstream 5-50 kb
MASP1	chr1	80611859	rs43250737	G	A	0.095	0.2525	6.30	intron
MASP1	chr1	80613393	rs43250748	C	T	0.09	0.245	6.29	intron
CGN1	chr28	35543676	rs209924982	T	C	0.0225	0.13	6.24	upstream 5-50 kb
CGN1	chr28	35548272	rs210383798	C	T	0.025	0.1325	6.23	upstream 5-50 kb
CGN1	chr28	35548416	rs209630668	A	G	0.025	0.1325	6.23	upstream 5-50 kb
CGN1	chr28	35548620	rs210811039	T	A	0.0275	0.14	6.23	upstream 5-50 kb
MASP1	chr1	80611537	rs43250734	T	A	0.0925	0.245614035	6.23	intron
CGN1	chr28	35546381	rs209271778	T	C	0.035	0.1525	6.08	upstream 5-50 kb
CGN1	chr28	35545152	rs209401407	A	C	0.0275	0.1375	6.04	upstream 5-50 kb
CGN1	chr28	35547884	rs211137037	A	C	0.03	0.14	6.04	upstream 5-50 kb
CGN1	chr28	35548911	rs208348702	C	T	0.0225	0.125	6.04	upstream 5-50 kb
MASP1	chr1	80611569	rs43250735	T	A	0.105	0.26	6.03	intron
MASP1	chr1	80613074	rs43250746	A	G	0.09	0.24	5.98	intron
MASP1	chr1	80620340	rs109975315	C	T	0.085	0.2325	5.98	intron
MASP1	chr1	80612541	rs110375944	T	C	0.0975	0.25	5.94	intron
MASP1	chr1	80614960	rs43250758	A	G	0.0975	0.25	5.94	intron
MASP1	chr1	80611723	rs43250736	A	G	0.0925	0.2425	5.92	intron
MASP1	chr1	80612519	rs109919595	C	T	0.1	0.2525	5.87	intron
CGN1	chr28	35548748	rs436645644	A	T	0.025	0.13	5.86	upstream 5-50 kb
FCN1	chr11	106776174	rs482440534	C	T	0.1025	0.0125	5.83	upstream 5-50 kb
MASP1	chr1	80616621	rs43250770	C	G	0.1	0.25	5.69	intron
COLEC11	chr8	112878108	rs452265753	T	C	0.245614035	0.4325	5.68	intron
CGN1	chr28	35543754	rs209082641	G	C	0.025	0.1275	5.66	upstream 5-50 kb
CGN1	chr28	35546761	rs385336894	T	C	0.0225	0.1225	5.66	upstream 5-50 kb
CGN1	chr28	35546899	rs210538163	C	T	0.0275	0.130325815	5.66	upstream 5-50 kb
CGN1	chr28	35548955	rs209661333	A	C	0.0225	0.1225	5.66	upstream 5-50 kb

MASP1	chr1	80614924	rs43250757	C	A	0.1	0.2475	5.53	intron
MASP1	chr1	80615246	rs43250760	A	G	0.095	0.24	5.51	intron
CGN1	chr28	35548161	rs209933151	C	T	0.0325	0.1375	5.48	upstream 5-50 kb
CGN1	chr28	35548552	rs454118093	C	T	0.0325	0.1375	5.48	upstream 5-50 kb
CGN1	chr28	35548121	rs207834730	C	T	0.03	0.135	5.48	upstream 5-50 kb
FCN1	chr11	106774678	rs378747274	C	T	0.421052632	0.240601504	5.47	upstream 5-50 kb
CGN1	chr28	35543521	rs210717733	G	A	0.0275	0.13	5.47	upstream 5-50 kb
CGN1	chr28	35559685	rs208751250	G	A	0.025	0.1225	5.45	upstream 5-50 kb
CGN1	chr28	35559696	rs210841169	A	T	0.0225	0.1175	5.45	upstream 5-50 kb
COLEC12	chr24	35720361	rs386007052	G	T	0.085	0.0075	5.40	intron
COLEC11	chr8	112868826	rs42393553	C	A	0.2175	0.3925	5.37	downstream 3 kb

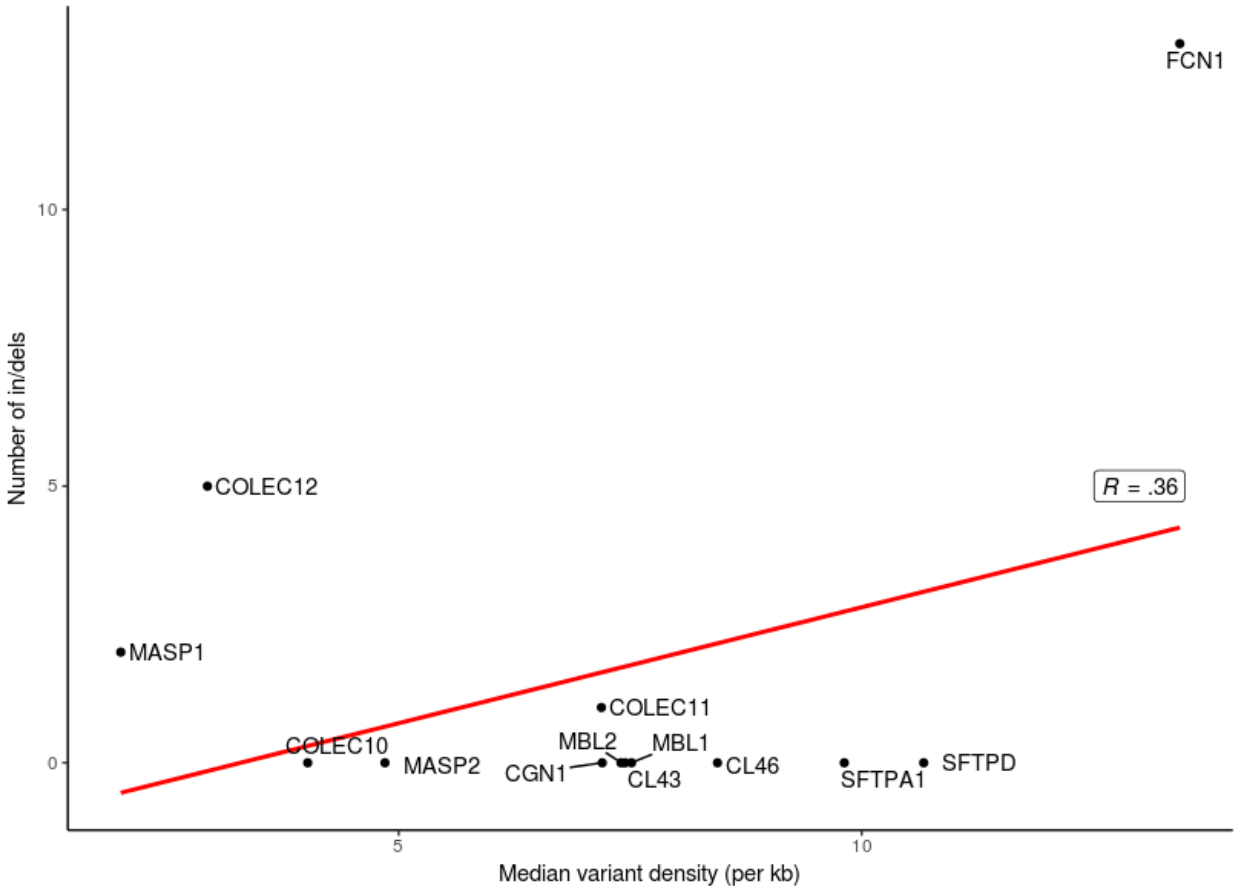
*The p-value has been adjusted according to the Benjamini-Hochberg procedure.

Appendix 3.5. The correlation between the percent GC content and the variant density in bovine collagenous lectins.



The correlation between the percent GC content and the variant density in the targeted genes was not significant ($p = 0.23$, Pearson's correlation).

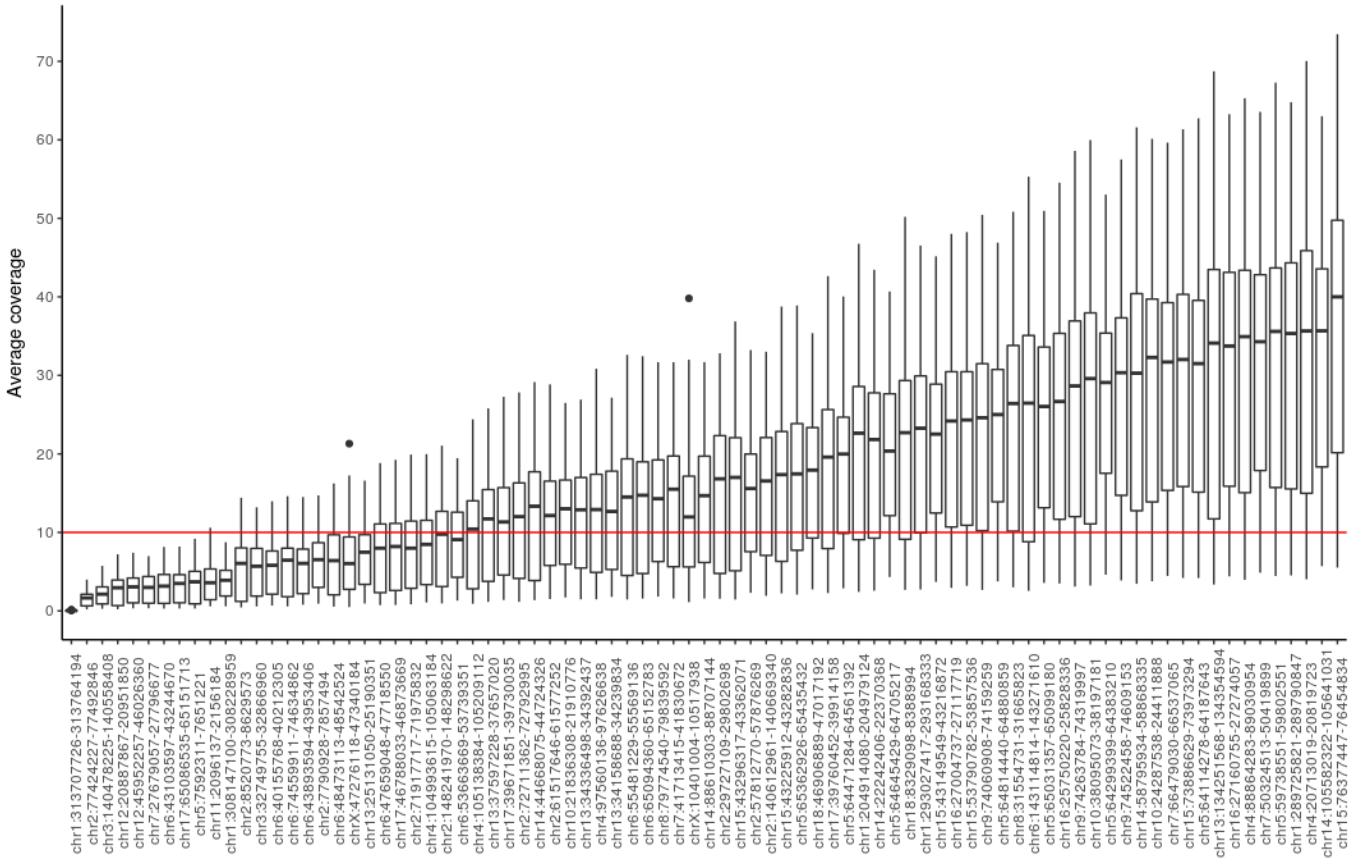
Appendix 3.6. The correlation between the median variant density and the frequency of indels in bovine collagenous lectins.



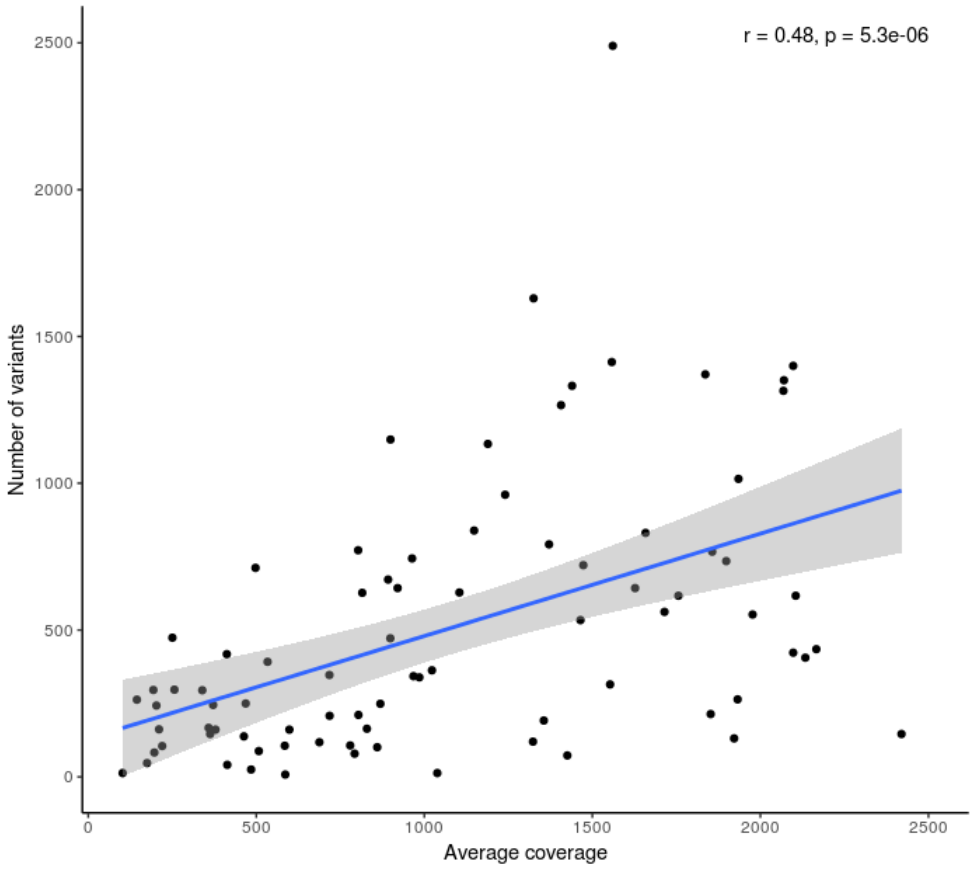
The correlation between the median variant density and the frequency of indels. Correlation was not significant (p = 0.23, Pearson's correlation)

Appendix 4.1. Median depth of coverage by target and relationship between depth and variant density.

a)



b)



a) The average depth of coverage across all pigs is shown for each targeted region. The blue line indicates a minimum depth of coverage of 10 which was required for variant calling. b) The average coverage for each targeted region is correlated to the total number of variants identified in that region (Pearson's coefficient).

Appendix 4.2. List of genes and coordinates (based on Sscrofa 10.2) targeted for re-sequencing.

Chr	Target start	Target end	Targeted gene	Ensembl gene ID
GL894246_2	1	23701	*C1R*	ENSSSCG00000000675
GL894265_1	1	14183	*NLRC5*	ENSSSCG00000020798
GL896416_1	1206	58201	*NOD2*	ENSSSCG00000038073*
JH118996_1	1	18737	*DEFB103A*	ENSSSCG00000020934
chr1	204914080	204979124	*LGALS3*	ENSSSCG00000005055
chr1	289725821	289790847	*TLR4*	ENSSSCG00000005503
chr1	293027417	293168333	*C5*	ENSSSCG00000005512
chr1	308147100	308228959	*FCN2*	ENSSSCG00000023333
			FCN1	ENSSSCG00000029414
chr1	313707726	313764194	*C8G*	ENSSSCG00000005840
chr2	7790928	7857494	*LGALS12*	ENSSSCG00000013056
chr2	8520773	8629573	*SCGB1A1*	ENSSSCG00000013060
chr2	29727109	29802698	*CD59*	ENSSSCG00000024791
chr2	57812770	57876269	*NLRP3*	ENSSSCG00000013940
chr2	61517646	61577252	*PGLYRP2*	ENSSSCG00000025560
chr2	71917717	71975832	*CLEC4G*	ENSSSCG00000013578
chr2	72711362	72792995	*C3*	ENSSSCG00000013551
chr2	77424227	77492846	*CFD*	ENSSSCG00000013418
chr2	140612961	140669340	*LEAP2*	ENSSSCG00000014290
chr2	148241970	148298622	*CD14*	ENSSSCG00000014369
chr3	32749755	32866960	*CLEC16A*	ENSSSCG00000026569
chr3	140478225	140558408	*COLEC11*	ENSSSCG00000025629
chr4	20713019	20819723	*COLEC10*	ENSSSCG00000029324
			SELE	ENSSSCG00000006286
chr4	88864283	89030954	*SELL*	ENSSSCG00000006287
			SELP	ENSSSCG00000006288
chr4	97560136	97626638	*ITLN2*	ENSSSCG00000028374
chr4	104993615	105063184	*PGLYRP4*	ENSSSCG00000006587
chr4	105138384	105209112	*PGLYRP3*	ENSSSCG00000027322
chr5	7592311	7651221	*LGALS1*	ENSSSCG00000000125
chr5	59738551	59802551	*C1S*	ENSSSCG00000039847*
chr5	64114278	64187643	*KLRA1*	ENSSSCG00000000636
chr5	64299399	64383210	*KLRC1*	ENSSSCG00000000640
			KLRK1	ENSSSCG00000000641
chr5	64471284	64561392	*CLEC1A*	ENSSSCG00000000649
			CLEC7A	ENSSSCG00000000648
chr5	64645429	64705217	*CLEC1B*	ENSSSCG00000000650
chr5	64814440	64880859	*KLRF1*	ENSSSCG00000000655
chr5	65031357	65099180	*KLRB1*	ENSSSCG00000030978

chr5	65362926	65435432	*KLRG1*	ENSSSCG00000000663
chr6	40155768	40212305	*HAMP*	ENSSSCG00000002886
chr6	43103597	43244670	*LGALS4*	ENSSSCG00000022017
			LGALS7	ENSSSCG00000024281
chr6	43893594	43953406	*LGALS13*	ENSSSCG00000002983
chr6	47659048	47718550	*PGLYRP1*	ENSSSCG00000003099
chr6	48473113	48542524	*C5AR1*	ENSSSCG00000029371
chr6	53663669	53739351	*NLRP2*	ENSSSCG00000003286
chr6	55481229	55569136	*NLRP5*	ENSSSCG00000003316
chr6	65094360	65152783	*MASP2*	ENSSSCG00000003410
			C1QA	ENSSSCG00000003524
chr6	74559911	74634862	*C1QC*	ENSSSCG00000003525
			C1QB	ENSSSCG00000003526
chr6	143114814	143271610	*C8B*	ENSSSCG00000003834
			C8A	ENSSSCG00000003835
			C4A	ENSSSCG00000001427
chr7	27679057	27796677	*CFB*	ENSSSCG00000024914
			C2	ENSSSCG00000001422
chr7	41713415	41830672	*TREM2*	ENSSSCG00000001614
			TREM1	ENSSSCG00000001617
chr7	50324513	50419899	*DEFB114*	ENSSSCG00000001741
			DEFB110	ENSSSCG00000001743
chr7	66479030	66537065	*CLEC14A*	ENSSSCG00000023760
chr8	31554731	31665823	*TLR10*	ENSSSCG00000027877
			TLR6	ENSSSCG00000026592
chr8	79774540	79839592	*TLR2*	ENSSSCG00000009002
chr9	74060908	74159259	*C4BPA*	ENSSSCG00000015662
chr9	74263784	74319997	*CD55*	ENSSSCG00000015664
chr9	74522458	74609153	*CD46*	ENSSSCG00000015293
chr10	21836308	21910776	*TLR5*	ENSSSCG00000010885
chr10	24287538	24411888	*CFH*	ENSSSCG00000010893
chr10	38095073	38197181	*DDX58*	ENSSSCG00000030408
chr11	2096137	2156184	*C1QTNF9A*	ENSSSCG00000009285
chr12	20887867	20951850	*DHX58*	ENSSSCG00000017416
chr12	45952257	46026360	*LGALS9*	ENSSSCG00000017754
chr13	25131050	25190351	*MYD88*	ENSSSCG00000011251
chr13	34158688	34239834	*PMAP-23*	ENSSSCG00000021877
chr13	34336498	34392437	*PMAP-37/NPG4*	ENSSSCG00000011349
chr13	37597228	37657020	*TLR9*	ENSSSCG00000011436
chr13	134251568	134354594	*MASP1*	ENSSSCG00000011806
chr14	22242406	22370368	*DDX60*	ENSSSCG00000009720
chr14	44668075	44724326	*SELPLG*	ENSSSCG00000024495

chr14	58795934	58868335	*LGALS8*	ENSSSCG00000010146
			SFTPD	ENSSSCG00000010334
chr14	88610303	88707144	*MBL1*	ENSSSCG00000024252
			SFTPA1	ENSSSCG00000010336
chr14	105582322	105641031	*MBL2*	ENSSSCG00000010427
chr15	43149549	43216872	*BD104-like*	ENSSSCG00000030758
chr15	43225912	43282836	*DEFB1 (PBD1)*	ENSSSCG00000027555
chr15	43296317	43362071	*DEFB1 (PBD2)*	ENSSSCG00000029990
chr15	53790782	53857536	*TLR3*	ENSSSCG00000015801
chr15	73886629	73973294	*CD302*	ENSSSCG00000015884
chr15	76377447	76454834	*IFIH1*	ENSSSCG00000015897
chr16	25750220	25828336	*C9*	ENSSSCG00000016856
chr16	27004737	27117719	*C7*	ENSSSCG00000016859
chr16	27160755	27274057	*C6*	ENSSSCG00000016861
chr17	39671851	39730035	*DEFB129*	ENSSSCG00000007218
			DEFB125	ENSSSCG00000007217
chr17	39760452	39914158	*DEFB122*	ENSSSCG00000030975
			DEFB121	ENSSSCG00000028629
			DEFB123	ENSSSCG00000007226
chr17	46788033	46873669	*LBP*	ENSSSCG00000028758
chr17	65086535	65151713	*ZBP1*	ENSSSCG00000007508
chr18	8329098	8388994	*CLEC5A*	ENSSSCG00000021933
chr18	46906889	47017192	*NOD1*	ENSSSCG00000016678
chrX	10401004	10517938	*TLR7*	ENSSSCG00000012117
			TLR8	ENSSSCG00000012118
chrX	47276118	47340184	*CFP*	ENSSSCG00000012278

*EnsemblID from Sscrofa11

Appendix 4.3. SNVs associated with changes in gene expression (eQTL) in 69 healthy, market weight pigs.

Gene	Chr	SNP ID	Ref	Alt	p-value*	Adjusted p-value	Effect on expression
MBL2	chr14	rs341757289	C	T	1.15E-009	6.68E-005	-2.91
MBL2	chr14	rs343818728	T	C	1.55E-009	6.68E-005	-2.89
MBL2	chr14	rs328135675	C	T	2.49E-009	7.13E-005	-2.88
MBL2	chr14	rs321481373	G	A	3.70E-009	7.96E-005	-2.86
MBL2	chr14	rs336690766	C	A	8.14E-009	0.000116688	-2.83
MBL2	chr14	rs343578555	C	G	6.97E-009	0.000116688	-2.81
MBL2	chr14	rs341136794	A	G	9.54E-009	0.000117257	-2.78
MBL2	chr14	rs326271681	G	A	1.54E-008	0.000165772	-2.77
SFTPD	chr14	rs319707230	G	C	4.53E-008	0.000433066	4.06
MBL2	chr14	rs327359428	C	A	8.62E-008	0.000741749	-2.72
MBL2	chr14	rs340532136	C	T	1.08E-007	0.000847711	-2.71
FCN2	chr1	rs320342887	A	G	1.28E-007	0.000916363	-0.96
FCN2	chr1	rs341336572	G	T	2.10E-007	0.001390932	-3.20
FCN2	chr1	rs341159094	A	G	2.58E-007	0.001585909	-3.19
FCN2	chr1	rs343849769	C	G	3.04E-007	0.001745172	-3.65
MBL2	chr14	rs318652830	C	A	3.84E-007	0.00206356	2.74
FCN2	chr1	rs333222079	C	G	5.00E-007	0.002490575	-3.57
LGALS3	chr1	rs329498254	G	A	5.42E-007	0.002490575	0.99
LGALS3	chr1	rs327236346	A	C	5.79E-007	0.002490575	0.96
MBL2	chr14	rs322157250	C	T	5.51E-007	0.002490575	-2.75
FCN2	chr1	rs330321112	A	G	8.77E-007	0.002901728	-0.90
FCN2	chr1	rs343422522	A	G	8.25E-007	0.002901728	-0.91
LGALS3	chr1	rs328028771	C	G	8.50E-007	0.002901728	1.00
LGALS3	chr1	rs81349838	G	A	8.66E-007	0.002901728	0.90
MBL2	chr14	rs345981827	A	T	7.45E-007	0.002901728	-2.42
MBL2	chr14	rs336606363	A	G	1.06E-006	0.003154065	-2.40
MBL2	chr14	rs337560267	C	A	1.10E-006	0.003154065	-2.60
C9	chr16	rs341476407	G	A	1.31E-006	0.003407647	-0.76
C9	chr16	rs338256612	G	A	1.35E-006	0.003412282	-0.77
C9	chr16	rs334283045	T	C	1.52E-006	0.003670055	-0.79
FCN2	chr1	rs342490172	T	C	1.59E-006	0.003670055	-1.05
FCN2	chr1	rs691977343	T	A	1.66E-006	0.003670055	-2.80
TLR7	chrX	rs318825117	T	C	1.67E-006	0.003670055	0.51
ITLN2	chr4	rs333031027	G	T	1.89E-006	0.003878036	3.92
TLR7	chrX	rs344488277	C	T	1.87E-006	0.003878036	0.50
LGALS3	chr1	rs322326231	A	T	2.04E-006	0.004086812	0.99
LGALS3	chr1	rs323704144	G	A	2.10E-006	0.004101975	0.87
FCN2	chr1	rs326158227	A	C	2.25E-006	0.004200054	-0.86

FCN2	chr1	rs318962795	T	A	2.39E-006	0.004374358	-0.89
FCN2	chr1	rs337034769	T	C	2.56E-006	0.004490783	-0.88
FCN2	chr1	rs342195622	G	A	2.93E-006	0.004495435	-0.89
LGALS3	chr1	rs345875831	A	T	2.91E-006	0.004495435	0.96
MBL2	chr14	rs344981933	T	A	2.67E-006	0.004495435	-2.25
SFTPD	chr14	rs322111022	T	G	2.88E-006	0.004495435	4.12
TLR7	chrX	rs324023759	G	A	3.05E-006	0.004529867	0.55
FCN2	chr1	rs324326399	G	A	3.25E-006	0.004588518	-0.87
MBL2	chr14	rs329132054	C	T	3.18E-006	0.004588518	-2.24
TLR7	chrX	rs336142494	T	C	3.25E-006	0.004588518	0.50
CD55	chr9	rs330938856	T	C	3.49E-006	0.004627123	0.58
CD55	chr9	rs340141583	G	A	3.49E-006	0.004627123	0.58
MBL2	chr14	rs329878332	A	C	3.49E-006	0.004627123	-2.23
C9	chr16	rs318612015	A	G	3.62E-006	0.004713943	-0.80
MBL2	chr14	rs342988863	A	G	3.69E-006	0.004733816	-2.23
FCN2	chr1	rs333340587	G	A	3.79E-006	0.004792884	-0.92
FCN2	chr1	rs342499704	A	T	4.18E-006	0.004991744	-1.19
MBL2	chr14	rs328673019	T	A	4.16E-006	0.004991744	2.96
MBL2	chr14	rs342733630	A	G	4.03E-006	0.004991744	2.94
SFTPD	chr14	rs331907708	C	T	4.07E-006	0.004991744	4.07
FCN2	chr1	rs338191496	T	G	4.36E-006	0.005070074	-0.84
SFTPD	chr14	rs327302123	G	A	4.31E-006	0.005070074	4.06
LGALS3	chr1	rs342505569	C	T	4.66E-006	0.005118364	0.96
SFTPD	chr14	rs341704971	A	C	4.47E-006	0.005118364	4.05
SFTPD	chr14	rs341850018	G	A	4.64E-006	0.005118364	4.05
SFTPD	chr14	rs319870426	A	G	4.88E-006	0.005243664	4.05
TLR7	chrX	rs322726571	G	A	5.03E-006	0.005276904	0.49
MBL2	chr14	rs320029735	C	T	5.33E-006	0.005469043	2.93
PBD-2	chr15	rs332209636	T	G	5.60E-006	0.005469043	2.10
PBD-2	chr15	rs339770947	T	C	5.60E-006	0.005469043	2.10
PBD-2	chr15	rs340048862	A	G	5.60E-006	0.005469043	2.10
TLR7	chrX	rs343573982	T	A	5.47E-006	0.005469043	0.48
LGALS3	chr1	rs323773237	A	G	5.72E-006	0.005525924	0.86
MBL2	chr14	rs338350707	G	A	5.97E-006	0.005621343	2.90
PBD-2	chr15	rs330073132	C	T	6.01E-006	0.005621343	2.19
PBD-2	chr15	rs343917580	G	A	6.01E-006	0.005621343	2.19
FCN2	chr1	rs334837690	T	G	6.25E-006	0.005723733	-0.92
C9	chr16	rs345164093	A	T	6.42E-006	0.005751541	0.76
C9	chr16	rs330251216	C	T	6.67E-006	0.005879843	-0.78
C9	chr16	rs330353029	C	T	6.84E-006	0.005879843	-0.72
CD55	chr9	rs344217470	C	T	7.19E-006	0.00605044	0.52
LGALS3	chr1	rs342263300	G	A	7.19E-006	0.00605044	0.98
MBL2	chr14	rs328826195	G	A	7.32E-006	0.00605044	-2.16

MBL2	chr14	rs344636925	G	A	7.32E-006	0.00605044	-2.16
CLEC5A	chr18	rs322190257	A	G	7.43E-006	0.006083721	1.03
TLR7	chrX	rs341913558	A	T	7.77E-006	0.006263546	0.48
CD55	chr9	rs329627322	G	A	8.05E-006	0.006410741	0.54
MBL2	chr14	rs325223250	A	T	8.42E-006	0.00664026	-2.17
MBL2	chr14	rs709263755	A	G	8.64E-006	0.006757042	-2.17
FCN2	chr1	rs330465066	A	G	8.75E-006	0.006780538	-1.28
C9	chr16	rs344706119	G	A	8.84E-006	0.006785317	-0.76
C9	chr16	rs328250865	A	G	9.25E-006	0.006924363	-0.74
LGALS3	chr1	rs344010591	C	A	9.55E-006	0.006924363	0.89
MBL2	chr14	rs323437099	G	A	9.82E-006	0.006924363	2.88
PBD-2	chr15	rs323532096	A	T	9.63E-006	0.006924363	2.09
TLR7	chrX	rs328438129	G	A	9.66E-006	0.006924363	0.97
CD55	chr9	rs338269240	A	C	1.06E-005	0.00737564	-0.55
PBD-2	chr15	rs345811814	G	T	1.10E-005	0.007533792	2.11
C9	chr16	rs321078517	T	C	1.13E-005	0.0075657	-0.75
FCN2	chr1	rs320593237	T	C	1.15E-005	0.007680772	-0.88
FCN2	chr1	rs334836779	C	A	1.17E-005	0.007737757	-0.91
C9	chr16	rs323491083	A	T	1.29E-005	0.008368123	-0.74
C9	chr16	rs340320773	G	A	1.30E-005	0.008368123	-0.75
LGALS3	chr1	rs341181428	G	A	1.38E-005	0.008731793	0.86
LGALS3	chr1	rs81246297	A	G	1.49E-005	0.009230992	0.89
MBL2	chr14	rs320522795	G	A	1.58E-005	0.009487026	2.83
MBL2	chr14	rs337644648	T	C	1.58E-005	0.009487026	2.83
PBD-2	chr15	rs330573086	C	T	1.59E-005	0.009487026	2.09
CLEC5A	chr18	rs336562067	G	A	1.62E-005	0.009585246	1.16
CD55	chr9	rs322614244	C	T	1.69E-005	0.00989789	0.52
C9	chr16	rs336979880	T	C	1.73E-005	0.010046007	-0.79
C9	chr16	rs325643836	A	C	1.78E-005	0.010251483	-0.78
C9	chr16	rs333867087	G	A	1.79E-005	0.01028897	-0.73
CD55	chr9	rs326945657	A	G	1.81E-005	0.010315405	0.53
C9	chr16	rs709466423	T	A	1.87E-005	0.010521112	-0.77
C9	chr16	rs324303965	T	C	1.89E-005	0.010530666	-0.75
C9	chr16	rs324205238	G	C	1.98E-005	0.010784428	-0.75
C9	chr16	rs336101345	A	G	1.97E-005	0.010784428	-0.75
MBL2	chr14	rs331367618	T	C	1.95E-005	0.010784428	2.80
MBL2	chr14	rs345313231	C	T	2.05E-005	0.011104634	-2.65
FCN2	chr1	rs327406019	T	C	2.09E-005	0.011175555	-0.91
FCN2	chr1	rs345392108	A	T	2.09E-005	0.011175555	-0.90
PBD-2	chr15	rs322713647	G	A	2.12E-005	0.011197402	1.91
C9	chr16	rs331047160	G	A	2.19E-005	0.011487535	0.69
C9	chr16	rs319697453	T	C	2.22E-005	0.011586261	-0.69
C8A	chr6	rs320836863	C	T	2.28E-005	0.011787678	-0.66

PBD-2	chr15	rs319868268	T	C	2.49E-005	0.012838668	2.00
FCN2	chr1	rs338072079	T	C	2.54E-005	0.013006507	-0.65
FCN2	chr1	rs323352210	G	T	2.59E-005	0.01309166	-0.95
MBL2	chr14	rs324034606	C	T	2.61E-005	0.013122332	2.80
LGALS3	chr1	rs328549353	C	T	2.75E-005	0.01366744	0.89
SFTPD	chr14	rs694606761	C	A	2.74E-005	0.01366744	3.73
PBD-2	chr15	rs341333021	G	T	2.79E-005	0.013800539	2.03
C9	chr16	rs344415563	A	G	2.84E-005	0.013824916	-0.70
FCN2	chr1	rs322664568	G	A	2.84E-005	0.013824916	-0.93
CLEC5A	chr18	rs331422587	C	A	2.89E-005	0.013960679	1.00
CLEC5A	chr18	rs327744812	T	C	2.94E-005	0.014139286	0.92
FCN2	chr1	rs334121458	C	A	3.08E-005	0.014481671	-1.32
PBD-2	chr15	rs343692142	G	A	3.13E-005	0.014481671	1.87
TLR7	chrX	rs319920863	T	G	3.13E-005	0.014481671	0.63
TLR7	chrX	rs330843903	C	G	3.13E-005	0.014481671	0.63
TLR7	chrX	rs702287436	T	C	3.07E-005	0.014481671	0.51
TLR7	chrX	rs712876697	G	T	3.07E-005	0.014481671	0.51
MBL2	chr14	rs335832823	G	T	3.17E-005	0.014571371	-2.14
SFTPD	chr14	rs337855332	T	C	3.33E-005	0.015243841	3.70
CLEC5A	chr18	rs344720428	G	C	3.45E-005	0.015720622	1.09
C9	chr16	rs328409005	G	A	3.60E-005	0.016194145	-0.77
SFTPD	chr14	rs328692709	T	A	3.72E-005	0.016584931	2.66
CLEC5A	chr18	rs707996549	G	A	3.77E-005	0.016696096	1.02
CLEC5A	chr18	rs336598808	G	C	3.85E-005	0.016994498	0.98
C9	chr16	rs325953863	C	T	3.92E-005	0.017046829	-0.84
CLEC5A	chr18	rs343784954	A	C	3.89E-005	0.017046829	1.09
SFTPD	chr14	rs330536494	C	T	3.92E-005	0.017046829	3.71
FCN2	chr1	rs319451320	C	T	3.97E-005	0.01715165	-0.89
SFTPD	chr14	rs330919108	G	C	4.11E-005	0.017657687	2.91
SFTPD	chr14	rs340763735	G	A	4.18E-005	0.017793337	3.66
SFTPD	chr14	rs696949030	G	A	4.18E-005	0.017793337	3.66
CLEC5A	chr18	rs326196078	T	C	4.27E-005	0.01784124	1.03
CLEC5A	chr18	rs346369612	C	T	4.27E-005	0.01784124	1.03
FCN2	chr1	rs698904605	A	G	4.24E-005	0.01784124	-1.34
TLR7	chrX	rs330030893	T	G	4.23E-005	0.01784124	0.50
CLEC5A	chr18	rs343093524	T	C	4.30E-005	0.017853044	0.93
SFTPD	chr14	rs345019883	A	G	4.43E-005	0.018297144	3.66
TLR7	chrX	rs320984193	A	C	4.47E-005	0.018297144	0.46
TLR7	chrX	rs329133338	A	G	4.47E-005	0.018297144	0.46
CLEC5A	chr18	rs332647244	T	G	4.49E-005	0.018319732	1.06
CLEC5A	chr18	rs331312213	G	C	4.56E-005	0.01850682	0.99
CD55	chr9	rs334872391	G	C	4.60E-005	0.018591392	0.52
C9	chr16	rs324895466	A	G	4.81E-005	0.019111269	0.65

FCN2	chr1	rs328159775	A	G	4.89E-005	0.019111269	-0.86
MBL2	chr14	rs320465107	C	A	4.86E-005	0.019111269	-2.11
MBL2	chr14	rs340601674	A	G	4.86E-005	0.019111269	-2.11
SFTPD	chr14	rs338005155	G	A	4.78E-005	0.019111269	3.64
TLR7	chrX	rs339177737	G	C	4.88E-005	0.019111269	0.45
C9	chr16	rs331521202	A	G	5.55E-005	0.021309502	-0.74
TLR7	chrX	rs333909078	G	A	5.61E-005	0.021459108	0.48
C9	chr16	rs332895350	T	A	5.65E-005	0.021498612	0.72
C9	chr16	rs339532403	G	C	5.70E-005	0.021590381	0.67
C9	chr16	rs319520928	C	T	5.80E-005	0.021861128	0.64
CD55	chr9	rs331876371	C	T	5.96E-005	0.02237377	-0.48
TLR7	chrX	rs320821001	A	G	5.99E-005	0.022408912	0.52
C9	chr16	rs334173551	A	C	6.22E-005	0.023102105	-0.79
FCN2	chr1	rs324560299	G	T	6.24E-005	0.023102105	-0.86
FCN2	chr1	rs329763687	A	G	6.78E-005	0.024910042	-0.89
C9	chr16	rs334586317	C	T	6.98E-005	0.025555309	-0.66
CD55	chr9	rs339342977	C	A	7.29E-005	0.02657907	0.47
ITLN2	chr4	rs700070722	G	A	7.36E-005	0.026638718	2.73
MBL2	chr14	rs334156767	G	T	7.37E-005	0.026638718	3.30
CLEC5A	chr18	rs330294991	T	C	7.50E-005	0.026974724	1.03
CD55	chr9	rs330080617	T	C	7.53E-005	0.026991278	0.47
FCN2	chr1	rs344247372	C	T	7.56E-005	0.026991278	-0.90
PBD-2	chr15	rs345490517	C	T	7.64E-005	0.027148313	1.99
MBL2	chr14	rs345964266	C	T	7.70E-005	0.027190485	-2.07
TLR7	chrX	rs322452453	T	G	7.71E-005	0.027190485	0.52
FCN2	chr1	rs326256366	C	T	8.00E-005	0.028023583	-0.86
LGALS3	chr1	rs333750167	C	G	8.02E-005	0.028023583	0.83
FCN2	chr1	rs327402132	T	C	8.06E-005	0.028081332	-0.88
SCGB1A1	chr2	rs81227376	A	G	8.10E-005	0.0281002	1.91
FCN2	chr1	rs321609361	G	A	8.23E-005	0.028210128	-0.86
FCN2	chr1	rs337144001	G	A	8.23E-005	0.028210128	-0.86
FCN2	chr1	rs695704175	G	A	8.23E-005	0.028210128	-0.86
TLR7	chrX	rs334257861	C	T	8.28E-005	0.028255893	0.60
FCN2	chr1	rs695124986	T	C	8.32E-005	0.028301164	-0.80
CLEC5A	chr18	rs343548293	T	C	8.36E-005	0.028324226	1.04
SFTPD	chr14	rs45434493	G	A	8.56E-005	0.028748654	3.55
TLR7	chrX	rs326874199	A	G	8.73E-005	0.029217351	0.46
TLR7	chrX	rs339740820	C	A	8.95E-005	0.029837348	0.56
FCN2	chr1	rs328171340	T	C	9.14E-005	0.03022503	-0.88
FCN2	chr1	rs335824751	G	C	9.11E-005	0.03022503	-0.86
FCN2	chr1	rs333952728	T	C	9.18E-005	0.030247761	-0.86
FCN2	chr1	rs322403126	G	A	9.39E-005	0.030705366	-0.89
FCN2	chr1	rs331316020	C	A	9.39E-005	0.030705366	-0.85

PBD-2	chr15	rs319252499	A	G	9.62E-005	0.031326383	2.00
FCN2	chr1	rs344970007	A	T	9.97E-005	0.031988901	-0.87
ITLN2	chr4	rs698060414	C	G	9.92E-005	0.031988901	-3.15
TLR7	chrX	rs325640555	G	A	9.94E-005	0.031988901	0.54
SFTPD	chr14	rs713965714	C	G	0.000100103	0.032006381	3.64
CLEC5A	chr18	rs320201658	C	T	0.00010287	0.032648629	0.95
CFB	chrX	rs322378897	G	A	0.000104906	0.033172341	-0.35
FCN2	chr1	rs341954405	G	A	0.000107154	0.033635897	-0.85
TLR7	chrX	rs326226114	C	T	0.000107102	0.033635897	0.45
FCN2	chr1	rs339250981	G	A	0.000107606	0.033654947	-0.84
C9	chr16	rs328617983	A	G	0.000109207	0.033922668	0.64
TLR7	chrX	rs346239293	G	A	0.000109251	0.033922668	0.44
FCN2	chr1	rs328892780	C	G	0.000113937	0.035124164	-0.80
FCN2	chr1	rs334477584	C	T	0.000116011	0.035635646	-0.85
CFB	chrX	rs320914345	G	A	0.000116848	0.03565324	-0.36
LGALS3	chr1	rs342708446	C	T	0.000116897	0.03565324	0.85
FCN2	chr1	rs343420433	G	C	0.000118839	0.036117291	-0.88
FCN2	chr1	rs340612906	G	A	0.000122148	0.036862573	-0.87
NOD1	chr18	rs337167571	G	A	0.000126582	0.037903567	-0.71
CLEC5A	chr18	rs711332767	C	T	0.000129179	0.038312404	1.02
NOD1	chr18	rs343304937	G	C	0.00012913	0.038312404	-0.80
TLR7	chrX	rs320490535	A	C	0.000130598	0.038405738	0.43
TLR7	chrX	rs326190137	T	C	0.000130834	0.038405738	0.44
TLR7	chrX	rs333865165	G	T	0.00013071	0.038405738	0.48
FCN2	chr1	rs344296124	G	T	0.000132205	0.038555904	-0.85
LGALS3	chr1	rs318464617	G	A	0.000132242	0.038555904	1.14
C9	chr16	rs342277995	A	C	0.000132909	0.038619413	0.69
C9	chr16	rs335752545	G	A	0.000135544	0.038884826	0.62
CFB	chrX	rs344211622	T	C	0.000135631	0.038884826	-0.32
SFTPD	chr14	rs342243796	G	C	0.000135518	0.038884826	3.46
TLR7	chrX	rs333335126	A	G	0.000134524	0.038884826	0.44
C9	chr16	rs341753259	C	A	0.000140444	0.03999817	0.62
TLR7	chrX	rs337075708	G	T	0.000141441	0.040149273	0.44
CLEC5A	chr18	rs81305879	G	A	0.000142173	0.040224226	1.07
C9	chr16	rs343424082	A	T	0.000144961	0.040429553	0.60
FCN2	chr1	rs319704295	C	A	0.000145249	0.040429553	-0.85
FCN2	chr1	rs329339982	A	C	0.000144781	0.040429553	-0.83
FCN2	chr1	rs330995398	C	A	0.000145249	0.040429553	-0.85
FCN2	chr1	rs340411758	A	C	0.000144781	0.040429553	-0.83
PBD-2	chr15	rs344398416	G	C	0.000146208	0.040434676	1.52
PBD-2	chr15	rs329928999	G	C	0.00014952	0.041190662	1.80
TLR7	chrX	rs319364959	A	G	0.000149899	0.041190662	0.52
TLR7	chrX	rs345901448	T	C	0.000151342	0.041454742	0.51

FCN2	chr1	rs337576635	G	T	0.000152366	0.041602816	-0.84
CLEC5A	chr18	rs340069284	C	T	0.000155479	0.042039476	1.07
TLR7	chrX	rs334772554	G	T	0.000155044	0.042039476	0.44
TLR7	chrX	rs342600664	A	G	0.000155921	0.042039476	0.44
C9	chr16	rs339497134	T	C	0.000156721	0.042123061	-0.83
KLRK1	chr5	rs326639224	G	A	0.000157433	0.042182593	-0.57
TLR7	chrX	rs340395921	G	A	0.000159397	0.042444632	0.42
LGALS3	chr1	rs337354369	G	C	0.000161011	0.042610564	0.82
TLR7	chrX	rs342613370	C	T	0.000160994	0.042610564	0.59
C9	chr16	rs330521468	G	C	0.000165193	0.043551792	0.61
CD55	chr9	rs339116407	T	G	0.00016643	0.043551792	0.45
DEFB122	chr17	rs319850710	C	T	0.0001671	0.043551792	1.44
TLR7	chrX	rs328671703	A	G	0.000166283	0.043551792	0.44
TLR7	chrX	rs341031207	G	A	0.000166859	0.043551792	0.53
CD55	chr9	rs333679324	C	T	0.000169303	0.043992779	0.48
SELL	chr4	rs343628241	C	T	0.000169839	0.04399903	-1.83
C9	chr16	rs327241196	T	G	0.000172292	0.044367195	0.58
ZBP1	chr17	rs328726867	C	A	0.000174565	0.044818383	-2.36
FCN2	chr1	rs335102706	G	A	0.000177604	0.045327947	-0.88
C9	chr16	rs345775384	C	A	0.000179421	0.045656306	0.69
C9	chr16	rs344037272	A	G	0.000180133	0.045702225	-0.71
TLR7	chrX	rs81294185	C	T	0.000185369	0.046892309	0.45
FCN2	chr1	rs339469390	C	T	0.000186535	0.047049054	-1.37
SFTPD	chr14	rs327156991	T	C	0.000189115	0.047560095	2.07
TLR8	chrX	rs334862436	A	C	0.000191603	0.048045467	-0.81
C9	chr16	rs321093005	T	C	0.000195611	0.04848504	0.69
LGALS3	chr1	rs326353586	G	A	0.000195505	0.04848504	0.78
C9	chr16	rs333628473	A	G	0.000197804	0.048596186	-0.66
C9	chr16	rs333672557	T	G	0.000196789	0.048596186	0.60
ITLN2	chr4	rs690716289	A	G	0.000199982	0.048596186	2.61
LGALS3	chr1	rs330720145	C	T	0.000200015	0.048596186	1.16
LGALS3	chr1	rs342885446	T	C	0.000200015	0.048596186	1.16
TLR7	chrX	rs324124940	G	T	0.000199699	0.048596186	0.45
C9	chr16	rs324563349	C	T	0.000203342	0.048742571	0.61
CLEC5A	chr18	rs339063186	C	T	0.00020211	0.048742571	1.07
FCN2	chr1	rs323814354	T	C	0.000202238	0.048742571	-2.80
FCN2	chr1	rs339179546	G	A	0.000203889	0.048742571	-0.97
FCN2	chr1	rs341236626	T	C	0.000203042	0.048742571	-1.04
FCN2	chr1	rs341477612	T	A	0.000205291	0.048880938	-0.86
FCN2	chr1	rs341485935	G	T	0.000207756	0.049090242	-1.36
FCN2	chr1	rs324618153	A	C	0.000207548	0.049090242	-0.90
PBD-2	chr15	rs328487402	A	T	0.00020955	0.049243592	1.65
PBD-2	chr15	rs335696039	C	T	0.00020955	0.049243592	1.65

FCN2	chr1	rs331664783	G	A	0.000212121	0.049576941	-0.92
TLR7	chrX	rs318720124	T	C	0.000211555	0.049576941	0.44

*The p-value was adjusted using the Benjamini-Hochberg procedure with an FDR = 0.05.

Appendix 4.4. List of eQTLs that were genotyped in healthy and diseased pigs.

Gene	rsID	Chr	Position	Ref	Alt	Location relative to impacted gene	Distance relative to impacted gene
LGALS3	rs341181428	1	204,916,854	G	A	upstream	39,777
LGALS3	rs342885446	1	204,927,302	T	C	upstream	29,329
LGALS3	rs81246297	1	204,934,912	A	G	upstream	21,719
LGALS3	rs322326231	1	204,935,789	A	T	upstream	20,842
LGALS3	rs323704144	1	204,940,667	G	A	upstream	15,964
LGALS3	rs326353586	1	204,949,440	G	A	upstream	7191
FCN2	rs339469390	1	308,176,244	C	T	upstream	20,857
FCN2	rs333222079	1	308,176,460	C	G	upstream	20,641
FCN2	rs342490172	1	308,185,559	T	C	upstream	11,542
FCN2	rs338072079	1	308,189,467	T	C	upstream	7634
FCN2	rs326158227	1	308,189,511	A	C	upstream	7590
FCN2	rs329763687	1	308,189,671	A	G	upstream	7430
FCN2	rs344970007	1	308,190,003	A	T	upstream	7098
FCN2	rs322403126	1	308,190,076	G	A	upstream	7025
FCN2	rs320342887	1	308,190,797	A	G	upstream	6304
FCN2	rs324618153	1	308,192,981	A	C	upstream	4120
FCN2	rs342195622	1	308,205,698	G	A	downstream	-2619
FCN2	rs328892780	1	308,210,356	C	G	downstream	-7277
FCN2	rs331664783	1	308,211,156	G	A	downstream	-8077
FCN2	rs333340587	1	308,213,409	G	A	downstream	-10,330
FCN2	rs337034769	1	308,225,195	T	C	downstream	-22,116
FCN2	rs330321112	1	308,225,322	A	G	downstream	-22,243
FCN2	rs322664568	1	308,226,384	G	A	downstream	-23,305
FCN2	rs335102706	1	308,228,389	G	A	downstream	-25,310
SCGB1A1	rs81227376	2	8,609,365	A	G	downstream	-10,601
SELL	rs343628241	4	88,904,530	C	T	upstream	26,948
ITLN2	rs690716289	4	97,609,194	A	G	downstream	-943
ITLN2	rs698060414	4	97,609,676	C	G	downstream	-461
C8A	rs320836863	6	143,270,088	C	T	upstream	3478
CD55	rs330938856	9	74,266,167	T	C	upstream	47,618
CD55	rs322614244	9	74,276,435	C	T	upstream	37,350
CD55	rs333679324	9	74,292,150	C	T	upstream	21,635
SFTPD	rs319870426	14	88,622,626	A	G	upstream	37,678
SFTPD	rs319707230	14	88,635,598	G	C	upstream	24,706
SFTPD	rs328692709	14	88,653,734	T	A	upstream	6570
SFTPD	rs327156991	14	88,694,407	T	C	downstream	-19,384

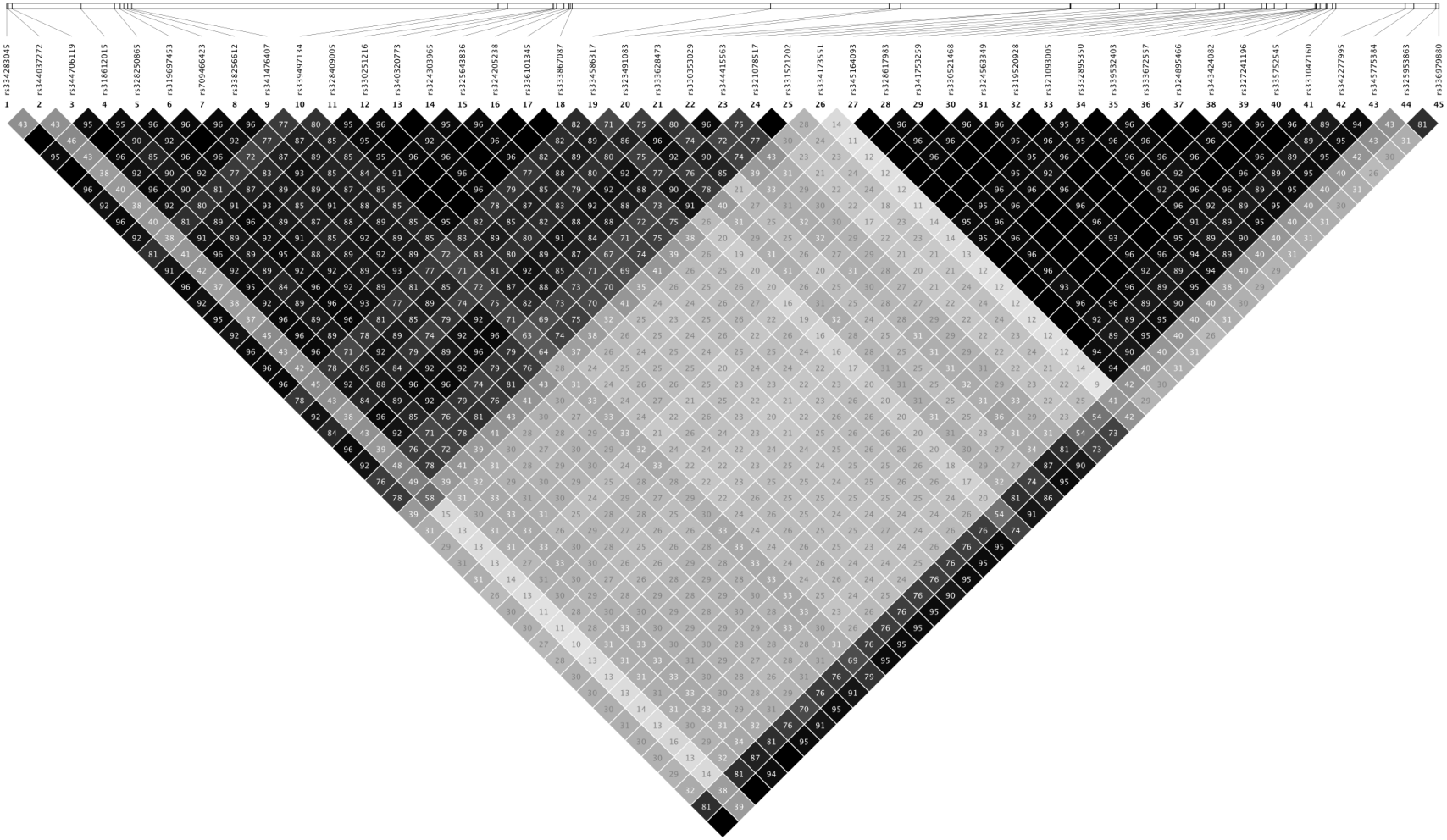
MBL2	rs344636925	14	105,584,184	G	A	upstream	47,085
MBL2	rs328673019	14	105,585,994	T	A	upstream	45,275
MBL2	rs327359428	14	105,586,853	C	A	upstream	44,416
MBL2	rs334156767	14	105,611,145	G	T	upstream	20,124
MBL2	rs318652830	14	105,636,724	C	A	downstream	-294
MBL2	rs341757289	14	105,638,542	C	T	downstream	-2112
DEFB1	rs330573086	15	43,309,656	C	T	upstream	36,662
DEFB1	rs319252499	15	43,315,923	A	G	upstream	30,395
DEFB1	rs335696039	15	43,356,023	C	T	intragenic	
C9	rs344037272	16	25,751,067	A	G	downstream	-49,154
C9	rs328250865	16	25,756,412	A	G	downstream	-43,809
C9	rs319697453	16	25,756,694	T	C	downstream	-43,527
C9	rs341476407	16	25,757,280	G	A	downstream	-42,941
C9	rs339497134	16	25,775,724	T	C	downstream	-24,497
C9	rs325643836	16	25,779,029	A	C	downstream	-21,192
C9	rs334586317	16	25,789,438	C	T	downstream	-10,783
C9	rs333628473	16	25,795,977	A	G	downstream	-4244
C9	rs330353029	16	25,804,503	C	T	intragenic	
C9	rs335752545	16	25,817,714	G	A	intragenic	
C9	rs342277995	16	25,821,376	A	C	intragenic	
C9	rs325953863	16	25,822,913	C	T	intragenic	
C9	rs336979880	16	25,823,068	T	C	intragenic	
ZBP1	rs328726867	17	65,086,594	C	A	downstream	-49,942
CLEC5A	rs343093524	18	8,348,273	T	C	upstream	30,826
CLEC5A	rs327744812	18	8,350,992	T	C	upstream	28,107
CLEC5A	rs343548293	18	8,352,119	T	C	upstream	26,980
CLEC5A	rs346369612	18	8,355,766	C	T	upstream	23,333
CLEC5A	rs339063186	18	8,357,288	C	T	upstream	21,811
CLEC5A	rs336562067	18	8,372,441	G	A	upstream	6658
NOD1	rs343304937	18	47,009,993	G	C	intragenic	
TLR7	rs319364959	23	10,430,593	A	G	upstream	20,066
TLR7	rs330843903	23	10,462,037	C	G	intragenic	
TLR7	rs344488277	23	10,477,315	C	T	downstream	-660
TLR7	rs334257861	23	10,483,622	C	T	downstream	-6967
TLR7	rs318825117	23	10,485,866	T	C	downstream	-9211
TLR7	rs320984193	23	10,485,945	A	C	downstream	-9290
CFB	rs320914345	23	47,277,665	G	A	upstream	17,370
CFB	rs344211622	23	47,314,430	T	C	downstream	-2284

Healthy (n = 592) and diseased (n = 421) pigs were genotyped. The distance relative to the impacted gene is given based on the distance from the annotated start of the gene (if upstream) or the annotated end of the gene (if downstream)

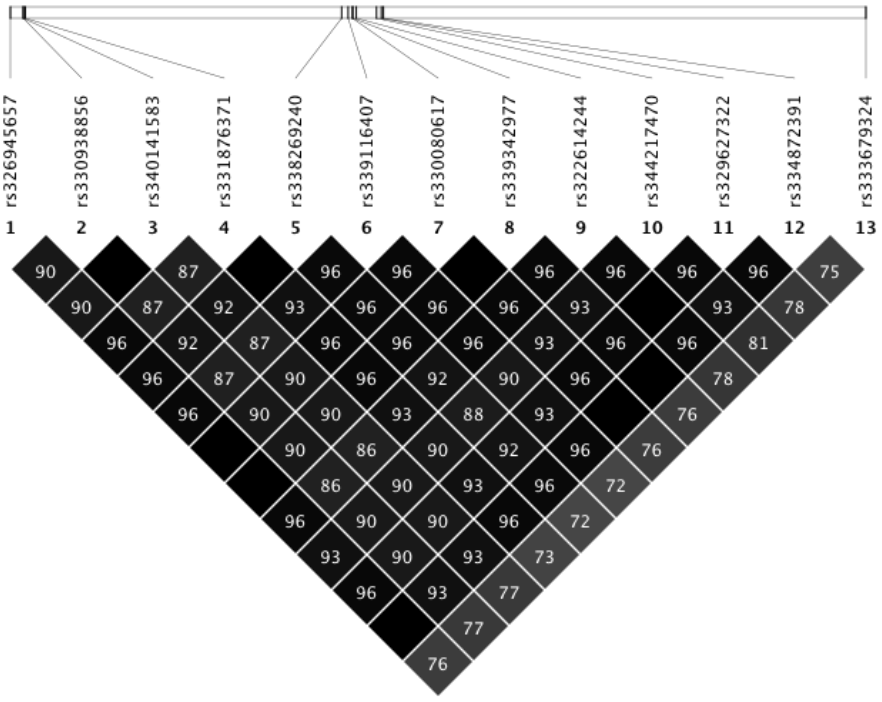
Appendix 4.5. Linkage disequilibrium analysis for significant eQTLs within 10 of 19 innate immune genes.

Note: The relative position of a SNV (identified by rsID) is illustrated in the bar at the top. The numbers within boxes represent the r^2 value between two SNVs. Empty black boxes indicate an r^2 of 100.

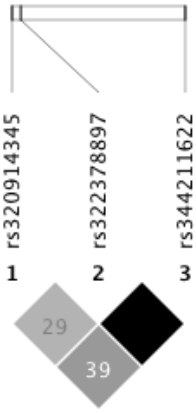
a) *C9A*



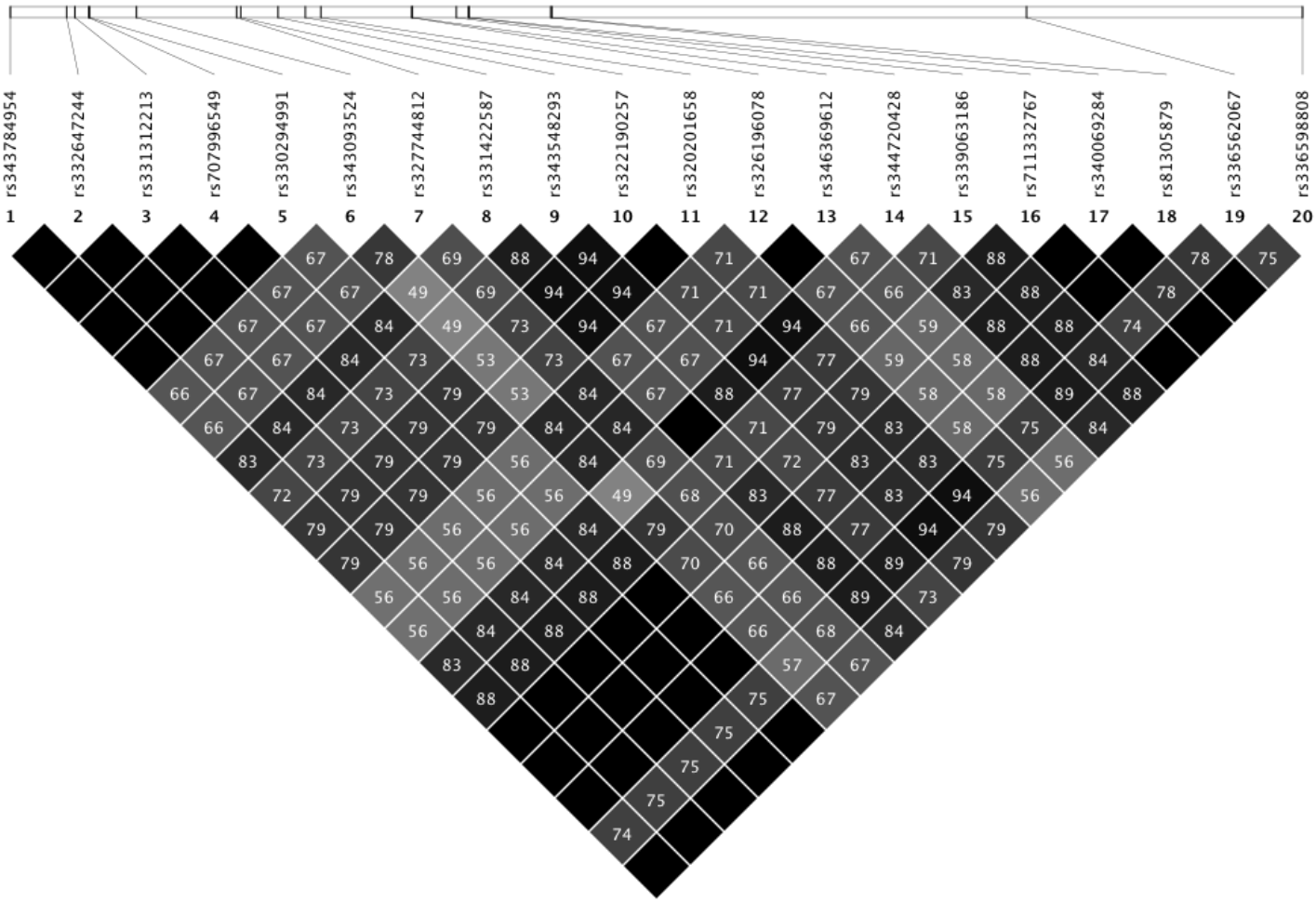
b) *CD55*


rs326945657
rs330938856
rs340141583
rs331876371
rs338269240
rs339116407
rs330080617
rs339342977
rs322614244
rs344217470
rs329627322
rs334872391
rs333679324
1 2 3 4 5 6 7 8 9 10 11 12 13
90
90 87
87 92 93
96 96 96 96
96 96 96 96 93
96 96 96 96 93 96
96 96 96 96 93 96 96
96 87 90 96 92 90 96 78
96 90 90 93 88 93 76
90 86 90 92 96 76
86 90 93 96 72
96 90 90 96 72
93 90 93 73
96 93 77
77
76
75
78
81

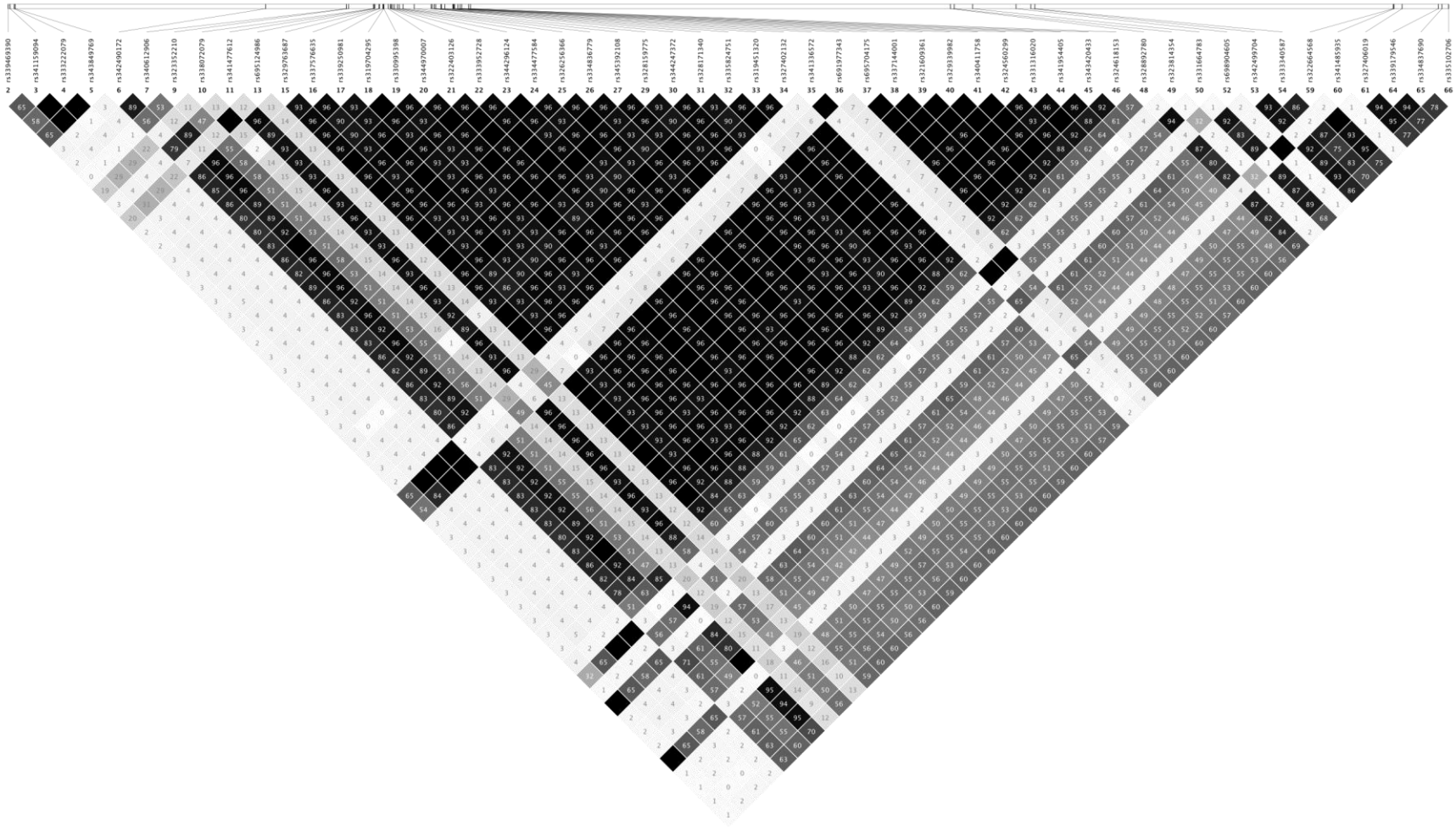
c) *CFB*


rs320914345
rs322378897
rs344211622
1 2 3
29
39

d) *CLEC5A*

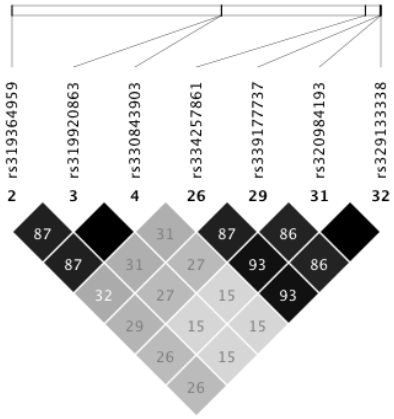

rs343784954
rs332647244
rs331312213
rs707996549
rs330294991
rs343093524
rs327744812
rs331422587
rs343548293
rs322190257
rs320201658
rs326196078
rs346369612
rs344720428
rs339063186
rs711332767
rs340069284
rs81305879
rs336562067
rs336598808
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
67 78 69 88 94 71 67 71 88 78 75
67 67 49 69 94 94 71 71 67 66 83 88 78
67 67 84 49 73 94 67 71 94 66 59 88 88 74
67 67 84 73 53 73 67 67 94 77 59 58 88 84
66 67 84 73 79 53 84 67 88 77 79 58 58 89 88
66 84 73 79 79 84 84 71 79 83 58 75 84
83 73 79 79 56 84 69 71 72 83 83 75 56
72 79 79 56 56 49 68 83 77 83 94 56
79 79 56 56 84 79 70 88 77 94 79
79 56 56 84 88 70 66 88 89 79
56 56 84 88 66 66 89 73
56 84 88 66 68 84
83 88 57 67
88 75 67
75
75
75
74

e) *FCN2*



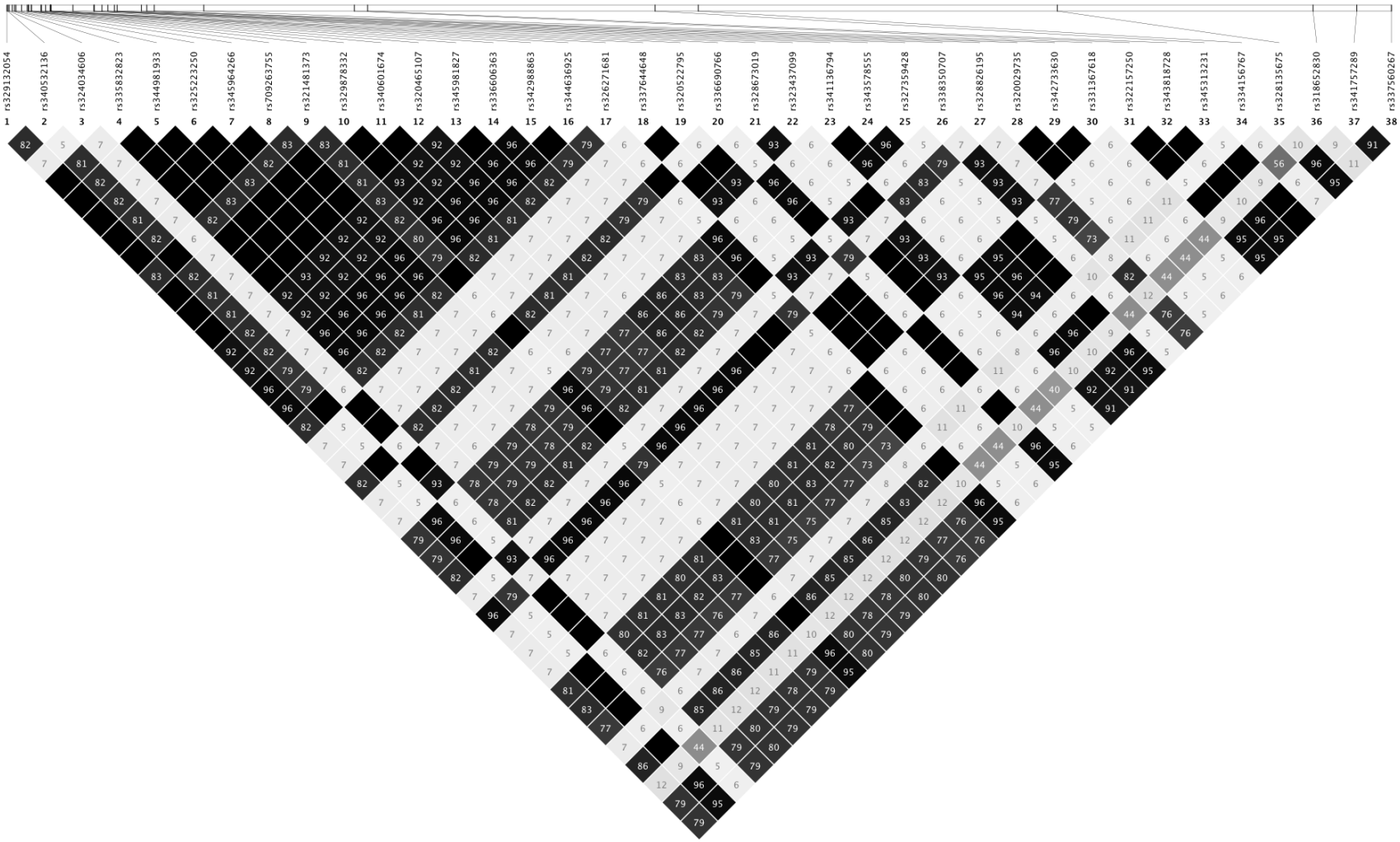
f) *ITLN2*

g) *TLR7*

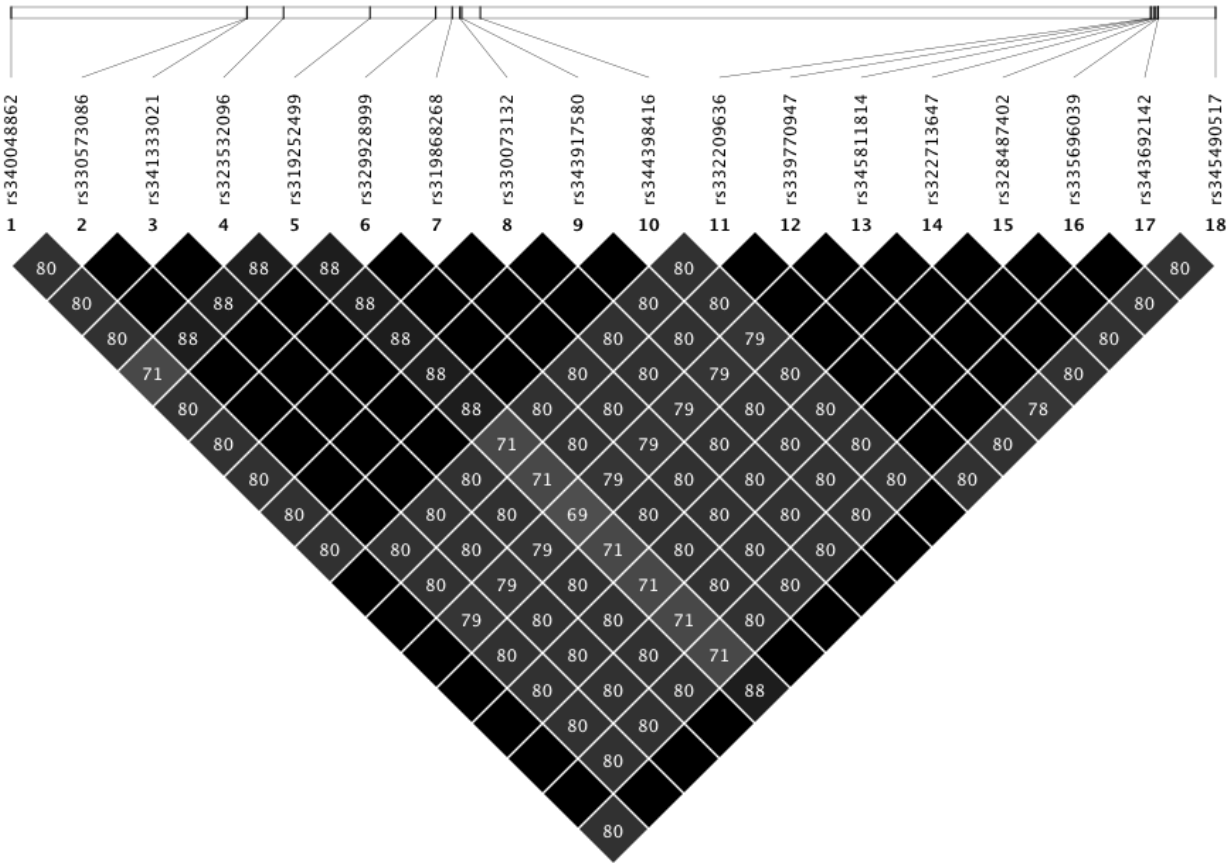


h) *NOD1*

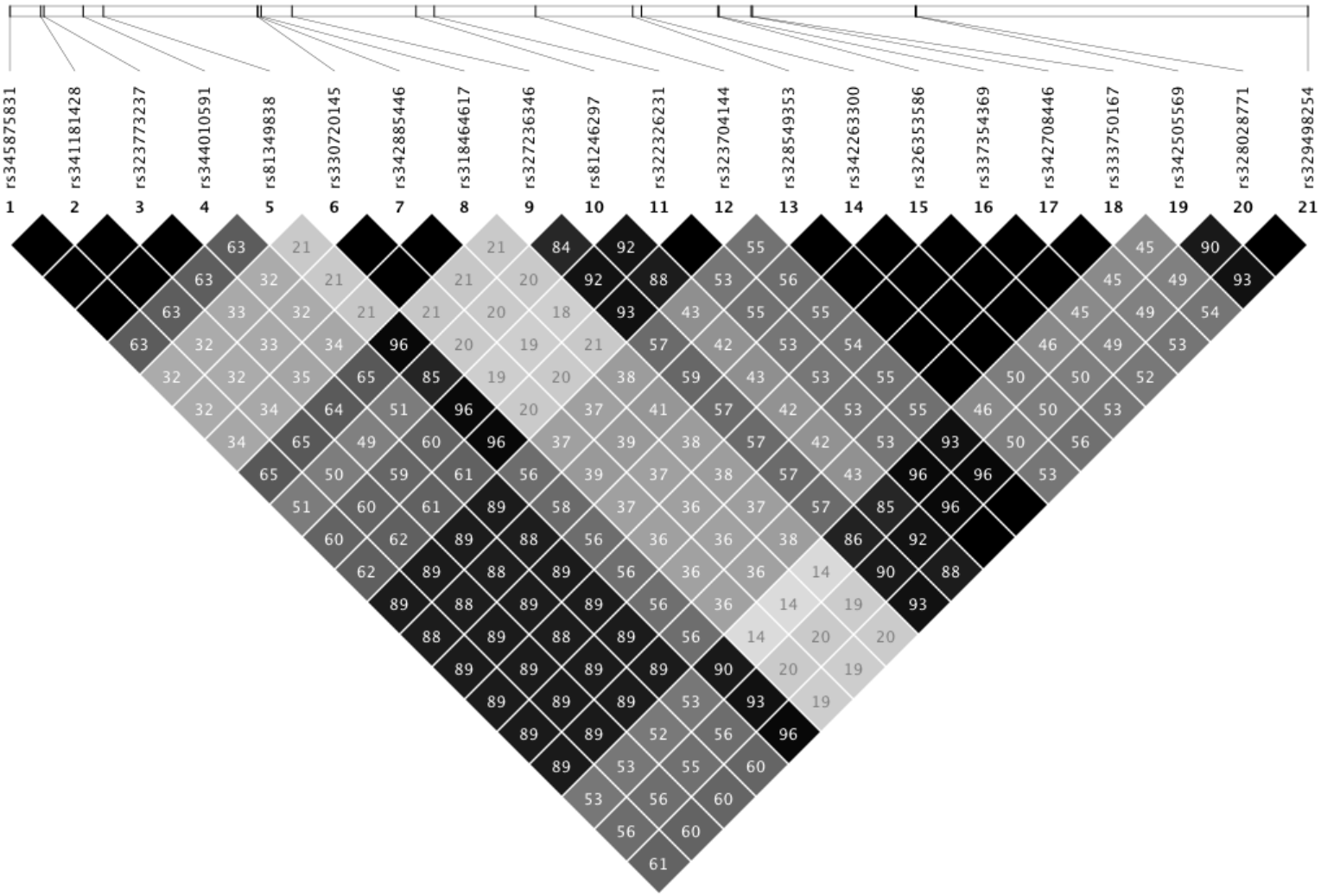
i) *MBL2*


rs329132054
rs340532136
rs324034606
rs335832823
rs344981933
rs325223250
rs345964266
rs709263755
rs321481373
rs329878332
rs340601674
rs320465107
rs345981827
rs336606363
rs342988863
rs344636925
rs326271681
rs337644648
rs320522795
rs336690766
rs328673019
rs323437099
rs341136794
rs343578555
rs327359428
rs338350707
rs328826195
rs320029735
rs342733630
rs331367618
rs322157250
rs343818728
rs345313231
rs334156767
rs328135675
rs318652830
rs341757289
rs337560267
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38

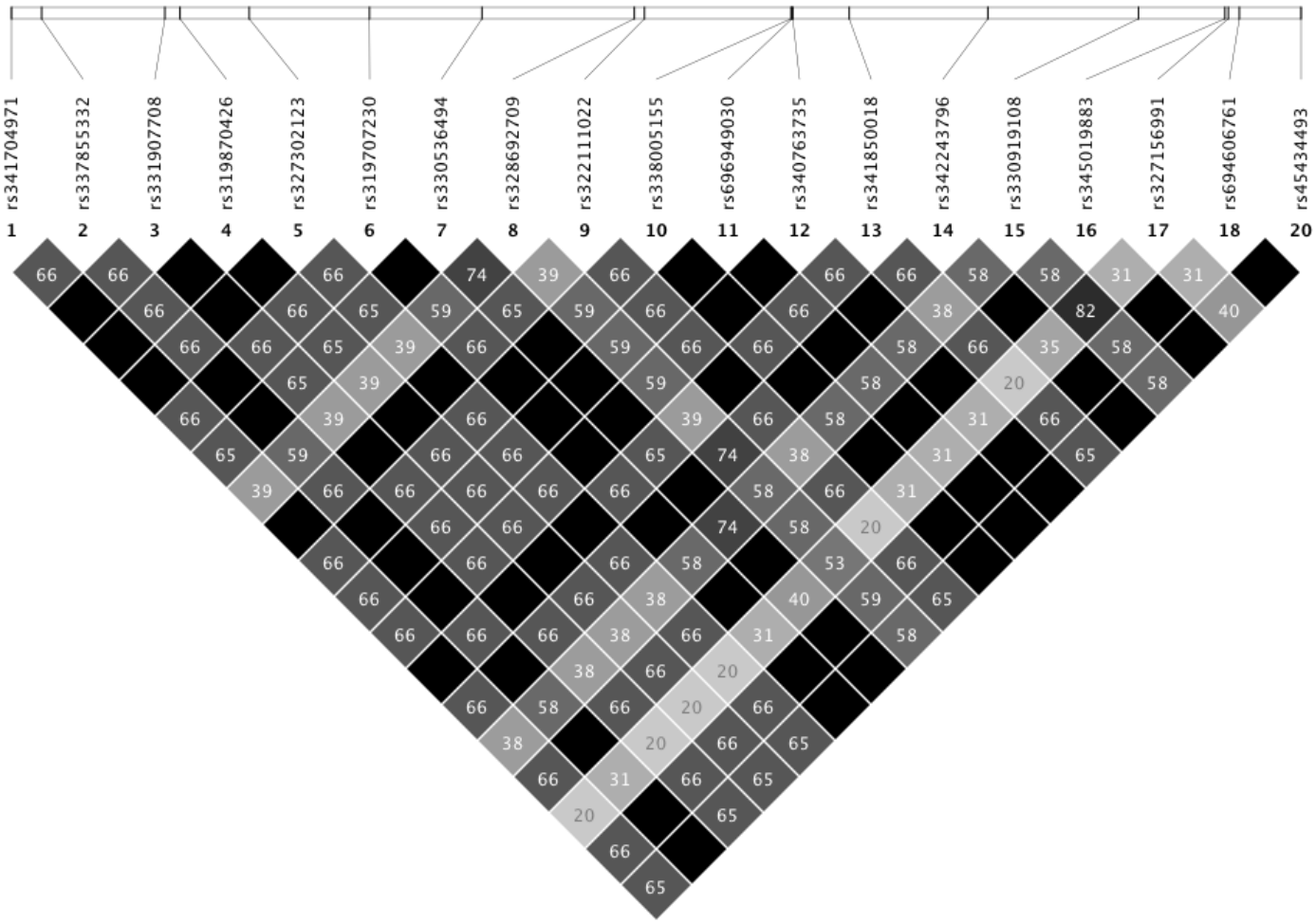
j) *PBD-2 (DEFB2)*


rs340048862
rs330573086
rs341333021
rs323532096
rs319252499
rs329928999
rs319868268
rs330073132
rs343917580
rs344398416
rs332209636
rs339770947
rs345811814
rs322713647
rs328487402
rs335696039
rs343692142
rs345490517
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

k) *LGALS3*


rs345875831
rs341181428
rs323773237
rs344010591
rs81349838
rs330720145
rs342885446
rs318464617
rs327236346
rs81246297
rs322326231
rs323704144
rs328549353
rs342263300
rs326353586
rs337354369
rs342708446
rs333750167
rs342505569
rs328028771
rs329498254
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

l) *SFTPD*


rs341704971
rs337855332
rs331907708
rs319870426
rs327302123
rs319707230
rs330536494
rs328692709
rs322111022
rs338005155
rs696949030
rs340763735
rs341850018
rs342243796
rs330919108
rs345019883
rs327156991
rs694606761
rs45434493
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20

Appendix 4.6. Porcine SNVs selected for genotyping are shown with the IDs of SNVs determined to be in LD.

Gene	SNV rsID	Linked SNV	Total # of Linked SNVs	Genotyped	Failed/Omitted
C8A	rs320836863	-	0	Yes	-
C9	rs335752545	rs332895350, rs343424082, rs324563349, rs341753259, rs339532403, rs319520928, rs333672557, rs345164093, rs327241196, rs328617983, rs330521468, rs321093005, rs324895466, rs331047160, rs345775384	15	Yes	-
C9	rs336979880	rs340320773, rs328409005	2	Yes	-
C9	rs319697453	rs709466423, rs341476407, rs319697453	3	Yes	-
C9	rs321078517	rs321078517	1	No	unresolved cluster
C9	rs325643836	-	0	Yes	-
C9	rs328250865	rs330251216, rs336101345, rs324303965, rs333867087, rs318612015, rs344706119, rs324205238, rs328250865	8	Yes	-
C9	rs330353029	rs330353029, rs344415563	2	Yes	-
C9	rs344037272	-	0	Yes	-
C9	rs342277995	-	0	Yes	-
C9	rs339497134	-	0	Yes	-
C9	rs333628473	-	0	Yes	-
C9	rs334586317	-	0	Yes	-
C9	rs325953863	-	0	Yes	-
C9	rs341476407	-	0	Yes	-
CD55	rs330938856	rs340141583	1	Yes	-
CD55	rs333679324	-	0	Yes	-
CD55	rs322614244	rs344217470, rs329627322, rs339116407	3	Yes	-
CFB	rs344211622	rs344211622	1	Yes	-
CFB	rs320914345	-	0	Yes	-
CLEC5A	rs322190257	rs320201658	1	No	poor extension
CLEC5A	rs346369612	rs346369612	1	Yes	-
CLEC5A	rs339063186	-	0	Yes	-

CLEC5A	rs343093524	-	0	Yes	-
CLEC5A	rs327744812	-	0	Yes	-
CLEC5A	rs336562067	-	0	Yes	-
CLEC5A	rs343548293	-	0	Yes	-
DEFB122	rs319850710	-	0	No	unresolved cluster
FCN2	rs329763687	rs334836779, rs324560299, rs319451320	3	Yes	-
FCN2	rs333222079	rs323814354, rs341336572, rs333222079, rs341159094, rs691977343	5	Yes	-
FCN2	rs322664568	rs327406019, rs322664568, rs334837690	3	Yes	-
FCN2	rs339469390	rs341485935, rs698904605	2	Yes	-
FCN2	rs338072079	rs338072079	1	Yes	-
FCN2	rs342499704	-	0	No	unresolved cluster
FCN2	rs342490172	-	0	Yes	-
FCN2	rs324618153	-	0	Yes	-
FCN2	rs328892780	-	0	Yes	-
FCN2	rs322403126	rs334477584, rs344296124, rs322403126, rs341954405, rs328159775, rs340612906, rs337576635	7	Yes	-
FCN2	rs344970007	rs333952728, rs344970007, rs345392108, rs329339982, rs340411758, rs330995398, rs695704175, rs335824751, rs328171340, rs327402132, rs321609361, rs337144001, rs326256366, rs343420433, rs331316020, rs319704295, rs341477612	17	Yes	-
FCN2	rs331664783	-	0	Yes	-
FCN2	rs335102706	-	0	Yes	-
FCN2	rs333340587	-	0	Yes	-
FCN2	rs339179546	-	0	No	unresolved cluster
FCN2	rs320342887	-	0	Yes	-
FCN2	rs326158227	-	0	Yes	-
FCN2	rs330321112	-	0	Yes	-
FCN2	rs337034769	-	0	Yes	-
FCN2	rs342195622	-	0	Yes	-

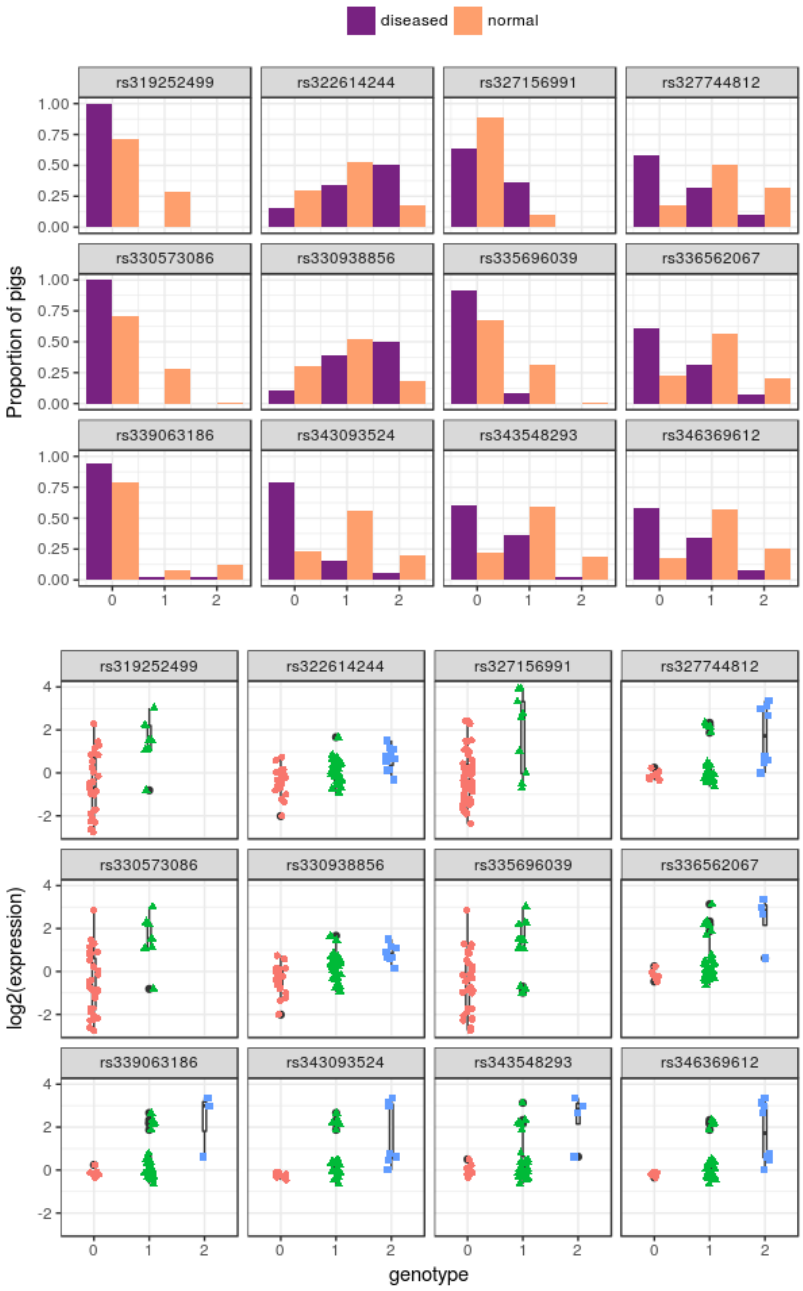
FCN2	rs318962795	-	0	No	unresolved cluster
ITLN2	rs690716289	-	0	Yes	-
ITLN2	rs698060414	-	0	Yes	-
ITLN2	rs700070722	-	0	No	unresolved cluster
KLRK1	rs326639224	-	0	No	poor extension
LGALS3	rs322326231	rs329498254, rs342505569, rs322326231, rs323704144	4	Yes	-
LGALS3	rs326353586	rs333750167, rs326353586, rs342263300, rs342708446, rs337354369	5	Yes	-
LGALS3	rs341181428	rs345875831, rs344010591, rs323773237	3	Yes	-
LGALS3	rs342885446	rs330720145, rs342885446	2	Yes	-
LGALS3	rs81246297	-	0	Yes	-
LGALS3	rs323704144	-	0	Yes	-
MBL2	rs344636925	rs325223250, rs345981827, rs709263755, rs320465107, rs329878332, rs336606363, rs340601674, rs345964266, rs328826195, rs329132054, rs342988863, rs335832823, rs344981933	13	Yes	-
MBL2	rs327359428	rs345313231, rs343578555, rs341757289, rs341136794, rs343818728, rs326271681, rs340532136, rs336690766, rs321481373, rs322157250, rs327359428, rs328135675	12	Yes	-
MBL2	rs328673019	rs320029735, rs337644648, rs342733630, rs320522795, rs324034606, rs328673019	6	Yes	-
MBL2	rs323437099	rs323437099	1	No	unresolved cluster
MBL2	rs334156767	-	0	Yes	-
MBL2	rs318652830	-	0	Yes	-
MBL2	rs341757289	-	0	Yes	-
NOD1	rs343304937	-	0	Yes	-
PBD-2	rs330573086	rs344398416, rs330073132, rs323532096, rs343917580, rs329928999, rs319868268, rs345490517, rs330573086	8	Yes	-
PBD-2	rs335696039	rs339770947, rs332209636, rs345811814, rs343692142, rs328487402, rs322713647, rs335696039	7	Yes	-

PBD-2	rs319252499	-	0	Yes	-
SCGB1A1	rs81227376	-	0	Yes	-
SELL	rs343628241	-	0	Yes	-
SFTPD	rs319870426	rs319870426, rs341850018, rs331907708, rs327302123, rs341704971	5	Yes	-
SFTPD	rs328692709	-	0	Yes	-
SFTPD	rs327156991	-	0	Yes	-
SFTPD	rs330919108	-	0	No	unresolved cluster
SFTPD	rs319707230	-	0	Yes	-
TLR7	rs330843903	rs330843903	1	Yes	-
TLR7	rs320984193	rs320984193	1	Yes	-
TLR7	rs334257861	-	0	Yes	-
TLR7	rs339177737	-	0	No	unresolved cluster
TLR7	rs319364959	-	0	Yes	-
TLR7	rs318825117	-	0	Yes	-
TLR7	rs344488277	-	0	Yes	-
TLR8	rs334862436	-	0	No	unresolved cluster
ZBP1	rs328726867	-	0	Yes	-

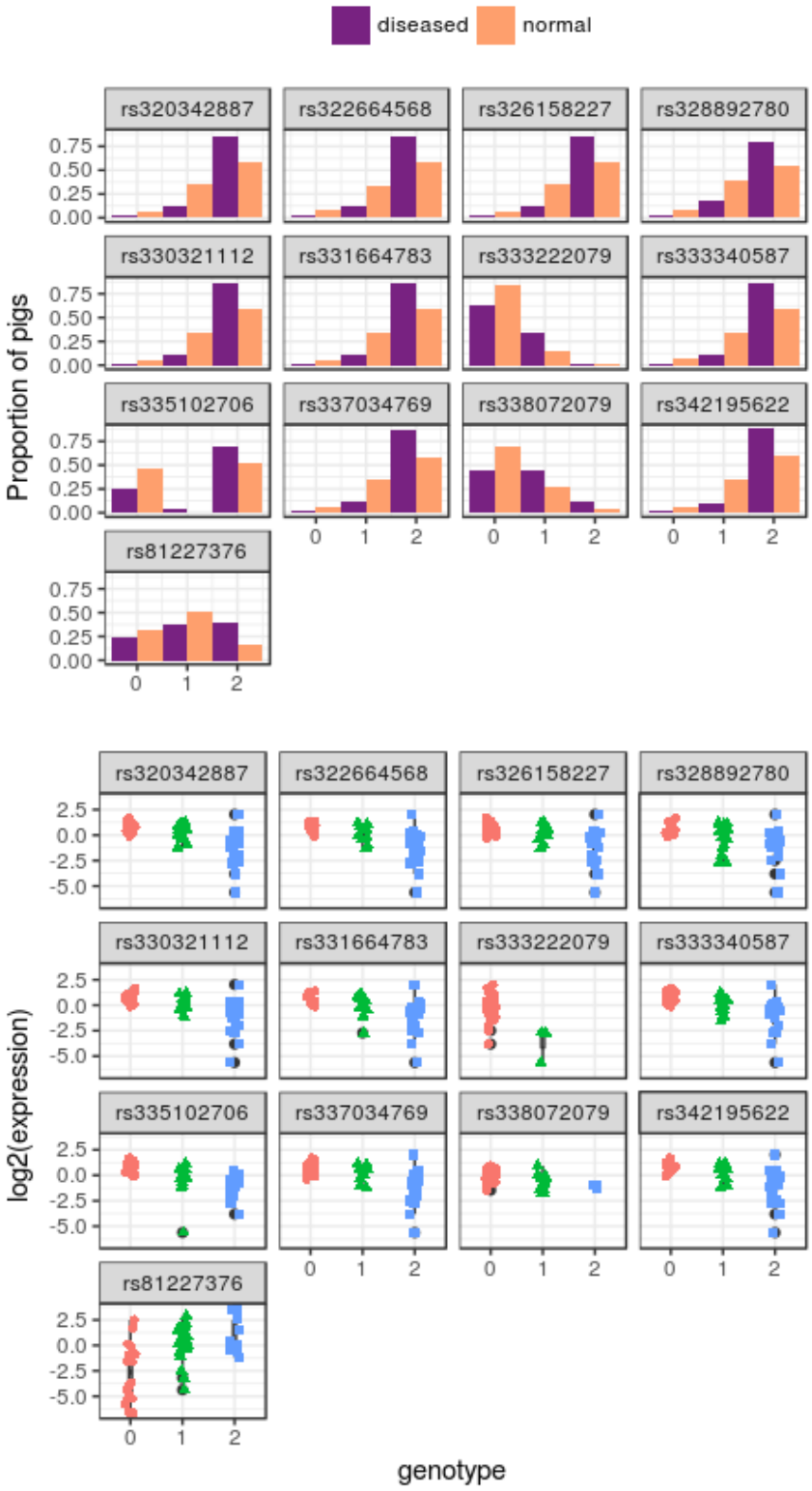
Appendix 4.7. Allele density and corresponding eQTL profiles of the 28 eQTL associated with infectious disease.

Each figure consists of two sets of graphs: a bar graph on the top, representing the proportion of alleles within pigs diagnosed with a specific pathogen or condition (diseased, purple bars) or within healthy pigs (normal, orange bars). The bottom figure represents eQTL data from the same SNV(s), illustrating the association of SNV(s) with gene expression. For both graphs, 0 = homozygous reference, 1 = heterozygous, 2 = homozygous alternate

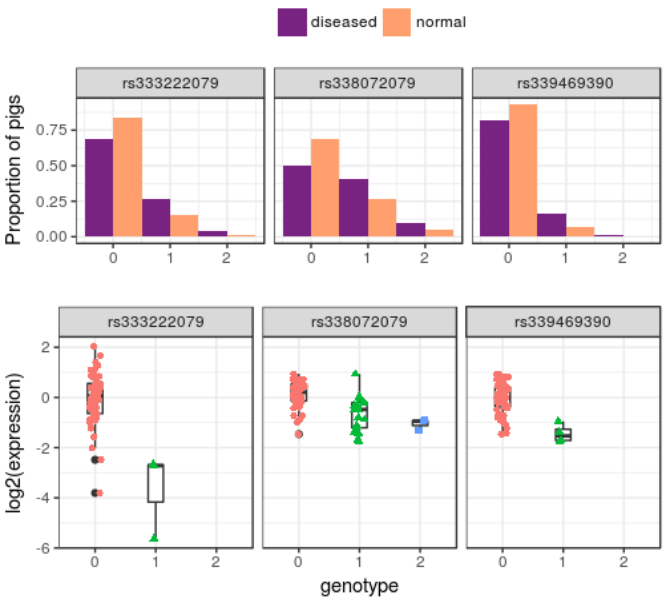
a) *E. coli*



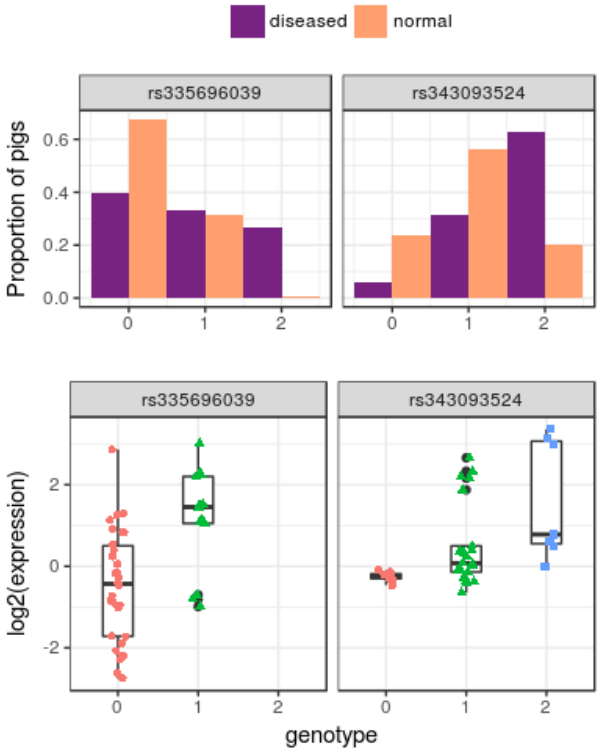
b) *Mycoplasma*



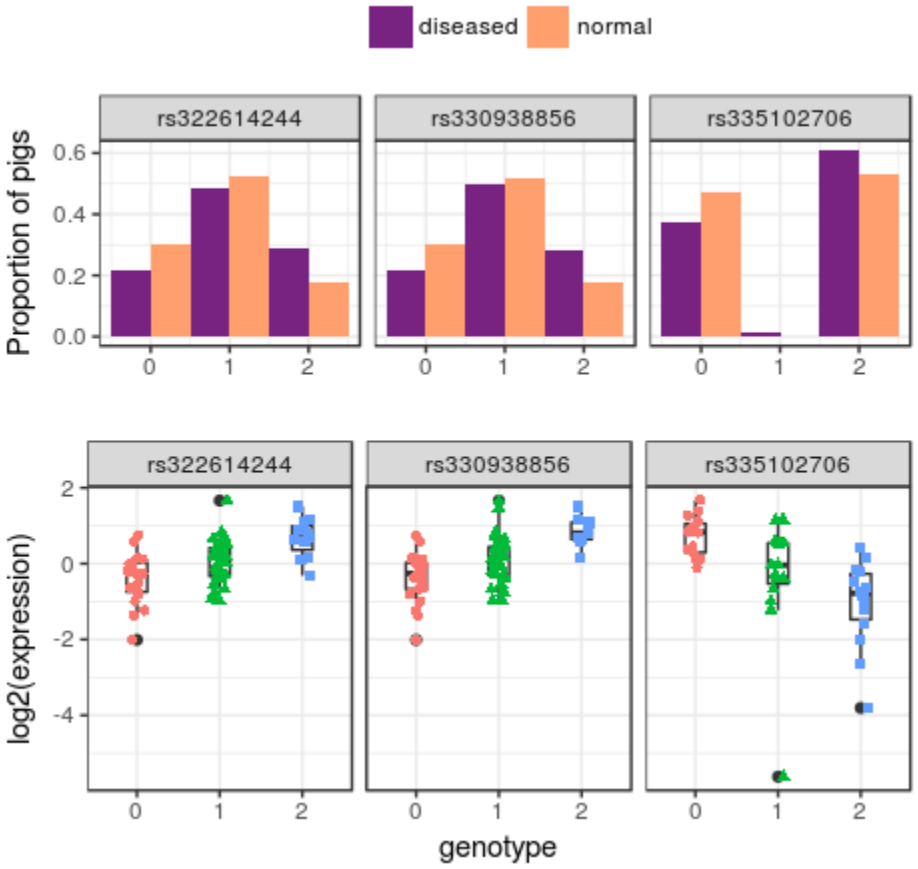
c) Porcine reproductive and respiratory syndrome virus



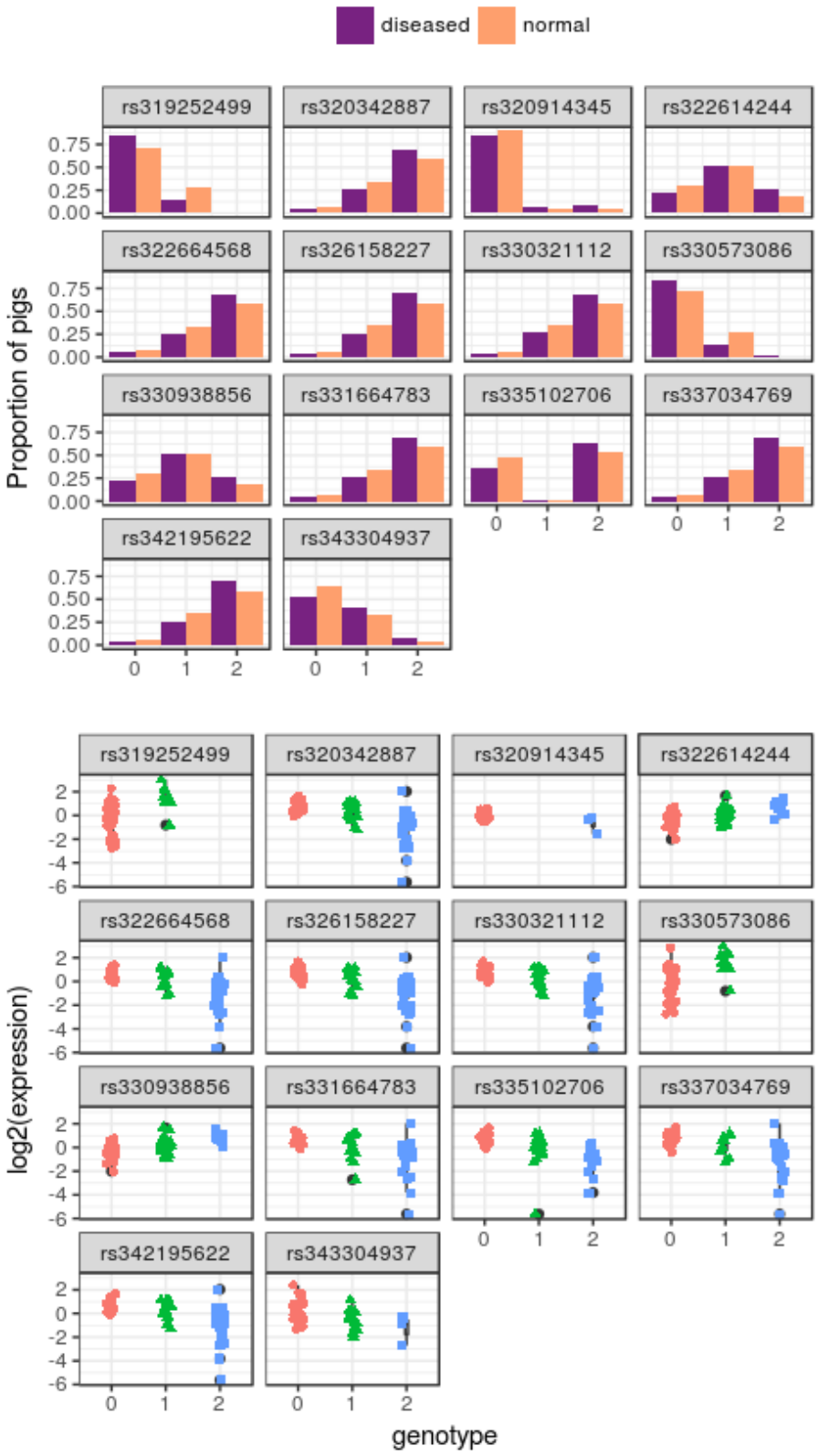
d) Influenza A virus



e) Enteritis


diseased
normal
rs322614244
rs330938856
rs335102706
Proportion of pigs
0.6
0.4
0.2
0.0
0
1
2
log2(expression)
2
0
-2
-4
genotype

f) Pneumonia


diseased
normal
rs319252499
rs320342887
rs320914345
rs322614244
rs322664568
rs326158227
rs330321112
rs330573086
rs330938856
rs331664783
rs335102706
rs337034769
rs342195622
rs343304937
Proportion of pigs
0.75
0.50
0.25
0.00
0
1
2
log2(expression)
2
0
-2
-4
-6
genotype

g) Septicemia