Enabling systems-level analyses of the host response to infectious diseases in bovine and other mammalian species

by

Amir Bahram Khosravizadeh Foroushani

M.Sc. (Computer Science), University of Texas at Dallas, 2009
Diplom Informatik, University of Tuebingen, 2003

Thesis Submitted In Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in the

Department of Molecular Biology and Biochemistry

Faculty of Science

SIMON FRASER UNIVERSITY

Summer 2014

Approval

Name: **Amir Bahram Khosravizadeh Foroushani**

Degree: **Doctor of Philosophy**

Title of Thesis: ***Enabling systems-level analyses of the host response to infectious diseases in bovine and other mammalian species***

Examining Committee: **Chair: Dr. Mark A. Brockman**
Associate Professor, Faculty of Health Sciences // Dept of Molecular Biology and Biochemistry

Dr. Fiona S.L. Brinkman
Senior Supervisor
Professor, Department of Molecular Biology and Biochemistry

Dr. David J. Lynn
Supervisor
Group leader, Computational Biology, AGRIC,Teagasc, Ireland
By video conference (Adelaide, Australia)

Dr. Jack Chen
Supervisor
Professor, Department of Molecular Biology and Biochemistry

Dr. Willie Davidson
Supervisor
Professor, Department of Molecular Biology and Biochemistry

Dr. Lisa Craig
Internal Examiner
Associate Professor, Department of Molecular Biology and Biochemistry

Dr. Paul Pavlidis
External Examiner
Associate Professor, *Psychiatry*
University of British Columbia

Date Defended/Approved: September 29, 2014

Partial Copyright Licence

Abstract

The innate immune response is a critical branch of immunity, providing a first line of defense against pathogens and shaping subsequent adaptive immune responses. The complexity of this system necessitates the application of systems-level approaches. InnateDB is an integrated web-accessible database and systems biology platform being developed to facilitate the systems level analysis of innate immunity pathways and networks. One of the aims of this thesis was to enhance InnateDB with bovine data, thereby providing a resource for investigation of this agriculturally important model organism. Using an orthology based approach, over 70% of InnateDB's human protein-protein interactions (PPIs), and a similar fraction of human pathways were reconstructed in cow and integrated into InnateDB.

Pathway analysis, the statistical association of observations at the molecular level with processes at a more systems level, plays a crucial role in the interpretation of high-throughput experimental datasets. A widely neglected challenge in pathway analysis relates to the handling of multifunctional genes. I therefore developed SIGORA, a novel pathway analysis method that identifies genes and gene-pairs that are unique signatures of a pathway and examines their over-representation in a given list of genes of interest (e.g. the list of differentially expressed genes in an infectious condition). With several biological datasets, SIGORA outperformed traditional methods, delivering biologically more plausible and relevant results. This was also reflected in significantly lower false positive rates for simulated datasets.

An additional challenge in high-throughput dataset interpretation concerns the lack of functional annotation for many genes. The guilt by association (GBA) principle was applied in a conservative manner to a large tissue expression dataset (105 Tissues, 13000 genes) to infer gene functions from co-expression data. Overall, 180 previously un-annotated bovine genes were assigned a putative function by this approach. In 20% of the cases, the inferred function was additionally supported by literature in other species.

microRNAs are emerging as important innate immune response regulators and as biomarkers of disease. Determining microRNA functions requires the identification of their targets, yet computational prediction of such targets is challenging. As part of a group investigating microRNA roles in bovine mastitis, I used a combination of prediction tools to compile a list of likely targets. Here, the overall emerging picture (including pathway enrichment) is consistent with our current understanding of this condition.

Collectively this work provides new tools and insights that may more broadly be used to improve systems-based analysis of bovine and other mammalian responses.

Keywords: Computational Biology; Bioinformatics; innate immunity; bovine infectious diseases; pathway analysis; miRNAs.

To Ebrahim, Fatemeh and Zaneta

Acknowledgements

I am very thankful for the invaluable guidance and contribution of my supervisory committee; Dr. Fiona Brinkman, Dr. David Lynn, Dr. Willie Davidson and Dr. Jack Chen. I would like to acknowledge all project collaborators at Teagasc, SFU, UBC and TCD, in particular Nathan Lawless, Peter Vegh, Karin Breuer, Matthew Laird, Geoffrey Winsor, Matthew Whiteside, Raymond Lo, Anastasia Sribnaia, Carol Chen as well as Dr. Bob Hancock, Dr. Cliona O'Farrelly and Dr. Mathew McCabe.

I would also like to acknowledge Lakshmi Matukumalli and other members of the Bovine Gene Atlas project for providing the bovine tissue expression data.

This work has been generously supported by a four year Walsh Fellowship Scheme from Teagasc and a Special Graduate Entrance Scholarship (SGES) from SFU.

Table of Contents

List of Tables

List of Figures

List of Acronyms

Term	Initial components of the term
BAM	Bovine alveolar macrophage
BGA	Bovine gene atlas
BME	bovine mammary epithelial cell
BSN	Bi-stochastic normalization
CLIP	cross-linked immuno-precipitation
CRISPR	clustered regularly interspaced short palindromic repeats
DAMP	danger associated molecular patterns
DC	dendritic cell
ECM	experimental cerebral malaria
GBA	guilt by association
GEO	Gene Expression Omnibus
GO	Gene Ontology
GO-BP	Gene Ontology Biological Process
GPS	Gene-pair signature(s)
GSEA	Gene Set Enrichment Analysis
HT	high throughput
hpi	Hours post infection
IL	Interleukin
KEGG	Kyoto Encyclopedia of Genes and Genomes
LPS	Lipopolysaccharide
MFE	minimum free energy
microRNA	miRNA
NGS	Next generation sequencing
NK	natural killer cells
NLR	NOD like receptor
NOD	nucleotide-binding oligomerization domain containing
ORA	over-representation analysis
ORF	open reading frame
PAMP	pathogen associated molecular patterns
PPAR	peroxisome proliferator-activated receptor

PPI	protein-protein interaction
RIG-I	retinoic acid-inducible gene I
RIP	RNA immunoprecipitation
RISC	RNA-induced Silencing Complex
RLR	RIG-I (retinoic acid-inducible gene I)-like receptors
SFU	Simon Fraser University
TB	Tuberculosis
TCD	Trinity College Dublin
TLR	Toll like receptor
TMM	Trimmed mean of M-values
TN	true negative
TNF	Tumour necrosis factor
TP	true positive
tpm	Tags per million
UBC	University of British Colombia
USDA	United States Department of Agriculture
UTR	untranslated region

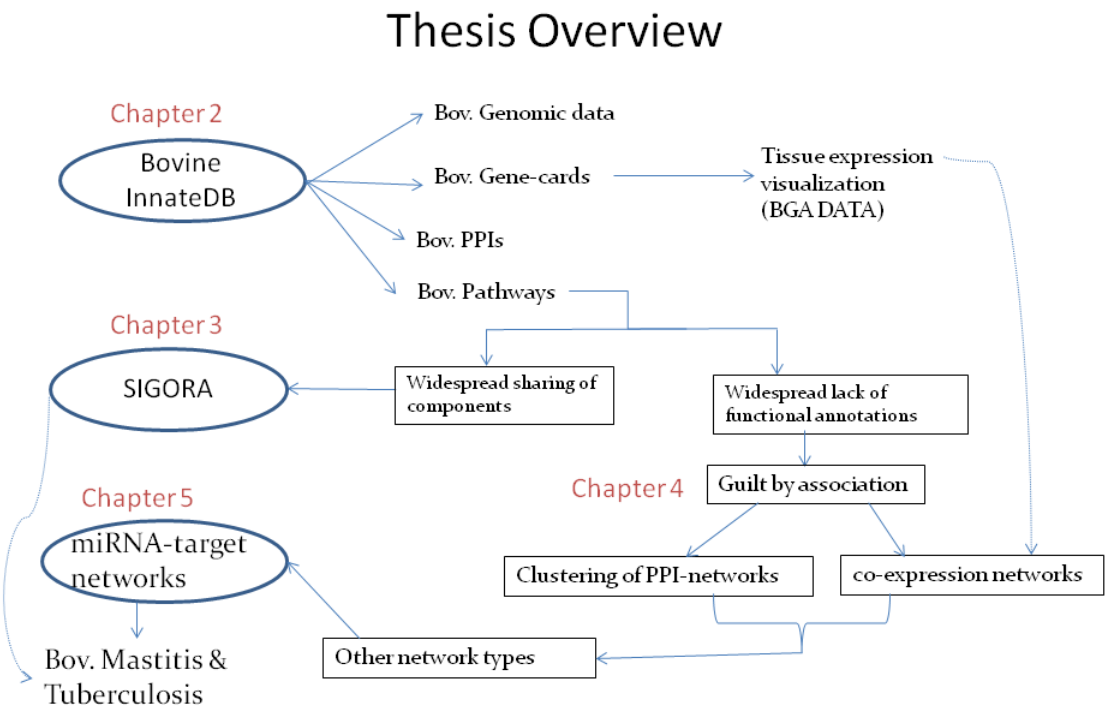
Glossary

Term	Definition
Adhesion Molecule	Any of a large number of cell-surface molecules of several different classes that affect the binding of one cell to another or to the extracellular matrix.
Antibody	A protein produced by the immune system in response to an antigen, often a virus or bacterium.
Autoimmune Disease	Disorder in which the immune system mistakenly attacks and destroys body tissue that it believes to be foreign.
B Cell	A cell produced by the bone marrow that becomes either a memory cell or a plasma cell that forms antibodies against a foreign substance.
Chemokine	Molecule that causes white blood cells such as neutrophils and monocytes to move throughout the body (e.g., toward an infected site) via the process of chemotaxis.
Chemotaxis	Movement of a cell toward or away from a chemical substance.
Complement System	Cleavage cascades involving a set of molecules (primarily produced by liver) in the blood; activated by the presence of bacteria, injury or other immune triggers, causing a range of responses associated with starting and maintaining inflammation.
Cytokine	Secreted proteins and signaling molecule that control cell-cell interactions in course of inflammation.
Endogenous	Arising within the body or derived from the body.
Endocytosis	Engulfment of molecules by a cell, leading to their absorption. An important subtype is phagocytosis.
Endotoxin	Poison in bacterial outer membranes that is harmful to the body, see lipopolysaccharide.
Eosinophil	Amoeba-like scavenger leukocyte (white blood cell) that disposes of cellular debris; often involved in allergic responses.
Epithelial Cell	One of the closely packed cells in a thin layer that covers the internal and external surfaces of the body, including body cavities, ducts and vessels.
Exogenous	Originating outside the body.
Genome	All the genetic material in the chromosomes of a particular organism.
Immune system	The body's system for protection against infection and disease; involves immune cells, antibodies, and other molecules.

Infection	Invasion of the body by harmful microorganisms such as viruses, bacteria, fungi, or parasites.
Infectious disease	Disease transmitted by microorganisms.
Inflammation	The immediate, stereotyped defensive reaction to any injury.
Inflammatory Mediator	Molecule inside or outside the body that plays a role in inflammation.
Interferon	Molecule (protein) produced by virally infected cells that helps the body fight off viral infections.
Interleukin	One of a class of cytokines that act as inflammatory mediators.
Leukocyte	White blood cell; acts as a part of the immune system by destroying invading cells and removing cellular debris.
Lipopolysaccharide	Poison in outer membranes of (gram negative) bacteria that is harmful to the body; see endotoxin.
Lymphocyte	Type of leukocyte (white blood cell) that mainly resides in lymphatic tissue (e.g., the lymph nodes) and is active in immune responses, including the production of antibodies; two types include B cells and T cells.
Macrophage	Type of large leukocyte (white blood cell) that uses a process called phagocytosis to eat bacteria and digest cellular debris; during inflammation, develops the ability to produce inflammatory molecules.
Mast cell	Type of leukocyte (white blood cell) found in connective tissues that produces histamine and other inflammatory molecules.
Monocyte	Type of leukocyte with large, kidney shaped nucleus; engulfs and breaks down debris and invading cells; can mature into macrophages or dendritic cells.
Neutrophil	The most abundant type of leukocytes in the blood. Neutophils are short lived first responders that travel through the blood to an infected or injured site via a process called chemotaxis.
Nitric Oxide	A highly reactive gas that is involved in a wide array of biological functions and functions as a part of the body's immune system.
Orthologs	Homologous genes that originated by vertical descent from a common ancestral gene in the last common ancestor when the species diverged.
Pathogen	Microorganism that causes disease.
Phagocytosis	A type of endocytosis in which solid molecules (such as bacteria or cell debris) are engulfed by phagocytic cells (e.g. macrophages).
Prostaglandin	Any of a class of hormone-like molecules that participate in diverse body functions including inflammation.

Protein	A large molecule encoded by a gene; they are required for the structure, function, and regulation of the body's cells, tissues, and organs; examples include hormones, enzymes, and antibodies.
Proteome	All the proteins made by a cell, organ, or organism at a particular time and under specific conditions.
Sepsis	Amplified systemic inflammation subsequent to infection or injury; typical symptoms include fever, mental confusion, and organ (lung and kidney) failure.
Systemic Inflammation	Inflammation throughout the body.
T Cell	A type of cell produced by the thymus that plays a major role in immune reactions.
Toll-Like Receptor [TLR]	Molecule on cell surfaces that helps the body sense the presence of endotoxin and other microbial products, and sends an alert to the immune system.
Tumor Necrosis Factor [TNF]	A member of a family of cytokines that induce cell death (apoptosis). Produced primarily by monocytes and macrophages.

Preface or Executive Summary or Introductory Image



Chapter 1.

Introduction

1.1. Burden of infectious disease

Infectious diseases are directly responsible for approximately a quarter of all annual human deaths (15 out of 57 million), with the majority of fatal cases occurring in the developing world (Morens, Folkers, and Fauci 2004). Even in non-lethal cases, the burden posed by infectious agents on productivity and quality of life for affected individuals is substantial and results in considerable economic loss. In addition to these direct and immediate threats, it is now increasingly clear that the dysregulation of the mechanisms responsible for detecting and combating infections is a contributing factor in the pathogenesis of non-infectious diseases like auto-immune disease, diabetes, and cancer (Lehuen et al. 2010; de Martel and Franceschi 2009; Marshak-Rothstein 2006). Last but not least, infectious diseases of domesticated plants and animals also pose additional serious global threats on several levels, including food security (e.g. limitation of available protein resources), food safety (e.g. antibiotics residues in milk and meat of treated animals), food borne pathogens (e.g. *E. coli, Salmonella*) and economics (e.g. milk yield loss due to bovine mastitis). In the US alone, there are an estimated 48 million cases of food borne illnesses per year, including 128,000 hospitalizations (Scallan et al. 2011). Worldwide, 1.3 billion cases of gastroenteritis and 3 million deaths due to *Salmonella* are reported each year (Gonose et al. 2012). Long term health consequences (sequelae) of infection by food borne pathogens may include kidney damage (E. coli O157:H7), neurological disorders (Listeria monocytogenes) and reactive arthritis (Salmonella Blockley) (Batz, Henke, and Kowalcyk 2013; McKenna 2012; I. G. Wilson and Whitehead 2006).

1.2. The importance of infectious disease in cattle

Of course, infectious disease is not only important for potential impacts on human health, but also has important consequences directly for animal health. In the late 19th century, trade introduced Rinderpest ('the plague of the cow') to Ethiopia. For centuries, Rinderpest had been endemic to Eurasia, where it killed about 30% of the cattle in affected herds. In its new environment, Rinderpest proved even more deadly: between 1892 and 1900, Rinderpest killed more than 90% of sub-Saharan cattle and decimated 40 other animal populations, including buffaloes, elands, wild swine, sheep, goats, antelopes, gazelles and giraffes. The ensuing, long lasting famine cost countless human lives (Morens et al. 2011). Rinderpest is a viral disease, which is caused by the Rinderpest virus (RPV), and there is strong evidence that the human Measles virus (a pathogen responsible for the death of over 100,000 children a year as of 2009) diverged from the Rinderpest virus in large Middle Eastern cities in the Middle Ages (Furuse, Suzuki, and Oshitani 2010), with a common ancestor possibly infecting both humans and cows. After a long, coordinated global vaccination campaign, Rinderpest was eradicated in 2010, making it only the second infectious diseases (after smallpox) in any species to ever have been formally and globally eradicated (de Swart, Duprex, and Osterhaus 2012).

The history of Rinderpest illustrates several important aspects of animal infectious diseases: their multi-host-character and potential to spread to other species, the threat they pose to human food security, the unforeseeable extent they take in new environments, and how manmade factors can exasperate these threats.

1.2.1. Zoonotic diseases and emerging diseases

An interesting question regarding emerging human diseases is 'where do these diseases emerge from?' Overall, only few human infectious diseases –old or new- are entirely human-specific: Most human pathogens also circulate in animals or else originated in nonhuman hosts (Lloyd-Smith et al. 2009; Greger 2008). Approximately 60% of all known human pathogens, and over 70% of emerging diseases of the past decades have been reported to be of zoonotic origin (Woolhouse and Gowtage-Sequeria 2005; Jones et al. 2008).

Testing complex hypotheses on disease emergence and chains of transmission is difficult, in part due to the multihost ecology of zoonotic infections and in part, because much less is known about infectious agents of wildlife, livestock and companion animals than of humans (Daszak et al. 2013; Parrish et al. 2008). Conceptually, it seems clear that emerging viral diseases are almost by definition of zoonotic origin, because viruses are obligate parasites (Wain-Hobson and Meyerhans 1999). Viral host-switching can involve several steps, including contact between the virus and the host, infection of an initial individual leading to amplification and an outbreak, and the generation (within the original or new host) of viral variants with the ability to spread efficiently between individuals in populations of the new host (Parrish et al. 2008). For livestock viruses, the ability of a virus to complete replication in the cytoplasm has been reported to be a strong predictor of cross-species transmission (Pulliam and Dushoff 2009).

A historic example of a human disease 'born on farm' is smallpox, which -after being introduced by landing Europeans- contributed to the loss of 90% of indigenous human populations in the Americas. Smallpox –at that time, an 'emerging disease' for native Americans- is believed to have arisen from camel domestication, with camelpox virus having a cowpox-like ancestor. The susceptibility of native Americans to this and other old-world diseases is attributed to the fact that unlike the landing Europeans, these populations had never been exposed to the virus before, because there were no domesticated cows or camels in Americas, i.e. no close contact between the old and the new host (Wain-Hobson and Meyerhans 1999; Greger 2008). Although smallpox has been eradicated, poxviruses have a broad host range and new cases of human infection by other poxviruses originating in cattle (including "Brazilian Cantagalo and Araçatuba Vaccinia viruses" , "Buffalopox virus" and "Cowpoxvirus") continue to (re-) emerge (Essbauer, Pfeffer, and Meyer 2010).

An example for a *bacterial* pathogen affecting both human and cattle is Mycobacterium *bovis.* The animals are infected by inhaling or ingesting the bacterium, which is then shed in their respiratory secretions, feces, and milk. Before the introduction of pasteurization, M. *bovis* contaminated milk was responsible for large outbreaks of 'consumption', an often lethal wasting disease which started with symptoms in the lungs. Clinically indistinguishable from M. *tuberculosis*, M. *bovis* is still estimated to be the causative agent of 3.1% of all human TB cases worldwide, with higher

prevalence in Asia and Africa (Ayele et al. 2004). As for its effects on the cattle industry, it has been estimated that bovine TB results in losses of approximately US$3 billion to global agriculture annually (Garnier et al. 2003). Recent outbreaks of bovine tuberculosis in the UK led to the culling of thousands of cows.

Other examples of infectious diseases affecting both human and cow include Anthrax, Brucellosis (causing spontaneous abortion and infertility in cow and long lasting undulant fever and chronic fatigue in human), Trypanosomiasis (vector-borne sleeping sickness), Cryptosporidiosis, Dermatophilosis, Listeriosis, Salmonellosis, Ringworm , Q Fever, Leptospirosis and Giardiasis.

E. *coli* O157:H7- first described as an emerging food-borne zoonotic pathogen in 1982- seems to be commensal in cattle, but highly pathogenic to humans. This and other emerging Shiga toxin-producing Escherichia coli (STEC) cause human illnesses ranging from bloody diarrhea and hemorrhagic colitis to the life-threatening hemolytic uremic syndrome. Rates of human infection have been repeatedly positively correlated with regional cattle density (Valcour et al. 2002; Kistemann et al. 2004; Hussein 2007), with contamination of water resources by animal manure and consumption of under-cooked meat being leading causes of human infections.

1.2.2. Animal infectious diseases in a changing world

Accelerating environmental and anthropogenic changes are altering the rates and nature of contact between human and animal populations, the modes of cross-species infection, as well as the occurrence-patterns and the geographic range of infectious diseases (Lloyd-Smith et al. 2009; Sutherst 2004). One example is the spread of Blue tongue disease (BTV) into Northern Europe (A. J. Wilson and Mellor 2009).

Methods of farm production are tremendously varied and bring with them their own particular risks in terms of the introduction and transmission of infectious diseases (Tomley and Shirley 2009): traditional, small animal holdings - still present in parts of developing world- put different livestock species in close contact with each other and with humans which can facilitate the exchange of diseases in both directions.

Large, industrial animal farming operations on the other hand, subject the animals to high levels of stress, which makes 'carrier' animals (infected but asymptomatic) more likely to shed the pathogens and non-infected animals more susceptible to infection. Examples include, long crowded livestock transports ('shipping fever') and densely packed feed-lots (IBR, BDV). The intensification of food industries also intensifies and perpetuates cycles of infection and cross-species infection, e.g. by usage of animal carcasses as animal food (BSE). Another problem associated with modern animal farming practices is the emergence and dissemination of multi-drug resistant pathogens, which is partly attributed to the use of antibiotics as growth promoters.

As the growing human population is projected to lead to a 50% increase in demand food and particularly livestock between 2000 and 2030, new strategies for sustainable industrial-scale farming practices become indispensible. As part of such strategies, host oriented approaches could help prevent (through vaccination, selection for breeding), detect (biomarkers for identification of subclinical, 'carrier animals') and possibly treat bovine infectious diseases.

1.3. The immune system

All multi-cellular organisms have evolved mechanisms to detect and combat infection. These mechanisms are collectively termed the immune system. Immunology, i.e. the study of the immune system at the cellular and molecular levels, has the potential to identify molecular diagnostic markers and novel therapeutic strategies and in the case of domesticated animals, can also lead to better breeding selection strategies.

Traditionally, immunology has primarily focused on the *adaptive* branch of immune system, which is specific to jawed vertebrates and is centered on the antigen specific mechanisms of pathogen recognition and clearance by T-cells and B-cells (Thymus derived cells). The adaptive (or acquired) immune system displays the hallmarks of an *immunological memory*, which underlies the great success of vaccination campaigns in decreasing the burden of infectious disease by priming the

host through exposure to dead or attenuated pathogens and antigen specific compounds.

1.3.1. The innate immune system

By the late 1980s, immunological studies seemed close to "reaching an asymptote". The main components of the adaptive immune system's recognition, response and immunological memory, including the structure and function of the antigen-specific receptors, the mechanisms of MHC (major histocompatibility complex) restriction, clonal gene rearrangements (that lead to creation of trillions of clones of B and T lymphocytes, each expressing a unique antigen receptor), lymphocyte development and activation, and the specificity of antibody responses seemed well understood and any remaining work was widely considered to be a matter of details. Yet it had also been observed that a vaccine's efficacy largely depends on additional, less specific ingredients like lipopolysaccharide (*LPS), methylated CpG* and *Freund's complex*. The question, as to why such *adjuvants* are necessary for an effective vaccine ('*vaccine's dirty little secret*') – and more broadly, how the adaptive response is *initiated*, lead CA Janeway to predict the existence of a class of innate immune receptors recognizing conserved microbial structures or "*patterns*". According to his "*pattern recognition theory*", the activation of the adaptive immune response is controlled by the evolutionary more ancient *innate* immune system which lacks antigen-specificity, and vaccine adjuvants act as stimulators of these innate mechanisms (Janeway 1989).

Over the past two decades, this *"pattern recognition theory"* has both revolutionized immunology and undergone several revisions along the way. On the more praxis oriented level, several families of germline-encoded pattern-recognition receptors (*PRR*s) have been identified. These PRRs include Toll-like receptors (*TLR*s), retinoid acid-inducible gene I (*RIG-I*)-like receptors (*RLR*s), Nuclear Oligomerization Domain (*NOD*)-like receptors (*NLR*s), C-type lectin receptors, and DNA receptors (cytosolic sensors for DNA). Originally, PRRs were modeled as disjoint mechanisms for the detection of broad, nonspecific antigenic patterns, that are *invariant* among entire classes of pathogens, *essential* for the survival of the pathogen, and well *distinguishable* from "self".

The presence of invading pathogens is commonly detected by tissue macrophages using families of pattern-recognition receptors. In a simplified view, each receptor focuses on a unique class of such pathogen associated elements: e.g. *TLRs 1,2,4,5,6* generally recognize microbial cell wall components while TLRs 3,7,8 and 9 are specialized for nucleic acid structures (Dunne and O'Neill 2005). More specifically, *TLR4* recognizes lipopolysaccharide (LPS) from cell membrane of *Gram-negative bacteria*, *TLR2* recognizes microbial lipoproteins and TLR5 detects *flagellin* from motile bacteria. TLR3 recognizes double stranded RNA from viruses, while TLR7 and TLR8 sense viral single-stranded RNA (which contain GU-rich or poly-U sequences) and TLR9 recognizes CpG DNA from bacteria or viruses (which -in contrast to mammalian CpG DNA- is unmethylated) (Mogensen 2009). The targeting of particular classes of pathogens by different sensors is achieved in part by differences in adaptor usage, cellular localisation and signaling cascades (Dunne and O'Neill 2005). Similarly, *RIG-I* is involved in recognition of short dsRNAs and is important for response to (enveloped) paramyxoviruses and influenza viruses, whereas a structurally similar receptor, *MDA5* seems to be critical for recognition of long dsRNAs and involved in recognition of (non-enveloped) picornaviruses and noroviruses (McCartney et al. 2008; Loo and Gale 2011). Engagement of receptors by microbial and fungal patterns stimulates the production of protein- and lipid-based inflammatory mediators such as IL-1β, TNFα (tumour necrosis factor α), IL-8, RANTES (regulated upon activation, normal T-cell expressed and secreted) and prostaglandins. Recognition of viral patterns additionally causes the production of IFN (interferon)-α/β, which in turn results in changes in expression profiles of hundreds of interferon inducible genes. Activated macrophages release inflammatory mediators that activate their surrounding cells such as epithelial and endothelial cells which then in turn also release pro-inflammatory mediators, cytokines and chemokines. The resulting inflammatory response increases vascular permeability and leads to the rapid recruitment of circulating leucocytes such as neutrophils and monocytes that work together with macrophages to clear the infection. In mammalian systems, dendritic cells (*DC*) process the antigen material and present it on their surface to the T cells of the adaptive immune system, thereby acting as messengers between the innate an adaptive system.

Janeway's original distinction between *self* and *microbial non-self* was subsequently expanded to include recognition of host-derived *danger signals* and *altered self* (signs of cellular stress, chronic inflammation, tissue damage signals, ATP and other self-molecules at aberrant locations or in abnormal molecular complexes), as well as signs of '*missing self*' (absence of 'do not destroy' / NK inhibitor) (Medzhitov 2009). On the other hand, it is now increasingly clear that, depending on cell type and tissue location (e.g. skin, intestine), non-self signals from commensal microorganisms need to be widely ignored or tolerated. This point brushes on another, more general recent trend in immunology: as many of the non-specific, "*first line of defense*" effector mechanisms of the innate response are extremely powerful (e.g. inflammation) and potentially harmful to the host itself, their activation has to be tightly regulated and fine-tuned, in order to limit the collateral damage to the host itself. It is now becoming increasingly clear that the study of immunology is as much about understanding the mechanisms of immune-modulation (i.e. how the system strikes a balance between activation and inhibition to avoid detrimental inflammatory responses) as it is about understanding immune system's recognition, communication and effector-mechanisms.

1.3.2. A glance at the complexity of the innate immune response

The immune systems of higher organisms are riddled with redundancies such as multiple subtypes of immune agents with complementary and mutually enhancing antimicrobial roles, alternative response pathways, and multitudes of regulatory layers. In many cases, these built-in redundancies can ensure the overall functionality of the immune system in spite of minor local defects. However, this robustness comes at the price of higher complexity. Our understanding of the molecular underpinnings of the innate immune system is still in its infancy, and we are only beginning to appreciate its staggering intricacies.

Complexity is inherent to the system, because the innate system has to fulfill several, at times conflicting, requirements and provide an effective, rapid and reversible response with limited resources. It has to be ready for combating an extraordinary range of pathogens without prior exposure, which necessitates the availability of broad, non-specific sensory and response mechanisms. Despite lacking the antigen diversity of the adaptive response, the innate immune response must be robust to immune evasion

strategies employed by pathogens, while accounting for the fact that different pathogens exploit different host mechanisms and attempt different evasion strategies. In the evolutionary arm race between pathogens and host, many pathogens have developed strategies to evade, manipulate or subvert *PRR*s or their crosstalk (Hajishengallis and Lambris 2011; Alto and Orth 2012; Diacovich and Gorvel 2010). An intriguing example is Salmonella, which takes the activation of *PRR*s as a cue that it has arrived in its intracellular niche and initiates the appropriate changes in its own virulence gene expression (Keestra and Bäumler 2011).

Once the host is infected, the system has to act rapidly to cope with the replication rates of pathogens, which necessitates simultaneous invocation of multiple pathways with synergetic effects. Intriguingly, although the individual sensory mechanisms seem nonspecific, the overall response (and the development of the symptomatic phenotype) is more distinct. This is partly due to the fact, that -in contrast to the simplified view presented above- PRRs cooperate, crosstalk and shape a combinatorial, more specific responses (Kawai and Akira 2011; Hirata et al. 2008; Loo and Gale 2011; Ozinsky et al. 2000). For instance, the attenuated yellow-fever (YF) virus used in the YF-vaccine triggers 4 different TLRs; another example is Salmonella *Typhimurium* infection, which triggers TLR2, TLR4, TLR9 and, to a lesser degree, TLR5, and TLR7. (Arpaia et al. 2011).

Downstream from recognition receptors, powerful inflammatory signals ultimately result in the activation of gene expression and synthesis of a broad range of molecules, including NF-κB, AP1, CREB, c/EBP, antimicrobial peptides (AMPs), immunoreceptors, defensins, interferons (which, in turn trigger the expression of hundreds of interferon inducible genes), nitric oxide, cell adhesion molecules, chemokines and cytokines (Kumar, Kawai, and Akira 2011; Newton and Dixit 2012). Cytokines, in turn, synergistically activate various cell types to induce the production of chemokines, which enhance the recruitment of leukocyte (such as neutrophils and monocytes), which infiltrate the affected site and further amplify the response (Gouwy et al. 2012). As for chemokines, the type and magnitude of their effects is extremely context sensitive and regulated by heteromerization. E.g. induced monocyte recruitment by *CCL7* is enhanced 100 times by *CCL19* and *CCL21*, but not *CCL2* (Blanchet et al. 2012).

Excessive production of inflammatory molecules contributes to the pathogenesis of inflammatory diseases such as rheumatoid arthritis and in the case of bacterial LPS, septic shock. Hence, the invoked mechanisms should act in coordinated and measured manner and be scaled back once the threat is contained to limit the collateral damage to the organism itself. Such immunomodulatory mechanisms include multiple checkpoints in form of several levels of regulation of gene expression like transcriptional regulation by Transcription Factors (*TF*s) (Zaslavsky et al. 2010) and posttranscriptional modulation by microRNA (*miRNA*) networks (O'Neill, Sheedy, and McCoy 2011), a range of posttranslational modifications (*PTM*s) of the sensory and communication systems (Moelants et al. 2013; Boone et al. 2004), cytokine networks (Blanchet et al. 2012) as well as crosstalk and feedback loops (Crozat, Vivier, and Dalod 2009; Mukhopadhyay et al. 2008) between these mechanisms within and across cells of different types.

In many cases, the signal to resolve inflammation has been embedded in the signals for its initiation, i.e. the necessary mechanisms for the subsequent dampening of the immune response are integrated into the initial response. One example is *TLR*-mediated regulation of both inflammatory and anti-inflammatory cytokine production (Hirata et al. 2008). At times, the same gene can play a role in both enhancing and inhibiting the inflammatory response, depending on the context. Notably, the very distinction between pro- and anti-inflammatory cytokines is not always applicable, as demonstrated by the dual roles of the cytokine IL-10 (Mocellin et al. 2003). Another example of extreme *pleiotropy* is *T cell intracellular Ag-1* (TIA-1), which has both a positive and negative role in the regulation of the powerful, multi-factorial inflammatory protein, TNF-α (Mazumder, Li, and Barik 2010). In the context of trafficking of leukocytes to inflamed sites by chemokine networks, there are well documented cases of chemorepulsion at high concentrations and chemoattraction at low concentrations (Blanchet et al. 2012). On a cellular level, both the innate and the adaptive immune system employ certain cells as modulators. For example, activated macrophages can be classified as the *M1* (pro-inflammatory and microbicidal) or *M2* (immunomodulators), based on characteristic gene expression profiles of e.g. ILs, TGFB, VEGF, MMPs, CXCL10 or TNF. Activation of macrophages by the innate inflammatory mediator interferon-γ (IFN-γ) leads to the *M1* response, while stimulation by interleukin-4 (IL-4) results in the *M2* phenotype (Xue et al. 2014). The activation has, however, been shown

to be plastic, rapid, and fully reversible, suggesting that macrophage populations are dynamic and may first take part in inflammation and then participate in its resolution (Benoit, Desnues, and Mege 2008). Further complicating matters, it was recently shown that the presence of additional stimuli not associated with either *M1* or *M2* activation (e.g. free fatty acids or high-density lipoprotein (HDL) or combinations of stimuli associated with chronic inflammation) leads to a spectrum of phenotypes beyond *M1* and *M2* (Xue et al. 2014).

In addition to this inherent complexity, several additional issues further confound the study of the innate immune system: first, although the 'hierarchical' distinction between adaptive and immune system has been very fruitful, there is considerable overlap between innate and adaptive immunity and increasing evidence for (some form of) regulation of the innate response by the adaptive response (Costantini and Cassatella 2011; Clarkson et al. 2012), as exemplified by the cross talk between DC and NK cells (Marcenaro et al. 2012; Harizi 2013). Second, due to their nature and function, innate immune mechanisms are both ubiquitous and deeply intertwined into other vital functions. As all tissues are potential targets of microbial invasion, professional innate-immune cell types (e.g. phagocytic cells, macrophages, NK-cells) reside in all tissues and have tissue-specific lineages with very different expression profiles that depend on the microenvironments in which they reside (Kowarsch et al. 2010; Hu and Pasare 2013). Furthermore, these cells are -*as cells*- subject to cell fate and cell cycle decisions and are regulated by the respective complex mechanisms (like MAPK and NOTCH signaling), but due to their special powers (to kill other cells), even more strictly regulated. The cross talk between such pathways and infection (Mitchell and Olive 2010) leads to a new assessment of traditionally non-immunological pathways, like autophagy (Oh and Lee 2013) as key regulators of the immune response. Going beyond such 'professional cells', virtually all cell types contribute to the innate immune recognition and response (via MHC expression and interleukin production). To a certain degree, the very organization of a compartmentalized cell serves immunological tasks (as barrier, lysosome). Finally, genetic factors, polymorphisms and alternative splicing (Pothlichet and Quintana-Murci 2013; Colobran et al. 2007), metabolism and nutrient excess (Schwartz et al. 2010; O'Neill and Hardie 2013), resident microbiota (Kitano and

Oda 2006; Hu and Pasare 2013), aging (Agrawal 2013) all affect susceptibility to infectious disease and/or modulate the inflammatory response.

Given the immense complexity of the dynamic networks of inter- and intra-cellular processes that determine the course of an infection and the innate immune response to it, use of systems level approaches has been proposed to enhance our understanding (Zak and Aderem 2009; Gardy et al. 2009; Lynn et al. 2008; Tegnér et al. 2006; K. D. Smith and Bolouri 2005). The next sections provide a short overview of Systems Biology approaches.

1.4. Systems level approaches and immunology

A common characteristic of the examples listed in the previous sections is the *emergence* of new, unexpected, context dependent properties that arise from interactions among different components. Emergent properties cannot be fully understood in terms of isolated components. Their existence limits the usefulness of models that seek to explain immunology in simple terms: *pleiotropy* makes it difficult to unambiguously classify cytokines as pro-inflammatory or anti-inflammatory, the dynamic *interactions* between regulators of expression add to plasticity of cellular behavior and show that there is more to gene expression than mere on/off switches, and *crosstalk* contradicts the traditional view of pathways as a series of linear events. Clearly, detailed component lists will not by themselves provide the needed insights.

One way of tackling some of the emergent properties and related complexities is the use of systems-oriented methods. Systems biology primarily considers the interplay between different components of a system in order to account for complexities that cannot be explained or predicted by examining the components in relative isolation. A famous example is ‘boids model’ (Reynolds 1987) of bird flocks: it is impossible to explain flocking constellations by looking at the flight patterns of one bird at a time, but three simple rules on how each bird aligns itself in relation to other birds in its local environment suffice for a full explanation. A more mechanistic analogy is a car engine: having a catalogue of parts is important, but not enough for assembling or repairing an

engine. There is a lot of additional information in the diagrams that explain how the parts fit and work together.

It is worth noting that there is an active and ongoing discussion about theoretical fundaments, scientific methodology and future direction of Systems biology (Kitano 2002; Gatherer 2010; Friend 2010; Lander 2010). However, (on a phenotypic level) there are some reoccurring themes in a wide range of studies that consider themselves systems oriented. These include: the use of network representations of large scale and or complex interaction data sets along with algorithms to explore the structure and topology of those networks; the application of mathematical and graph theoretical concepts to predictively model the immune response and large-scale experimental approaches. Conceptually, Systems biology is founded on the synergistic interplay between biological, technological and computational sciences. Biological questions drive technological advances, which require new computational tools. Similarly, technological and computational advances provoke biological insight and new models for biological systems (Smith and Bolouri 2005).

1.4.1. Types of systems studies

Broadly speaking, one can distinguish between three types of systems approaches to immunology: the experimental approach, the integrative approach and the modeling approach. In the experimental approach, the phenotypic effects of different possible combinations of factors are explored systematically, by devising experiments that cover large parts of the feature space without compromising the quality of measurements. Modeling and simulation oriented approaches [box 1] try to harvest our current knowledge and understanding of immunological principles and mechanisms for predicting cell fate, disease trajectory and therapy response. Finally, in the integrative, network based approach [box 2], vast amounts of large-scale, genome-wide and often noisy data from different layers of a molecular process are consolidated, analysed and mined to uncover novel, pathways and/or regulatory mechanisms as well as the key regulators of these responses. Ideally, elements from different approaches are combined in a study. An example would be a study that experimentally and systematically explores candidate gene knockdown effects by siRNA assays and then uses the

integration of this data with pathway and protein-protein interaction networks to facilitate the interpretation of these experiments.

Experimental systems immunology

The main idea in systems oriented experiments is systematic, but targeted perturbation of biological mechanisms or manipulation of biological agents and observation of the phenotypic outcomes. An example is (Hsueh et al. 2009). This study takes a systemic experimental approach to examine crosstalk and non-additivity effects. They measure simultaneously the secretion of six cytokines (G-CSF, IL-6, IL-10, MIP-1α, RANTES, and TNF-α) by RAW 264.7 cells, a macrophage-like cell line, in response to the simultaneous application of multiple stimuli. One of their conclusions is that although synergy or anergy in response to input-pairs is very common, higher order combinations of ligands (three or more stimuli at a time) rarely have additional non-additive effects. This study is particularly interesting because it rejects a 'combinatorial explosion' of complexity that one might suspect from analyzing high throughput (*in vitro*) interaction data alone.

Many single manipulations are likely to lack observable *in vivo* effects due to built in redundancies (Friend 2010). On the other side, it is often not feasible to experimentally explore all possible perturbations or combinations. Hence, experimental designs have to be selective about their targets and the order of performed manipulations. Additionally, experimental results need to be interpreted: in a situation, where gene knockdown experiments have indicated that several genes belong to a certain pathway, it is still important to understand the pathway structure, i.e. the upstream/downstream relations (Frohlich et al. 2008; Markowetz and Spang 2007).

Computational methods, discussed below, can be supportive in predicting and interpreting experimental results as well as in selecting appropriate targets. The support comes in the form of modeling and simulation, as well as providing statistical links between molecular-level observations (of high throughput experiments) and systems (cell, tissue, organism) level processes, thereby identifying possibly relevant pathways in case-control studies, and finally by providing knowledge discovery from integrated biological networks.

Box1: Modeling and simulation

In a *model* based approach, existing knowledge is incorporated into a simplified but coherent view of the system that is subject to a starting configuration (a set of basic parameters describing the initial *states* of the models' components), basic *transition* conditions (rules that describe how the states are updated over time) and any necessary additional *constraints* (for example maximum concentration thresholds). The next step is computational *simulation*, which allows changes of the configuration to be tracked over time due to the transition conditions, and thereby enables predictions about the real outcomes. Models can be *binary or continuous*, depending on what type of numeric values the settings can assume. In a binary model, for example, a gene is expressed or not, whereas a continuous model can capture the expression rate. Models can be *deterministic or stochastic*, depending on the type of transition rules. A deterministic model always results in the same outcome (cycle or steady state) for a given initial configuration and after certain number of time points, whereas stochastic models account for the effects of noise. Models can be *synchronous or asynchronous*, depending on whether the state of all nodes is updated simultaneously after applying all rules, or if individual states are updated on the fly. Models can be *population-based or individual-based*, depending on whether they are interested in tracking the total numbers of each type of cells or species or in tracking each individual cell's history. Finally, models can differ on whether or not they allow different events to take place over *different time scales* (globally fixed vs. individually variable durations).

Quantitative models

Quantitative models try to determine the individual contributions of different factors to the development of dynamic processes over time. Many quantitative models describe the system of interest as a set of ordinary differential equations (ODEs). For example, in a known regulatory pathway containing a set of co-expressed genes, one could solve ODEs to fit observations from a time series experiment. This would amount to quantifying how the expression levels of each gene at a time point *t* effects the expression levels of that gene and other genes at the time point *t+1*. For instance, (Sorathiya, Bracciali, and Lio 2010) give a deterministic and a stochastic model for the interplay of HIV and TB infections in the context of highly active antiretroviral therapy. The deterministic model fits well with long term outcomes, while the stochastic model captures the short term and noise-related fluctuations.

Qualitative and rule based models

Two main groups of rule based and qualitative models are Boolean Network Models and Cellular Automata & Mobile Agents. In a Boolean network model (Albert et al. 2008), the system is represented as a directed graph, where nodes represent entities (cell types, cytokines, antibodies, antigens, etc), edges represent processes (presentation, activation, modulation, etc) and edge directions correspond to information flow. These types of models aim to reflect the topology of regulatory networks; inhibitory signals and signal combinations can be incorporated using logical operators. By simulating the model over a range of different initial configurations and comparing the outcomes, even a simple deterministic, binary, qualitative Boolean model can be useful in identifying the key driving components, e.g. the number of steps it takes to reach a steady state. The case studies presented in (Albert et al. 2008) use a mixture of qualitative models and differential equations. They include simulations of cytotoxic T lymphocytes' expansion and apoptosis in TLG-leukemia and pathogen-host interactions for *B. bronchieseptica*.

The main concepts in Cellular Automata (CA) and Mobile Agents (MA) models are agents (for example cells) and environments (for example local chemokine profiles). In contrast to population based methods like ODEs, simulations of these types of models provide a history for each individual agent, making them particularly useful in cell fate studies. The conceptual distinction between agents and environments facilitates the simultaneous description of processes on cellular and molecular levels in an intuitive way (e.g. immune cells as agents and chemokine profiles as environments). In contrast to other modeling paradigms that require a prior knowledge of high level rules, complex rules can emerge in CA/MA models from simpler rules by *self organization* of the agents. CA/MA models are also realistic in the sense that variations to initial constellations (e.g. concentration and local distribution of cells and chemicals) can have profound effects on trajectory of simulations.

An extensive introduction and review can be found in (Chavali et al. 2008). Due to their disposition for a multi-scale modeling approach, agent-based based models are particularly popular in cancer research. There, understanding how mutated cells interact with other cells in their microenvironment could explain how tumour cells' intracellular

processes become independent from external growth signals (Laubenbacher et al. 2009). In an immunological context, this approach is starting to gain momentum. For instance, (Rapin et al. 2010) use a combination of an agent-based model and a set of molecular binding prediction methods for computational simulation of the immune response. This allows for a neat separation of processes on molecular level and cellular levels.

The previous sections gave an overview of modeling and simulation based methods and their applications in immunology. A precondition for building models is availability of existing knowledge. The following sections describe the important contribution of another branch of computational systems biology, data driven knowledge discovery and reverse engineering of immunological systems.

Box2: Networks and network analysis

A) Descriptive power of network representation

Networks are a popular way of representing and analyzing complex systems. They are versatile frameworks, both intuitive to humans and accessible to computational analysis. A network consists of a set of *nodes* that represent entities and a set of *edges* that represent relations between those entities. Nodes can be all of the same (homogenous) or of different (heterogeneous) type. The edges can have optional attributes, most important among them: *directions*, *weights* and/or *capacities*. If present, the weights of edges convey information about the relative intensity of the relationship, the capacities describe possible thresholds and constraints and the directions clarify the relative order of processes. Most deeply studied bimolecular networks fall into 3 categories: a) Protein-Protein Interaction (PPI) Networks. PPI are modeled as undirected graphs. An edge in a PPI asserts the possibility of physical interactions between the corresponding proteins (nodes). PPI's are usually derived from error-prone high throughput methods. *Clustering* PPI's (delineating highly interconnected regions of the network) is used as a means of protein module identification. B) Metabolic Networks describe a web of reactions catalysed by enzymes and can be modeled as weighted directed graphs with edge capacities. Using FBA (flux balance analysis, a linear program/optimization problem for an objective function such as growth rate, subject to certain irreversibility and other constraints) biochemists can describe the relations between metabolite concentrations and reaction fluxes. C) Transcriptional regulatory networks, directed graphs describing dynamic regulations between genes, proteins and other agents, are derived from combinations of high throughput methods and available literature. Their study is central to cell fate prediction, for example in cancer research.

B) Network analysis

Although network representations are powerful descriptive tools, their popularity is largely due to the availability of formalisms to analyze and mine them. A node in a biological network is said to be essential if the organism cannot survive without it. Use of centrality measures, structural properties of nodes in a network, as a predictor of essentiality has several potential applications e.g. in drug target identification. Node degree, the number of connections of a node, is the simplest measure for centrality and influence of a node and widely used in analyzing non biological (e.g. social) networks. However, in contrast to social network, most biological networks are disassortative: Highly connected nodes (hubs) avoid connecting directly to each other. Node degree gives some indication of a protein's likelihood to be essential, but the relationship is not simple. A more global view of centrality of a node is the notion of "bottlenecks" and "articulation points", nodes whose removal disrupts or significantly prolongs the information flow among a high portion of other nodes. Mutations in such nodes have been shown to be more likely to be lethal for an organism than mutations in nodes with high degree alone. A central concept to network based analysis is similarity or its dual, distance. It is widely used to classify nodes or cluster networks by identifying regions of similar nodes. Guilt by association approaches infer unknown features of certain nodes based on known features of other nodes in the same cluster. A number of different similarity measures have been proposed and used in network analysis, including: Graph distance (negated shortest path), Common neighbors (CN), normalized CN (Jaccard), Preferential attachment score, Katzβ and SimRank. (Getoor and Diehl 2005; Mason and Verwoerd 2007; Barabási 2007; Sharan, Ulitsky, and Shamir 2007).

1.4.2. Integrative systems biology

The development of a range of high-throughput (HT) technologies has resulted in the collection of many large scale data sets that deal with different aspects of biological processes and encode different levels of biological information. The mere size and high dimensionality of these data sets prohibits their manual interpretation. Networks (see box 2) have become a convenient tool for representing and analyzing these data sets, because they are versatile frameworks that are both intuitive to humans and accessible for computational analysis. Many current efforts focus on reverse engineering biological mechanisms from these networks.

One example are large scale protein-protein interaction datasets driven from yeast two-hybrid (Y2H) experiments. These datasets can be translated into an interaction network, where edges indicate an observed *in vitro* interaction. Determining densely interconnected regions in such networks (*clustering*) can help to identify functional complexes and to predict the unknown function of a protein from the known functions of its interaction partners (Sharan, Ulitsky, and Shamir 2007). Time-series transcriptional snapshots by microarrays that contain mRNA expression status of thousands of genes, represent another common example. Using correlation analysis, this data can be translated into a coexpression network, where edges indicate coordinated changes in expression behavior, possibly subject to common regulatory mechanisms.

There are, however, major quality issues and caveats regarding the reliability and comprehensiveness of high-throughput data sets. Such large scale measurements are known to include a high fraction of misleading information (*false positives, FP*) while failing to provide some important information (*false negatives, FN*). Major contributors to the FP and FN rates are the noisy nature of high-throughput technologies and discrepancies between *in vitro* and *in vivo* situations. Furthermore, each data set regarded in isolation is weakly informative: it can only describe one single aspect of a dynamic molecular process, neglecting others. As an explanatory example, real protein-protein interactions (PPIs) have contexts, timing, duration, location and rates, which are not usually captured by current technologies in a high-throughput manner.

Therefore, reverse engineering efforts often have to combine information from several layers of omic data (genomic sequence information, transcriptomic abundance measurement, proteomic interaction data) and additional pathway membership and annotation information.

In the first example above, many *in vitro* observed interactions can be discarded by additionally taking the subcellular location of respective proteins into account, which can result in more significantly enriched modules. Concerning the second example, (Ramsey, Gold, & Aderem, 2010) describe how expression data and interaction data can be combined to identify a common transcription factor that is itself not differentially expressed. They also refer to a study by Gilchrist et al in 2006 that used a combination of transcriptional profiling and promoter sequence data to identify ATF3 as a negative regulator of macrophage responses to LPS and also of TLR4-induced expression of IL6. Using a similar approach, the authors of (Ramsey et al., 2008) identified TGIF1 (TGFB-induced factor homeobox 1) as a master regulator of a cluster of TLR-responsive genes including cytokines Csf2 and Gm1960.

Regarding the third issue (limitations on the range of aspects that can be simultaneously captured by HT experiments), overlaying PPI data with sub-cellular localization data of respective proteins can help clarify the spatial and by extension temporal order of a chain of events (as in a signaling pathway) (Barsky et al. 2007). In related examples, combining PPI, transcriptional regulatory and pathway data has been useful for expanding existing pathways, uncovering novel pathways or regulatory mechanisms and identifying the key modulators of such pathways and mechanisms. Such (integrated) data can build a framework for a better interpretation of measurements from targeted experiments. For example, pathway overrepresentation analysis of differentially expressed (DE) genes in a particular condition often gives a clearer picture of active processes in that condition than fold change data alone. Moving beyond pathways, connectivity analysis of the interaction networks of DE genes can further clarify the underlying mechanisms. Similarly, by investigating networks that include interactions between DE genes and their non-differentially expressed interacting partners, one has the potential to identify key regulators of gene expression, even though these regulators themselves may not be differentially expressed but regulated at the posttranscriptional level.

One of the first platforms for systems level analysis of human and mouse innate immune response is InnateDB (Lynn et al. 2008), which allows users to investigate their data (e.g. gene expression) in interaction network or pathway context. InnateDB is a collaboration between the Brinkman Bioinformatics Group at Simon Fraser University, the Hancock Laboratory at the University of British Columbia and the Lynn Systems Biology Group at the Teagasc Animal Bioscience Department, Ireland. One of its primary goals is to provide a manually-curated knowledgebase of the genes, proteins, and particularly, the interactions and signaling responses involved in mammalian innate immunity (Lynn et al. 2010). Enhancing InnateDB to incorporate and analyze orthology-inferred bovine pathways and PPIs was a major goal of this thesis (chapter 2).

1.5. Pathways, pathway databases and pathway analysis

It is clear that from a systems biology perspective, cellular processes governing health and disease are viewed as phenotypic outcomes of changes to highly intertwined, dynamic networks of interactions between heterogeneous, context sensitive agents. This is a quite different view than the standard text book presentation of pathways as relatively simple linear cascades. Nevertheless, pathways are useful abstractions allowing for a compact representation of main elements and key interactions in common, recurring, biological processes with particular phenotypic outcomes. A pathway's outcome can be e.g. the assembly of new molecules, the propagation of information within or across cells or changes in the status of cell. As such, pathways can be thought of as (often conserved) modules in cellular interaction networks working towards a common goal. A typical signaling pathway, for example, can represent receptor-binding events, phosphorylation reactions, protein complexes, translocations and transcriptional regulation, with only a minimal set of symbols, lines and arrows.

Pathway repositories, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome and the Pathway Interaction Database (PID) (Minoru Kanehisa et al. 2011a; Matthews et al. 2009; C. F. Schaefer et al. 2009), try to condense the available biological knowledge, providing some level of mechanistic detail, while making the source data available in a computationally accessible format.

Pathway analysis is a crucial first step in interpreting results of high-throughput omics experiments. As a methodological approach, pathway analysis is now a little over a decade old and the field has seen an array of proposed methods and tools. The current predominant pathway analysis methods can be categorised into two major branches: Over-representation Analysis (ORA) methods (reviewed in (Khatri and Drăghici 2005)) and methods related to Gene Set Enrichment Analysis (GSEA) (Mootha et al. 2003; Subramanian et al. 2005). Pathway ORA methods investigate whether the observed fraction of genes belonging to a specific pathway in a user specified list is more than one would expect by chance. The genes in the input list are usually determined based on an arbitrary threshold for significance, for example, genes that are significantly differentially expressed or genes that are significantly associated in a genome-wide association study (GWAS). GSEA methods, on the other hand, do not apply such a threshold and instead consider the collective rank of all genes in a given set (e.g. a pathway). GSEA purports to avoid some type II (false negatives) errors of ORA approaches by accounting for subtle but coordinated differences in a given pathway between conditions of interest.

GSEA based methods are limited to experimental designs consisting of two groups of samples (e.g. infected vs. control). In many cases, the upstream data processing and comprehensive gene selection statistics cannot be simply avoided or replaced by GSEA, the effects of additional factors (e.g. gender, age) besides the assignment to one of the two groups cannot be taken into consideration, and the results of such pre-processing often don't conform to GSEA-required input data structures (D. W. Huang, Sherman, and Lempicki 2009).

Both ORA and GSEA based methods view pathways as collections of individual genes, and treat all genes annotated in a pathway as equally informative indicators for activation/perturbation of that pathway. As will be described in chapter 3, an important question for the work presented in this thesis was the potential unintended consequences of such egalitarian perspective for the results of pathway analysis. This question lead to the design and implementation of an alternative pathway analysis method (signature over-representation analysis, SIGORA) that seeks to avoid some of the issues by focusing on statistical over-representation of weighted pathway specific combinations of gene pair signatures (termed “Pathway-GPS”). In tests on simulated

and published datasets, SIGORA outperformed several popular pathway analysis methods by delivering more plausible and relevant results. SIGORA inherits the versatility of the ORA statistical framework and is applicable to lists of genes of interest obtained in any type of high throughput experimental set-up, including lists of predicted targets of a set of differentially expressed miRNAs (exemplified in chapter 5).

1.6. miRNAs as novel regulators of innate immunity

MicroRNAs (miRNAs) are an abundant class of highly conserved, small (19–24 nt long), non-protein coding RNA molecules. They act as important post-transcriptional regulators and they function, via imperfect base-pairing with complementary sequences within mRNA molecules, as repressors (or less frequently, enhancers) of specific target genes at the post-transcriptional level. Repression of gene expression by microRNAs is achieved by translational repression or degradation of the target mRNA (Baek et al. 2008).

1.6.1. Biogenesis

Approximately half of all known human miRNA-coding genes reside in intergenic regions outside known genes while others are contained in the intronic sequences of protein-coding genes or in the exons of untranslated genes. Some miRNA primary transcripts encode only a single mature miRNA (e.g.: mir-203). *Polycistronic* miRNAs, on the other hand, are clusters of several miRNA that are transcribed together from one transcription unit (e.g. mir-17-92 cluster). Most miRNAs are first transcribed as long, mRNA-like polyadenylated primary transcripts (pri-miRNA), by RNA polymerase II enzyme. The pri-miRNA, which can be surprisingly long (up to several kilo bases) is then cleaved in the cell nucleus by the ribonuclease enzyme, Drosha, to a shorter (~70 nucleotide) hairpin structure known as the precursor (pre-) miRNA. The hairpin is then exported to the cytoplasm, where it is further cleaved by an RNase III enzyme (Dicer) to a short (19-24 nt) double stranded miRNA duplex. One strand of the duplex (the “guide” strand) is then incorporated into the RNA-induced silencing complex (RISC), while the other (the “*” or “passenger” strand) is degraded. The guide strand is selected by the argonaute proteins, the catalytic components in the RISC complex. The choice of the

guide strand is believed to be related to the lower stability base pairing of the 2–4 nt at the 5_ and 3_ end of the duplex (Schwarz et al. 2003), but the guide strand is more generally operationally defined as the more abundant strand.

There are several important exceptions to the above simplified description. Some intronic miRNAs, known as mirtrons, bypass the Drosha processing and are spliced from the intron (Westholm and Lai 2011). Additionally, Argonaute-2 (Ago2)-mediated pre-miRNA cleavage can induce Dicer-independent miRNA processing. Furthermore, some passenger strands (like miR-19*) have also emerged as functionally active microRNAs (J.-S. Yang et al. 2011). Some miRNA loci yield multiple functional products, from both hairpin arms or from both DNA strands, i.e. a single genomic miRNA locus may produce up to four miRNAs, each with distinct targets (Stark et al. 2007). Finally, very recently it has been shown that the guide-to-passenger strand expression and activity ratio is not constant and can be shifted in favor of the passenger strand by Argonaute-3 (Ago3) (Winter and Diederichs 2013).

1.6.2. Function and relevance to innate immunity

Aside from their well documented role in developmental timing, cellular differentiation, signalling pathways and apoptosis, microRNAs are now also emerging as key components of the immune system. In (studies of) the innate immune system, they assume multiple roles: as regulators and modulators of the inflammatory response within cells (Quinn and O'Neill 2011; Rossato et al. 2012), as drivers of innate immune cell differentiation, proliferation and activation (Bi, Liu, and Yang 2009; Cichocki et al. 2011), as crucial routes of pathogen-host interaction (Marcinowski et al. 2012; Pfeffer et al. 2004), as predictive markers of auto-immune diseases (Zhu, Pan, and Qian 2013), and - encapsulated in secreted exosomes - as intercellular messengers that facilitate innate-to-innate and adaptive-to-innate intercellular communication and coordination (Mittelbrunn et al. 2011; Valadi et al. 2007).

The list of miRNAs with experimentally verified roles in innate immunity includes: *miR-146a/b, miR-132, miR-155, miR-21, miR-147, miR-125b* (negative regulation of the TNF pathway)*, miR424, miR-511, miR-223, miR-187 (*negative regulation of the TNF pathway*), miR-181b, mir-29a* as well as members of the *let-7* family (*let-7a, Let-7e, let-*

7i) (Staedel and Darfeuille 2013; Bi, Liu, and Yang 2009). Innate as well as adaptive immune cells possess specific miRNA expression patterns regulating both cell fate and function. Granulocyte-monocyte progenitors (derived from common myeloid progenitors) produce neutrophils (regulated by *miR-223*) and monocytes (regulated by members of *miR-17-92* cluster as well as *miR-155* and *miR-106a*). Monocytes further differentiate into myeloid-derived dendritic cells (DCs) or macrophages. Activated macrophages respond by up-regulation of e.g. *miR-155* and/or *miR-146* (Bi, Liu, and Yang 2009).

MiRNAs are also emerging as one of the molecular mechanisms involved in resolving inflammation. During bacterial infection, peptidoglycan (PGN)-mediated *TLR2* signaling induces *miR-132/-212* to down-regulate IRAK4, an early component in the *MyD88*-dependent pathway (Nahid et al. 2013), whereas *LPS/TLR4*-induced *miR-146a* down-regulates downstream components of the same *MyD88*-dependent pathway (Williams et al. 2008).

A liver-specific miRNA, miR-122, interacts with sequences in the 5′ noncoding region of the hepatitis C virus (HCV) RNA, and this interaction is required for viral replication and maintains high viral RNA abundance in liver cells. The tissue specificity of miR-122 expression also helps HCV to establish its tissue selectivity (Jopling 2012).

MicroRNA Biology remains full of unsolved questions, exceptions and paradoxes. For instance, miR-511, which is highly expressed in monocyte-derived DCs and macrophages, up-regulates its putative direct target TLR4 in DCs through a yet unknown mechanism. In other cases, the direction of miRNA function (i.e. repression or up-regulation) is context sensitive. For instance, miR-155 has been shown to both positively and negatively regulate inflammatory pathways (Quinn and O'Neill 2011). As a positive regulator, it suppresses the suppressor of cytokine signalling 1 (SOCS1) and SHIP1, while the negative mechanism is again, unknown. Notably, some miRNAs, like let-7 can induce translation up-regulation of target mRNAs on cell cycle arrest, while repressing translation in proliferating cells (Vasudevan, Tong, and Steitz 2007). Additional layers of context sensitivity have been documented for combinations of miRNAs: For instance, until very recently, it was believed that *mir-146* and *mir-155* are always co-activated, but now they have been shown to display very different dose-

dependent expression profiles in responses to environmental stimuli (Schulte, Westermann, and Vogel 2013).

Examining which genes are affected by a given microRNA is the best way for understanding the miRNA's role. In some cases, the control of single key target genes appears to largely explain the functions of individual miRNAs. For instance, it was observed that mir-29 targets interferon gamma mRNA and thereby changes intercellular communications during the course of infection (F. Ma et al. 2011). However, in general, miRNAs do not switch off the expression of their target genes but only reduce the amount of mRNA and protein (Baek et al. 2008). Estimates on the average number of targets per miRNA in vertebrates range from 100 to roughly 200 transcripts each (Krek et al. 2005; Brennecke et al. 2005).

The influence of a single miRNA on the expression level of a single target mRNA can remain undetectable by using current measurement methods. On the other hand, because miRNAs often target several mRNAs from the same pathway, as well as the fact that a single mRNA can be targeted by multiple miRNAs, the influence of miRNAs becomes indispensable (O'Neill et al., 2011). Larger groups of microRNA's seem to have overlapping functions, because of the conservation of the seed within a given miRNA family, and, therefore, similar target profiles. Outside such families, unrelated miRNAs (that do not show sequence similarity) may be activated by the same transcription factor and thus overlap in their activity in targeting (different parts of) a given cellular pathway.

Understanding the functions of miRNAs requires means of identifying their target genes. Although databases of experimentally verified miRNA:mRNA interactions exist and are rapidly growing (Vergoulis et al. 2012), the real number of such interactions are believed to be much higher. Numerous computational methods for predicting targets are currently available. The next section describes a few of the guiding principles of computational target predictions, prominent methods and their limitations.

1.6.3. Computational prediction of miRNA targets

The most common mechanism of action for microRNAs is to target mRNAs for degradation or suppress their translation by binding to the 3'-UTR, other, less widespread mechanisms include targeting in an ORF (Lewis, Burge, and Bartel 2005; Stark et al. 2007). Unlike in plants, in vertebrates there are very few cases with absolute complementarity between an entire microRNA and (part of) its target (Miller and Waterhouse 2005). Accordingly, most target prediction programs search primarily for complementarities between the six to eight nucleotides at the 5' end of the mature miRNA sequence (the 'seed region' of a microRNA) and the 3'UTR of putative targets. The match in the seed region does not have to be perfect, as G.U base pairing in the seed region is tolerated. It has been shown that imperfect base pairing of the target with the seed segment can be compensated by additional matches in the 3' out-seed segment of the miRNA–mRNA duplex. The focus on the 3' UTR leads to some false negatives, as miRNA can target mRNA outside of the 3' UTR, but this mechanism is neglected by most prediction methods. A more serious problem is the very high false positive rates of computational prediction methods (Rajewsky 2006). As the seed region is extremely short, the likelihood of chance-matches within the genome is relatively high; hence, the 6- to 8-base-pair perfect seed pairing is not a generally reliable predictor for targeting (Didiano and Hobert 2006). To address the false positive problem, most methods rank and filter the predicted targets by some additional criteria, described below.

Phylogenetic conservation:

The premise here is that evolutionarily conserved target sites are more likely to be true targets. An example for conservation based criteria is the probability of conserved targeting (PCT), employed by TargetScanS versions 5 and above. PCT is a comparative genomics based measure which reflects the Bayesian estimate of the probability that a site is conserved due to selective maintenance of miRNA targeting. The PCT incorporates knowledge of the conservation level of a particular site, the seed-match type, the number of selectively maintained seed matches for the particular miRNA, the background conservation level for the k-mer across 23 vertabrates, and the UTR conservation context (taking into account that a site falling within a UTR with high

overall conservation is likely to be conserved due to reasons other than miRNA targeting).

This type of 'phylogenetic conservation' based filtering can increase the false negatives while it decreases false positives. Furthermore, the alignment is sensitive to the number and order of the aligned species, and cannot address novel (non-conserved) miRNAs. It has been shown that about 30% of mammalian targets are not conserved in alignment.

Genomic context:

An example of conservation-independent criteria is context+ scores (Garcia et al. 2011), that are provided in TargetScanS versions 6 and above. Context+ scores are an enhancement of Context scores. Context scores (Lewis, Burge, and Bartel 2005) account for features in the surrounding mRNA, including local A-U content, location of the miRNA-mRNA match (sites near either end of the 3′ UTR are preferred), site number, and 3′-supplementary pairing (Lewis, Burge, and Bartel 2005). Context+ scores additionally take the target site abundance and seed pairing stability (both of which influence the sRNA proficiency // robustness of targeting) into account.

Importantly, although both PCT and context+ scores are provided in current versions of TargetScanS, the two ranking criteria are based on completely orthogonal types of considerations and provide independent and complementary information on biological relevance and efficacy of each site (Friedman et al. 2008). Hence, taking the intersection of top results based on these two criteria would be rather counterproductive, as it significantly increases the false negative rates.

Thermodynamic stability:

The use of "thermodynamic stability calculations" in miRNA-target prediction is based on the premise that formation of miRNA:target hybrids is more likely if the system gains more energy by forming the duplex than it expends on the creation of that duplex. Hence, miRNA-mRNA duplexes with very small minimum free energy of hybridization (MFE) are considered favorable. One algorithm that uses thermodynamic stability to filter the results of a weighted nucleotide complementariness analysis is miRanda

(Enright et al. 2003). Thermodynamic stability was also used in early versions of TargetScan (Lewis et al. 2003), but removed in TargetScanS.

In praxis, MFE calculations depend on arbitrary estimates for several parameters including temperature, the relative concentration of miRNA and mRNA molecules, and the presence of proteins that facilitate or impede the reaction (Hammell 2010). Furthermore, experimental observations suggest that a strong secondary structure formed by 3' UTR itself will prevent the binding of miRNA.

Independent prediction by several methods:

Some tools, including mirSystem, mirGator and TargetCombo (Sethupathy, Megraw, and Hatzigeorgiou 2006; Lu et al. 2012; Cho et al. 2013) try to combine predictions of existing algorithms, based on the idea that targets that are predicted by several independent methods are more likely to be true positives. Maintaining such methods can be difficult and predictions by such method are not robust to changes in the source methods or changes in cut-off thresholds. For instance, (X. Xu 2007) tried to reproduce the target predictions of a previously published method (Sethupathy, Megraw, and Hatzigeorgiou 2006) by combining the same three computational methods (but possibly different releases of the prediction programs). The two sets of predicted targets for *miR-155* contain 16 and 10 putative targets, respectively and have only one target in common (X. Xu 2007). As mentioned above, naively using the intersection of several methods can be counter-productive, particularly if the underlying methods are based on complementary information. Conversely, the union of predictions by several methods can improve the true positive rate while greatly increasing the false positive rates. These observations lead Alexiou et al to conclude that “In most cases an accurate algorithm is better than a combination of predictions” (Alexiou et al. 2009).

Profiling based target identification

Where simultaneous miRNA mRNA time series expression profiles are available, context-specific, miRNA induced suppression of targets can be inferred by examining anti-correlations between miRNA and mRNA profiles (J. C. Huang, Morris, and Frey 2007).

Similarity to known examples / Machine-learning based methods:

Establishing generalizable rules of miRNA-target interactions is extremely difficult, and such rules are being progressively challenged by genetic and biochemical studies. Changes to such rules have at times dramatic effects on the prediction results. For instance, widely cited estimates for the fraction of protein coding human genes that are subject to regulation by miRNA range from 30% (Lewis, Burge, and Bartel 2005) to 60% (Friedman et al. 2008). Note that both of these estimates originate from the same group (the developers of the TargetScan family of methods). While the discrepancy is partly due to the increased number of identified miRNAs, changes to the putative targeting rules play a major role.

A few algorithms (Yousef et al. 2007; Sturm et al. 2010) try to forgo the establishment of general/hard targeting rules altogether and base their predictions on the "learning from examples principle". Such machine-learning (ML) based methods extract dozens to hundreds of features from experimentally validated miRNA-mRNA pairs and then search for new pairs that exhibit similar features. A notable method is TargetSpy (Sturm et al. 2010), which according to its authors can predict dozens of functional target sites without a seed match per microRNA that are missed by all other currently available algorithms. However, machine learning based methods depend on the quality and amount of data set. In particular, while experimentally verified miRNA-target interactions can be obtained from databases like TarBase (Vergoulis et al. 2012), ML algorithms for miRNA target prediction suffer from the lack of experimentally verifiable negative examples (i.e. cases of non-targeting), as systematic identification of non-target mRNAs is challenging (e.g. current experimental methods cannot detect subtle changes).

1.7. Goal of Present Research

The work presented in this thesis was generously supported by a 4 year Walsh Fellowship from Teagasc, the Irish Food and Agriculture Authority. As such, a primary goal of this work was the creation of Systems Biology tools for investigation of bovine infectious diseases. The completion of a draft of the bovine genome along with falling costs of HT technologies is starting to make 'omics' datasets available for use in animal

health studies. However, the appropriate bioinformatics tools and frameworks for integrating and interrogating these datasets, which would accelerate our understanding of bovine infectious diseases, are still largely missing. A prerequisite for application of integrative SB/Network-Biology based methods to bovine datasets is the availability of bovine protein-protein interaction (PPI) and pathway annotations. Due to the lack of such data, the first section of my research was dedicated to the computational (orthology based) reconstruction of bovine PPI networks and pathways. After completion of this task, I integrated the inferred data into InnateDB, a Systems Biology platform and knowledge base that is being jointly developed at SFU, UBC and Teagasc (Breuer et al. 2013; Lynn et al. 2008; Lynn et al. 2010).

During my analysis of the inferred bovine pathways, I observed that the actual average number of genes per pathway (16) is not reconcilable with the average number of pathway participants that one would expect if pathways were disjoint (4.8). This lead me to examine the implications of component-sharing among pathways for the validity of results of current pathway analysis methods, and resulted in the development of the second thesis goal, a new pathway analysis tool (SIGORA) that focuses on the statistical overrepresentation of pathway specific gene combinations. In a comparative evaluation of several simulated and previously published datasets, SIGORA outperformed several existing, popular pathway analysis methods by delivering biologically more plausible and relevant results (Foroushani, Brinkman & Lynn, 2013).

Another observation during the creation of the bovine InnateDB was the seemingly general lack of knowledge regarding the functional significance of the vast majority of genes in higher organisms. In chapter 4, I present the results of a preliminary (unpublished) attempt at enhancing the functional annotation of bovine genes by applying the guilt by association (GBA) principle to a large bovine issue expression data set that we had obtained through a collaboration with the US Department of Agriculture. A subsequent literature search provided supporting evidence for a considerable portion of the newly predicted annotations, suggesting that despite its limitations, this can be a viable approach.

Lastly, I had the opportunity to partake in the analysis of changes to bovine microRNA expression profiles in response to infectious agents (Lawless et al. 2013;

Vegh et al. 2013). In chapter 5, I used the intersection of the results of one method (miRanda) with the union of two orthogonal approaches (TargetScanS PCT and context+ score) to arrive at target predictions for a set of differentially expressed microRNAs. The subsequent functional analysis of combined targets (by SIGORA and InnateDB's manually curated list of immunity relevant genes) showed significant associations to specific cellular/signaling pathways and concurred reasonably well with known biology of bovine mastitis, suggesting that the aforementioned pitfalls of straightforward combinations of miRNA target prediction methods can be partly avoided by more sophisticated strategies. In chapter 6, I apply the same target prediction strategy to a larger catalog of all microRNAs encountered in the profiling of unchallenged bovine alveolar macrophages, thereby potentially facilitating the interpretation of the follow-up challenge and time-series experiments that are currently in preparation.

The work and analyses reported in chapters 3 and 4 has been carried out by me. In three cases (chapters 2, 5 and 6), a thesis objective is presented within the context of the results of a larger cooperative effort. In those sections, I outline my contributions at the start of the chapter and provide the list of other contributors.

Chapter 2. *InnateDB: systems biology of innate immunity and beyond*

Portions of this chapter have been published in the article "InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation", co-authored by Breuer, Karin; Foroushani, Amir K; Laird, Matthew R; Chen, Carol; Sribnaia, Anastasia; Lo, Raymond; Winsor, Geoffrey L; Hancock, Robert E W; Brinkman, Fiona S L & Lynn, David J in Nucleic acids research, Database issue © The Authors 2012. Sections 2.5, 2.6, 2.7 (and all their subsections) and Appendix A are exclusively my work. Sections 2.5.2, 2.6, 2.7 had not been in the NAR article (of which I am a joint first author).

2.1. Abstract

InnateDB (http://www.innatedb.com) is an integrated analysis platform that has been specifically designed to facilitate systems-level analyses of mammalian innate immunity networks, pathways and genes. In this article, we provide details of recent updates and improvements to the database. InnateDB now contains >196 000 human, mouse and bovine experimentally validated molecular interactions and 3000 pathway annotations of relevance to all mammalian cellular systems (i.e. not just immune relevant pathways and interactions). In addition, the InnateDB team has, to date, manually curated in excess of 18 000 molecular interactions of relevance to innate immunity, providing unprecedented insight into innate immunity networks, pathways and their component molecules. More recently, InnateDB has also initiated the curation of allergy- and asthma-related interactions. Furthermore, we report a range of improvements to our integrated bioinformatics solutions including web service access to InnateDB interaction data using Proteomics Standards Initiative Common Query Interface, enhanced Gene Ontology analysis for innate immunity, and the availability of new network visualizations tools. Finally, the recent integration of bovine data makes InnateDB the first integrated network analysis platform for this agriculturally important model organism.

2.2. Introduction

The innate immune response is a critical branch of immunity, which not only provides a first line of defence against pathogens, but also regulates and shapes subsequent adaptive responses. Innate immunity, however, can also do great harm by

driving inappropriate inflammatory cascades. Therefore complex molecular networks are required to regulate innate immunity and maintain appropriate and specific responses to different pathogens, while limiting potential harm from dysregulated inflammation (Delano et al. 2011; Karin, Lawrence, and Nizet 2006; Lin and Karin 2007; Shi, Ljunggren, and Sarvetnick 2001; Wen et al. 2008). The intricate interplay of a multitude of regulatory layers that initiate and coordinate the innate immune response has led to an ever-increasing interest in applying systems-oriented approaches to better understand innate immunity and its modulators (Gardy et al. 2009).

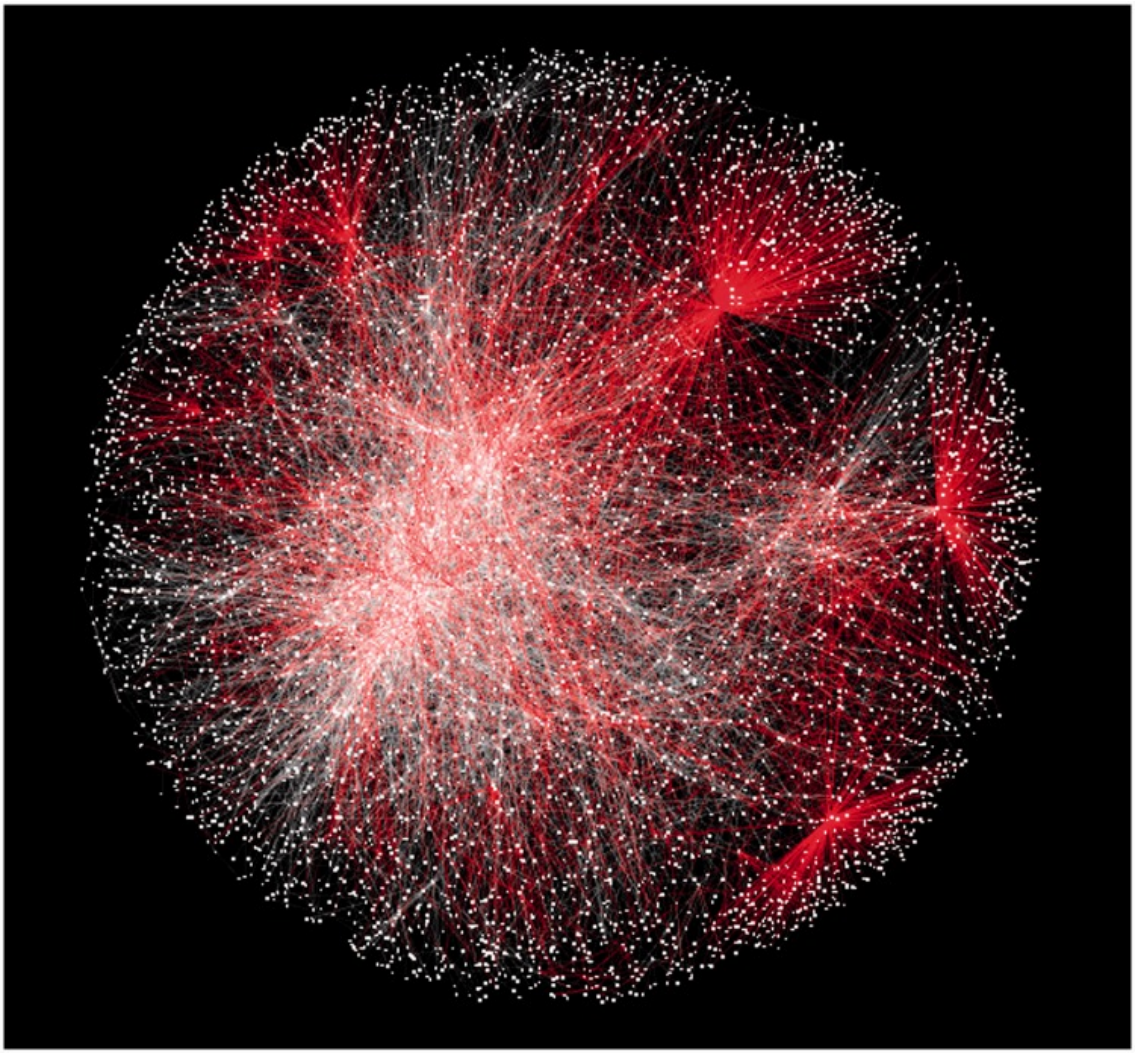
InnateDB (publicly accessible at http://www.innatedb.com and mirrored at http://innatedb.teagasc.ie) is a knowledge base and analysis platform that was specifically designed to provide a system-oriented yet user-friendly tool for integrative analyses of the mammalian innate immune response(Lynn et al. 2008).

2.3. InnateDB Curation

A key component of the InnateDB project is the contextual manual curation of innate immunity interactions, pathways and their component molecules. The curation process has previously been described in detail (Lynn et al. 2010). InnateDB was first publicly released in 2008 (Lynn et al. 2008). At that time, ~3500 molecular interactions had been curated. By 2010, the database contained 11 786 InnateDB-curated molecular interactions from the review of >3 000 published articles. As of September 2012, our curation team has reviewed >4 300 publications, and >18 000 interactions of relevance to innate immunity have been annotated (Figure 1; for detailed statistics see http://www.innatedb.com/statistics). More recently, as part of the AllerGen Networks of Centres of Excellence (NCE) (http://www.allergen-nce.com), InnateDB curators have also begun to annotate interactions and pathways of relevance to allergy and asthma. All interactions in InnateDB are provided with detailed contextual information according to the minimum information required for reporting a molecular interaction experiment (MIMIx) standards (Orchard, Salwinski, et al. 2007) ,including the evidence supporting each interaction, the tissue or cell type the interaction was reported in, the type of interaction and the method of detection. New interactions are added to the database weekly, providing up-to-date annotation on the innate immunity interactome. This

resource can be mined to identify new relationships between innate immunity and other processes, to identify potential novel regulators of innate immunity and to interpret a user's own data (e.g. gene expression data) from a network biology perspective.

Figure 2-1 The InnateDB curated interactome in July 2012.



Red edges represent interactions that have been added in 2011 and 2012. Note: this figure was featured on the NAR database issue cover.

2.3.1. Building a comprehensive list of innate immunity genes

Aside from annotating molecular interactions, InnateDB now also annotates which genes have a published role in innate immunity, providing a brief summary of that

role and links to the relevant publications. This data set, available at http://www.innatedb.com/curatedGenes, presently contains >1 500 genes (957 human, 527 mouse and 46 bovine) and is the most comprehensive list of genes involved in innate immunity that is available. This list was recently used by a group of researchers to show that human proteins which are targeted by viruses are highly enriched for having a role in innate immunity (Pichlmair et al. 2012).

2.3.2. Contribution to the International Molecular Exchange Consortium

In 2010, InnateDB became a member of the International Molecular Exchange Consortium (IMEx) (Orchard et al. 2012). This organization is dedicated to developing rules for describing molecular interaction data, actively curating these interactions from the scientific literature and making them available through a common website.

Within IMEx, InnateDB has committed to curating every issue of *Nature Immunology* from September 2010 onwards using IMEx curation standards (Orchard, Kerrien, et al. 2007). Because IMEx curation requires more annotation detail than the MIMIx level (Orchard, Salwinski, et al. 2007) currently supported by InnateDB's submission system, InnateDB curators are submitting these IMEx interactions through the IntAct interaction database (Kerrien et al. 2012). On submission, each IMEx interaction is thoroughly reviewed by an IntAct curator before it is accepted and released. In addition to submitting to IntAct, all InnateDB acceptable interactions (i.e. interactions of relevance to innate immunity) from *Nature Immunology* are also deposited into InnateDB.

2.4. Integrating data from external resources

To supplement our manual curation efforts and to provide a snapshot of the entire interactome beyond known innate immunity interactions, InnateDB imports data from a wide range of genome, interaction and pathway databases (http://www.innatedb.com/resources). Currently, InnateDB contains 178 000+ imported experimentally validated interactions, 3000+ pathways and 300 000+ interactions based

on Ortholuge (Fulton et al. 2006) orthology predictions (interologs) in addition to the 18 000+ InnateDB manually curated interactions.

2.5. Integration of Bovine Data - Orthology based Pathway and Network Reconstruction

In February 2012, a new version of InnateDB was released that included the incorporation of bovine gene, pathway and molecular interaction annotation in addition to the existing data for human and mouse. This new version of the platform now also facilitates a systems biology approach to the investigation of the bovine innate immune response and is poised to deepen our understanding of important bovine infectious diseases associated with significant economic losses (e.g. bovine tuberculosis and mastitis), as well as enabling cross-species comparisons of innate immunity.

As bovine experimentally validated interactions and pathways are virtually non-existent, InnateDB uses an orthology-based approach to predict bovine pathways and interactions primarily from human data. One should be aware that this approach results in a humanized and frequently incomplete representation of the bovine interactome, but in the absence of widespread experimental data it provides at least a network biology framework to build on and to generate hypotheses that can be subsequently experimentally validated. InnateDB experimentally validated and predicted interactions are clearly labelled. As of September 2012, InnateDB contained >70 000 bovine interologs (interactions based on orthology) involving 10 717 bovine genes. In each case, one can link back to the orthologous human interaction to review evidence for the interaction.

The latest release of InnateDB also uses orthology predictions to transfer human and mouse pathway annotations to bovine genes in real time. Currently, pathway annotations can be assigned to 7032 bovine genes by orthology to human genes. Notably, although only ~70% of all human genes (14 316 genes) have a predicted bovine ortholog, a significantly higher proportion (85%) of human genes with pathway annotations have a bovine ortholog. This higher prevalence of conserved genes among

pathway-annotated genes indicates that many of the associated processes may be well preserved.

2.5.1. Variability of pathway conservation

To further examine the appropriateness of the orthology-based annotation transfer on a per-pathway basis, we determined the 'conservation rate' (*cons*) of each pathway, defined as the ratio of pathway participants in the source organism (human/mouse) that have a putative counterpart in the target organism (cow) to the total number of participants in the source organism. As of September 2012, InnateDB contains 1536 human pathways with five or more pathway participants, 80% (1257 pathways) of these have a conservation rate of 0.8 or better. The corresponding number for a conservation rate of ≥0.7 is 93% (1442 pathways). The high prevalence of strongly conserved pathways seems to largely justify an orthology-based approach for inferring bovine pathways. Appendix A lists the remaining 107 pathways with a relatively low conservation rate (*cons* <0.7). Notably, the list of pathways for which an orthology-based reconstruction is challenging includes >30 immunologically important pathways. In some cases, the low conservation rate can be attributed to real divergence of the underlying processes. The bovine Type I Interferon family, for example, has been shown to have undergone widespread expansion, including the divergence of a new Type I interferon (IFN) family (IFNX) in the cow from IFN alpha (Walker and Roberts 2009). In other cases, the conservation rate might further increase with future improvements to the quality of the bovine draft genome.

2.5.2. Pathway conservation and cellular localization

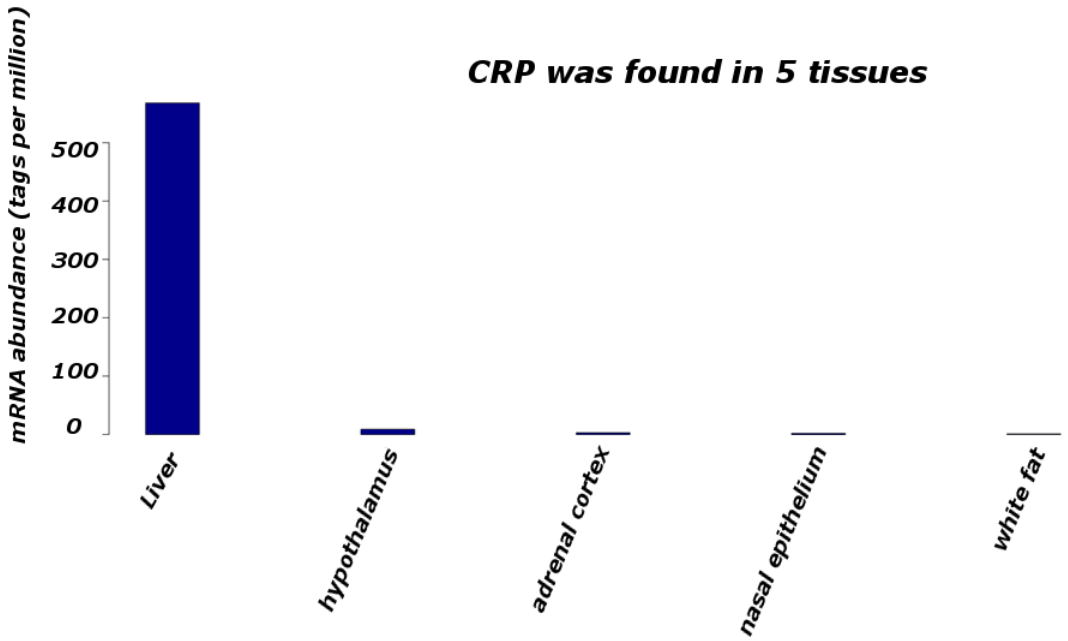
A closer look at the moderately or poorly conserved pathways suggests a distinct trend in the cellular localization of the non-conserved members of these pathways. Overall, there were 1,513 human genes with pathway annotations but without any predicted bovine orthologs. Gene Ontology (Ashburner et al. 2000) Cellular_Component (GO_CC) analysis of these genes revealed a strong enrichment for the terms "plasma-membrane" (509 genes, FDR 4.26E-50) and "extra-cellular region" (260 genes, FDR 3.33E-28). This trend is even more marked for individual signaling pathways. For instance, the human "Toll-like receptor signaling pathway" in KEGG (M Kanehisa and

Goto 2000) contains 102 distinct members, 68 (67%) of which have a predicted bovine ortholog. The vast majority (30 out of 34) of the non conserved members of this pathway localize to either the “plasma membrane” or the “extracellular region”. Furthermore, 16 of these 30 genes are interferon receptors, which in turn are likely to diverge from human receptors due to aforementioned changes in the bovine interferon repository (reported by (Walker and Roberts 2009)). Overall, the above observation is in concordance with the results of a very recent (June 2014) study (M. H. Schaefer et al. 2014) which determined that for signaling pathways, the input layer (ligands and receptors) tends to be significantly less conserved across species.

2.5.3. Tissue expression and function

In addition to orthology-based annotation transfer, the tissue expression profile of a gene can provide some insight into its potential function (Dezso et al. 2008). Through collaboration with colleagues at the United States Department of Agriculture, InnateDB now integrates bovine tissue expression data for >13 000 genes. These data were sourced from the Bovine Gene Atlas (Harhay et al. 2010), which has profiled gene expression across 87 different bovine tissues using a next generation sequencing approach. A graphical tissue expression profile is available on the Gene Card page of bovine genes (Figure 2-2).

Figure 2-2 An example for graphical tissue expression profiles in InnateDB



According to the corresponding gene card on InnatDB, C-reactive protein (CRP) is a major acute phase protein. It is involved in response to inflammatory stimuli (complement pathway, macrophage differentiation) and also plays a role in metabolism (regulation of lipid storage/ leptin pathway). This information concurs with the expression profile.

2.6. Analysis of InnateDB's protein interaction networks

Genome scale biological networks leave many powerful details of individual interactions aside, but despite or maybe because of these simplifications, they can help in identifying possibly important aspects and general trends of these interactions (Vidal, Cusick, and Barabási 2011). A biological network can be analyzed with respect to several local and global topological properties, including degree, clustering coefficient, topological coefficient and path lengths (Assenov et al. 2008). If substantially different from randomized models, such properties can then be related back to a better understanding of biological processes. The structure and evolution of biological networks has been shown to follow basic organizing principles (Vidal, Cusick, and Barabási 2011). The following sections discuss some notable topological properties of InnateDB's interactome.

2.6.1. Connectivity analysis of the human interaction network in regard to conservation in cow

As of January 2014, InnateDB contains about 160, 000 unique binary human interactions involving ca. 17, 700 human genes. Close to 14, 000 (78%) of the interacting genes have a predicted bovine ortholog and approximately 120, 000 of the interactions (75%) can be reconstructed by orthology.

The proportions of conserved nodes (78%) and (putatively) conserved interactions (75%) seem misleadingly similar, but it is worth noting that the proportion of (putatively) conserved edges is decidedly higher than what one would expect in a randomized model: under the assumption that node degrees are normally distributed and nodes are randomly selected for conservation, one would expect only 60% (0.78^2) of the binary interactions to be conserved. As described below, the assumptions of a randomized model are unjustified:

A) The median degree of nodes in the source (human) network is 5, while the mean is 18.05. (If the number of interaction partners in human interaction network was normally distributed, the mean and the median of node degrees would be both close to 18.) In the actual network, about a third of the nodes have only one or two interaction partners, while 6 genes have a degree larger than 1000. This is consistent previous observations on degree distribution in human protein interaction networks (Stelzl et al. 2005) and in line with the postulated power-law distribution of degrees in biological networks (Barabási, Gulbahce, & Loscalzo, 2011; Barabási & Oltvai, 2004).

B) Human genes with a predicted bovine ortholog (i.e. conserved, evolutionary older genes) tend to have more interaction partners within the human PPI than non-conserved genes (Welch t-test p value of 0.0016). The median number of interaction partners for conserved genes is 7, while the median degree of non-conserved genes is 2. This is consistent with the idea of network growth by preferential attachment, which postulates that recently added nodes (e.g. lineage-specific and duplicated genes) tend to preferentially interact with already established, well-connected nodes.

2.6.2. Confounding issues in application of interologs

Notably, although the subset of human genes without a predicted bovine ortholog is highly enriched in nodes having small degrees in the (original) human interaction network, it also contains Ubiquitin C (UBC) and Small Ubiquitin-Like Modifier 2 (SUMO2). In the original network, UBC is the by far most prominent hub (involved in more than 9,600 interactions) and SUMO2 is the fourth largest hub (involved in over 1,200 unique interactions). The prominence of UBC in the human interaction network is consistent with the widely accepted notion that ubiquitin is one of the evolutionary most conserved proteins, but the lack of a predicted bovine ortholog for UBC accounts for about a quarter of all non-reconstructible interactions.

The case of UBC illustrates some of the confounding issues in high-throughput application of interologs that relate to the incoherencies of the "namespace maze" of biological identifies. All analysis results reported in this chapter are based on Ensembl genes and orthology predictions by Ortholuge. Using the Ensembl sequence for UBC as an input, currently neither Ortholuge (Fulton et al. 2006) nor reciprocal best-BLAST-hit method (RBH) (Altschul et al. 1990) can detect a bovine ortholog of this gene. The HGNC database (Gray et al. 2013) (http://www.genenames.org/cgi-bin/hcop?species_pair=Human+and+Any+species&column=symbol&Search=Search&query=UBC), on the other hand, reports a bovine ortholog for UBC, based on the Entrez gene for human UBC (Entrez gene 7316) and RBH. The Entrez ID of this putative bovine ortholog (Entrez 444874), however, maps to two Ensembl bovine genes on two different loci: ENSBTAG00000032436 (Chromosome 17: 53142511-53143425, forward strand) and ENSBTAG00000017246 (Chromosome 19: 33853788-33855685, reverse strand).

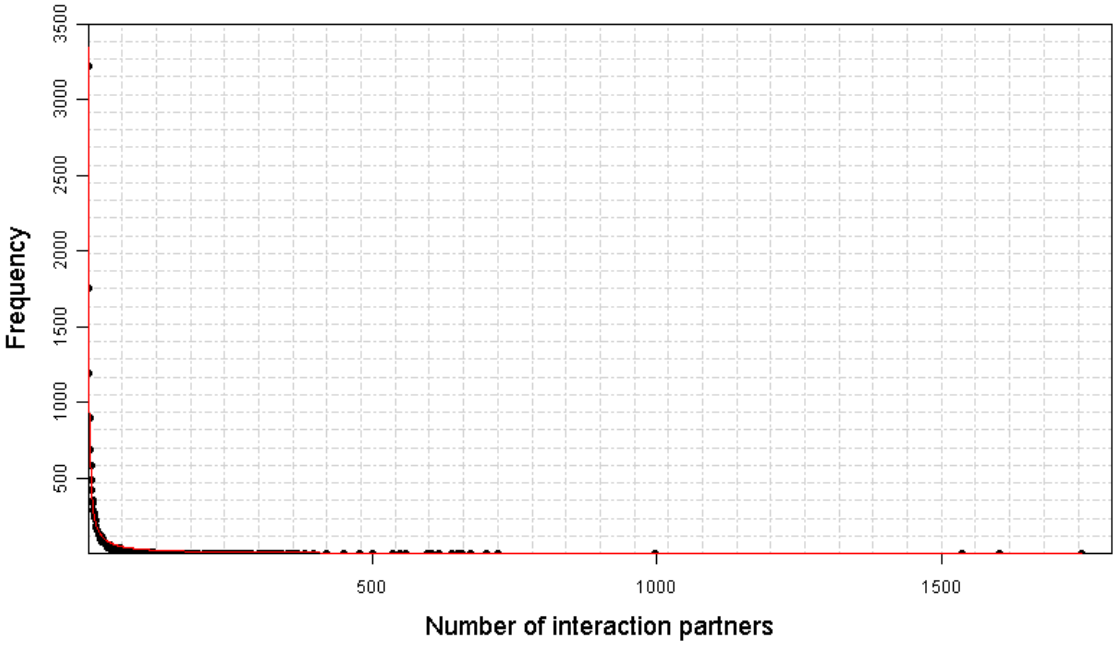
Further complicating matters, there are some inconsistencies within the same repository. For instance, most repositories list UBC as a synonym of UBB, but do not list UBB as a synonym of UBC. Despite these issues (inconsistency of nomenclature for some individual genes), the general coherence of the above results suggest that the inferred network can be an overall useful starting point.

2.6.3. Analysis of the inferred bovine interaction network

The inferred bovine interaction network contains a giant component (largest connected sub-graph) of 13, 904 nodes. This component contains the vast majority of the nodes and edges (all but 29 of the nodes and all but 28 of the edges), and exhibits several known properties of biological networks.

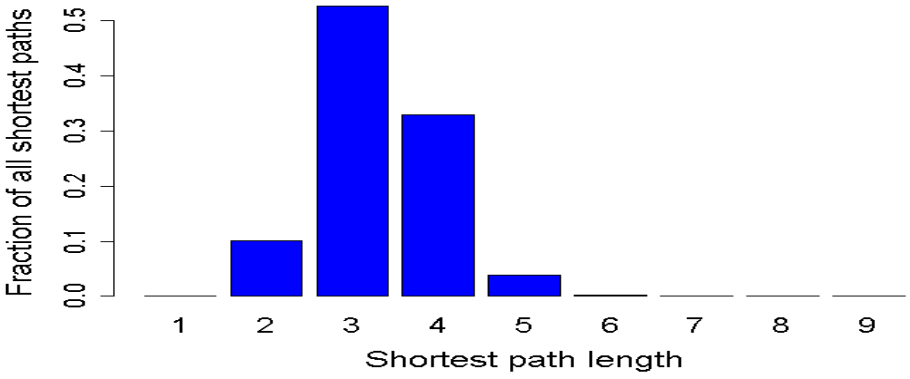
The network is scale-free, i.e. the number of interaction partners per node follows a power-law distribution (the vast majority of the nodes have only few interaction partners, while a few hubs are extremely well-connected) (Figure 2-3). Further, the network exhibits the “small word” property: every node in the giant component can be reached from every other node by crossing only few intermediate nodes, with an average shortest path length of 3.14 (Figure 2-4). Finally, the topological coefficient of a node (a relative measure for the extent to which a node shares neighbors with other nodes) in this network decreases exponentially with the increase in its degree (the number of its neighbors) (Figure 2-5). All of these finding are directly comparable to previous observations in human interaction networks (cf. Figure 3 in (Stelzl et al. 2005)). These similarities are to be expected, since the bovine network is reconstructed from a human one.

Figure 2-3 Degree distribution in the inferred interaction network



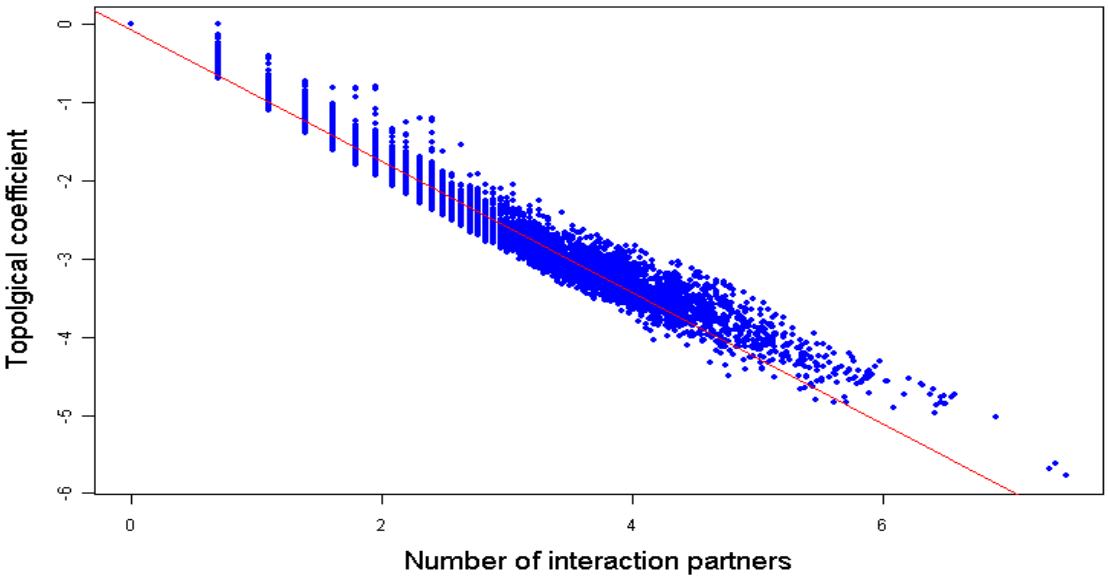
Number of nodes with a given number of interaction partners (k) in the network approximates a power-law. The red line shows the fitted curve frequency(k)=a* k^{γ}; with a=3431 and γ = -1.04, coefficient of determination (R^2=0.99), p-value<2*e-16.

Figure 2-4 Distribution of the length of shortest paths in the inferred network



In over 96% of cases, there are less than 3 intermediate nodes between any two nodes in the network.

Figure 2-5 Dependency of topological coefficient on node degree



Degree and topological coefficient are shown to logarithmic scale. The topological coefficient of a node *n* is calculated as average(*J(m,n)*)/degree(*n*), where J(*m,n*) is the number of neighbors that node m has in common with n, plus one if m and n are also directly connected (J(*m,n*) is only defined only if share at least one neighbor). The red line has an slope of -0.85, i.e. the decline in the coefficient is approximately proportional to the reciproc of the degree.

2.6.4. Analysis of the interaction network of innate-immunity related genes in the conserved network

At the time of this writing, InnateDB curation team has identified 947 human genes with high relevance to innate immunity. Bovine orthologs of approximately 72% of these genes (641 genes) occur in the inferred interaction network and are involved in 31,974 predicted interactions (25% of all inferred interactions). The number of interactions only involving bovine orthologs of innate immunity related genes is 4,259. Compared to the background of all genes in the inferred network, these conserved innate-immunity related genes tend to have a higher number of interaction partners: median degree of innate immunity related genes in the inferred network is 25, while the median number for all genes in this network is 5. They also have smaller average

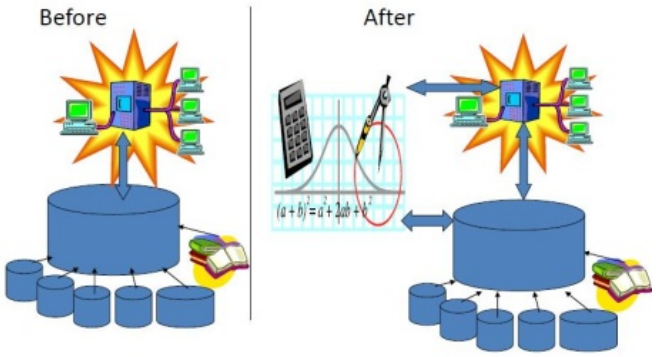
topological coefficients (0.06 vs. 0.2) and smaller clustering coefficients (0.13 vs. 0.17). As these two coefficients are measures for the modularity of the network modularity and functional commitment of the nodes, respectively, these findings might reflect the context sensitivity of interactions involving innate immunity related genes.

2.7. The mirror at innatedb.taegasc.ie

Since 2011, a new European mirror of InnateDB (based at Teagasc facilities in Ireland and publicly available at innatedb.teagasc.ie) has been on line, to provide faster access for users outside North America. Like the original server, this mirror runs on two dedicated multi-core linux servers (one hosting the backend database and one hosting the web-server) and deploys of a variety of technologies, including Apache Tomcat servlet container, Apache HTTP web server, MySQL database server, JavaServer Pages, Apache Struts Framework ad CakePHP.

As part of the work on integration of bovine data into InnateDB, an additional, experimental version of the mirror was also created, which differs from the public version by the integration of the R statistical framework. On this experimental version, bovine tissue expression barplots are not stored offline, but are rendered on demand and embedded into bovine gene-cards '*on the fly*'. When a user requests a bovine gene card, the server first determines if a tissue expression barplot for the gene in question already exists. If this is not the case, the web-server sends a request to an R-daemon to run an R script which generates this graphic and stores it as a BLOB (binary large object) in the MySql database. Once the R-script has completed, the daemon notifies the web-server, which then retrieves the image along with the other necessary data from the database, generates the gene card with the embedded graphic and delivers it to the user. The integration between Java code (web-server) and R is established through the Rserve package (http://www.rforge.net/Rserve), which is currently only available as a beta-release. Due to the still experimental nature of Rserve, this architecture is not currently deployed in the publicly available mirror, but this work demonstrates the potential that future releases of InnateDB could harness the vast array of available R/Bioconductor analysis packages to significantly enhance the range of analysis services that InnateDB can offer (Figure 2-6).

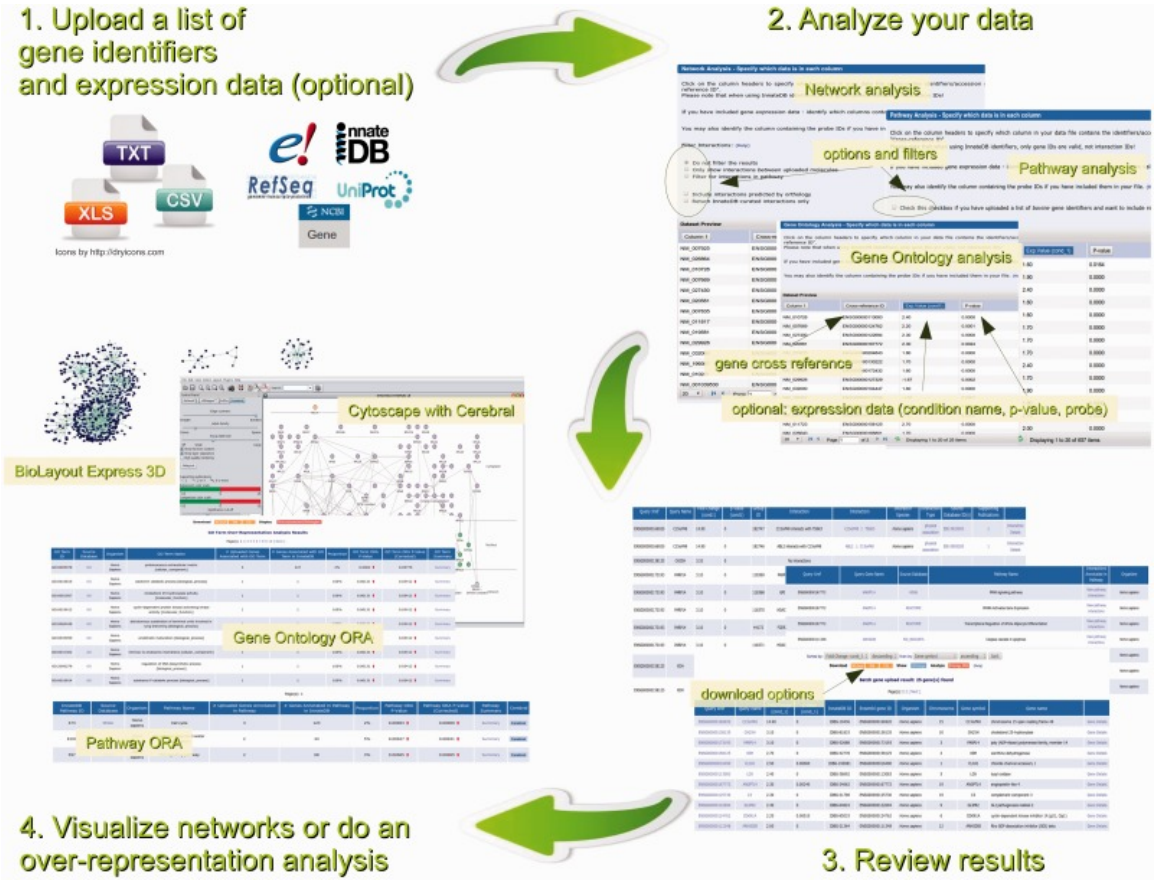
Figure 2-6 R integration on the test-version of the mirror



2.8. InnateDB Data Analysis and Visualization

InnateDB can serve as a knowledge base where users can search for particular genes or proteins of interest and their associated interactions and pathways, using a variety of search fields. Alternatively, InnateDB can be queried in a more high-throughput fashion, where users can upload a list of genes/proteins and associated quantitative data from up to 10 different conditions (e.g. gene expression data) and carry out more complex searches and analyses (Figure 2). After uploading a list of gene IDs (human, mouse and bovine Ensembl, RefSeq, Entrez Gene, UniProt and InnateDB IDs are all accepted), users can quickly find which pathways, Gene Ontologies (including enhanced innate immunity gene annotation) or transcription factor binding sites are statistically over-represented in their dataset. Users can also use InnateDB to build, visualize and analyse molecular interaction networks consisting of their uploaded genes and their encoded products. One can, for example, construct a network of how differentially expressed genes interact with one another. Quantitative data uploaded by the user is automatically overlaid on these networks. Recent improvements to InnateDB include the option to incorporate interactions based on orthology in the construction of molecular interaction networks and the option to restrict the networks to contain only InnateDB manually curated interactions. Further enhancements to the web-interface include more intuitive page layouts, faster searches and analyses, and a variety of other changes (see http://www.innatedb.com/news).

Figure 2-7 Data analysis workflow in InnateDB.



2.8.1. Network visualization tools

All interactions in InnateDB may be downloaded in several standardized formats including text-based formats (tab, csv, xls), the simple interaction format (sif) and the PSI-MI XML 2.5 and MITAB formats (Kerrien et al. 2007). Additionally, interaction networks may also be visualized in our Cerebral program (Barsky et al. 2007), a Java plugin for the Cytoscape network visualization software (Shannon et al. 2003; Smoot et al. 2011) , which uses subcellular localization information to orientate interaction networks in a more biologically intuitive pathway-like layout. Networks can also be visualized in other third-party software including the CyOog plugin (Royer et al. 2008), which uses Power Graph analysis to reduce network complexity by explicitly representing re-occurring network motifs. Recently, we have also integrated BioLayout

Express 3D 2.2 (Theocharidis et al. 2009), an application designed for the visualization, clustering and analysis of large networks in 2D and 3D space.

2.8.2. Proteomics Standards Initiative Common Query Interface implementation

Interaction data in InnateDB can now also be queried using web services implementing The Proteomics Standards Initiative Common Query Interface (PSICQUIC). PSICQUIC is an effort from the Human Proteome Organization Proteomics Standards Initiative (http://www.hupo.org/research/psi/) to standardize programmatic access to molecular interaction databases based on the PSI standard formats (PSI-MI XML and MITAB) (Kerrien et al. 2007). It defines standard web services and also a query syntax for powerful and flexible searches.

All data sources implementing PSICQUIC can be queried in the exact same way, i.e. the same query can be used to retrieve the relevant data from many different interaction data sources. Independently published observations of an experimental system, curated by independent databases, are then integrated in response to a single user query (see http://www.ebi.ac.uk/intact/imex). PSICQUIC web services are RESTful (REpresentational State Transfer) but can also be accessed through SOAP (Simple Object Access Protocol). A list of available services for InnateDB can be found at http://imex.innatedb.com/psicquic-ws/webservices. InnateDB updates the data files for the PSICQUIC web services weekly and additionally provides them for download in a compressed format at http://www.innatedb.com/downloads.

2.9. Ongoing Developments

InnateDB will maintain its curation efforts to annotate interactions and genes of relevance to innate immunity, with weekly updated annotation, thus continuing to provide a comprehensive platform for systems and network biology analyses of innate immune-associated responses. Continued incorporation of data from external resources, encompassing the wider human, mouse and bovine interactomes, will also continue to facilitate analyses beyond innate immunity by a wide range of researchers. Additionally,

InnateDB intends to expand beyond the curation of innate immunity relevant networks, incorporating more adaptive immunity information. We are currently developing a first version of an Allergy and Asthma Portal that will further integrate data on allergy and associated immune interactions from both the literature and researchers from AllerGen. This portal will be built on InnateDB and will provide an analysis platform for more sophisticated network biology-based investigations of allergy and asthma responses. These interactions will be identifiable from innate immunity interactions, so that users can continue to have focused analyses on the innate immunity interactome.

Further future developments will include improvements to InnateDB pathway analysis tools. The over-representation–based methods for pathway analysis that are currently available through InnateDB's data analysis interface are widely established and considered a 'gold standard'; yet, they neglect the fact that many components are shared between seemingly unrelated pathways. To address this issue, we have used the InnateDB collection of pathway annotations as a basis to identify pairs of genes that co-occur only in a single pathway and developed a novel pathway analysis method [signature over-representation analysis (SIGORA)] that focuses on the over-representation of such gene pairs in a list of genes of interest (Foroushani, Brinkman, and Lynn 2013). SIGORA is currently implemented as an R package [available from http://sigora.googlecode.com/svn/] and will be integrated into the future releases of InnateDB.

Finally, together with the PSICQUIC development team and other IMEx members, we are working on an improved reference implementation of PSICQUIC. We are also preparing to export our data in MITAB 2.7 format.

Chapter 3. *Pathway-GPS and SIGORA: identifying relevant pathways based on the over-representation of their gene-pair signatures*

Portions of this chapter have been published in the article "Pathway-GPS and SIGORA: identifying relevant pathways based on the over-representation of their gene-pair signatures", co-authored by Foroushani, Amir K; Brinkman, Fiona S L & Lynn, David J in PeerJ © The Authors 2013. I performed all analyses for this paper, with the support of my thesis co-supervisors which are co-authors.

3.1. Abstract

Motivation: Predominant pathway analysis approaches treat pathways as collections of individual genes and consider all pathway members as equally informative. As a result, at times spurious and misleading pathways are inappropriately identified as statistically significant, solely due to components that they share with the more relevant pathways.

Results: We introduce the concept of Pathway Gene-Pair Signatures (Pathway-GPS) as pairs of genes that, as a combination, are specific to a single pathway. We devised and implemented a novel approach to pathway analysis, Signature Over-representation Analysis (SIGORA), which focuses on the statistically significant enrichment of Pathway-GPS in a user-specified gene list of interest. In a comparative evaluation of several published datasets, SIGORA outperformed traditional methods by delivering biologically more plausible and relevant results.

Availability: An efficient implementation of SIGORA, as an R package with precompiled GPS data for several human and mouse pathway repositories is available for download from http://sigora.googlecode.com/svn/.

3.2. Introduction

Pathway analysis identifies biological pathways that are statistically enriched in a given dataset and plays a crucial role in the interpretation of high-throughput experimental datasets including gene or protein expression profiles (Khatri and Drăghici 2005; D. W. Huang, Sherman, and Lempicki 2009) and genome-wide association

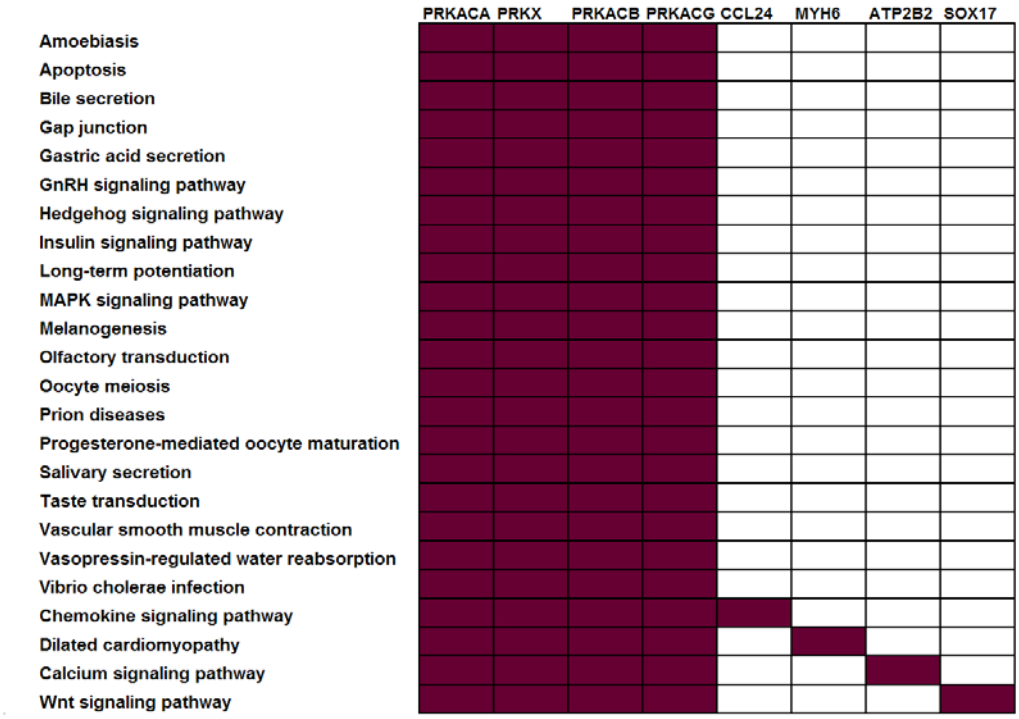
studies (GWAS) (K. Wang, Li, and Hakonarson 2010). Pathway analysis can guide the understanding of complex biological datasets through the statistical association of observations at the molecular level to processes at the systems level. Such analysis can, for example, highlight processes that are dysregulated in certain pathological conditions, such as cancer (Copeland and Jenkins 2009) or infection (Wherry et al. 2007) .

Currently, two types of pathway analysis methods are widely used: Over-representation Analysis (ORA) methods (reviewed in Khatri and Drăghici 2005) and methods related to Gene Set Enrichment Analysis (GSEA) (Mootha et al. 2003; Subramanian et al. 2005; Dinu et al. 2009).

Despite major differences between ORA and GSEA methods (see e.g. (Emmert-Streib and Glazko 2011) for a discussion), these approaches share a notable limitation: most current methods treat all genes in a given pathway as equal indicators that that pathway is significant. This assumption, that each gene in a pathway has the same power to distinguish one pathway from another, and that genes assume their roles without consideration of the context and expression of other genes, is undoubtedly flawed (Gillis and Pavlidis 2011; J. Ma, Sartor, and Jagadish 2011; Khatri, Sirota, and Butte 2012).

To illustrate this point, consider four protein kinases, PRKACA, PRKACB, PRKACG, and PRKX. Within the KEGG (Minoru Kanehisa et al. 2011b) pathway repository, these genes are members of 24 different pathways, i.e. they co-occur in roughly 10% of KEGG human pathways (Figure 3-1). Consider a dataset where all four of these genes were observed to be differentially expressed - many pathway analysis tools would identify all 24 different pathways as statistically significant leaving the biologist perplexed as to which of these pathways are the most biologically relevant to their study. The underlying problem (that genes may be associated with multiple pathways and, as such, that all genes are not equivalent “Signatures” of a given pathway) is widespread and not limited to kinases. Within KEGG, 52% of genes are annotated in more than one pathway (Table 3-1).

Figure 3-1 Not all genes have the same power to distinguish between different pathways.



In this example, all current KEGG annotations of seven selected genes are shown. Red: annotated in pathway; white: not annotated in this pathway.

As a result, many pathway analysis methods return misleading statistically significant pathways that are significant solely due to shared components with other pathways (e.g. "*Prion Disease*" is identified as a significant pathway in a dengue fever microarray study (Hoang et al. 2010) simply because many of the genes annotated in this "pathway" are co-annotated in inflammation-related pathways).

Table 3-1 Annotation and co-annotation of human genes in current pathway repositories.

Repository	annotated genes	Genes with a single pathway annotation in the Repository	Genes with multiple pathway annotation in the Repository	co-annotated gene-pairs	Gene Pairs that co-occur in a single pathway
KEGG	5,660	48%	52%	1,205,807	90%
REACTOME	5,046	62%	38%	197,034	87%
PID_BIOCARTA	1,368	54%	46%	32,361	78%
PID_NCI	2,374	51%	49%	116,852	87%

The number of human genes annotated in pathways in the KEGG, Reactome and PID databases. On average more than 40% of genes are annotated in more than one pathway whereas gene-pairs rarely co-occur in multiple pathways.

Here, we report a novel approach to address this problem, which involves the identification of statistically over-represented Pathway Gene-Pair Signatures (Pathway-GPS) (i.e. weighted pairs of genes which uniquely occur together in a single pathway). The use of such gene pairs is also motivated by the data in Table 3-1: in contrast to single genes, co-annotated gene pairs tend to be specific to a single pathway. We provide an implementation of this approach in R (SIGORA; downloadable from http://sigora.googlecode.com/svn/). We describe this approach and demonstrate how SIGORA significantly reduces the identification of spurious pathways in analyses of simulated and real biological datasets.

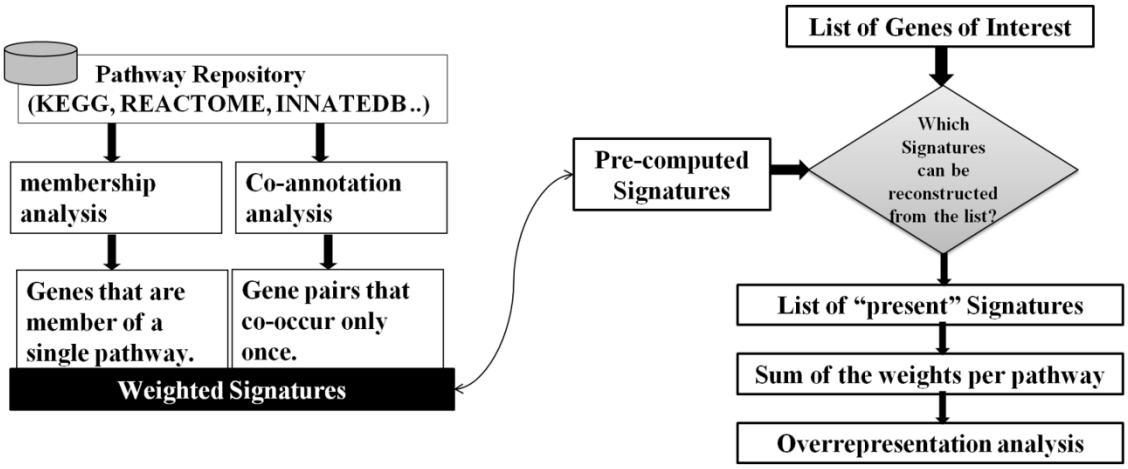
3.3. Materials and Methods

3.3.1. Algorithm

As illustrated in Figure 3-2, our approach to the problem consists of two phases: In an offline phase, we compile a set of weighted markers/Signatures for each pathway in a repository, which we call Pathway Gene-Pair Signatures (*Pathway-GPS*). Subsequently, in an online phase, the method identifies the statistical over-

representation of such Signatures in a user-specified gene list using an adapted version of the hypergeometric test.

Figure 3-2 SIGORA's two phases.



In the off-line phase (left) a pathway repository is transformed to disjoint sets of weighted GPS. These precompiled signatures are used in the on-line phase (right) to evaluate a user specific input gene list.

Given a pathway repository (e.g. KEGG), for each gene-pair in a pathway, SIGORA investigates the co-appearance of the two genes in other pathways of the repository. A gene-pair that uniquely occurs in a single pathway is considered a Signature of that pathway and is assigned a weight. The weight of a Signature (from [0,1]) quantifies the average commitment of the components of the GPS towards the common pathway, i.e. the weight scores the reliability of the Signature as evidence for the associated pathway. For hierarchically organized repositories (like REACTOME (Matthews et al. 2009)), this process is repeated iteratively after the removal of pathways on the top level of the repository, i.e. in each iteration, new weighted Signatures are identified for the pathways on the lower, more specific levels of the hierarchy. Once this offline stage is completed, the resulting sets of weighted gene-pairs that represent each pathway are non-overlapping and can be *re-used* for pathway analysis of any user-specified gene lists.

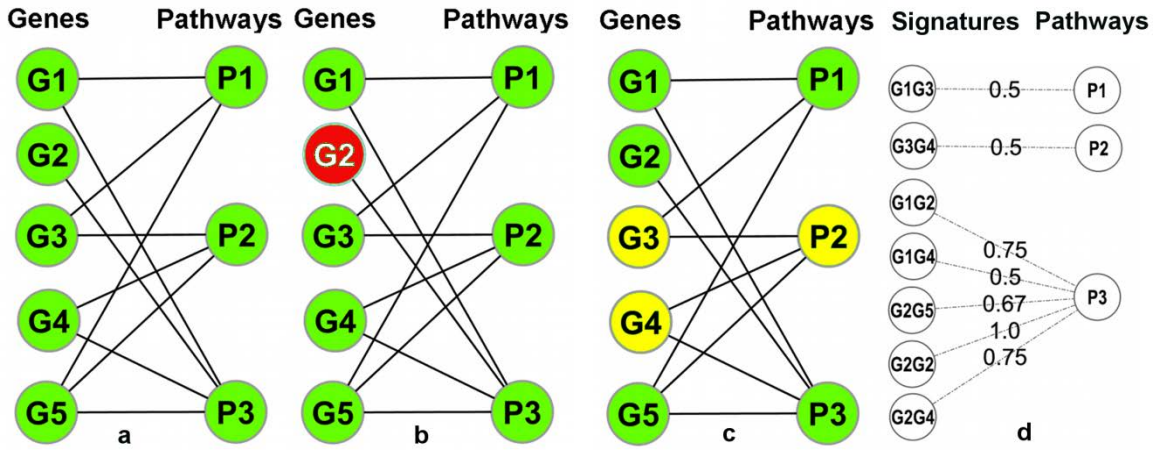
When presented with a gene list of interest (e.g. genes that are differentially expressed), SIGORA determines which of the pairs from its (pre-compiled) Signature repository can be reconstructed from the genes in the list. A Signature is considered

"present" only if both of its constituent genes are found in the user-specified query list. This inherently leads to the selection of the more relevant roles of a gene in the experimental context, as SIGORA relies on the status of the other genes in the pathway for the reconstruction of the Signatures. For each pathway, the weights of present Signatures are summed up and hypergeometric probabilities are used to assess the statistical significance of the observed Signature sets.

Pathway Gene-Pair Signatures (Pathway-GPS)

A pathway database/repository contains (at least) two types of entities: pathways and genes. This can be represented by a bipartite graph (or bipartite network) *B* = (*Vg*, *Vp*, *E*) with two distinct sets of nodes (*Vg*: gene nodes *and Vp*: pathway nodes) where the edges in *E* connect the genes to the pathways and signify the annotation of a gene in a particular pathway (Figure 3-3, panel a). In this graph, the degrees (number of incident edges) of the pathway nodes correspond to pathway sizes (i.e. the number of genes annotated in the pathway) and the degree of the gene nodes corresponds to the number of different pathways a gene is annotated in. In particular, genes with degree one are exclusively annotated within a single pathway (‘*Pathway Unique Genes (PUGs)*’).

Figure 3-3 Overview of the signature transformation.



A schematic pathway repository as a bipartite graph (B); G1…G5: genes; P1…P3: pathways. (B) A pathway unique gene. (C) A Gene-Pair Signature (GPS): G3 and G4 co-occur only in P2. (D) Each GPS is associated with a single pathway and has a weight equal to the average inverse degree (in B) of its constituent genes.

Using the igraph (Csardi and Nepusz 2006) package of the R-Statistical Framework (R Development Core Team 2008) , *B* can be manipulated and transformed. A weighted one-mode projection (M. E. Newman 2001) of *B* into the gene dimension, yields a new graph *Proj*=(*Vg, E', W*) with only one type of nodes (only gene nodes *Vg*), where two genes are connected by an edge in *E'* if (and only if) they are co-annotated in one or more pathways, and each edge is associated with a weight (in *W*) which signifies the number of pathways that are shared between the two incident genes (i.e. the two genes connected by the edge). Two genes connected by an edge of weight one in *Proj* are, *as a combination,* unique to a single pathway. We call such pairs Pathway Gene-Pair Signatures (*Pathway-GPS*) of that pathway. The process of identifying Pathway-GPS for all pathways in a given repository is termed the *'Signature Transformation'* of the repository.

Need for combinations

A simplistic approach to the challenges posed by shared components would be to discard genes with multiple pathway annotations and to limit the analysis to the investigation of single characteristic genes that have only one pathway annotation (we call these *pathway unique genes*). This would, however, drastically reduce the discovery power of the analysis, because most genes are not pathway unique genes and many pathways do not have such markers. Furthermore, genes with multiple pathway annotations do contain valuable information on the underlying biological processes.

Hypothetically, one could go beyond gene-pairs and also consider n-tuples (with $n > 2$) of genes that co-occur in a single pathway *P* as signatures of *P*. As explained in the next section, the possible benefits of such extensions, however, do not currently seem to justify the associated complications to the computational and methodological framework.

Viability of higher order combinations as signatures (Sufficiency of Gene pairs)

The move from *PUG*s to *GPS* is motivated by the observation that if the method were to focus exclusively on *PUG*s, pathway-membership information for a substantial fraction of human genes would be lost. Our analysis suggests that GPS deal with this

issue rather effectively: for instance, all human genes with a KEGG pathway-annotation participate in at least 9 (in average, 382) KEGG-GPS, i.e. there are no 'orphan' genes.

A naturally arising question is: "*Would an extension of the signature concept to sets of more than two genes not provide even more information about the relevant pathways?"*

To answer this question, let us first recall that for a pathway containing *n* genes, there are (*n choose k*) distinct subset of size *k* and a total of 2^n -1 distinct non-empty subsets. (e.g. for a pathway with 102 genes –like TLR – there are around 5000 gene pairs , 171700 possible triplets, and over 4,2 million possible 4-tuples). Overall, there are over 256 Million co-annotated triplets and approximately 62 Billion co-annotated quartets of human genes in KEGG.

The vast majority of these higher order combinations of co-annotated genes are bound to be specific to a single pathway, but this is due to a rather trivial effect: Expanding any GPS *g1,g2* by an additional gene *g3* from the same pathway would automatically create a new 'triplet signature' *g1,g2,g3* (otherwise *g1,g2* could not be a GPS). Only in an estimated 4.8 million out of 256 million or 1.9% of all cases, these triplets would potentially contain novel information: (There are ~100,000 non-GPS co-annotated gene pairs, the average human KEGG pathway contains around 50 genes, which means that each non-GPS pair can be expanded by -in average- 50-2=48 different genes to build a triplet).

In other words, the GPS capture at least 98.1% of the information that can be coded by triplet signatures, at a much smaller computational cost.

A more refined strategy would seek to identify triplets that genuinely carry novel information. In order for a set of more than two genes to provide new evidence (i.e. information that is not already captured by the GPS), such set would have to include three or more genes such that a) all the genes in the set co-occur in a single pathway *and* b) no combination of two of the genes in the set (e.g. triplet Signature) is a GPS. Although the second criterion reduces the number of candidate set, the resulting gene sets still do not add an actionable amount of additional information.

To further elucidate this point, we identified 619 'triplet signatures' for the TLR pathway fulfilling the above criteria and involving AKT1. AKT1 is annotated in TLR and several other pathways. All TLR-GPS involving AKT1 contain a TLR-PUG, (i.e. no combination of AKT1 with another multi-pathway gene results in a TLR GPS). All 619 TLR-triplets contained AKT1 and two additional genes *g1, g2*, such that *g1,g2* was not a GPS of TLR (or in fact a GPS of any KEGG pathway). In total, 66 out of the 102 TLR genes were involved in these triplets. Using the list of these 66 genes (the building blocks of the signature triplets) as (the query list) input, SIGORA identified 169 TLR-GPS as present and declared TLR as significant. To understand this behaviour, consider two triples *g1,g2,g3* and *g1,g4,g5*. Although by construction we postulate that *g2g3*, *g1g3*, *g4g5* and *g1g5* should not be GPS, *g4g2*, *g5g2*, *g4g3* and *g5g3* still might be. For a concrete example, consider the triplet signatures (AKT1, TRAF3, MAPK12) and (AKT1, CXCL11, IL10RB) of TLR-signaling: although within each of these triplets, there is no gene-pair specific to any single pathway, (MAPK12, CXCL11) and (TRAF3, CXCL11) are TLR-GPS.

Assignment of weights to GPS

As, by definition, each Pathway-GPS is uniquely associated with a single pathway, identifying such Signatures in a gene list of interest (e.g. observing that both constituent genes of a Pathway-GPS are in the list of differentially expressed genes) can serve as an indicator of the activation/perturbation of the associated pathway. Each pathway can (and usually does) have multiple possible Pathway-GPS, and (as discussed below) the method does not rely on the observation of an individual GPS but rather on the statistical over-representation of multiple GPS in comparison to the expected proportion.

Yet before the over-representation of GPS can be used for the identification of relevant pathways, we need to emphasize that different GPS vary in their reliability as indicators of a particular pathway. This is due to the fact that, while the two genes comprising a Pathway-GPS can only be co-annotated in a single pathway, each of the two genes –*considered individually*- can be a member of several distinct pathways.

Consider GPS of the form (*g1,g2*) for a pathway *P*, where *g1* is annotated in *i* pathways and *g2* is in *j* pathways. Intuitively, a GPS that consists of two PUGs (i.e. a

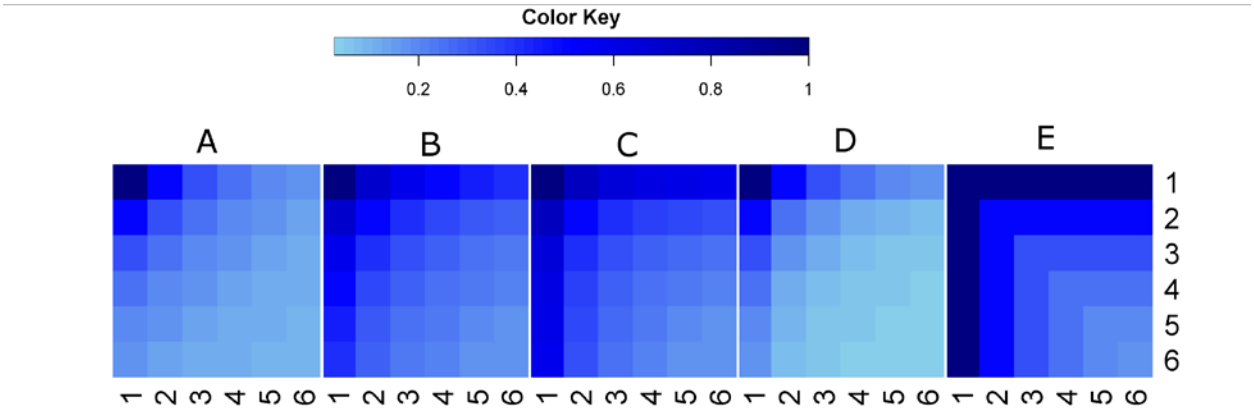
case where *i*=*j*=1) is a more appealing signature than a GPS that consists of two 'multifunctional' genes (say, *i* =4 and *j*=3), where the simultaneous observation of the two genes might be due to other factors (e.g. simultaneous activation of two different pathways). To address this issue, a weight is assigned to each GPS to quantify its reliability as an indicator of its associated pathway.

Choosing an appropriate weighting scheme

As it is often the case in harnessing information from projections of bipartite networks (Padrón, Nogales, and Traveset 2011; Allali, Magnien, and Latapy 2013) , there are many different plausible and 'natural' ways to quantify this intuitively clear notion that with increasing *i* and *j*, the reliability of (*g1,g2*) as an indicator of *P* monotonically decreases.

We have explored five different such weighting functions (Figure 3-4).

Figure 3-4 The monotonic decline of five alternative GPS-weighting schemes with increasing i and j. Only the values for i and j up to six are illustrated



A: Jaccard B: cosine C: inverse harmonic mean D: independent decisions E: topological overlap.

A) $\frac{1}{i+j-1}$: The Jaccard similarity of *g1* and *g2.* (Number of common annotations of *g1* and *g2*, which is by definition of *GPS* always 1, divided by the total number of annotations of *g1* and *g2*.) The inverse of total number of pathways annotations of *g1* and *g2* (considered *individually*).

B) $\frac{1}{\sqrt{i * j}}$: The cosine normalization of *i* and *j.*

C) $\frac{1}{2} * \left(\frac{1}{i} + \frac{1}{j} \right)$: The inverse of the harmonic mean of *i* and *j.*

D) $\frac{1}{i * j}$: The reciprocal of the product of *i* and *j.*

E) $\frac{1}{\min(i, j)}$: The topological overlap of *g1* and *g2*: number of common pathways (by definition of *GPS* always 1) divided by *min (i,j).*

All of these functions are plausible weighting schemes and each has its own strengths and limitations. For example, the weighting scheme D corresponds to the probability that, assuming that *g1* and *g2 uniformly and independently* 'decide' to engage in one of their annotated pathways, they both choose *P*. At the same time, the independence assumption in this scheme seems biologically unsupportable.

In practical terms, the functions A and D do not seem very useful as they decline rapidly with increasing *i* and *j* (Figure 3-4). In these schemes, for most possible values of *i* and *j*, the resulting weights are very close to zero and hardly distinguishable from each other. Similarly, weighting scheme (E) could be interpreted as the possibility that any of the two constituent genes 'regulates' *P*. In this scheme, however, all GPS with the same value of *min(i,j)* obtain the same weight, regardless of their respective values for *max(i,j),* which again does not necessarily result in extracting most information out of the query genes in a mathematically sound manner (Figure 3-4).

Among the functions listed above, the weighting function (C) seems to offer a more gradual and fine grained monotonic decline (Figure 3-4). It corresponds to a normalized voting scheme in (Allali, Magnien, and Latapy 2013) or to the shared visits model in (Padrón, Nogales, and Traveset 2011). Figuratively, one could think of *g1* and *g2* (the two genes in the GPS) as actors collaborating towards a common goal (the common pathway *P*): *g1* commits 1/*i* of its resources to *P*, while *g2* assigns 1/*j* of its resources to *P*. The function under C is the average commitment of *g1, g2* to *P.*

In addition to the above practical considerations, there are also some epistemic and biological arguments in favor of a less stringent penalty for GPS that involve genes with higher *i* and *j*:

1) The "*correlated expression problem*": Goeman and Bühlmann (Goeman and Bühlmann 2007) have argued that the statistical framework of Over-representation based methods is flawed: In their view, Over-representation analysis does not account for the fact that changes in the expression levels of a gene are not random and independent of expression levels of other genes, as genes are often subject to common regulatory mechanisms. While this critique seems particularly convincing in the case of the individual gene ORA, one can argue that if *g1* is member of *i* pathways and *g2* member of *j* pathways, and *g1* and *g2* have a single common pathway, then there are *i*+*j*-1 counter-examples to the assumption that *g1* and *g2* are subject to exactly the same transcriptional regulatory mechanisms. Notably, this type of "evidence for transcriptional independence of *g1* and *g2" strengthens with increasing i and j.*

2) The "*knowledge bias problem*": Some well studied genes might be annotated in a relatively large number of pathways, in part because these genes have been known for a longer time and been subjected to more intensive scrutiny, and conversely, other genes might be annotated in only a few pathways, simply because they have not been a research focus.

Paradoxically, these considerations suggest that a GPS with rather large *i* and *j* could be a relatively *reliable* indicator of *P,* because the co-annotation of *g1* and *g2* in *P* (and only *P*) is less likely to be due to gaps in the state of our knowledge about pathway annotation of *g1* and *g2*, and less likely to be due to the transcriptional co-regulation of *g1* and *g2*.

Hence, a gradually declining weighting scheme (like B or C) seems overall more appropriate than the more steeply declining alternatives. Weighting scheme C is the default in the implementation of SIGORA, and all results presented here are based on this scheme. Nevertheless, in the implementation, the user also has the option to use any of the other weighting schemes mentioned above or assign a constant weight of 1 to all GPS. Certain user-defined weighting schemes are also supported.

In summary, let (*g1, g2*) be a GPS associated with a pathway *P* and let *i* and *j* be the number of individual pathways annotations of *g1* and *g2*, respectively. The weight of the GPS is

$$w_{i,j} = \frac{1}{2}\left(\frac{1}{i} + \frac{1}{j}\right) = \frac{i+j}{2*i*j}$$

Identifying statistically over-represented Pathway-GPS

Analogous to traditional (individual gene) ORA (IG-ORA) methods, the distribution function of hypergeometric probabilities is used to calculate p-values indicating the statistical enrichment of Pathway-GPS in a user-specified gene list, which is given by:

$$p(k,n,m,N) = p(x \geq k) = 1 - \sum_{x=0}^{k-1} \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$$

In individual gene ORA, *k* denotes the number of query genes in the tested pathway, *n* the number of the genes in the pathway, *N* the number of all (assayed or annotated) genes, and *m* the length of the query list (i.e. the number of the genes with interesting status). In contrast to these traditional approaches, however, the parameters of the hypergeometric function in SIGORA are calculated as sums of GPS weights rather than frequency statistics of individual gene annotations. The Signature ORA parameters for p-value calculations are summarized in Table 3-2.

Table 3-2 Interpretation of the hypergeometric distribution test parameters used in SIGORA.

Parameter	Interpretation in Signature ORA
k (success)	Rounded sum of the weights of all *present* GPS of the tested pathway.
n (success states)	Rounded sum of the weights of all possible GPS of the tested pathway.
N (universe)	Rounded sum of the weights of all possible GPS of the repository.
m (sample size)	Rounded sum of the weights of all *present* GPS

A GPS is present if *both* of its component genes are in the query list.

Note that -*strictly speaking*- hypergeometric probabilities are only defined on natural numbers while the sums of GPS weights are positive floating point values, which is why the table refers to *rounded* values of the sums (to the closest integer). Theoretically, this rounding could lead to a *'blurring of the weights for pathways with few GPS'*, however, in practice this issue hardly materializes, as such pathways are very rare (e.g. 219 out of 226 KEGG human pathways in our repository are associated with 10 or more possible GPS). In tests with simulated and real biological data sets, different choices of the rounding strategy (*floor*, *ceiling* or *nearest integer*) did not substantially affect the significance or the rank order of the identified pathways. As an aside, in the widely popular statistical framework R, the *phyper* function (which computes the distribution function of the hypergeometric distribution) does accept non-integer (floating point) parameters and handles such input by applying the following rounding strategy: the number of successes is rounded down, all remaining parameters are rounded to the closest integer value. In our implementation, the user has the option to restrict the GPS-sets for the universe (*N*), and the success states (*n*) by providing a list of assayed genes (background). Furthermore, all PUGs are by default considered to represent a GPS of weight 1 (as a combination of the PUG with itself), but the user has the option to restrict the analysis to pairs of genuinely distinct genes.

Multiple Testing Correction

Undertaking pathway analysis generally involves a large numbers of significance tests. As testing a multitude of hypotheses will inevitably lead to *some* 'significant' results, adjustment of p-values for multiple testing is a crucial feature of any pathway analysis tool. Bonferroni's method is used by default in SIGORA for multiple testing

correction (MTC). It is, however, relatively easy to change the MTC procedure, if a user prefers to explore other adjustment methods (see the implementation section).

Selection of cut-off threshold for statistical significance of pathways

The choice of a reasonable cut-off threshold for statistical significance of the hypergeometric test results is an open methodological question in IG-ORA. In practice, values smaller than 0.1 or 0.05 after correction for multiple testing are commonly considered significant. Shifting the perspective from individual genes to the weighted Gene-Pair Signatures brings an additional challenge: as the size of the universe for the hypergeometric test dramatically increases (we move from a few thousand genes to up to a few hundred thousand weighted gene-pairs), the calculated p-values become by several orders of magnitude smaller than those observed in a typical IG-ORA analysis.

Based on our experience with simulated and biological datasets, we recommend a significance threshold of 0.001 after MTC (by Bonferroni). In the implementation, the default output of the analysis is the ranked list of pathways that achieve a corrected p-value up to this value. The user can also export the entire results table (including the p-values, corrected p-values and the parameters of the hypergeometric test) and the evidence (lists of present PUGs, list of the genes involved in present GPS or list of all present GPS along with their weights).

Dealing with redundancies of semantic origin

Thus far, we have described the motivation for Signature Transformation from a *biological* perspective. However, in some cases, there are additional *semantic* reasons for sharing of components among pathways. In particular, some repositories (e.g. REACTOME (Matthews et al. 2009), INOH (Yamamoto et al. 2011) and Gene Ontology (Ashburner et al. 2000)) are organized in a hierarchical structure, where all genes associated with one pathway (child) are also included in a more general pathway (parent). This poses a general challenge for pathway analysis tools that ideally should identify the most relevant level of the hierarchy (Alexa, Rahnenführer, and Lengauer 2006; Grossmann et al. 2007; Jupiter, Sahutoglu, and VanBuren 2009).

The hierarchical nature of such repositories poses a special challenge for our method. As any gene-pairs from a child category also co-occur in the parent pathway,

they would thus be excluded from being identified as a possible *Signature*. This would have the undesirable effect that all child pathways on the lower levels of the hierarchy would be left without any Signatures at all in the offline stage and hence be undetectable by SIGORA in the online stage.

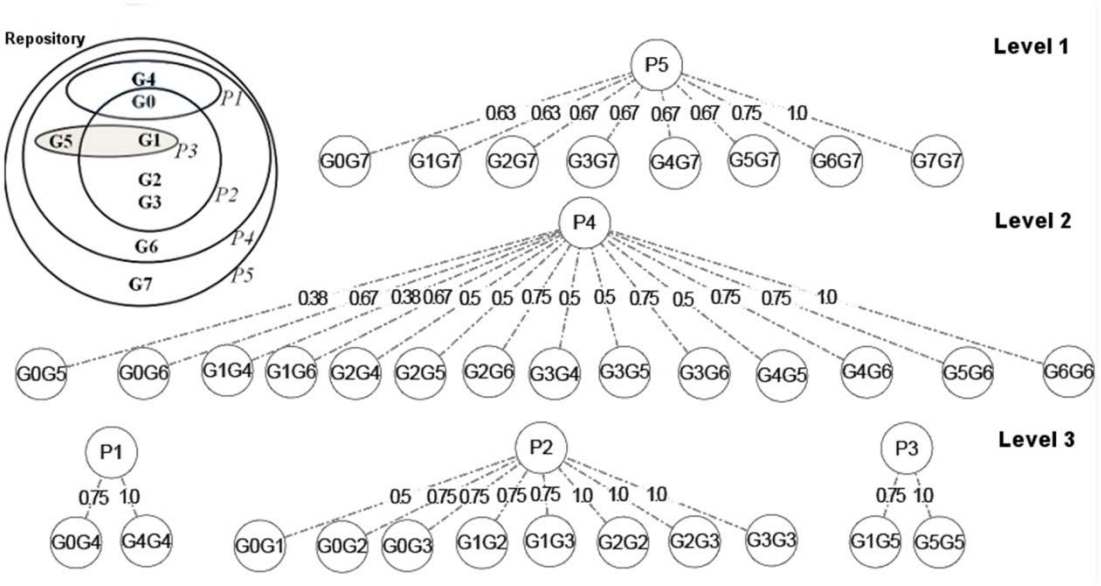
To address this issue, we deploy the following iterative top-down strategy in the offline stage (Signature Transformation):

1) Set level=1. Compile the repository Signatures as described for the non-hierarchical case. Assign the compiled signatures to level 1.

2) Remove all pathways from the top level of the hierarchy, increase the level and recompile Signatures for the remaining pathways. Assign the GPS to the current level. Iterate this step until no further hierarchical levels can be removed.

Figure 3-5 illustrates this procedure on a simplified hierarchical repository. In the *online* stage (the identification of significantly over-represented *Signatures* in a user-specified gene list) the user can specify how many levels of the hierarchy should be considered in the analysis. Any *GPS* (and any pathways) that are deeper down the hierarchy (i.e. are at a higher threshold level) are left out of the universe (and the analysis). If the user (for instance) asks for analysis up to the second level, then only *GPS* from levels one and two are considered.

Figure 3-5 Signature Transformation of a hierarchically organized pathway repository as an iterative process.



Here, G0 to G7 are genes that are annotated in the hierarchically organized pathways P1 to P5, as shown in the Vann-Diagram (inset). In the first iteration, only signatures for the outer-most level of the hierarchy(P5) are determined. The GPS for the higher levels are the ones associated with the less general pathways, and are only visible after removal of the more general terms. For example, the GPS of P1, P2 and P3 are only visible at level 3, after removal of P5 and P4.

The effect of this simple modification (i.e. the iterative strategy for Signature Transformation) is similar to a combination of the *Elim* (Alexa, Rahnenführer, and Lengauer 2006), and *TreeHugger* (Jupiter, Sahutoglu, and VanBuren 2009) algorithms in dealing with GO's structure. *Elim* essentially excludes genes in more specific categories from consideration in more general categories, whereas *TreeHugger* weakens the contribution of such genes to the higher levels.

Complexity and computational cost of the method

The complexity of the transformation step (in an implementation based on the outlined bipartite network interpretation of gene-pathway membership) is dominated by the complexity of the bipartite projection, which is given as:

$$O(\|V\| * d^2 + \|E\|)$$

With ||*V*||: number of nodes in the network (number of genes + number of pathways), *d* the average degree of nodes, ||*E*|| the number of edges in the bipartite network. For most inputs, the online phase (i.e. analysis of user specific input lists using the precompiled Signatures) completes within 15 seconds on a standard laptop (1.8 GHZ, 2 GB of RAM).

Implementation

SIGORA is implemented as an R package. The package and a detailed manual are available for download from http://sigora.googlecode.com/svn/. The following highlights the most important steps in a typical work-follow in R (requires R version ≥2.10):

```
## install and load the downloaded package
> install.packages('sigora_0.9.8.tar.gz', type='source', repos = NULL)
> library('sigora')
## (please note that all of the following commands require that the
        package is already loaded)
## import the query list from a file (assuming the list is given as
        Ensembl gene IDs):
> myquerylist <-ens_converter(scan('myfile.txt',what='character'))
## Alternatively, if the file consists of Entrez gene IDs
> myquerylist <-entrez_converter(scan('myfile.txt'))
## perform signature over-representation analysis, using KEGG GPS
> sigs(myquerylist,'k',markers=1,level=2)
## multiple testing correction is done by Bonferroni and FDRs are also
        provided. In order to add Hommel's method:
> cbind(summary_results,p.adjust(summary_results[,5],'hommel'))
## export the results into a file
>  export_results(filename='my_results.csv', genes=T)
## help (on Windows systems, help is shown in the web-browser)
> help(sigs)
## the following demo is also available
> demo(sigora)
```

3.3.2. Evaluation methods

We evaluated the performance of SIGORA by comparison to several other analysis tools on simulated and published biological datasets. Three of the methods compared are based on individual gene over-representation (*DAVID* (Jiao et al. 2012; D. W. Huang, Sherman, and Lempicki 2009)*, gProfileR* (Reimand et al. 2007) *, InnateDB* (Lynn et al. 2008)) and the remaining three (*GSEA_Preranked* ,*GSEA* (Subramanian et al. 2005), *GSEA_AF* (J. Ma, Sartor, and Jagadish 2011)) are GSEA based.

Acknowledging the inherent challenges associated with comparison of p-values across different statistical frameworks, for each tool we follow the recommendations of the authors of that tool regarding the choice of the most appropriate significance threshold and multiple testing correction (MTC) method (Table 3-3).

Table 3-3 Overview of pathway analysis methods that are compared to SIGORA in this chapter.

Method	Reference	MTC and Threshold
DAVID v.6.7	(Jiao et al. 2012; D. W. Huang, Sherman, and Lempicki 2009)	FDR < 0.05
gProfiler	(Reimand et al. 2007)	g:SCS < 0.05
GSEA_Preranked	(Subramanian et al. 2005)	FDR< 0.05
GSEA	(Subramanian et al. 2005)	FDR <0.25
GSEA_AF (AF)	(J. Ma, Sartor, and Jagadish 2011)	FDR <0.25
InnateDB	(Lynn et al. 2008)	FDR < 0.05

Among the methods listed above, we use *DAVID, gProfileR, GSEA_PRERANKED* for the simulation study, as these tools (like *SIGORA*) can be run on any pre-selected gene list and in 'batch mode'. *GSEA* and Appearance frequency modulated GSEA (*AF*) are limited to particular experimental designs and have specific input data format requirements, and are used here only in the analysis of three biological dataset for which data in the required input format was available. *InnateDB* is used as a reference point in evaluation of the biological datasets, because SIGORA's GPS are based on pathway annotation data as present in InnateDB. The rationale for selecting *GSEA* is its popularity; while *AF* was chosen because it attempts to address similar

issues as SIGORA. A short summary of the relevant characteristics of each of these methods is given below.

***InnateDB* (www.innatedb.com** (Lynn et al. 2008; Breuer et al. 2013)**)**

InnateDB's pathway analysis interface provides traditional IG-ORA using the standard hypergeometric test. Its recommended MTC method is Benjamini-Hochberg. The pathway GPS in SIGORA's current implementation are calculated using the pathway annotation as present in the latest release of InnateDB. In other words, any observed differences in analysis results between SIGORA and InnateDB are solely due to the differences between Signature-over-representation and individual-gene over-representation, and there are no additional confounding issues regarding the gene identifier mapping or different update status of the repositories across tools. We compare InnateDB to SIGORA using three biological datasets.

***gProfileR* (http://biit.cs.ut.ee/gprofiler/** (Reimand et al. 2007)**)**

gProfileR is the R package associated with the web-server of same name. Like InnateDB, it deploys a traditional individual gene over-representation based method using the standard hypergeometric test. In gProfileR, p-values are corrected by default using a unique multiple testing correction method (MTC), called the Set Counts and Sizes (SCS) procedure, which is analytically derived from extensive simulation experiments and purports to account for "the actual structure behind functional annotations". In other words, issues relating to the overlapping structures within annotation repositories are believed to be addressed indirectly and implicitly, as a special case of MTC. We use gProfileR in the simulation experiment. For completeness, we also list gProfileR's results on the three biological datasets evaluated here; however, some of the pathways listed by gProfileR are very recent additions to the KEGG repository that are as yet not available in other tools, including our current implementation of SIGORA.

DAVID (http://david.abcc.ncifcrf.gov/ (D. W. Huang, Sherman, and Lempicki 2009; Jiao et al. 2012)

DAVID provides individual-gene pathway analysis over-representation analysis using the EASE Score, a modified Fisher Exact p-value that is designed to raise the bar

for statistical significance of smaller pathways. This avoids situations in which observation of only a few genes from a small pathway would make it equally (or even more) significant than observing dozens of genes from a larger pathways. The EASE Score is generally more conservative than the Fisher Exact p-values.

Apart from this modification, DAVID's *functional annotation charts* implement a traditional individual gene over-representation based method that treats all genes equally. DAVID has also introduced the concept of *functional annotations clusters* that are motivated by the idea that rather than focusing on significance of individual pathways, the true nature of a phenotype should be examined by considering the overall emerging picture of interrelated pathways. In some situations, clusters of interrelated pathways can be considered collectively significant while some (or most) of the individual pathways in those clusters might fall slightly beyond the significance threshold. Although this is undoubtedly a sensible statement, the measure of *interrelatedness* used in DAVID's functional clusters is in diametrical contrast to the reasoning behind SIGORA: DAVID's authors postulate that similar pathways tend to contain similar gene members. In DAVID's functional annotation charts, the more common genes annotations share, the higher chance they will be grouped together as interrelated pathways, and the better the chances of the emerging cluster to become (collectively) significant.

For the simulation experiment, we will focus on DAVID's functional charts. In our evaluation of analysis results on biological datasets, we will also briefly exemplify the pitfalls of DAVID's Functional clusters.

GSEA (Subramanian et al. 2005)

The three methods discussed above (InnateDB, gProfileR and DAVID) are over-representation based methods that (like SIGORA) operate on a pre-filtered list of genes of interest (query list). The list of genes of interest is often determined using a (combination of) threshold(s) (e.g. fold change and p-value of differential expression). The p-values are in essence derived from a contingency table.

GSEA, in contrast, is the most prominent representative of a very different category of pathway analysis tools that do not operate on a pre-selected list and do not

use contingency tables. In GSEA, all genes in the dataset are first ranked by their difference regarding a single biological metric (e.g. signal to noise ratio) between the two conditions. This ranked gene list is then used to assign a normalized enrichment score (NES) - defined as the maximum deviation of a running sum statistic from zero, adjusted for the number of genes in the pathway - to each pathway. The statistical significance of the NES is determined by sample permutation (i.e. randomly exchanging the phenotype class labels).

We compare GSEA and SIGORA in the analysis of three biological datasets. Some of the observed differences in the results of GSEA analysis to SIGORA are inevitably due to the fundamental differences between ORA and GSEA methods. In particular, regardless of their biological relevance to the examined dataset (or lack thereof), the significance of some of pathways observed by GSEA is due to 'subtle but coordinated changes' in expression levels of genes that are not in the list of differentially expressed genes.

This issue equally applies to the two remaining methods, GSEA-PRERANKED and AF, which are described below.

GSEA-PRERANKED

As the name suggests, this is a variant of GSEA where the input format is not an expression matrix, but a pre-ranked list of genes. Accordingly, as there is no sample information available, the statistical significance is derived from gene set permutation instead of sample permutations. Technically, applying GSEA and GSEA-PRERANKED to the same dataset can lead to identical NES, but very different FDRs (http://www.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html). Hence, the recommended threshold for statistical significance is different (0.05 instead of 0.25). We compare GSEA-PRERANKED to SIGORA in analysis of simulation datasets. Like standard GSEA, GSEA-PRERANKED does not provide a mechanism for dealing with shared components of pathways.

Appearance Frequency modulated GSEA (AF) (J. Ma, Sartor, and Jagadish 2011)

AF is a recently proposed variant of GSEA that is explicitly designed to deal with issues posed by shared components. In this respect (the intended benefit), *AF* is the

most similar method to SIGORA among all methods compared here. *AF* assigns weights to individual genes based on number of associated pathways and performs a GSEA analysis. Methodologically, *AF* inherits most of GSEA's characteristics and is quite distinct from SIGORA. We compare *AF* and SIGORA in the analysis of three biological datasets.

Simulation experiment

Creation of simulated input lists

As a preliminary measure to quantify the effect of shared components on the number of spurious pathways, we conducted a simulation experiment over 1,000s of simulated gene lists that are created by applying the following procedure: From a set of 175 human KEGG human pathways that are in the repository of all four compared tools (SIGORA, DAVID, gProfiler and GSEA_preranked), *n* pathways are chosen at random and a fraction (*alpha*) of genes in each selected pathway are marked as differentially expressed (DE). The restriction to 175 common pathways is intended to reduce the effects of diverging update-status across analysis tools. The list of DE genes from five selected pathways is used as a query list for SIGORA, gProfiler and DAVID. To create an input list for GSEA_preranked, a score of 2 is assigned to the selected DE genes and a score of 1 to all remaining human genes. This procedure is repeated 1,000 times at fixed values for *alpha* and *n*.

Biological datasets

We further compare each of the different pathway analysis tools by examining their results when applied to three different gene expression datasets. These datasets incorporate the results of microarray studies investigating the host response to a parasite infection (Experimental Cerebral Malaria), a viral infection (Dengue Fever), and a bacterial infection (Tuberculosis) (Lovegrove et al. 2007; Hoang et al. 2010; Thuong et al. 2008). We map the lists of differentially expressed genes in each experiment (as provided by the authors of the respective studies) to unique Ensembl/Entrez IDs and use the resulting sets as input lists for four over-representation based methods (InnateDB, DAVID, gProfiler and SIGORA), ensuring that all methods are run on identical input lists.

Additionally, two GSEA-based methods (GSEA and AF) are also applied to the corresponding expression datasets obtained from the gene expression omnibus (GEO): *GSE25001*, *GSE11199*, *GSE11199*.

3.4. Results

3.4.1. Results on simulated gene lists

To evaluate the performance of SIGORA, we compared it to three other popular pathway analysis methods (DAVID, gProfiler and GSEA_Preranked) applied to simulated data, where we know *a priori* which are the significant pathways. The simulated input data was created by randomly choosing five KEGG pathways and selecting a fraction (alpha, 50% or 15%) of the genes in each of these pathways as being "differentially expressed". Each of the methods was then applied to the selected gene list to determine the statistically significant pathways (using the respective recommended significance threshold and MTC approach, see Table 3-3).

If the sharing of genes between different pathways was not a factor, we would expect that each method should identify only the five preselected pathways as significant. As can be seen in Table 3-4, in our experiments (using 1,000 simulated datasets with alpha=50% and 1,000 datasets with alpha=15%), this was not the case: gProfileR, for example, identified more than 60 pathways on average as being significant at alpha=50%, despite only 5 pathways being simulated as significant in the input data. SIGORA performed best by this measure and identified on average 8 pathways as significant across the different datasets.

Table 3-4 Performance metrics for several pathway analysis methods run on 1,000 simulated gene lists at two different alphas (15% or 50%).

Alpha	Method	Average number of significant pathways	Average Recall	Average Precision	F1 score (harmonic mean of precision and recall)	Average rank of the original pathways within the analysis results
15%	DAVID	11.22	0.71	0.32	0.44	6.8 (sd: 8.55)
	gProfiler	34.17	**0.95**	0.14	0.24	7.8 (sd:10.71)
	GSEA_Preranked	0.09	0.01	**0.74**	0.03	29.6 (sd:26.73)
	SIGORA	**6.41**	0.72	0.56	**0.63**	**3.6 (sd:2.08)**
50%	DAVID	30.25	0.98	0.16	0.28	6.8 (sd:9.07)
	gProfiler	62.48	**0.99**	0.08	0.15	8 (sd:11.87)
	GSEA_Preranked	13.91	0.87	0.32	0.46	5.6 (sd:7.82)
	SIGORA	**8.87**	0.89	**0.50**	**0.64**	**3.7 (sd:2.31)**

Each of the simulated datasets contained either 15 or 50% of genes in 5 randomly chosen pathways. For each analysis method and each input gene list, the identified (statistically significant) pathways were recorded and the Precision *(true positive results/all significant pathways)*, Recall *(true positive/chosen)* and F1 score (*harmonic mean of Precision and Recall*) were calculated by comparing the list of the (*n*) chosen pathways with the list of identified pathways. More specifically, for the purpose of this analysis, a statistically significant pathway is considered a *true positive* if it is among the five chosen pathways and a *false positive* otherwise; furthermore, any chosen pathways that are not identified by a method as statistically significant are considered *false negatives*. The values in parentheses in the last column show the standard deviations for the ranks. The entries in bold show the method with the best performance according to each measure.

The Recall and Precision metrics in the third and fourth columns of Table 3-4 capture the relationship between originally preselected 'target' pathways and the identified pathways. More specifically: Recall describes the fraction of the target pathways that were identified as significant and Precision signifies the fraction of statistically significant pathways that were among the originally selected pathways. Neither of these two metrics by itself is decisive, and there is a certain trade-off between the two metrics. This trade-off is captured by the F1-score, the harmonic mean of Precision and Recall. As can be seen in Table 3-4, SIGORA had the best F1 scores when compared to the other methods.

Finally, note that although this analysis is based on recommended significance thresholds in each method, Precision, Recall and F1 score are all dependent on the –ultimately arbitrary- choices of significance thresholds. The last column in Table 3-4 describes the results according to a less threshold-dependent measure. Ideally, the preselected pathways would occupy the first five positions in the list of the identified pathways, resulting in an average rank of 3. As can be seen in the last column of Table 3-4, at both choices for alpha (15% or 50% of the genes in each pathway selected) the average rank of the originally preselected pathways ('target pathways') in SIGORA's results (3.6 and 3.8 respectively) is very close to this ideal value, whereas the other methods tend to identify several additional pathways as more significant than the target pathways, resulting in higher average ranks of target pathways in these methods.

3.4.2. Results on published datasets

In addition to the simulated data, we also compared SIGORA to five different methods (InnateDB, DAVID, gProfileR, GSEA and AF) applied to real biological data, in this case three different gene expression datasets. Full details of the input genes, p-values and highlighted pathways obtained by each method can be found in the Appendices B to D.

Tuberculosis:

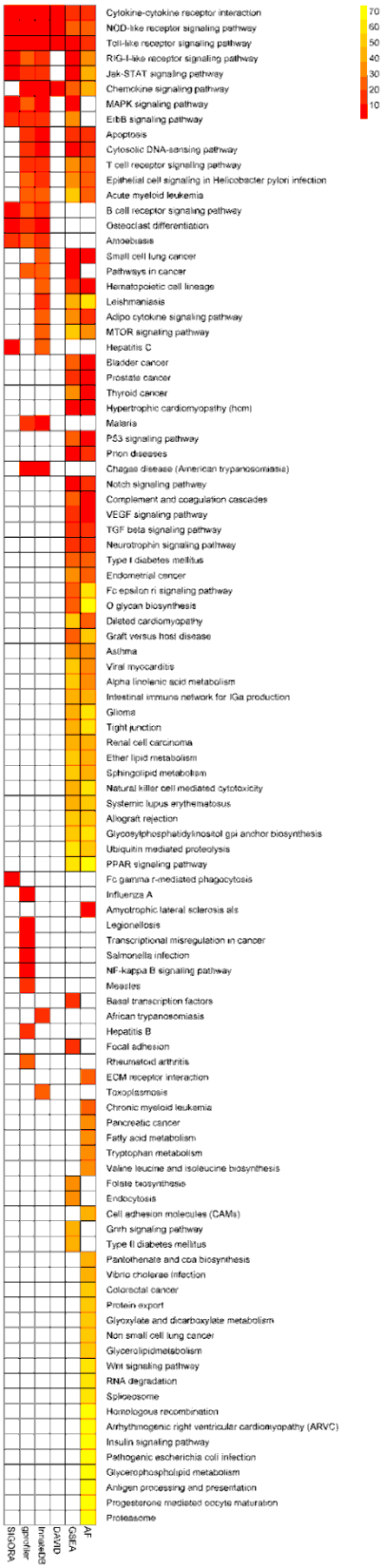
SIGORA was compared to five different pathway analysis methods (InnateDB, DAVID, gProfileR, GSEA and AF) applied to a gene expression dataset (GSE11199) which measured the host transcriptional response in human macrophages infected with *Mycobacterium tuberculosis* (Thuong et al. 2008). 1,250 transcripts (corresponding to 1,100 distinct Ensembl genes) were identified by Thuong et al. as being induced in response to this infection. Figure 3-6 shows the pathways that were identified as statistically significant by each of the six methods. The first thing that one notes is that the GSEA-based methods tended to predict large numbers of pathways as statistically significant (the AF method predicted nearly a third of the KEGG database as significant in this example). This is somewhat by design, as GSEA methods attempt to identify subtle but coordinated changes in gene expression. This may be very helpful in investigating cases where there are only subtle differences between conditions but in a

dataset like this one, it leaves the biologist bewildered as to which pathways should be followed-up on experimentally. In the other extreme is DAVID, which predicted only 4 pathways as statistically significant. SIGORA, on the other hand, identified 12 pathways as statistically significant; 10 of which were also identified by at least two other methods. Comparing the pathways identified by SIGORA as significant to the significant pathways identified by the other methods, one can see (Figure 3-7) that many of the pathways identified by other methods as significant but not by SIGORA share many genes with the SIGORA pathways. Notably, after removing the multifunctional genes that are involved in the pathways identified by SIGORA from the input list, the individual gene over-representation based methods (DAVID, InnateDB and gProfileR) did not return any significant pathways at all. This reinforces our observation from the simulated data that SIGORA will identify truly significant pathways but avoid identifying pathways that are significant because they share genes with other more relevant pathways.

Interestingly, SIGORA may also be able to identify some important pathways that are not significant using other methods. One pathway was identified as significant in this dataset only by SIGORA; *Fc gamma R−mediated phagocytosis*. Fcγ receptors regulate immune activation and susceptibility during *Mycobacterium tuberculosis* infection (Maglione et al. 2008; Maertzdorf et al. 2011) and it has been implied that “entry through Fcγ receptors may specify a distinct intracellular trafficking pathway for virulent *M. tuberculosis”*(Ernst 1998). Finally, DAVID’s top *functional cluster* for this dataset (Enrichment score 0.99) contained 16 pathways, 12 of which are different cancer subtypes (Appendix B).

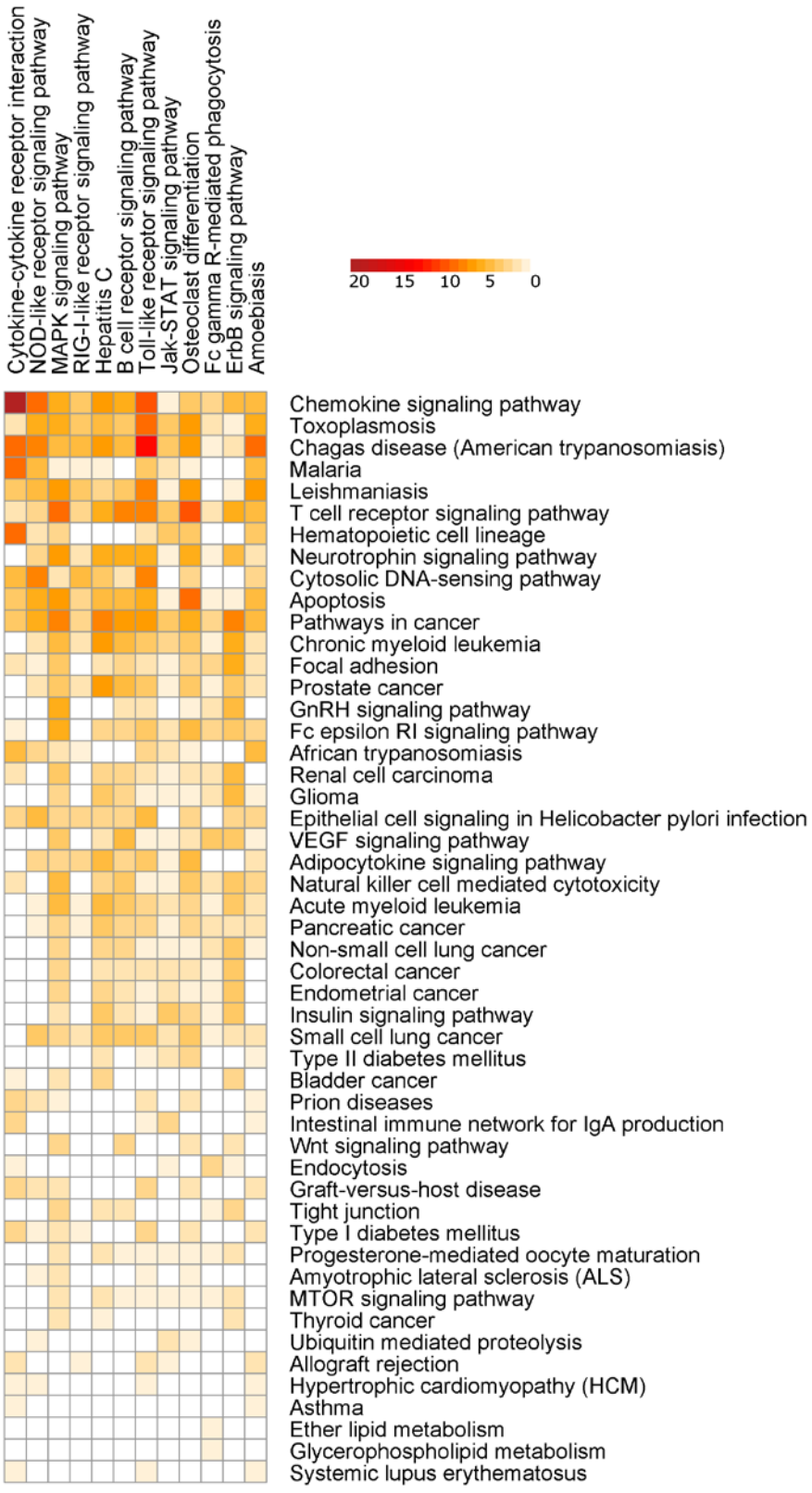
Appendix B lists all Pathways identified as statistically significant by each of the considered methods and their respective ranks (by p-value) in analysis of this dataset.

Figure 3-6 Results of six different pathway analysis methods applied to a gene expression dataset measuring the host transcriptional response to *M. tuberculosis* infection of human macrophages.


Cytokine-cytokine receptor interaction
NOD-like receptor signaling pathway
Toll-like receptor signaling pathway
RIG-I-like receptor signaling pathway
Jak-STAT signaling pathway
Chemokine signaling pathway
MAPK signaling pathway
ErbB signaling pathway
Apoptosis
Cytosolic DNA-sensing pathway
T cell receptor signaling pathway
Epithelial cell signaling in Helicobacter pylori infection
Acute myeloid leukemia
B cell receptor signaling pathway
Osteoclast differentiation
Amoebiasis
Small cell lung cancer
Pathways in cancer
Hematopoietic cell lineage
Leishmaniasis
Adipocytokine signaling pathway
MTOR signaling pathway
Hepatitis C
Bladder cancer
Prostate cancer
Thyroid cancer
Hypertrophic cardiomyopathy (hcm)
Malaria
P53 signaling pathway
Prion diseases
Chagas disease (American trypanosomiasis)
Notch signaling pathway
Complement and coagulation cascades
VEGF signaling pathway
TGF beta signaling pathway
Neurotrophin signaling pathway
Type I diabetes mellitus
Endometrial cancer
Fc epsilon RI signaling pathway
O glycan biosynthesis
Dilated cardiomyopathy
Graft versus host disease
Asthma
Viral myocarditis
Alpha linolenic acid metabolism
Intestinal immune network for IGa production
Glioma
Tight junction
Renal cell carcinoma
Ether lipid metabolism
Sphingolipid metabolism
Natural killer cell mediated cytotoxicity
Systemic lupus erythematosus
Allograft rejection
Glycosylphosphatidylinositol gpi anchor biosynthesis
Ubiquitin mediated proteolysis
PPAR signaling pathway
Fc gamma r-mediated phagocytosis
Influenza A
Amyotrophic lateral sclerosis als
Legionellosis
Transcriptional misregulation in cancer
Salmonella infection
NF-kappa B signaling pathway
Measles
Basal transcription factors
African trypanosomiasis
Hepatitis B
Focal adhesion
Rheumatoid arthritis
ECM receptor interaction
Toxoplasmosis
Chronic myeloid leukemia
Pancreatic cancer
Fatty acid metabolism
Tryptophan metabolism
Valine leucine and isoleucine biosynthesis
Folate biosynthesis
Endocytosis
Cell adhesion molecules (CAMs)
Gnrh signaling pathway
Type II diabetes mellitus
Pantothenate and coa biosynthesis
Vibrio cholerae infection
Colorectal cancer
Protein export
Glyoxylate and dicarboxylate metabolism
Non small cell lung cancer
Glycerolipidmetabolism
Wnt signaling pathway
RNA degradation
Spliceosome
Homologous recombination
Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Insulin signaling pathway
Pathogenic escherichia coli infection
Glycerophospholipid metabolism
Antigen processing and presentation
Progesterone mediated oocyte maturation
Proteasome
70
60
50
40
30
20
10
SIGORA
gprofiler
Innatedb
DAVID
GSEA
AF

The heatmap shows all pathways that were identified as statistically significant by at least one of the six different pathway analysis methods. The more red the color the higher the rank of that pathway for a particular method. The heatmap is sorted by the number of methods identifying a particular pathway as significant.

Figure 3-7 Number of differentially expressed genes that are shared between SIGORA's pathways (columns, ordered by rank) and additional pathways identified as significant by other methods on the TB dataset.


Cytokine-cytokine receptor interaction
NOD-like receptor signaling pathway
MAPK signaling pathway
RIG-I-like receptor signaling pathway
Hepatitis C
B cell receptor signaling pathway
Toll-like receptor signaling pathway
Jak-STAT signaling pathway
Osteoclast differentiation
Fc gamma R-mediated phagocytosis
ErbB signaling pathway
Amoebiasis
20 15 10 5 0
Chemokine signaling pathway
Toxoplasmosis
Chagas disease (American trypanosomiasis)
Malaria
Leishmaniasis
T cell receptor signaling pathway
Hematopoietic cell lineage
Neurotrophin signaling pathway
Cytosolic DNA-sensing pathway
Apoptosis
Pathways in cancer
Chronic myeloid leukemia
Focal adhesion
Prostate cancer
GnRH signaling pathway
Fc epsilon RI signaling pathway
African trypanosomiasis
Renal cell carcinoma
Glioma
Epithelial cell signaling in Helicobacter pylori infection
VEGF signaling pathway
Adipocytokine signaling pathway
Natural killer cell mediated cytotoxicity
Acute myeloid leukemia
Pancreatic cancer
Non-small cell lung cancer
Colorectal cancer
Endometrial cancer
Insulin signaling pathway
Small cell lung cancer
Type II diabetes mellitus
Bladder cancer
Prion diseases
Intestinal immune network for IgA production
Wnt signaling pathway
Endocytosis
Graft-versus-host disease
Tight junction
Type I diabetes mellitus
Progesterone-mediated oocyte maturation
Amyotrophic lateral sclerosis (ALS)
MTOR signaling pathway
Thyroid cancer
Ubiquitin mediated proteolysis
Allograft rejection
Hypertrophic cardiomyopathy (HCM)
Asthma
Ether lipid metabolism
Glycerophospholipid metabolism
Systemic lupus erythematosus

Experimental Cerebral Malaria:

Example 2 is a mouse cerebral malaria (ECM) dataset, comparing the whole-brain transcriptional responses of genetically susceptible (C57BL/6) and resistant (BALB/c) inbred mouse strains 6 days after infection with Plasmodium *berghei* ANKA (NCBI GEO: GSE7814) (Lovegrove et al. 2007). We first performed a differential expression analysis using Geo2R to obtain a list of up-regulated genes at FDR < 0.01 (637 Ensembl genes, Appendix C). We use this list as input for SIGORA, InnateDB, gProfileR and DAVID, and apply GSEA and AF to the corresponding expression matrix.

Similar to the previous example, the number of KEGG pathways that different methods identified as statistically significantly enriched in this dataset varied widely: AF identified 59 pathways (out of 185 in its repository), while DAVID highlighted just two pathways. SIGORA identified 14 pathways as significant, 13 of which were also reported by at least two other methods (Figure 3-8, Appendix C). The remaining pathway is *PPAR signaling pathway* (discussed further below).

Aside from this big-picture view, the strong changes in the rank orders of the following individual pathways are also worth mentioning:

Complement and coagulation cascades: This pathway is the 6th ranked pathway in SIGORAs results. Three additional tools (InnateDB, GSEA and AF) also identify this pathway as significant, but only at considerably lower ranks (the17th, 25th and 27th position, respectively). Complement and coagulation pathways have been shown to be critically involved in the development of ECM (Francischetti, Seydel, and Monteiro 2008; van der Heyde et al. 2006; Ramos et al. 2012).

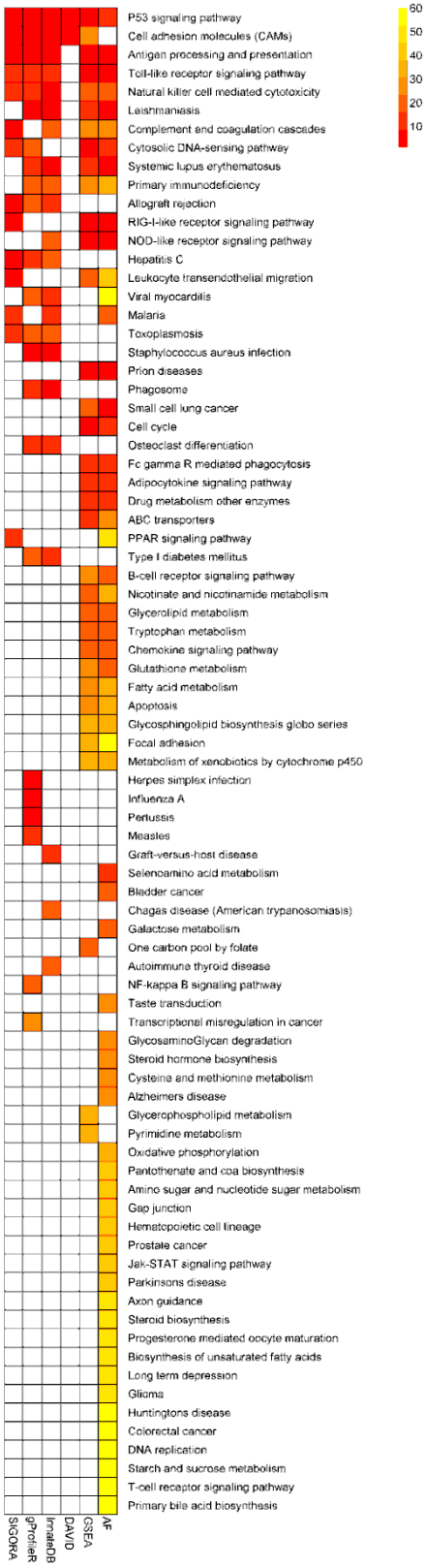
Leukocyte transendothelial migration: This pathway is the 7th ranked pathway in SIGORA's results, the 23rd ranked pathway in GSEA and the 44th ranked pathway in AF. The remaining methods do not identify this pathway as statistically significant. Polymorphonuclear leukocyte recruitment has been shown to be responsible for increased permeability of the blood-brain-barrier, and is strongly associated with fatality rates in ECM (Senaldi et al. 1994; Bell, Taub, and Perry 1996).

PPAR signaling pathway: This pathway is at position 14 of SIGORA's results. The only other method to identify PPAR signaling is *AF*, at position 47. Targeting of PPAR is currently being explored as a novel adjunctive therapy for cerebral malaria (Balachandar and Katyal 2011; Serghides 2012). Notably, PPARγ has been reported to be one of only two genes in a cerebral malaria-resistance locus identified using a genome-wide analysis of 32 different inbred mouse lines (Bopp et al. 2010) and modulation of the inflammatory response to *P. berghi* infection by an antagonist of this gene has greatly enhanced the survival rates in mice (Serghides et al. 2009).

At the same time, SIGORA avoids a few biologically implausible pathways that are considered highly significant by at least two other methods: e.g. *"Staphylococcus aureus infection"* (a bacterial infection) is the most significant pathway in InnateDB's results, and the second highest ranked pathway in gProfileR's results, but is not significant in SIGORA's results. Similarly, GSEA and AF both identify *"Prion diseases"* as significant (position 2 and 4, respectively) and again, this pathway is not significant in SIGORA's results. Other examples include: *"Small cell lung cancer"* (GSEA, AF), *"Viral myocarditis"* (InnateDB, gProfileR , AF), *"Type I diabetes mellitus"* (InnateDB, gProfileR) (Figure 3-8 and Appendix C).

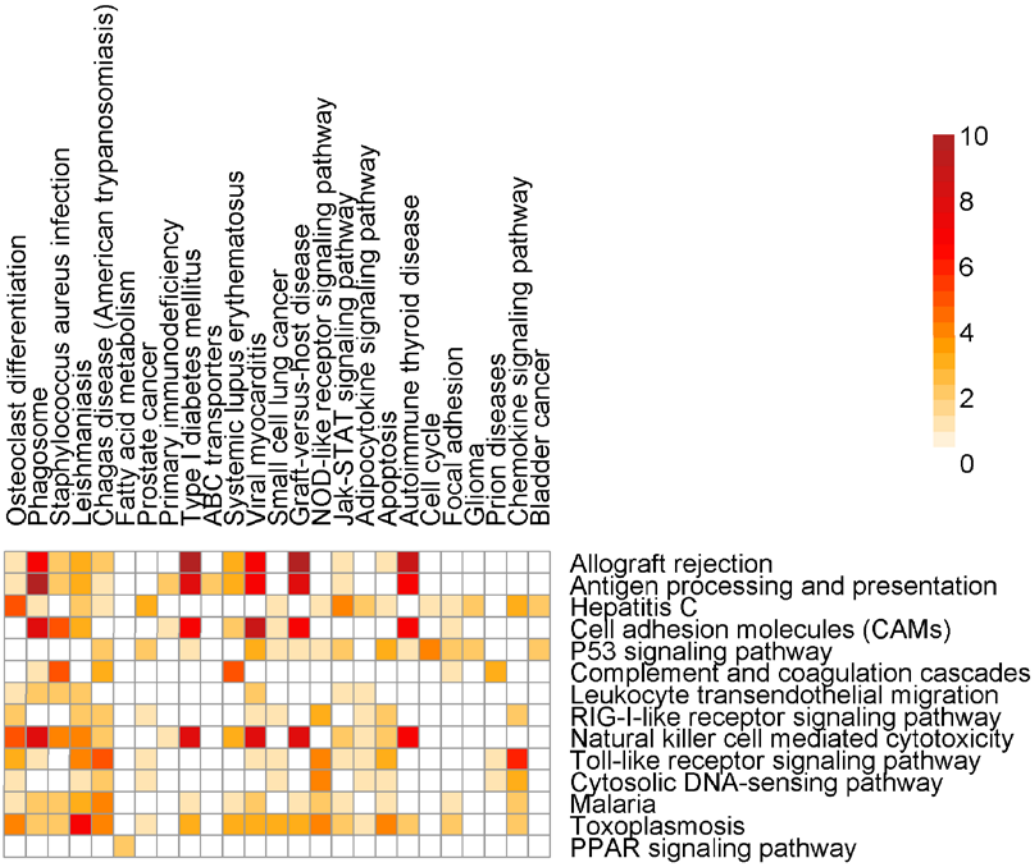
DAVID's top ranking functional cluster for this dataset (Enrichment Score: 2.7) groups *"Antigen processing and presentation", "Viral myocarditis", "Allograft rejection", "Graft-versus-host disease", "Type I diabetes mellitus"* and *"Autoimmune thyroid disease"* together (Appendix C). All of these pathways having highly overlapping annotations. It seems plausible that *"Type I diabetes mellitus"* , *"Autoimmune thyroid disease" and "Viral myocarditis"* are considered statistically significant by current methods due to the fact that they each share several (8 or more) differentially expressed genes with the natural killer cell pathway (Figure 3-9), which has been shown to be a determinant of murine malarial fatalities (Hansen et al. 2003).

Figure 3-8 Comparison of results of six different methods on a mouse experimental cerebral malaria dataset.


P53 signaling pathway
Cell adhesion molecules (CAMs)
Antigen processing and presentation
Toll-like receptor signaling pathway
Natural killer cell mediated cytotoxicity
Leishmaniasis
Complement and coagulation cascades
Cytosolic DNA-sensing pathway
Systemic lupus erythematosus
Primary immunodeficiency
Allograft rejection
RIG-I-like receptor signaling pathway
NOD-like receptor signaling pathway
Hepatitis C
Leukocyte transendothelial migration
Viral myocarditis
Malaria
Toxoplasmosis
Staphylococcus aureus infection
Prion diseases
Phagosome
Small cell lung cancer
Cell cycle
Osteoclast differentiation
Fc gamma R mediated phagocytosis
Adipocytokine signaling pathway
Drug metabolism other enzymes
ABC transporters
PPAR signaling pathway
Type I diabetes mellitus
B-cell receptor signaling pathway
Nicotinate and nicotinamide metabolism
Glycerolipid metabolism
Tryptophan metabolism
Chemokine signaling pathway
Glutathione metabolism
Fatty acid metabolism
Apoptosis
Glycosphingolipid biosynthesis globo series
Focal adhesion
Metabolism of xenobiotics by cytochrome p450
Herpes simplex infection
Influenza A
Pertussis
Measles
Graft-versus-host disease
Selenoamino acid metabolism
Bladder cancer
Chagas disease (American trypanosomiasis)
Galactose metabolism
One carbon pool by folate
Autoimmune thyroid disease
NF-kappa B signaling pathway
Taste transduction
Transcriptional misregulation in cancer
GlycosaminoGlycan degradation
Steroid hormone biosynthesis
Cysteine and methionine metabolism
Alzheimers disease
Glycerophospholipid metabolism
Pyrimidine metabolism
Oxidative phosphorylation
Pantothenate and coa biosynthesis
Amino sugar and nucleotide sugar metabolism
Gap junction
Hematopoietic cell lineage
Prostate cancer
Jak-STAT signaling pathway
Parkinsons disease
Axon guidance
Steroid biosynthesis
Progesterone mediated oocyte maturation
Biosynthesis of unsaturated fatty acids
Long term depression
Glioma
Huntingtons disease
Colorectal cancer
DNA replication
Starch and sucrose metabolism
T-cell receptor signaling pathway
Primary bile acid biosynthesis
SIGORA
gProfileR
InnateDB
DAVID
GSEA
AF
60
50
40
30
20
10

The heatmap shows all pathways that were identified as statistically significant in at least one of five different pathway analysis methods. The more red the color the higher the rank of that pathway for a particular method. The heatmap is sorted by the number of methods identifying a particular pathway as significant.

Figure 3-9 Number of differentially expressed genes that are shared between SIGORA's pathways (rows, ordered by rank) and additional pathways identified as significant by other methods on the ECM dataset.

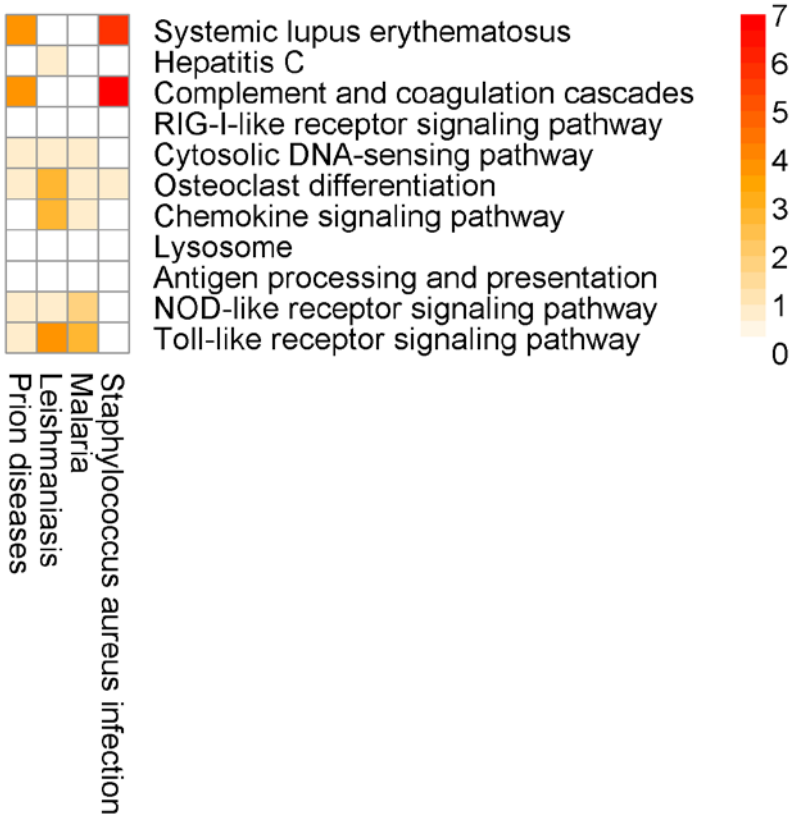


Dengue fever

As a third evaluation set, we re-examined a list of 483 up-regulated genes in the whole blood transcriptome of patients infected with dengue virus (NCBI GEO: GSE25001) (Hoang *et al.*, 2010) More specifically, we compared expression profiles of hospitalized patients with uncomplicated Dengue during acute phase (≤72 h of illness history) to follow up samples of such subjects two weeks after discharge (n=72).

Here, for the most part, the SIGORA results contain well-defined immunity related pathways. As before, some additional, potentially spurious pathways that are identified by other methods are not significant in the SIGORA analysis. Examples include "*Staphylococcus aureus infection*" (the second ranked pathway in InnateDB results) and the "*Prion Disease*" pathway (identified by both InnateDB and AF). These two pathways share components with the complement pathway (4 and 6 up-regulated genes respectively, Figure 3-10), which has been shown to have a role in neutralising Dengue (Shresta 2012).

Figure 3-10 Number of differentially expressed genes that are shared between SIGORA's pathways (rows, ordered by rank) and additional pathways by other methods (columns) in the analysis of the Dengue dataset.



Again, this example contains a possibly relevant pathway that was significant in SIGORA's results, but overlooked by the other methods: *"Lysosome*", at position eight of SIGORA's results (Table 3-5, Appendix D). Recent experimental evidence suggests that manipulation of the host's autophagolysosomes by the dengue virus is an important part of the virus's life cycle (Khakpoor et al. 2009; Heaton and Randall 2010).

DAVID does not return any functional clusters for this dataset.

Table 3-5 List of all pathways identified as statistically significant by each method compared in this study and their respective ranks (by p-value) in the analysis of a Dengue fever gene expression dataset.

	DAVID	GSEA	AF	gProfileR	InnateDB	SIGORA
Systemic lupus erythematosus	1	3	2		1	1
Hepatitis C					10	2
Complement and coagulation cascades	2			3	3	3
RIG-I-like receptor signaling pathway		1	3		9	4
Cytosolic DNA-sensing pathway		2	1		4	5
Osteoclast differentiation					7	6
Chemokine signaling pathway			6			7
Lysosome						8
Antigen processing and presentation			9			9
NOD-like receptor signaling pathway		4	4			10
Toll-like receptor signaling pathway		6	8		5	11
Staphylococcus aureus infection					2	
Long term potentiation			10			
Leishmania infection		5	11			
Ribosome			5			
Malaria					6	
Allograft rejection			7			
Prion diseases			12		8	
Measles				1		
Influenza A				2		
Herpes simplex infection				4		

	DAVID	GSEA	AF	gProfileR	InnateDB	SIGORA
Pertussis				5		

The entries in bold are significant largely due to sharing genes with other more relevant pathways.

3.4.3. Alternative evaluation criteria

So far, we have discussed that compared to existing methods, SIGORA performs favorably on biological and simulated datasets. There are, however, inherent challenges in developing and evaluating methods in a situation like this, where the ground truth is simply unknown. Comparisons of analysis results on biological datasets remain ultimately anecdotal, and the relevance of simulated datasets to real biological situations is questionable. The choice of datasets and metrics deployed in a comparative analysis might affect the outcome. In order to provide some additional evidence for the viability of signature over-representation analysis, in the next section we examine the performance of SIGORA in the context of evaluation criteria and datasets chosen by authors of other methods.

Reproducibility of results across independent datasets

One criterion for measuring the reliability of a method is the consistency of its results for two (or more) biologically similar yet independently created datasets. In such a setup, the method is applied to both datasets independently and the number of (relevant) significant pathways that are identified in both datasets is determined. The idea is that the higher the number of such common pathways, the more robust the method. Ma et al used this criterion to compare GSEA and AF (J. Ma, Sartor, and Jagadish 2011). More specifically, they applied both GSEA and AF to histological grade 1 *vs.* grade 3 ER+ tumors from GSE3494 (Miller et al. 2005) and histological grade 1 *vs.* grade 3 ER+ tumors from GSE2990 (Sotiriou et al. 2006) and reported that while GSEA identifies only three cancer related pathways in both datasets, AF identifies four. (Table 1 in (J. Ma, Sartor, and Jagadish 2011)). Furthermore, they reported one overlapping pathway in the top 5 results for GSEA, and two overlapping pathways in the top 5 results of AF (Figure 2b of their paper). We applied SIGORA to the same two datasets (DE genes for grade 1 *vs*. 3 samples were obtained from Genomic portals (Shinde et al. 2010) at a p-value of 0.001) and observed more consistent results than AF and GSEA.

There were four common cancer related pathways within the top five results *("Cell cycle", "RNA transport", "Proteasome"* and *"Spliceosome"*), and a fifth pathway (*"DNA replication"*) within the top seven results for the two datasets. Additional consistently significant pathways identified by SIGORA include "*Colorectal cancer*", "*Oocyte meiosis*", "*Cysteine and methionine metabolism*", "*Aminoacyl-tRNA biosynthesis*" and "*Base excision repair*".

Identification of 'target pathways' on a large collection of datasets.

The authors of PADOG (Tarca et al. 2012) proposed a different, presumably objective criterion to measure the performance of their method: they selected 24 expression datasets where the corresponding disease is the name of a KEGG pathway. The KEGG pathway describing that disease was then considered to be the 'target pathway' for this dataset. The analysis methods were compared in terms of their ability to identify the target pathway as statistically significant in the analysis of each data set. They report that PADOG was able to identify one (4.2%) of the 24 target pathways as significant (after adjusting for multiple testing) whereas GSEA and GSA did not identify any of the target pathways. In order to compare SIGORA to PADOG using this benchmark, SIGORA was applied to the 24 lists of differentially expressed ($p < 0.0001$) genes from these 24 datasets. In two cases (GSE9348 and GSE9476), SIGORA identified the target pathway as significant after adjusting for multiple testing. This is twice the number of such hits by PADOG on these datasets.

3.4.4. Coexpression and co-annotation

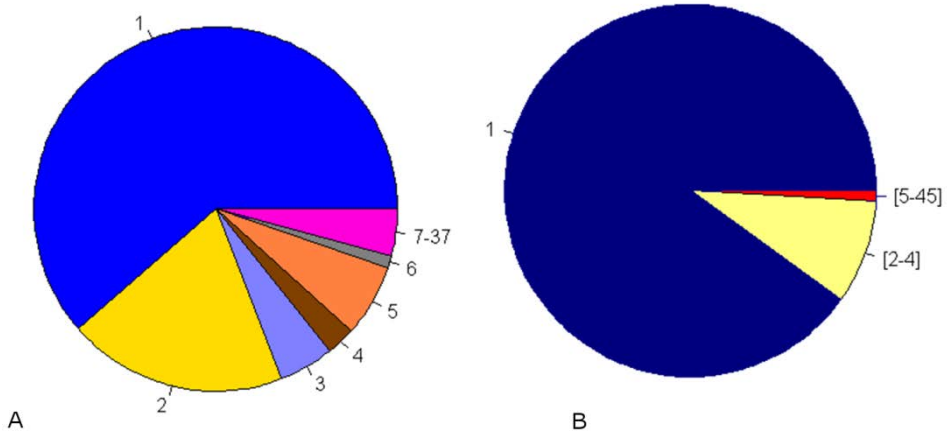
As mentioned on page 60 (the section discussing possible GPS-weighting schemes), correlated expression of genes has been long considered a methodological challenge to ORA-based methods, where the statistical framework presupposes sampling from an independent and identically distributed gene pool (Goeman and Bühlmann 2007).

To determine whether Pathway-GPSs were more or less likely to be compounded by co-expression biases than traditional methods, we investigated the relationship between the number of shared pathway annotations and correlated gene expression on a repository wide scale (in KEGG), using the 'human top three highest

correlated genes' list from COXPRESdb (Obayashi and Kinoshita 2011). COXPRESdb annotates the top three most correlated genes for each of over 19,000 genes, across hundreds of samples. The results can be summarized as follows: as is to be expected, highly co-expressed gene pairs tend to share pathway affiliations at much higher rates than randomly selected pairs of gene. From the 33,000 unique gene pairs of the COXPRESdb highly correlated list, 1402 (4.2%) co-occur in at least one KEGG pathway, while the corresponding fraction among randomly selected pairs of human genes is less than 0.07% (1,205,807 out of more than 180,000,000 possible pairs).

However, if we limit the analysis to gene-pairs that have at least one pathway annotation in common, then highly co-expressed gene-pairs are more likely to be found among gene-pairs that share multiple annotations than those having just one shared annotation. From the 1402 highly correlated gene pairs that share a pathway annotation, 542 (39%) share more than one annotation (Figure 3-11, panel A) , while the corresponding fraction among all gene pairs that share any pathway annotation is 10% (Figure 3-11, panel B). In other words, Pathway-GPSs are less likely to display highly correlated expression behavior.

Figure 3-11 Coexpression and co-annotaion in KEGG



A) Distribution of number of shared pathway affiliations for highly co-expressed, co-annotated human gene pairs in KEGG B) Distribution of number of shared pathway affiliations for all co-annotated human gene pairs in KEGG.

We were able to replicate these results in a case by case examination of several published lists of differentially expressed (DE) genes for individual pathological conditions. We observed that although such lists often contained many (hundreds) gene pairs from the above COXPRESdb-list, only a fraction (i.e. less than 10%) of such pairs were comprised of genes that shared a single pathway annotation. For example, DE genes in GSE781 (the dengue fever data set) contained 304 correlated gene pairs, only 22 of which were 'Signatures'. In the implementation, the user can examine the proportion of highly correlated genes and gene-pair signatures in a dataset (by executing *coexpress-sigs* after *sigs*).

3.5. Discussion, related and future work

The existence of shared components between pathways poses a challenge for pathway analysis methods: which of the statistically significant pathways associated with such components are the most biologically significant? In 2005, Khatri and Drăghici surveyed the state of the art analysis methods and tools of the time and outlined several limitations as the challenges for the next generation of analysis tools. One such challenge "*is related to genes that are involved in several biological processes. For such genes, all current tools weight all the biological processes equally. At the moment, it is not possible to single out the more relevant one by using the context of other genes [of interest] in the current experiment*."(Khatri and Drăghici 2005)

The years since have seen many interesting developments in the functional analysis of high throughput biological data: Several methods (Alexa, Rahnenführer, and Lengauer 2006; Grossmann et al. 2007; Jupiter, Sahutoglu, and VanBuren 2009) have been described to highlight processes at the appropriate level of specificity and to reduce the redundancy of the results in the context of Gene Ontology (GO) analysis, where the overlap between categories are often due to the hierarchical organization of the ontology. These methods, however, fail to account for overlap between pathways that don't involve the full inclusion of all members of a pathway in another pathway. To deal with such cases, (Antonov et al. 2008) proposed the creation of new functional categories as complex Boolean combinations of available GO terms. Unfortunately,

such combinations are often not easy to interpret and a comprehensive search over all possible combinations is computationally infeasible.

Outside of GO, a few methods have been proposed that indirectly tackle the issue by either discriminative treatment of individual genes or alternative representation of the pathway repository in specific scenarios. An example of the later approach (alternative representation of the pathway repository) is Bayesian Pathway Analysis, BPA (Isci et al. 2011). BPA transforms each pathway in a pathway repository into a separate Bayesian Network (BN) and scores the fit of each model with the experimental (expression) data. BPA is expected to leverage the expression status of other genes in the experimental context, as BNs, in contrast to simple lists of genes, are deemed capable of accommodating local interactions between genes. Although highly sophisticated, BPA is computationally intensive and by design limited to the interpretation of expression datasets.

An early example of the former approach (non-egalitarian treatment of individual genes) is impact-analysis (Draghici et al. 2007). Impact analysis integrates the magnitude of each gene's expression change along with the type (e.g. receptor, transcription factor) and position of each gene within the given pathways and their interactions into the statistical framework, however, the authors do not explicitly address the problems related to component sharing among pathways.

More recently, it has been proposed to add an appearance frequency based parameter to the statistical framework of GSEA. This additional parameter is intended to weaken the contribution of genes with multiple pathway memberships to the statistical significance of all of their associated pathways (J. Ma, Sartor, and Jagadish 2011). While this is a significant step in the right direction, the addition of such a parameter does not exploit the status of other genes in the experiment for the selection of the most relevant function of a gene in the experimental context. As exemplified by the gene, *BRCA1* in (Khatri and Drăghici 2005), even key players of one process (maintaining genomic stability) can have several less prominent roles in other unrelated pathways (response to nutrient and brain development). Nor can appearance frequency distinguish between the causes of appearance of a gene in several pathways, which,

aside from functional pleiotropy of genes, can be partly due to the hierarchical organization of some pathway repositories like REACTOME and GO.

Here we introduced the concept of Pathway-GPS, as genes or gene-pairs that (as a combination) are specific to a single pathway, and we described Signature Over-representation Analysis (SIGORA) as a novel (and comparably efficient) approach to pathway analysis. SIGORA uses Pathway-GPS to bridge the gap between the context sensitive and collaborative nature of biological processes on one side and the universal and discrete statistical framework of over-representation analysis on the other side. Although each Signature's weight is fixed in advance, the net contribution of an individual gene *G* to the measured success (parameter *k* in Table 3-2) of each of its associated pathways is not fixed and depends explicitly on the status of its partners (genes that together with *G* form a GPS).

In contrast to individual gene ORA, SIGORA seems relatively robust to biases that are introduced by correlated expression of genes. In contrast to the GSEA-based solutions, SIGORA inherits the versatility of the ORA statistical framework and is applicable to lists of genes of interest obtained in any type of high throughput experimental set-up (e.g. copy number variations from cellular profiling, lists of epigenetically silenced genes from promoter methylation analysis, differential gene expression data from NGS and microarrays or SNPs from GWAS experiments) without the need for adaptation of the computational method. This is especially notable in situations where ranking of the entire dataset by a single biological parameter (as required by GSEA) is not feasible (see (Huang *et al.*, 2009) for a few examples).

As described in 3.4.4, the independence assumption of hypergeometric test is already controversial for individual gene ORA (as used in InnateDB, DAVID, gProfileR): the fundamental assumption of the hypergeometric test, sampling without replacement from an independent and identically distributed gene pool (Goeman and Bühlmann 2007) is biologically questionable, as gene expression is a coordinated and correlated process (page 60 and 91ff). The move from single genes to gene pairs to some degree mitigates that (biologically motivated) issue, in part because (according to the notion of *present* signatures) a single gene's (G) contribution is not fixed but amplified or annihilated by presence or absence of other genes that build a signature with G. (In the

context of natural language processing / text labeling the use of GPS instead of single genes would correspond to using a certain class of bi-grams instead of a 'bag of words' model of language: Apple+ Windows -> computers; Apple + Banana-> fruits ; Apple alone-> abstain from using this word as evidence for any classification).

Among the four parameters of the hypergeometric test, a growth in the size of universe (or background) and the number of successes result in smaller p-values, where as growing sample size or larger success state sizes decrease the significance (i.e. lead to larger p-values). Due to the nature of the test, these effects are non-linear, i.e. doubling (for instance) the size of the universe does not result in halving the p-value. In the context of signature-overrepresentation, all four parameters are affected by the move from genes to gene-pairs, which might seem to further complicate a direct comparison of SIGORA's p-values to p-values from IG-ORA.

Desirable as it may seem, *mathematically*, one cannot compare methods across different statistical frameworks, tool-update statuses, significance thresholds and MTC. Even leaving SIGORA out of the comparison, the methods compared in this chapter use different MTCs, thresholds and statistically frameworks (Table 3.3 : FDR 0.25 or FDR 0.05, or g:SCS 0.05, permutation based versus urn-model/differential expression based, gene permutation vs. sample permutation). This has indeed been a confounding issue through-out the creation of this manuscript.

It is however important to remember that p-values do not make any claims about the **truth** of a hypothesis, and as such, their actual values are not very relevant to a meaningful comparison of the results of different methods. In my humble opinion, the only meaningful comparison is one from the **pragmatic** point of view. Will using a tool (viewed as a black-box, with parameter-settings according to recommendations made by the tool's authors) result in hints towards planning biologically meaningful follow-up experiments? Are "Prion diseases" in a Dengue dataset, "Cholera infection" in a breast cancer dataset and "small cell cancer" in Tuberculosis examples of such meaningful hints? This said, the chapter mentions several alternative criteria for comparison, including robustness across independently collected biological datasets (3.4.3).

One could argue that e.g. the sample size is still the number of genes, not the sum of weights of present GPS (the cumulative amount of encountered evidence). Presumably, this is the case because e.g. microarrays measure the expression levels of genes, not of GPS. Let us, however, recall that even for the same dataset and same traditional ORA method, the choice of biological identifier-type to which the probes are mapped (e.g. Entrez vs. Ensembl genes) can affect p-values, significance and ranking of the identified pathways (a good example is the Dengue fever dataset in this chapter, for which Entrez and Ensembl mappings lead to different pictures using the standard hypergeometric test in InnateDB). In light of this observation, would it be desirable that all ORA based analyses should cease until further notice?

This being said, it is reasonable to assume that if the sample size were to be considered to be the number of genes, then the p-values should be calculated by gene re-sampling. This is a valuable insight and an avenue that I did follow as part of development of SIGORA. Unfortunately, there are several confounding issues with this presumably methodologically sound approach, which would make a good research topic in their own right: Should the simulated samples correspond to the original sample only in their size? In size and overall co-expression behavior under 'normal' circumstances? In size, correlation behavior and annotation distribution? The results are indeed each time different depending on the respective underlying definition of 'similarity' for gene samples.

The comparably small suggested significance threshold (of 0.001) might raise concerns about increased power of the analysis. Examining the Appendices B-D (where all results for all methods are listed, regardless of significance threshold) shows that even at a significance threshold of 0.05, there would be no more 'significant' pathways in SIGORA's results than there are in InnateDB's results (which uses the traditional IG-ORA). The same observation is true for the simulated datasets: E.g. using a significance threshold of 0.05 and FDR as MTC, the entries for SIGORA in the first column of Table 3-4 would be '10.42' and '7.93' (instead of '8.87' and '6.41'), which is sill considerably smaller than the number of pathways returned by any of the other methods.

Getting pathway analysis right continues to remain a hotbed of fierce debate. A few examples were mentioned in the previous sections, additional examples for debates on methodological soundness of presumably established methods can be seen in (Tamayo et al.; Tripathi, Glazko, and Emmert-Streib 2013). In this light, clearly the method introduced here won't be the last word in pathway analysis or the only possible or 'correct' way of using pathway-GPS to highlight relevant pathways. For the inclined readers who accept that -given the limitations of 'bag of genes' pathway models- the concepts of 'GPS' and 'present GPS' that were introduced in this chapter might be worthy of their consideration, but object to the way the p-values (or maybe more appropriately: 'p-value like scores') of SIGORA are computed, it is relatively easy to obtain the list of all precompiled GPS for all repositories from SIGORA in order to develop their own (improved) methods.

3.6. Conclusions

This chapter highlights the level of component sharing between pathways and demonstrates how this can lead to misleading/spurious results in current pathway analysis approaches that treat all pathway members as equally informative. Here we introduce a novel approach that accounts for the overlapping structure of pathway annotation by focusing on unique features ('*Signatures*') of pathways. To our knowledge, this is the first over-representation based method to do so.

Applied to several published datasets, our approach highlights biologically meaningful processes that would otherwise fall below statistical significance thresholds, and avoids some of the biologically implausible processes highlighted by other methods. This suggests that our approach delivers a useful complementary tool for pathway analysis.

Chapter 4. Application of Guilt by Association to a bovine tissue-expression dataset

4.1. Abstract

Background: Orthology based annotation transfer (as used for reconstruction of bovine pathways in InnateDB) is a useful approach, but it cannot be applied to genes without known orthologs or genes that don't have a functional annotation in any species. An alternative, complementary approach for predicting the biological roles of such genes is the application of the *"guilt by association (GBA)"* principle to expression or interaction datasets. The bovine gene atlas (BGA) is a large tissue expression dataset containing the count statistics for 175,203 unique digital gene expression tag sequences in 105 tissue samples.

Objective: This chapter examines the application of guilt by association to the BGA dataset, to formulate hypotheses about potential gene functions from similarity of expression behaviour.

Results: Application of two different normalization strategies, followed by calculation of pair-wise Pearson's correlation coefficients (PCC) across all tissues, resulted in construction of two very different co-expression networks from the BGA dataset at the same correlation cut-off threshold (PCC> 0.9). Concordance with bovine protein-protein interaction networks (PPIs) and KEGG-GPS (Gene-Pair Signatures), as well as additional considerations on overall robustness to both normalization methods, were used to select one of these two networks for further analysis. The selected network was then subjected to clustering and functional analysis. Functionally enriched clusters satisfying additional quality criteria were selected to generate hypotheses on possible biological roles of (functionally) un-annotated genes that were contained in them.

4.2. Introduction

In higher organisms, the biological roles and functions of many genes and proteins are as yet unknown. In Chapter 2, we reported that - using orthology – more than 80% of human pathways can be reconstructed in cow. The genes annotated by that approach, however, correspond to only about a third of all bovine genes. Orthology based annotation transfer (as used for reconstruction of bovine pathways in InnateDB) is a useful approach, but it cannot be applied to genes without known/predicted orthologs or genes that don't have functional annotation in any species. An alternative, complementary, approach for predicting the biological roles of such genes is the application of the *"guilt by association (GBA)"* principle. An early step in most GBA approaches is the construction of a *"functional linkage"* network. In such networks, two nodes (genes) are connected by an edge if there is some indirect 'evidence' that they might share a common function. The underlying assumption is that the position of a given node within the network and known functions of its neighbors can help in generating new hypotheses about its function. A further assumption is that connectivity analysis or functional enrichment analysis can be used to further prioritize the hypothetical functions (Wolfe, Kohane, and Butte 2005).

Co-expression networks are a special case of functional linkage networks, where two genes are connected by an edge if they exhibit similar expression behaviour across conditions or tissues. Here, one assumes that correlated, seemingly coordinated, expression behavior across (subsets of) tissues or conditions potentially provides some evidence for common regulatory mechanisms like transcription factors, shared promoter regions or regional transcriptional domains (N. Chen and Stein 2006) and/or collaboration towards a common purpose, i.e. membership in a common module or pathway.

The similarity of expression behavior of genes is often measured by their pair-wise Pearson's correlation coefficient (PCC), which is defined as the covariance of two variables divided by the product of their standard deviations and provides a measure for the linear dependence of the genes being compared. After calculation of pair-wise PCC for all gene pairs in a dataset, the network is constructed either by connecting gene pairs that have a pair-wise PCC above an arbitrary threshold, or by linking each gene to the

top k most similar genes (k>1) (i.e. using the top k highest PCC scores for each gene). Regardless of method used to construct the network, it is important to note that the intuitive idea -that '*similar expression behaviour suggests similar function*' is both extremely vague (hard to formalize, hard to exploit and hard to evaluate) and in many cases not true (Box 4-1).

Box 4-1: Known limitations of GBA for expression datasets

In co-expression networks, functional linkage is established through *similarity of expression behavior.* There is, however, some inherent vagueness in the concept of 'similarity'. Different similarity metrics e.g. Pearson's correlation coefficient vs. Spearman's rank correlation - can result in substantially different co-expression networks. If pairwise Pearson's coefficient correlation (PCC) is applied, the choice of the samples that are used for the co-expression analysis can affect the correlation score for a given gene pair (Usadel et al. 2009). This is in part due to the sensitivity of PCC to single sample outliers: two genes that both reach their highest expression levels (within a given set of samples) in the same sample will automatically score a high correlation coefficient regardless of their differences in the remaining samples.

An important confounding issue for both the inference of functions and evaluation of inferred functions concerns the depth, quality and extent of existing functional annotations: the notion of '*genes with known function*' used in GBA to predict the functions of 'genes of unknown function' needs to be considered critically, as only a minority of genes has been exhaustively studied and the set of known functions of a gene are in most cases incomplete. Hence, a substantial portion of genes with known function might have many additional, as yet 'unknown functions' (Peña-Castillo et al. 2008). In contrast to '*molecular functions'*, '*biological process'* roles are systems-level properties (McGary et al. 2010). As a result, in case of genes with multiple 'known functions', the subset of relevant functions are often context sensitive (c.f. the *BRCA1* example in chapter 3). When using GO annotations in GBA, there are additional caveats regarding the quality of annotations with an unreviewed computational evidence code (Rhee et al. 2008), which are marked as "Inferred from Electronic Annotation (IEA)". Many '*biological process'* annotations for higher organisms are unreviewed: As of release of March 2014, over 87% of GO_BP annotations for Bos *taurus* genes carry the evidence cod IEA. Multi-functionality also affects the evaluation of GBA methods. A commonly used evaluation procedure for assessing the quality of the inferred functions is *cross-validation*, a procedure in which the known functions of a random subset of nodes is masked, and one determines how well a GBA method retrieves this masked functions. However, the authors of (Gillis and Pavlidis 2012) recently reported that the existence of '*exceptional links*' between highly multifunctional nodes can also skew the results of cross-validation.

Finally and most importantly, the dynamics of mRNA reflected in co-expression data are only *one* aspect of information flow from genome to protein and correlated expression is neither necessary nor sufficient for functional linkage (Y. Huang et al. 2007; Usadel et al. 2009). Co-expression GBA has been reported to work better for plants than in animals, partly because the more complex tissue organization and regulatory mechanisms in animals, as well as the higher frequency of alternative splicing in mammals hinder a precise evaluation of the strength of gene coexpression (Obayashi and Kinoshita 2011).

Many online resources including COXPRESdb (Obayashi and Kinoshita 2011), Gemma (Zoubarev et al. 2012), and GeneMania (Mostafavi et al. 2008) can be used to obtain co-expression (and/or other functional-linkage) networks or to lookup a gene within such networks.

GBA has been extensively applied to mRNA expression, protein interaction and genomic sequence datasets as well as networks constructed by integrating such datasets (Wolfe, Kohane, and Butte 2005; Eisen et al. 1998; S. K. Kim et al. 2001; Marcotte et al. 1999; H. K. Lee et al. 2004; Usadel et al. 2009; Chua, Sung, and Wong 2006; I. Lee et al. 2008). Early network-based GBA methods used simple majority votes in local neighborhoods (Marcotte et al. 1999): a gene with unknown function would be assigned the function of the majority of its neighbors in a co-expression or interaction network. Using a compendium of gene expression datasets in *C. elegans*, (S. K. Kim et al. 2001) constructed a three-dimensional expression map that displays correlations of gene expression profiles as distances in two dimensions and gene density in the third dimension, and showed how terrain map mountains in this map can be interpreted as clusters of genes with similar function. Network based neighborhood based methods were later extended to include second degree neighbors by (Chua, Sung, and Wong 2006) who noted that in protein-protein interaction networks “a substantial number of proteins are observed to share functions with level-2 neighbors but not with level-1 neighbors”, while extending the radius beyond second degree neighbors was generally shown to reduce the quality of the results.

A different family of methods (called “label propagation algorithms”) generalize local neighborhoods into a more global ‘diffusion’ process (P. I. Wang et al. 2012). Label propagation is an iterative process in which starting with a set of *seed* nodes (nodes with ‘known function’), a node’s function at each step “spills over” to its immediate neighbors. The function assignments for all nodes are then updated as the weighted average of the flow into the node and the node’s previous status. Eventually, label propagation methods return continuous values for ‘probability’ of a node being associated with particular functions.

As a general trend, over the past decade, positions on viability of the assumptions behind the application of guilt by association to animal co-expression

datasets have shifted. While (Wolfe, Kohane, and Butte 2005) claimed that the principle is 'generally and systematically applicable' to human datasets, more recent publications (Gillis and Pavlidis 2012) conclude that 'guilt by association is rather the exception than the rule'. This is in part due to the fact that as demonstrated *in vivo* for *C. elegans* (Z. Zhao et al. 2005) even in simple animals, co-expression does not necessarily imply co-regulation, and similar expression patterns can arise from distinct regulatory mechanisms. The neighborhood relationship in the co-expression network (i.e. seemingly coordinated expression patterns, as measured by correlation analysis of expression data) may or may not reflect a neighborhood relationship in the genomic sequence. Clearly, a combination of shared genetic loci, common direction of transcription and coordinated transcription, as described for *C. elegans* in (N. Chen and Stein 2006), is a stronger indicator of common regulatory mechanisms and functional relationships than possibly spurious similar expression patterns alone.

A related, concurrent, trend is the rise of probabilistic functional gene networks (PFGN), that combine heterogeneous types and sources of biological information into a single, predictive model, in order to increase both the reliability and coverage (number of genes) of the network (I. Lee et al. 2010). The individual datasets might -for instance- contain gene expression patterns, protein-protein interactions, genetic-interactions, gene-disease association and mutant phenotype data from several different species. In order to combine these very heterogeneous lines of evidence into a single model, each individual network is assigned a weight according to the level of its concordance with known co-annotation networks. Integrating different layers of information in this way is expected to enhance the reliability of the model, because links detected by several methods are considered more likely to be of functional relevance. For instance, observation of conserved correlated expression across several species reduces the possibility of experimental/technical artifacts. This type of GBA has been applied to *C. elegans* (I. Lee et al. 2008), mouse (Peña-Castillo et al. 2008), and *Arabidopsis thaliana* (Horan et al. 2008).

Before such combined models can be created for cow, bovine-specific gene co-expression networks are needed. These approaches in cow have been substantially limited to date by a lack of gene expression from a large number of conditions or tissues.

Here, we report the construction and analysis of a bovine gene co-expression network from a large bovine tissue expression dataset.

4.3. Material and Methods

4.3.1. The Bovine Gene Atlas (BGA) Dataset

Through a collaboration with USDA, we obtained the dataset behind the Bovine Gene Atlas (BGA) (Harhay et al. 2010). BGA is a genome-wide transcriptomic study across 105 tissue samples (87 unique tissue types) from a cow (L1 Dominette 01449, the Hereford cow whose genome serves as the bovine reference sequence), her calf, fetus and sire. The BGA dataset is the result of sequencing 20-base tags (including the GATC restriction site) from the 3'- most restriction site that were obtained by restriction digestion of bovine cDNA with the enzyme DpnII. BGA is currently "the deepest and broadest transcriptome survey of any livestock genome" (Harhay et al. 2010).

The profiling technology used for the creation of the Bovine Gene Atlas dataset is called Digital Gene Expression (DGE). DGE was a relatively short lived NGS technology and there are far fewer published studies using this technology than there are e.g. RNA-seq studies. Compared to RNA-seq, there are both some limitations and advantages: the required sequencing depth is smaller and there is no need for assembly of short reads, because –as per protocol- every gene should be represented by at most one tag (there is at most one most 3' restriction site per gene). On the other hand, the coverage (the number of genes whose expression can be profiled by this technology) is smaller in DGE than in RNA-seq studies (not all cDNAs have a DpnII restriction site) and partial digestion and potential isoforms might distort the quality of the final transcript counts for a portion of genes (Asmann et al. 2009; Harhay et al. 2010; Nicolae and Măndoiu 2011).

The BGA dataset contained count statistics for 175,203 unique digital gene expression tag sequences in 105 tissue samples (libraries). In total, 102,901 sequences had been mapped to 17,392 Wikigenes-IDs (Hoffmann 2008). For the analysis presented here, the tissue-expression data for a subset of 13,447 genes from this dataset, that fulfilled the following criteria, was used:

- The tag could be mapped to a unique Ensembl gene.
- The gene had a total transcript count (summing up across all 105 samples) above 10 tags per million.
- If multiple transcripts were mapped to a gene, then only the tag comprising at least 70% of the total expression over all tissues was considered to represent the gene and all other tags were discarded. If such a tag was not present, all tags were discarded and the gene was excluded from the subsequent analysis (because as per experimental design, there should be one tag type per gene: the GATC restriction site followed by 16 bases from the 3'-most DpnII binding site).

4.3.2. Normalization and network construction

Expression profiling by Digital Gene Expression technology has not been used in many studies. In contrast to other NGS technologies, (e.g. RNA-seq ((Bullard et al. 2010)) technology-specific guidelines for the statistical handling of this type of datasets are virtually non-existent. The prevailing assumption for analysis of DGE data is that, due to the experimental protocol (which in theory results in at most one tag per gene, independent of the length of the gene), "more complex normalization of the data is not necessary" (Asmann et al. 2009).

For my analysis, two different normalization methods (described below) were applied to the BGA dataset. For each normalization strategy, pairwise Pearson's correlation coefficients (PCC) were calculated for all 13,447 genes across all 105 tissue samples, resulting in approximately 9,000,000 pairwise correlation scores per normalization method. For each normalization method, gene pairs displaying strongly correlated expression behaviour (gene pairs with PCC > 0.90) were selected for construction of a co-expression network.

1) **Tag per million (TPM) normalization**: In TPM, each sample is independently scaled so that the expression values sum up to one million (i.e. the expression value of each gene in a given sample is multiplied by one million divided by the sum of all transcript counts for that sample). Tags per million is the method used by (Harhay et al. 2010) to examine the relations between different samples (tissues) in BGA.

2) **Bi-stochastic normalization (BSN)**: is an approach successfully used for bi-clustering of microarray data (Kluger et al. 2003), and has been reported to improve clustering performance in other contexts (F. Wang et al. 2011). Bi-clustering (Oghabian et al. 2014) allows for simultaneous clustering of genes and conditions, thereby acknowledging that two different types of similarity contribute to a meaningful analysis of correlated gene expression: a) the expression levels of co-regulated genes are expected to show correlated behavior and at the same time b) the expression profiles of closely related samples (tissues) are expected to be correlated. A bi-stocastically normalized expression matrix is a matrix in which all rows (each representing the expression level of a single gene across all samples) sum to a constant and all columns (each representing the expression levels of all genes in a single sample) sum to a different constant.

4.3.3. Network analysis, network Clustering and functional analysis

To obtain an estimate of the quality of the constructed co-expression networks, both networks, as well as their intersection, were examined with regard to their overlap with a) inferred bovine protein-protein interactions (from chapter 2), and b) the KEGG-GPS (from chapter 3).

The agreement of co-expression networks with the PPI or GPS network was measured in terms of precision and recall. For the purpose of this analysis, precision was defined as the fraction of suggested functional links between conserved genes (in the co-expression network) that corresponded to an edge in the PPI or the GPS network, ("confirmed divided by suggested"). Similarly, recall was defined as the number of retrieved functional links (between conserved genes in the co-expression network) divided by the number of all known connections between the same genes in the PPI or GPS network ("found divided by known").

The network with the more favorable performance (the BSN network) was then subjected to network clustering using the Cytoscape (Smoot et al. 2011) plug-in MINE (Rhrissorrakrai and Gunsalus 2011), to identify groups of densely interconnected genes (set of genes that are more strongly connected to each other than to the rest of the

network). MINE is a soft network clustering tool, meaning that an individual gene can be a member of multiple clusters, thereby potentially allowing for pleiotropy. Clusters identified by MINE were analysed to identify statistically enriched Gene Ontology biological processes (GO-BP), using the Hypergeometric test. The enrichment analysis was performed using the Cytoscape plug-in BINGO (Maere, Heymans, and Kuiper 2005) and custom bovine gene ontology and annotation files (downloaded from http://www.geneontology.org in May 2013).

To prioritize ensuing hypotheses about possible biological functions of genes without functional annotations in these clusters, additional size, enrichment and proportional filters were applied: clusters were discarded if they contained less than four genes or contained less than three 'genes with known function' (GKF), if they were not enriched in any biological processes (FDR>0.05) or if none of the top 10 statistically significant functions for the cluster was associated with at least 40% of the GKF in the cluster. Approximately 40% of the highly prioritized hypotheses were evaluated by literature search in PubMed.

4.4. Results

4.4.1. Comparison of the BSN and the TPM network

The application of the two normalization methods (bi-stochasic vs. tags per million) to the BGA dataset resulted in two very different co-expression networks: in the TPM network, there were over 104,000 suggested functional links between pairs of distinct genes at PCC>0.9, whereas at the same cut-off threshold (PCC>0.9), there were only 28,011 suggested functional linkages in the BSN network. The number of nodes (genes) in the two networks was less dramatically affected (3,100 vs. 2,841), i.e. on average, the application of tags per million normalization resulted in over three times more functional linkages per gene than the application of bi-stochasic normalization (67 vs. 20). The intersection of BSN and TPM networks contained 20,964 gene pairs, i.e. 75% (20,964 out of 28,011) of BSN links (or 20% of TPM links) were robust with regards to the effects of the two normalization methods (Table 4-1).

4-1 Properties of the coexpression network at fixed PCC threshold for two different normalization schemes

Network/ Attribute	BSN	TPM	Intersection of BSN and TPM
1. V (genes)	2,841	3,100	2,239
2. E (links)	28,011	104,461	20,964
3. Number of genes with conserved human ortholog	1,765	1,859	1,339
4. Number of edges (in E) that correspond to conserved PPI	142	131	82
5. Conserved PPIs between genes from V	2,023	2,326	1,219
6. *recall(PPI)*	*7%*	*5.6%*	*6.7%*
7. Number of edges that connect two conserved genes in the network	9,329	52,124	5,864
8. precision(PPI)	*1.5%*	*0.025%*	*1.3%*
9. Number of edges (in E) that correspond to (human) KEGG-GPS	530	628	462
10. KEGG-GPS consisting of genes from V	14,408	15,636	9,829
11. recall(KEGG-GPS)	*3.6%*	*4.0%*	*4.7%*
12. precision(KEGG-GPS)	*5.6%*	*1.2%*	*7.8%*

BSN: bi-stochasic normalization, PCC > 0.9. TPM: tags per-million normalization, PCC> 0.9. For both networks, approximately 60% of the nodes were conserved genes (having a human ortholog) and approximately 12-14% of the nodes were KEGG-PUGs. The values in the 6th row are calculated by division of the 4th row by the 5th row. The values in the 8 row are the result of division of the 4th row by the 7th row. Row 11 is obtained by dividing row 9 by row 10. Row 12 equals row 9 divided by row 7.

Comparison to co-annotation networks

As previously mentioned, congruence of correlated expression and co-annotation is sometimes used to estimate the quality of a co-expression network (I. Lee et al. 2008). Exploiting the fact that KEGG-GPS (chapter 3) represent a special class of co-annotated pairs, and that GPS come in the same format (pairs of genes) as co-expression data, it was determined how many of the proposed functional links in each network correspond to GPS.

In terms of agreement with KEGG-GPS at the fixed PPC threshold of 0.9, the BSN network had a 4.7 fold better precision than the TPM-network (5.6% vs. 1.2%, Table 4-1), and a comparable recall (3.6% vs. 4.0%). The proportion of GPS in the BSN network (the precision regarding GPS) was also six fold larger than the proportion of GPS in randomly constructed networks from the same genes (in average, 0.09%), whereas the performance of the TPM network was only 1.3 folds better than random.

For both networks, the fraction of links corresponding to GPS seems very small and this is certainly a further empirical confirmation of the caveats regarding inference of functional annotations from co-expression data (Box 4-1). KEGG-GPS are, however, a conservative measure of co-annotation (only considering co-annotation of orthologs in manually curated pathways, and only counting pairs that co-occur in a single pathway). For comparison, in section 3.4.4 of this thesis, we reported that for the list of most highly correlated human gene-pairs obtained from COXPRESdb, there were 860 KEGG-GPS among 33,000 unique gene pairs – corresponding to a precision score of 2.6%.

Co-expression and protein-protein interactions

The co-expression networks were also compared to the inferred bovine protein-protein interaction network (as present in InnateDB). Again, for both networks (TPM and BSN at PCC 0.9), co-expression seemed to be a poor predictor of known protein-protein interactions, for which there are several possible explanations: many interacting protein pairs are not necessarily co-expressed, co-expressed genes do not necessarily interact and the known interactome is incomplete. Again, relative to its size, the BSN network contained six times more conserved protein-protein interactions than the TPM network (1.5% vs. 0.25%), with both networks recognizing a comparable portion of all PPIs between their respective genes (7% of protein-protein interactions between the genes in BSN network corresponded to links in BSN network and 5.6% for TPM) (Table 4-1). Of the co-expressed gene pairs from the BSN network that did correspond to protein-protein interactions, 28 (20%) also corresponded to KEGG-GPS; and half of these (13) were annotated to complement and coagulation cascades. To be sure, complement components work as a functional complex and must be co-expressed together. The prominence of this sub-network in the overlap of the interaction- , the co-expression- and

the co-annotation-network possibly reflects the bias in the interactome of InnateDB towards innate immunity related interactions.

Amplification of differences in tissue specificity

The discrepancies in the number of strong pairwise correlations between the two networks (28,000 vs. 104,000 gene pairs) were in part due to the fact that bi-stochastic normalization amplifies differences in tissue-specificity. An illustrative example is the case of C2 (complement C2 precursor) and CYP1A2 (Cytochrome P450 1A2). In the BGA dataset, CYP1A2 is detected only in the liver at 87 tags per million (tpm), while C2 is detected in the liver at 29 tpm and (at much lower abundances) in 24 additional tissue samples, including *'perirenal adipocytes', 'spleen', 'ileum'* and *'thyroid'*. Since both genes achieve their highest expression level in the liver and both genes are not detected in 80 additional samples, in the TPM-normalized dataset, the two genes have a PCC of 0.92, leading to a functional link between the two genes in TPM-network. After bi-stochasic normalization, the correlation coefficient between these two genes drops to 0.67, hence the BSN network does not include a link between C2 (a gene with primarily immunological functions) and CYP1A2 (a gene with primarily metabolic roles).

MicroRNA-mRNA linkages in BSN and TPM networks

The BGA data set contains 31 miRNAs. Although these miRNA-genes (as to be expected) in general exhibit slight to moderately negative correlation with most genes (including their putative/verified targets), there are also a few strong positive correlations involving MIRNAs in both the BSN and the TPM network.

Interestingly, despite its smaller size, the BSN-network contained more microRNA-gene functional linkages than the TPM-network: In BSN, there are 190 correlations (PCC>0.9) between 11 distinct microRNAs and 190 distinct genes. Using tags per million normalization, there are only 6 links between genes and miRNAs among the top 28,000 correlations. The number of such links in the entire TPM network (over 100,000 edges) is 155.

MIR-122 is the miRNA gene with the highest number of strong positive correlations (PCC> 0.9) in the BSN network (71 correlations), followed by MIR223 (34 correlations) and MIR-140 (27 correlations). These positive miRNA-mRNA correlations

might encode some meaningful information. The subsequent clustering of the BSN network (see next section) suggests for example a relationship between MIR-122 and coagulation, which agrees with the results of a recent study in human sepsis patients, that identified serum levels of MIR-122 as a predictor of aberrant coagulation (H.-J. Wang et al. 2014).

4.4.2. Module identification in the BSN network

Analysis of high-throughput expression datasets is often confounded by "the curse of dimensionality" or the "problem of under-determination" (De Smet and Marchal 2010): there are usually far more genes than available samples. Clustering of expression data is a way of reducing dimensionality by grouping mutually similar genes together. In the case of co-expression data, network clustering (i.e. identifying groups of genes that are densely interconnected with each other, but only loosely connected to the rest of the network) can reveal functionally coherent groups of genes (H. K. Lee et al. 2004).

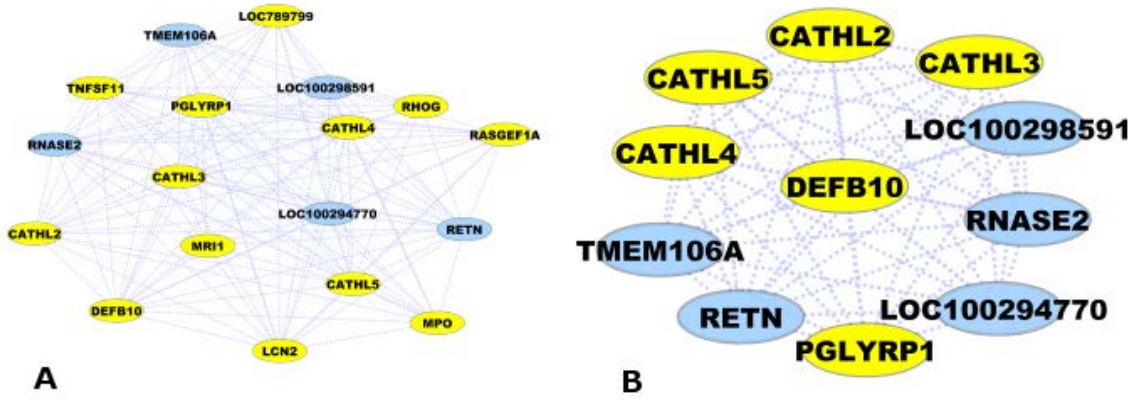
The BSN network (2,841 genes, 28,011 edges) was subjected to clustering by the Cytoscape (Smoot et al. 2011) plug-in MINE (Rhrissorrakrai and Gunsalus 2011), resulting in 127 modules of size 3 to 96 (Thereof 97 modules containing at least 4 elements). MINE is a soft network clustering tool, meaning that an individual gene can be member of multiple clusters, thereby potentially allowing for pleiotropy.

4.4.3. Function prediction and literature evaluation

In total, 2,120 genes were member of at least one MINE cluster. Among these, 1,500 had an associated Gene Ontology-Biological Process (GO-BP) function (Ashburner et al. 2000) , while 612 were not previously annotated. To infer the potential functions of the unannotated genes, first any clusters with less than 4 total genes, as well as clusters with less than 3 (GO-) annotated genes and clusters without any unannotated genes were discarded. Next, each of the remaining 45 modules were subjected to GO-BP enrichment analysis, using the Hypergeometric test, and ontology and bovine annotation files downloaded from http://www.geneontology.org in May 2013. For each module, only GO-BP categories achieving a False Discovery Rate (FDR) < 0.05 (Benjamini and Hochberg 1995) were considered statistically significant.

Examining the results of this ORA analysis showed that for 16 out of the 45 modules, the highest ranking GO-BP category (i.e. the category with the smallest p-value) contained more than 40% of all of the annotated genes. For such modules, the GO-BP category with the smallest p-value was assigned as the putative function of all un-annotated genes in the module. One example of such modules is shown in Figure 4-1:

Figure 4-1 an example for assignment of function by GBA.



This cluster contains 18 genes, 13 of which are annotated in GO- BP (panel A). Out of 13 annotated genes, 6 (46%) are annotated in "defense response to bacterium" (yellow in panel B), which is the category with the smallest p-value for this module (FDR 1.12E-07). Hence, the remaining 5 un-annotated genes in the cluster (blue in both panels) were also assigned to this category (i.e. "defense response to bacterium"). Notably, a subsequent literature search provided support for involvement of 4 of these genes (or their orthologs) in this category (Table 4-3).

For modules where the highest ranking GO-BP category contained 40% or less of all of the annotated genes, it was examined if any of the top ten significant GO-BP categories satisfies this criterion. This resulted in the putative assignment of (more general) GO-BP categories to un-annotated genes in 13 additional modules (Appendix E).

Overall, 180 previously uncharacterized bovine genes could be assigned a putative biological process by the approach outlined above. Table 4-1 lists existing literature evidence for 36 (20%) such predictions.

Table 4-2 Example of GBA based gene function predictions that are supported by literature.

Gene	Predicted putative function	PubMed ID of supporting publication
C5AR1	immune system process	23402022
CABP5	visual perception	18586882
CD180	immune system process	15852007
CD22	immune system process	1007929
CD300A	regulation of phagocytosis	22302738
CD53	immune system process	8335905 20407468
CLEC12A	regulation of phagocytosis	14739280
DAPP1	immune system process	21930970
FAIM3	immune system process	22675200
FAM65B	immune system process	23241886
FAM81B	cellular component assembly (cilia)	17971504
GMFG	immune system process	23677465
IL10RA	regulation of phagocytosis	10433356
LOC100294770	defense response to bacterium	8454635 (cow)
LOC100298591	defense response to bacterium	22138257
LRRC23	cellular component assembly (cilia)	17971504
LYZ	immune system process	23578963
MIR140	single-organism developmental process	21576357
MIR146A	immune system process	23028621
MIR223	immune system process	22937006
PDZK1IP1	transmembrane transport	19447883
PIK3R5	immune system process	21277760
PILRA	regulation of immune system process	21241660
POPDC2	circulatory system development	22290329 (zebrafish)
PSORS1C2	epidermis development	12664160
RETN	defense response to bacterium	12387885
RNASE2	defense response to bacterium	23711849 15032578
RP1L1	visual perception	22466457
SAMD9	immune system process	23758988

Gene	Predicted putative function	PubMed ID of supporting publication
SBSN	epidermis development	17330888
Six6	pituitary gland development	10473118
SLAMF6	immune system process	18501771
SLC7A13	transmembrane transport	19184091
SPAG6	cellular component assembly	12167721
VSNL1	synaptic transmission	18989702
WDFY4	immune system process	20169177

4.4.4. Tissue specificity

Overall, 106 genes were detected in only one tissue sample (at TPM > 10). The vast majority of these genes (91) are only expressed in testis (cluster 1 in Appendix C). There are also five liver-specific genes (CYP1A2, INMT, UGT2B4, LOC100138908, LOC100140261), seven retina-specific genes (CABP5, GRK7, KCNV2, NR2E3, OPN1SW, PDE6H, TEX28), as well as one bone specific (ACAN), one kidney-specific (LOC506670) and one abomasums-specific gene (PGC).

A substantial portion of the testis specific genes (74 genes) do not have any predicted human orthologs. These findings are consistent with the result of an independent study (Brawand et al. 2011), that had examined the evolution of gene expression in six tissues from 10 mammalian species and birds, and reported accelerated changes in evolution of the transcriptome in testis.

In the BGA dataset, MIR-122, which has been previously reported to be a liver-specific miRNA (Jopling 2012) is detected at comparable levels also in the hypothalamus, and at much lower levels in three additional tissues.

4.5. Discussion, limitations and future work

As mentioned before (Box 4-1), there are a range of known limitations to any co-expression based GBA for inferring function. Additional, dataset specific challenges

arise for the work in this chapter, as publicly available 'best practices' guidelines in dealing with DGE datasets are missing.

After comparing two different co-expression networks derived from applying two different normalization methods to this dataset, I chose the one with fewer predicted links but (proportionally) better agreement with inferred bovine PPI (from chapter 2) and KEGG-GPS (from chapter 3). This network (the BSN network) was then subjected to clustering and functional analysis, and a set of ad-hoc rules were used to select clusters that result in (presumably) most promising hypotheses.

To be sure, there are other possible candidate approaches for the normalization step which should be further explored and compared in future work. Examples of such alternative methods include log-transformation, square root transformation, and row-transformation (division of all values in a sample by the count value for a house-keeping gene in that sample, assuming that this gene's expression will be near constant in all cell-types). Preliminary experiments, however, indicate that each normalization method will come with its own benefits and limitations: log-transformation, for instance, seemed to slightly improve the agreement with the PPI-network, but severely reduce the coverage (the number of genes in the co-expression network at the PPC>0.9 threshold dropped by over 50%).

The reasoning behind the use of density based network-clustering (identification of sub-networks of highly interconnected nodes) instead of seed based diffusion-processes (starting from nodes with known function and iteratively propagating their function through the network) was the idea that the latter might be more sensitive to single missing or spurious exceptional links. Once the clusters are identified, however, one is still left with the problem of selecting the most relevant functions within the cluster. This is particularly true for larger clusters that are enriched in several –at times seemingly unrelated- functions. To be sure, a thorough manual review of all clusters and their functional enrichments - as well as a case by case selection of hypotheses that should be prioritized for verification - would be a better approach to leveraging this dataset than the *ad-hoc* criteria used here (FDR + proportional filter). Any such ad-hoc criteria for hypotheses-selection will inadvertently lead to a substantial number of non-verifiable predictions, while missing other, possibly relevant hypotheses. One illustrative

example is cluster 3 in Appendix E, in which MIR-122 (from section 4-5) is a prominent member (a hub). The cluster consists of 133 highly interconnected genes, 80 of which have known functions. The most significant GO-BP category (FDR < E-19) for this cluster is '*coagulation'; however,* this category is associated with only 17 genes in this cluster. Despite the low p-value, it seems unlikely that all 53 un-annotated genes in this cluster are coagulation related. On the other hand, serum levels of MIR-122 have recently been shown to be a predictor of abnormal coagulation in human sepsis patients (H.-J. Wang et al. 2014).

In future studies, it would be interesting to integrate the co-expression network with the inferred interaction network from chapter 2, in order to identify potential tissue-specific interactions. Once more comprehensive --and reliable-- bovine transcription factor binding site (TFBS) data become available, one could also investigate the clusters of co-expressed genes with respect to transcriptional co-regulation – an approach previously successfully applied to other organisms (Gasch and Eisen 2002). A related, but more topological question is the identification of potential 'date and party hubs' in the inferred PPI network by comparing the expression patterns of hubs (i.e. nodes having a remarkably high number of interaction partners) with that of their interaction partners. Such an approach can potentially distinguish hubs that coordinate specific cellular processes within functional modules ('party hubs') from hubs that organize the interactome by linking different processes ('date hubs') (Barabási, Gulbahce, and Loscalzo 2011).

Chapter 5. *Next Generation Sequencing Reveals the Expression of a Unique miRNA Profile in Response to a Gram-Positive Bacterial Infection*

This chapter is based on a modified version of the article "Next Generation Sequencing Reveals the Expression of a Unique miRNA Profile in Response to a Gram-Positive Bacterial Infection.", co-authored by Lawless, Nathan; Foroushani, Amir B K; McCabe, Matthew S; O'Farrelly, Cliona and Lynn, David in PloS one © The Authors 2013 . My contribution has been the prediction of putative targets for the differentially expressed miRNAs, the analysis of overrepresentation of innate immunity related genes among these targets, pathway analysis of the targets, and the discussion of these results.

.

5.1. Abstract

MicroRNAs (miRNAs) are short, non-coding RNAs, which post-transcriptionally regulate gene expression and are proposed to play a key role in the regulation of innate and adaptive immunity. Here, we report a next generation sequencing (NGS) approach profiling the expression of miRNAs in primary bovine mammary epithelial cells (BMEs) at 1, 2, 4 and 6 hours post-infection with *Streptococcus uberis*, a causative agent of bovine mastitis. Analysing over 450 million sequencing reads, we found that 20% of the approximately 1,300 currently known bovine miRNAs are expressed in unchallenged BMEs. We also identified the expression of more than 20 potentially novel bovine miRNAs. There is, however, a significant dynamic range in the expression of known miRNAs. The top 10 highly expressed miRNAs account for >80% of all aligned reads, with the remaining miRNAs showing much lower expression. Twenty-one miRNAs were identified as significantly differentially expressed post-infection with *S. uberis*. Several of these miRNAs have characterised roles in the immune systems of other species. This miRNA response to the Gram-positive *S. uberis* is markedly different, however, to

lipopolysaccharide (LPS) induced miRNA expression. Of 145 miRNAs identified in the literature as being LPS responsive, only 9 were also differentially expressed in response to *S. uberis*. Computational analysis has also revealed that the predicted target genes of miRNAs, which are down-regulated in BMEs following *S. uberis* infection, are statistically enriched for roles in innate immunity. This suggests that miRNAs, which potentially act as central regulators of gene expression responses to a Gram-positive bacterial infection, may significantly regulate the sentinel capacity of mammary epithelial cells to mobilise the innate immune system.

5.2. Introduction

MicroRNAs (miRNAs) are an abundant class of highly conserved, small (19–24 nt long), non-coding, double-stranded RNA molecules. They act as post-transcriptional regulators of gene expression, altering mRNA stability and translation efficiency by hybridizing to the 3′ untranslated regions (UTRs) of certain subsets of mRNAs (collectively as many as 60% of all mRNA transcripts) (Bi, Liu, and Yang 2009). Since their initial discovery in *Caenorhabditis elegans* in 1993 (R. C. Lee, Feinbaum, and Ambros 1993), researchers have gained much insight into the prevalence of miRNAs in other species. The latest miRBase database (release 19) contains 21,264 precursor miRNAs, expressing 25,141 mature miRNA products, in 193 species (Kozomara and Griffiths-Jones 2010). miRNAs have been shown to play key roles in the regulation of innate and adaptive immunity in humans and mice (O'Connell et al. 2010). miR-146a, for example, regulates the innate immune response to bacterial infection, targeting TNF receptor-associated factor 6 (TRAF6) and Interleukin-1 receptor-associated kinase 1 (IRAK1) (Williams et al. 2008), while miR-150 regulates the production of mature B cells (Xiao et al. 2007). Studies elucidating the regulatory roles of miRNAs in bovine infection and immunity, however, are more limited. Bovine miRNAs are expressed in a wide range of tissues, including immune-related ones (Coutinho et al. 2006), but only a handful of studies have investigated how the expression of bovine miRNAs are altered in response to infection. A recent RT-qPCR study, for example, highlighted the differential expression of five inflammation related miRNAs (miR-9, miR-125b, miR-155, miR-146a and miR-223) in response to *E. coli* lipopolysaccharide (LPS) and *S. aureus* enterotoxin B stimulation of bovine monocytes (Dilda et al. 2012). Two other recent studies have

used a similar approach to identify several miRNAs that were differentially expressed in the mammary gland tissue of cattle with mastitis (Naeem et al. 2012; Hou et al. 2012). These and other studies suggest roles for individual miRNAs in regulating bovine immunity, however, according to Ensembl v66 (Flicek et al. 2011; Hubbard et al. 2009) there are over 1,300 annotated miRNAs in the bovine genome. Therefore, studies which adopt genome-wide approaches are required to gain greater insight into the repertoire of bovine miRNAs involved in immunity and infection.

Although microarray technologies to profile miRNA expression have been around for some time (V. N. Kim, Han, and Siomi 2009), next generation sequencing (NGS) based technologies are revolutionising the field and provide the opportunity to profile the expression of known miRNAs with discriminating resolution and accuracy, and also to identify novel miRNAs (Buermans et al. 2010). Furthermore, these technologies allow one to differentiate between the expression of alternative mature miRNAs from the same precursor and to identify the differential expression of miRNA isomiRs (L. W. Lee et al. 2010). To date, a limited number of studies have applied these approaches to profile miRNAs in different bovine tissues (J. Huang 2011; X. Chen et al. 2010; Guduric-Fuchs et al. 2012) and only one study has used an NGS approach to investigate the expression of bovine miRNAs in response to infection (Glazov et al. 2009).

In this study, we implemented a NGS approach to profile the expression of bovine miRNAs at multiple time-points in primary mammary epithelial cells infected *in vitro* with *Streptococcus uberis*, a causative agent of bovine mastitis. This inflammatory disease of the mammary gland has significant economic impact on the global dairy industry. To the best of our knowledge, this study represents the most comprehensive NGS study to date that profiles the host miRNA response to infection, in any species. In comparison to previous studies, we have sequenced un-pooled miRNA libraries to a previously unprecedented sequencing depth from multiple replicates and controls across multiple time-points, allowing us to explore the statistically significant temporal changes in miRNA expression in response to infection.

5.3. Materials and Methods

5.3.1. Bovine Mammary Epithelial Cell Culture

Primary bovine mammary epithelial cells, which had been isolated from mammary parenchyma, were purchased from AvantiCell (AvantiCell Science Ltd., Ayr, UK) (Blatchford et al. 1999). The source animal was in her third trimester of first pregnancy, was aged between 26–30 months, and was negative for bovine viral diarrhoea and Bovine spongiform encephalopathy. Cells were plated (seed density of 1×10^6) directly onto collagen coated plastic flasks (Greiner-Bio-One GmbH, Frickenhausen, Germany) and immersed in AvantiCell medium – (199/Ham's F12 (50:50) pH 7.4 containing 5% (v/v) horse serum, 5% (v/v) fetal bovine serum, 5 µg/ml bovine insulin, 1 µg/ml hydrocortisone, 3 µg/ml cortisol, 10 ng/ml epidermal growth factor (EGF), 2 mM sodium acetate, 10 mM Hepes, U/ml penicillin/streptomycin and single strength Fungizone™).

Media was initially replaced after 48 h. Cells were split twice (75 cm^2 and 175 cm^2) and were then seeded at a concentration of 1.8×10^5 cells/well into collagen coated 6-well plates. Media was then changed after 24 h, and cells were inspected under microscopy for confluence. Cells were harvested by washing with Hanks Balanced Salt Solution (HBSS) pH 7.4 and treated with 4 ml of 0.25% trypsin for approximately 5 min at 37°C. An equal volume of medium to trypsin (1:1) neutralised trypsin.

Infection of Cells with Streptococcus uberis 0140J

Streptococcus uberis 0140J was purchased from the American Type Culture Collection (ATCC), Virginia, USA (Cat# BAA-854). *S. uberis* 0140J was first isolated in milk obtained from a clinical case of bovine mastitis in the United Kingdom in 1972. *S. uberis* was cultured as per ATCC instructions. BMEs were challenged with *S. uberis* 0140J at a multiplicity of infection (MOI) of 50, over a time course of 1, 2, 4, & 6 h. Three replicates were infected at each time point and three replicate uninfected controls were also maintained for each time point.

miRNA Extraction

Total RNA and small RNA were extracted from each of the 24 samples using the mirVana™ miRNA Isolation Kit (Life Technologies, Carlsbad, CA, USA). Procedures were performed according to the manufacturer's protocol. Briefly, cells were lysed using 500 µl lysis/binding solution directly on a culture plate. 50 µl of miRNA homogenate was added; solution was mixed by vortexing and left on ice for 10 min. 500 µl of acid-phenol chloroform was added and solution was mixed by vortexing for ~60 sec. The solution was then centrifuged for 5 min at 10,000×g at room temperature to separate phases. Aqueous phase was removed and transferred to a separate tube. 1/3 volume of 100% ethanol was added to the aqueous phase and mixed by vortexing. Samples were passed through a filter cartridge (glass-fiber filter) by centrifuge for ~30 sec at 10,000×g. The filtrate was collected (residue on filter contained RNA <200 nt, was retained for later use) and 2/3 volume 100% ethanol was added and mixed by vortexing. Filtrate was passed through a second filter cartridge by centrifuge for ~30 sec at 10,000×g. The flow through was discarded, and the filter was washed with 700 µl wash solution 1 and 500 µl wash solution 2 (twice). After discarding all flow through after each step, the filter was centrifuged for a further 1 min. 50 µl of pre-heated (95°C) nuclease free water was applied to the filter for 1 min, and the filter was centrifuged for 30 sec. Eluate was collected and stored at −80°C. Total RNA integrity was measured by the Agilent RNA 6000 Nano Kit using the 2100 Bioanalyzer (Agilent Technologies, Colorado Springs, CO, USA). The Agilent Small RNA Kit (Agilent Technologies) was used to quantify miRNA.

Small RNAseq Library Preparation and Sequencing

Twenty-four indexed miRNA libraries were prepared using the ScriptMiner™ Small RNAseq Library Preparation Kit (Epicentre, Madison, WI, USA). Procedures were performed according to the manufacturer's protocol. Briefly, a 3′-tagging sequence was added to the 3′- end of the RNA followed by treatment with a degradase enzyme to reduce excess 3′ adaptor oligo. A tagging sequence was then added to the 5′- end of the RNA. The RNA, now tagged at both ends (di-tagged), was purified using the Zymogen RNA Clean and Concentrator (Zymogen, Irvine, CA, USA). The di-tagged RNA was then reverse transcribed into cDNA, and the remaining RNA was removed using RNase. The PCR step used by the ScriptMiner™ kit, is a two stage process, firstly an analytical PCR step was carried out to optimise the number of cycles necessary for

amplification. Once this was determined, the libraries were amplified and adaptors were added. Size fractionation of the miRNA libraries to separate them from adapter dimers was achieved by electrophoresis on an 8% TBE polyacrylamide gel (Life Technologies, Carlsbad, CA, USA) (1.00 mm×10 well). The libraries were then purified from the gel and the Agilent High Sensitivity DNA Kit (Agilent Technologies, Colorado Springs, CO, USA) was used to quantify the molarity and size of finished miRNA-seq libraries. miRNA libraries were randomised across three lanes of a flowcell, with eight indexed samples on each lane. Libraries were sequenced on an Illumina HiSeq 2000 by the Norwegian Sequencing Centre with TruSeq v3 reagents. Fastq files were produced using the CASAVA pipeline v1.8.2. Barcodes (indexes) and adaptor sequences for multiplexed samples are provided (Appendix F).

5.3.2. Small RNAseq Analysis

Preliminary quality control analysis of the 24 fastq files was carried out with FASTQC software v0.10.0 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Cutadapt v1.1 (http://www.cutadapt/) was then used to trim 3′ adaptor sequences. Reads which were shorter than 18 nucleotides after trimming were discarded. Trimmed reads were then further filtered using the fastq quality filter (http://hannonlab.cshl.edu/fastx_toolkit/) v0.0.13. Reads where at least 50% of the bases had a Phred score <20 were removed (Cock et al. 2009). Finally, reads passing all the above filters were also trimmed at their ends to remove low quality bases (Phred score <20). Reads which successfully passed filtering were aligned to the bovine genome (UMD3.1) using novoalign version 2.07.11 (http://www.novocraft.com) using the "-m" miRNA mode. Reads that did not uniquely align to the genome were discarded. HTSeq version 0.5.3p3 (http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html) using the union model was used to assign uniquely aligned reads to Ensembl (v66) bovine gene and miRNA annotation (separately).

miRNAseq fastq files have been submitted to the NCBI Gene Expression Omnibus (GEO) database (Barrett et al. 2010) with experiment series accession number GSE41278.

Differential Expression Analysis

Prior to assessing differential expression, count data were first normalised across libraries using either the trimmed mean of M-values (TMM) normalisation method (Robinson, McCarthy, and Smyth 2009) or upper-quantile normalisation (Bullard et al. 2010). Differential expression analysis of miRNAseq data has been shown to be sensitive to the normalisation approach implemented (Garmire and Subramaniam 2012). To address this issue, we identified differentially expressed miRNAs in three alternatively normalised datasets; TMM-normalised, upper-quantile normalised and no normalisation. Only miRNAs which were identified as differentially expressed across all three datasets were considered further i.e. the differential expression of these miRNAs was robust to the normalisation procedure. As an aside, we found that the two different normalisation approaches resulted in very similar miRNAs being detected as differentially expressed.

The R (version 2.14.1) Bioconductor package EdgeR (v2.4.6) (Robinson, McCarthy, and Smyth 2009), which uses a negative binomial distribution model to account for both biological and technical variability was applied to identify statistically significant differentially expressed miRNAs. Only miRNAs that had at least 1 count per million in at least 3 samples were analysed for evidence of differential gene expression. The analysis was undertaken using moderated tagwise dispersions. Differentially expressed miRNAs were defined as having a Benjamini and Hochberg (Benjamini and Hochberg 1995) corrected P value of <0.05.

Novel miRNA Discovery

In addition to profiling the expression of known miRNAs, miRNAseq data can also be used to identify the expression of potentially novel miRNAs. To do this, miRNAseq data from this study was analysed using the software package miRDeep2 v0.0.5 (Mackowiak 2011). The miRDeep2 algorithm mines high-throughput sequencing data for the presence of multiple sequenced RNAs corresponding to predicted miRNA hairpin structures in the genome. It then uses Bayesian statistics to score the fit of sequenced RNAs to the biological model of miRNA biogenesis. MiRDeep2 predicted a large number of potentially novel miRNAs from our miRNAseq data. We further parsed this data using a number of different parameters to identify those novel miRNAs that have the highest likelihood of being true positives. Specifically, we identified those

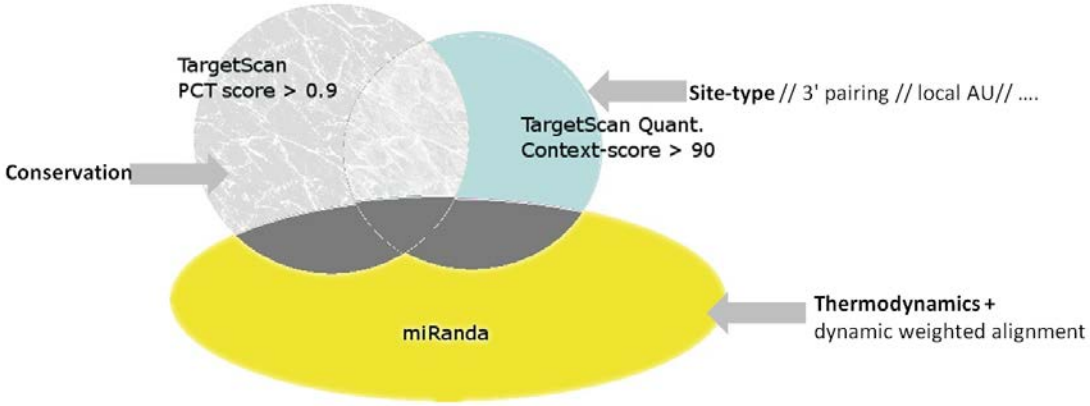
predictions where both the mature and star stands were expressed with a minimum of 5 reads each; where miRDeep2 predicted that the miRNA had >90% probability of being a true positive; where the hairpin structure had a significant Randfold p-value and where the novel miRNA was independently predicted in two or more different miRNAseq samples.

Customised Perl scripts were also written in house to examine the miRDeep2 output for the presence of miRNA isomiRs. These scripts were used to identify isomiRs that were expressed at a level of at least 100 reads and to identify cases where the expression of the isomiR was higher than the expression of the miRBase consensus mature sequence. Furthermore, we identified whether isomiRs were modified at the 5′ or 3′ ends (first and last 5 nucleotides). IsomiRs with >1 mismatch to the reference sequence were excluded from the analysis.

miRNA Target Predictions

Target genes that are potentially regulated by differentially expressed miRNAs were predicted using the consensus of two computational approaches, miRanda v3.3a (Betel et al. 2007) and TargetScan v6.2 (Lewis, Burge, and Bartel 2005; Friedman et al. 2008; Grimson et al. 2007). Given the high false positive rates for miRNA target prediction, we identified only those potential target genes that were predicted by both methods. More specifically, we first established a broad pool of potential targets by applying miRanda to bovine mature miRNA (miRBase v18) and cDNA sequences (UMD3.1, Ensembl v66) under default threshold settings. This resulted in the prediction of thousands of possible target genes per differentially expressed miRNA. To narrow down this pool of potential targets, we used TargetScan to independently identify conserved targets with a PCT-score above 0.9 and/or non-conserved targets with a context+ score above the 90th percentile of all targets of the respective miRNA. Target genes that were not corroborated by one of the two methods (PCT or context+ score) were discarded (Figure 5-1). Pathway analysis of predicted gene targets was undertaken using the SIGORA R package (http://sigora.googlecode.com/svn/) with KEGG pathway annotations (Minoru Kanehisa et al. 2012).

Figure 5-1 miRNA target prediction



Genes that have an annotated role in innate immunity were identified using www.innatedb.com (Lynn et al. 2008), a curated database of innate immunity genes, pathways and molecular interactions.

5.3.3. Results

Isolation of Small RNA from Bovine Mammary Epithelial Cells

Small RNA was isolated for NGS sequencing from *S. uberis* infected primary bovine mammary epithelial cells at 1, 2, 4 and 6 hours post-infection (hpi) (n = 3 infected and n = 3 controls at each time-point). Total RNA and small RNA were examined for quantity and integrity in each of the 24 samples (Appendix G). Total RNA was assessed to be of high quality based on both Bioanalyzer and 28S/18S analysis. RNA integrity numbers (RIN) for total RNA were >8 for each sample. The concentration of miRNA in each sample was also assessed (Appendix G). Sufficient quantities were present to proceed with RNAseq library preparation.

High-throughput Sequencing of Small RNA Libraries Prepared from Bovine Mammary Epithelial Cells

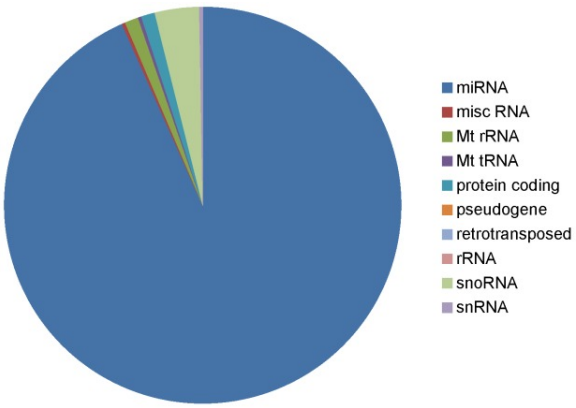
Small RNA libraries were prepared from size selected RNA (<200 nucleotides). Libraries were prepared using the ScriptMiner™ protocol with indexing before cluster generation, sequencing and imaging on an Illumina Hiseq 2000. Samples were

randomly multiplexed over 3 flowcell lanes for sequencing. Sequencing of small RNA libraries yielded more than 450 million raw sequence reads from mammary epithelial cells. Following a pipeline of adaptor removal, quality filtering and the removal of sequences that were too short, more than 213 million reads were retained for further analysis (78,604,161 and 134,850,887 for control and infected replicates, respectively). These filtered reads were then aligned to the reference *Bos taurus* UMD 3.1 genome. Over 116 million reads aligned uniquely to the genome (Appendix F). Reads that aligned to more than one position in the genome were discarded. Uniquely aligning reads were then assigned to known miRNAs using HTseq (http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html) based on Ensembl v66 (Hubbard et al. 2009; Flicek et al. 2011) annotation of the bovine genome.

Repertoire of RNA Species in Small RNA Libraries

The proportion of reads (averaged across 24 samples) uniquely aligning to different RNA biotypes demonstrates that miRNAs are the dominant ncRNA species sequenced in our small RNA libraries (Figure 5-2). The vast majority (>90%) of reads that align uniquely to known ncRNAs align to known miRNAs. There was no significant difference in the proportion of reads aligning to different RNA biotypes in the infected and control samples. The majority of the remaining reads primarily mapped to snoRNAs (Figure 5-2) (Guduric-Fuchs et al. 2012). Although the vast majority of reads align to known bovine ncRNAs, a low density of reads can be observed along each chromosome (Appendix K). These possibly represent mRNA degradation products.

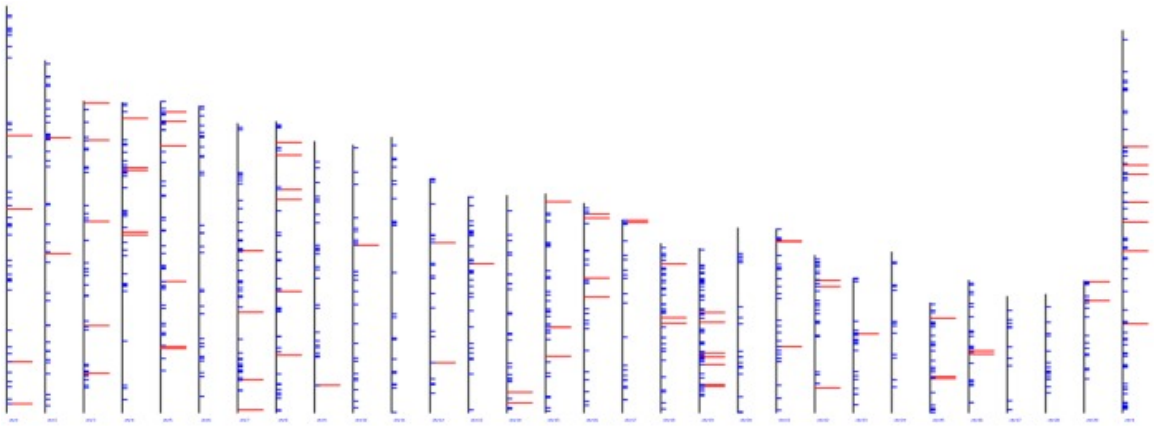
Figure 5-2 The proportion of reads aligning uniquely to bovine ncRNAs (averaged across 24 samples).



The Expression of miRNAs in Primary Bovine Mammary Epithelial Cells

To characterise the bovine mammary epithelial cell microRNome, miRNAs that were expressed at an appreciable level (based on mapped read counts in tags per million sequenced (tpm)) were identified. 276 miRNAs had a count of greater than 1 tpm (Appendix H). Of these, 114 miRNAs were expressed at a level >100 tpm. To determine whether these miRNAs were expressed from related genomic regions, we examined all miRNAs with >100 tpm for genomic clustering (Figure 5-3). There was no evidence of a substantial genomic bias from which these miRNAs were encoded.

Figure 5-3 The genomic position of bovine mammary epithelial cell expressed miRNAs with >100 tpm (red).



Highly expressed miRNAs were relatively evenly distributed across the genome. Vertical bars represent chromosomes, blue horizontal bars known miRNA locations. Red horizontal bars show highly expressed (>100 RPM) miRNA locations.

The top 10 highly expressed miRNAs, which accounted for >80% of all aligned reads (Figure 5-4), were evolutionarily conserved across multiple species. These miRNAs represent seven different miRNA families; miR-let-7 (bta-let-7i & bta-miR-3596), miR-21 (bta-miR-21), miR-27 (bta-miR-27a & bta-miR-27b), miR-28 (bta-miR-151), miR-184 (bta-miR-184), miR-200 (bta-miR-200a & bta-miR-200b), and miR-205 (bta-miR-205). Many of these miRNAs have been shown to have pleiotropic roles in other species (Table 5-1). miR-21 and miR-205, have been shown to have role in cancer,

regulating tumour suppressor genes such as VEGF-A and TGFI-R2 (R. Yang et al. 2009; Wu, Zhu, and Mo 2009; Y. Yu et al. 2011; Yue et al. 2012).

Figure 5-4 The top 10 most highly expressed miRNAs in bovine mammary epithelial cells.

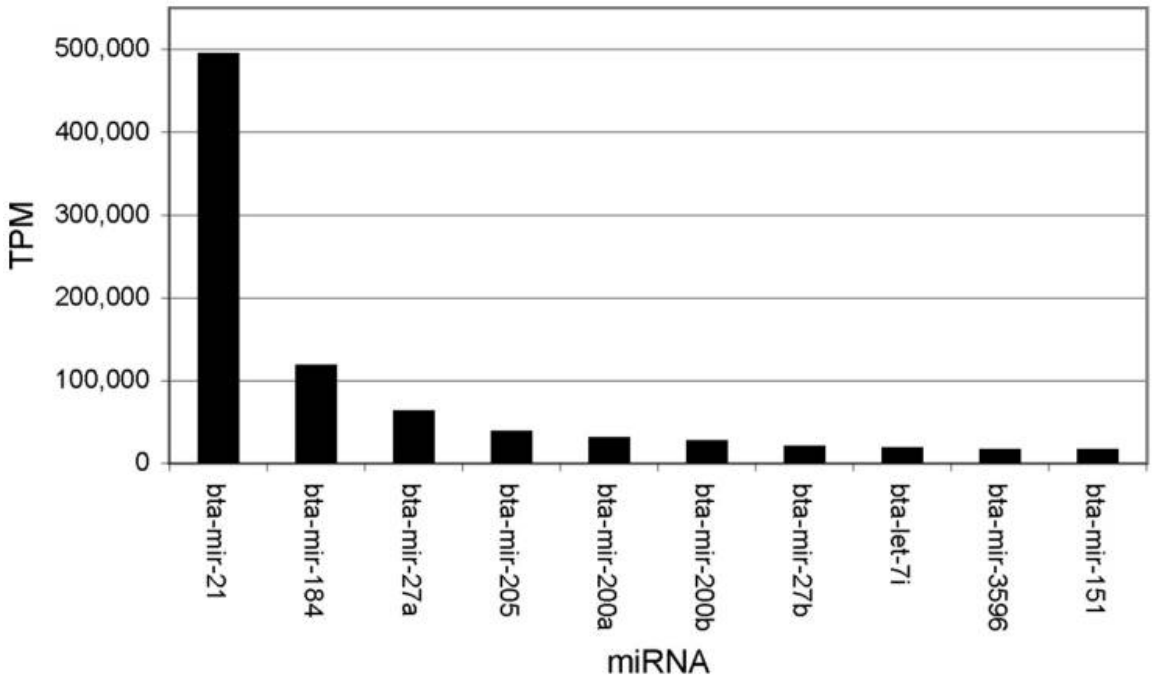


Table 5-1 Highly expressed miRNAs in bovine mammary epithelial cells have been shown to have pleiotropic functions in other species.

miRNA	Species	Tissue	Target	Function	Reference
miR-21	Human	Monocytes	CAMP/DEFB4A	Immune	(P. T. Liu et al. 2012)
miR-21	Human	Colon Cancer Cell	TGFI-R2	Cancer	(Y. Yu et al. 2011)
miR-184	Human	HeLa/HEK	SHIP2	Immune	(J. Yu et al. 2008)
miR-205	Human	MCF-7, MDA-MB-231, MDA-MB-453 and MDA-MB-468 cells	VEGF-A	Cancer	(Yue et al. 2012; Wu, Zhu, and Mo 2009)
miR-27b	Human	Monocytes	PPARgamma	Immune	(Jennewein et al. 2010)

Of particular interest to our study is the fact that several of the most highly expressed miRNAs in BMEs have been shown to have a role in immunity. miR-27b, for example, has been shown to negatively regulate the mRNA stability of peroxisome proliferator-activated receptor gamma (PPARgamma), a transcriptional regulator of the inflammatory response (Jennewein et al. 2010). Interestingly, miR-27b has also been found to be degraded by a viral transcript in lytic murine cytomegalovirus (MCMV) infection, further highlighting its role in immunity (Marcinowski et al. 2012). miR-21 has also recently been shown to be the most highly expressed miRNA in *Mycobacterium leprae* infected monocytes and to negatively regulate the Vitamin D-dependent antimicrobial pathway (P. T. Liu et al. 2012). Evidence from previous studies also suggests that highly expressed miRNAs in BMEs may also regulate each other. miR-184, for example, has been demonstrated to antagonise miR-205 to maintain SHIP2 levels in epithelia (J. Yu et al. 2008).

Multiple miRNAs are Differentially Expressed in Response to S. uberis Infection

Once we had characterised which miRNAs were expressed in unchallenged bovine mammary epithelial cells, we then utilised the EdgeR statistical package (Robinson, McCarthy, and Smyth 2009) to determine which miRNAs were significantly differentially expressed in response to *S. uberis* infection at 1, 2, 4 and 6 hpi. It has been suggested that differential expression analysis of miRNAseq is sensitive to the normalisation approach implemented (Garmire and Subramaniam 2012). To address this issue, we identified differentially expressed miRNAs in three alternatively normalised datasets; TMM-normalised (Bullard et al. 2010), upper-quantile normalised and no normalisation. Only miRNAs which were identified as differentially expressed across all three datasets were considered as significantly differentially expressed i.e. the differential expression of these miRNAs was robust to the normalisation procedure. We found that the two different normalisation approaches actually resulted in very similar miRNAs being detected as differentially expressed.

Fifteen different miRNAs were identified as being significantly up-regulated in response to the *S. uberis* challenge. No miRNAs were identified as differentially expressed at 1 hpi. At 2 hpi, 2 miRNAs, bta-mir-29e and bta-mir-708, were found to be up-regulated (Figure 5-5). bta-mir-29e was subsequently observed to be down-regulated at 6hpi. At 4 hpi, bta-let-7b and bta-miR-98 were up-regulated (Figure 5-6). Additionally, bta-miR-let-7c and bta-miR-708 were observed to be up-regulated at 4 hpi when the miRNAseq count data were normalised (both methods), but this was not observed in the un-normalised data. At 6 hpi, 12 miRNAs were found to be up-regulated (Figure 5-7), including bta-let-7b, which was also up-regulated at 4 hpi.

Figure 5-5 Differentially expressed miRNAs at 2 hours post-infection (hpi).

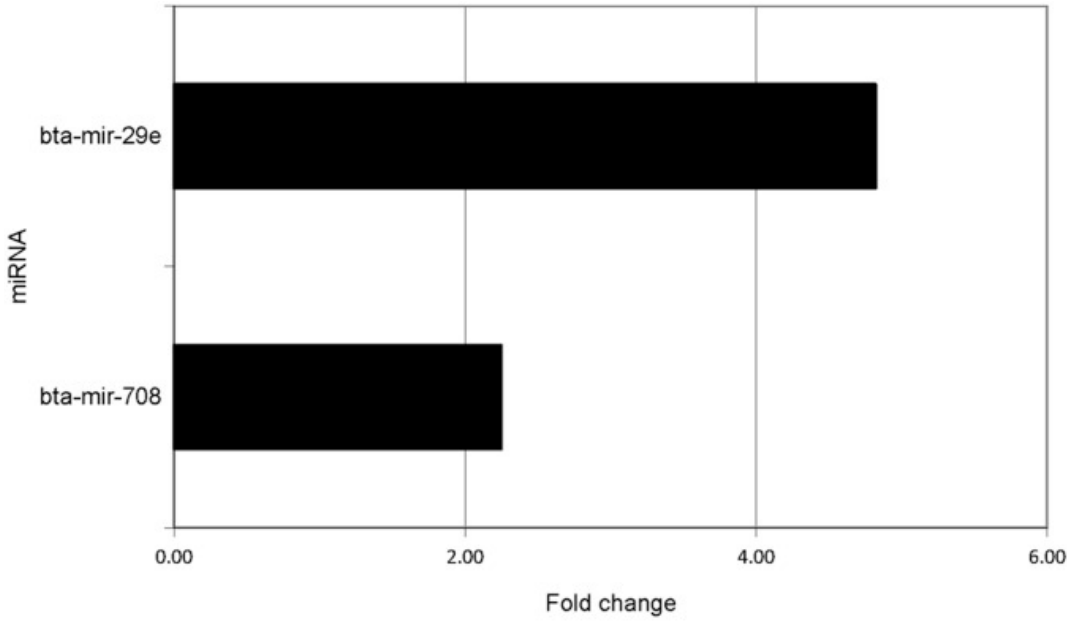


Figure 5-6 Heatmap of miRNA expression (tpm) across infected and control replicates for each 4 hpi differentially expressed miRNA.

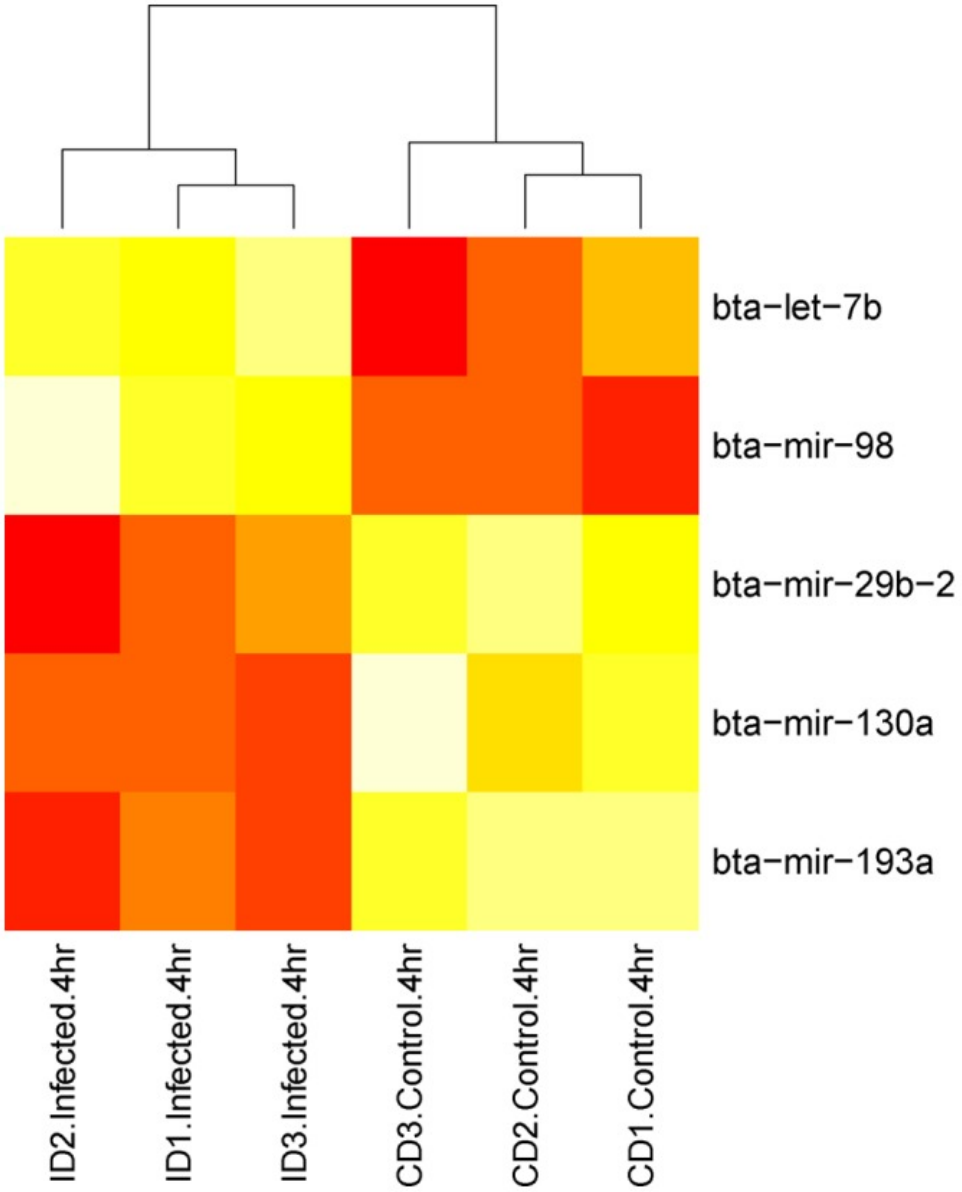
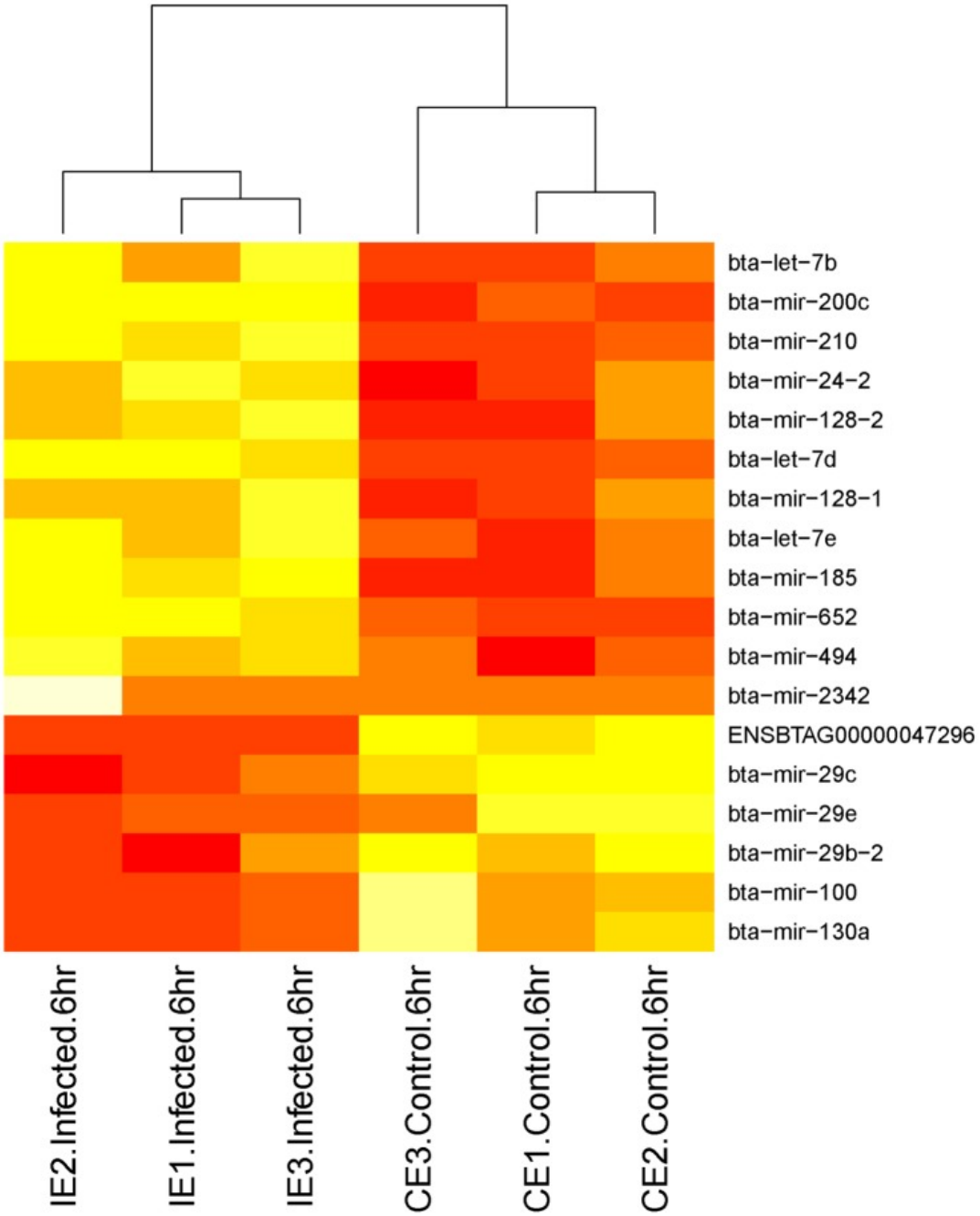


Figure 5-7 Heatmap of miRNA expression (tpm) across infected and control replicates for each 6 hpi differentially expressed miRNA.



Seven different miRNAs were identified as down-regulated in response to the *S. uberis* challenge. No miRNAs were down-regulated at 1 or 2 hpi. At 4 hpi, bta-miR-29b-2, bta-miR-193a, and bta-miR-130a were down-regulated. At 6 hpi, bta-miR-29b-2, bta-miR-29c, bta-miR-29e, bta-miR-100, bta-miR-130a and Ensembl predicted miRNA

ENSBTAG00000047296, were down-regulated. Two miRNAs, bta-miR-29b-2 and bta-miR-130a were down-regulated at both 4 and 6 hpi. Additionally, bta-miR-15a, bta-miR-17, bta-miR-26a-2, bta-miR-29a, bta-miR-29b-1, and bta-miR-193a were identified as down-regulated in the normalised data (both methods), but not in the un-normalised data. Fold changes in expression for miRNAs that are differentially expressed at 4 and 6 hpi are shown in Appendix L and Appendix M.

These results indicate that there are rapid temporal changes in the expression of miRNAs in response to a Gram-positive infection, with different miRNA repertoires being identified as differentially expressed at time-points that are 2 hours apart.

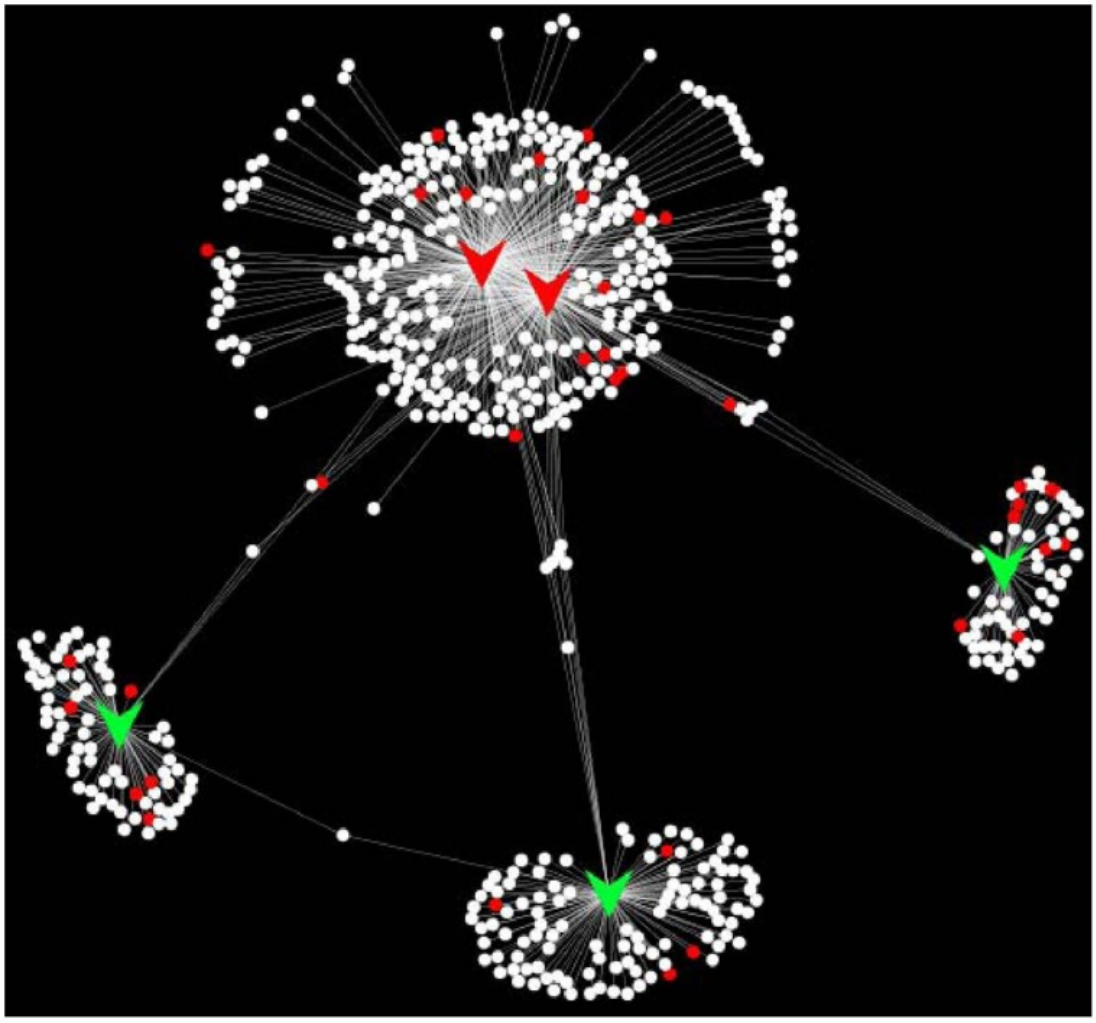
The miRNA Response to the Gram-positive S. uberis is Markedly Different to the LPS miRNA Response

To date, many immune-relevant miRNAs have been identified as part of the host response to lipopolysaccharide (LPS) stimulation (J. Qi et al. 2012; L. A. O'Neill, Sheedy, and McCoy 2011), which is frequently used to mimic a Gram-negative bacterial infection. We have completed a literature survey and identified over 145 miRNAs that have been shown to be differentially expressed in response to LPS across multiple different species and tissues (Appendix I). Eighty-four of the 145 LPS inducible miRNAs were found not be expressed above 1 tpm in BMEs. Of the 21 miRNAs that we identified as being differentially expressed in response to the Gram-positive *S. uberis*, only 9 of these (bta-let-7d, bta-let-7b, bta-mir-98, bta-miR-100, bta-mir-130a, bta-miR-193a, bta-miR-210, bta-miR-494, bta-miR-652) have also been reported to be differentially expressed in response to LPS in other species. Furthermore, 5 of these 9 (bta-miR-98, bta-miR-100, bta-miR-193a, bta-miR-210, bta-miR-494) show an inverse response to *S. uberis* infection in comparison to LPS. Most notably, bta-miR-100 and bta-miR-494, which were previously identified as up- and down-regulated, respectively, in mouse lung 6h post-stimulation with LPS, showed the inverse response at the same time-point in response to *S. uberis* infection (Moschos et al. 2007; Hsieh et al. 2012). This would suggest that the miRNA response to Gram-positive bacteria may be markedly different to Gram-negative.

Predicted Targets of Down-regulated miRNAs are Enriched for Genes with a Role in Innate Immunity

Target genes that are potentially regulated by differentially expressed miRNAs in response to *S. uberis* infection at 2, 4 and 6 hpi were predicted using two computational approaches, miRanda (Betel et al. 2007; Enright et al. 2003) and TargetScan (Lewis, Burge, and Bartel 2005; Grimson et al. 2007; Friedman et al. 2008), where the predicted targets by TargetScan were in turn selected by two independent criteria (PCT or Context+ scores). Given the high false positive rates for miRNA target prediction, we identified only those potential target genes that were predicted by both methods (Table 5-2). Target genes that were not corroborated by both methods were discarded. In total 1,417 unique genes were predicted to be targeted by differentially expressed miRNAs (Appendix J). This resulted in 2,491 miRNA-target interactions; 477 of these were targeted by down-regulated miRNAs; 1,921 were targeted by up-regulated miRNAs; and 93 were targeted by both up and down-regulated miRNAs. Because of the difficulties in accurately predicting miRNA targets, it is more appropriate to examine whether broad functional categories of genes are statistically over-represented among predicted target genes, rather than focusing on individual gene predictions. Statistical analysis (Hypergeometric test) revealed that the predicted target genes of down-regulated miRNAs at 4 and 6 hpi were significantly enriched ($P = 0.01$) in genes annotated by www.innatedb.com (Lynn et al. 2008) as having a role in innate immunity (Figure 5-8). The predicted target genes of up-regulated miRNAs were not enriched for a role in innate immunity suggesting that up and down-regulated miRNAs target different processes in response to *S. uberis* infection.

Figure 5-8 A network of miRNAs (arrow shapes) that were identified as being differentially expressed in BMEs at 4 hours post-infection with S. uberis and their predicted target genes (circles).



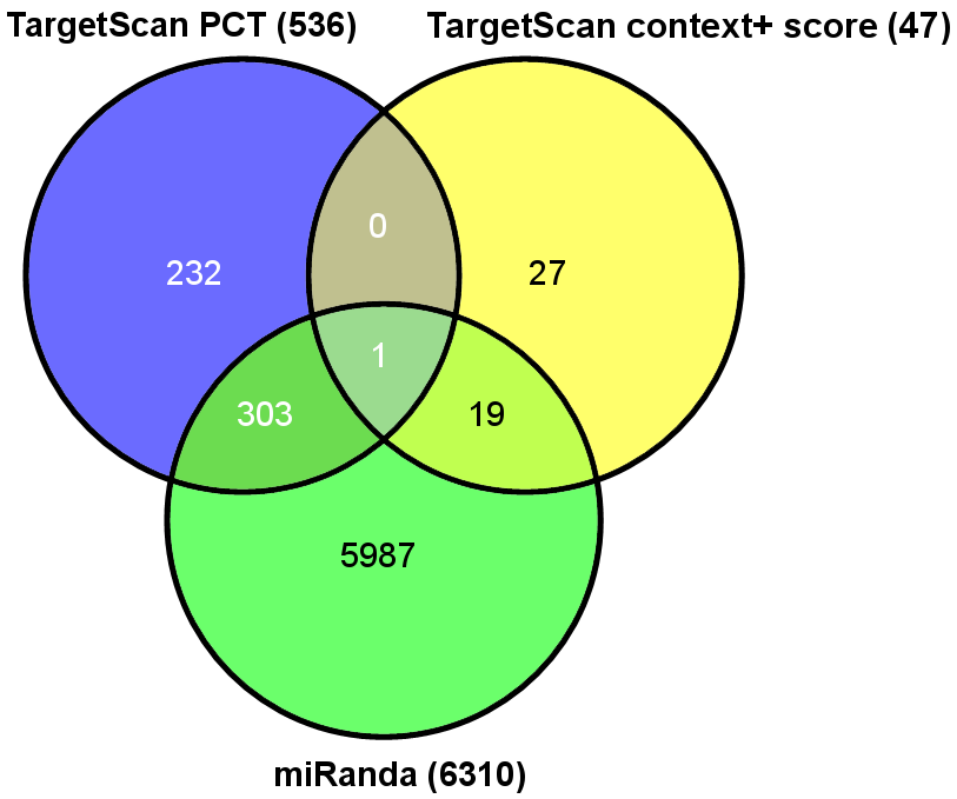
The combination of the target prediction methods used here results in an overall target list that is substantially different from the results by any of the underlying methods alone or any simple combination of those methods. This reflects the fact that the underlying criteria in the individual methods are independent and either complementary (genomic context in context+ score and phylogenetic conservation in PCT) or orthogonal (thermodynamics and a particular multiple alignment in miRanda and site type and a different multiple alignment in TargetScan). For instance, at 4 hours post infection, the combined target list for the up-regulated miRNAs contains 323 genes, compared to 536 genes by PCT, 47 genes by context+ score and 6310 genes by miRanda (Figure 5-9).

At this time point, the target list from the three-way intersection of miRanda and the two TargetScan criteria contains a single gene (SLC35C1), while the three way union would result in 6570 predicted targets (Figure 5-9). .

Table 5-2 miRNA target predictions by miRanda and TargetScan and their intersect.

Name	MiRanda Targets (default settings)	TargetScan (PCT>0.9)	TargetScan (Context+ Score above the 90th percentile)	Number of IntersectingTargets (miRanda and either one of the two TargetScan criteria)
bta-let-7b	6377	576	58	311
bta-let-7d	5637	576	53	290
bta-let-7e	5447	576	60	280
bta-mir-98	5095	576	58	274
bta-mir-185	5696	0	338	177
bta-mir-494	2952	0	336	151
bta-mir-200c	3362	74	201	123
bta-mir-29c	5131	179	76	115
bta-mir-29b-2	5394	179	76	114
ENSBTAG00000047296	5523	0	204	108
bta-mir-29e	5488	0	174	93
bta-mir-708	6414	0	179	84
bta-mir-210	4012	0	206	83
bta-mir-193a	3337	0	160	76
bta-mir-130a	2630	87	76	63
bta-mir-24-2	2157	0	107	45
bta-mir-2342	4296	0	64	37
bta-mir-128-2	4266	83	0	33
bta-mir-128-1	4266	83	0	33
bta-mir-100	946	0	13	1
bta-mir-652	0	0	0	0

Figure 5-9 Predicted targets of up-regulated miRNA at 4 hours post infection by each method and their relation.



The overall list of predicted targets for up-regulated genes at this time point contains 19+1+303=323 unique genes.

Pathway analysis of the predicted gene targets of up-regulated miRNAs (at 4 and 6 hpi), revealed that pathways which have been previously implicated in mastitis are statistically enriched among the predicted gene targets of up-regulated miRNAs (Table 5-3). These pathways include MAPK signalling; Cytokine-Cytokine Receptor Signalling and the JAK-STAT Signalling Pathway. The MAPK signalling pathway, for instance, has been identified as one of the top canonical pathways highlighted in a microarray study examining the bovine mammary tissue response to mastitis, 20 hpi with *S. uberis* (Moyes et al. 2009). Many of the other pathways that we identified as statistically enriched among the predicted gene targets of up-regulated miRNAs were also highlighted as significant in this previous microarray study.

Table 5-3 Pathway analysis of the predicted target genes of up-regulated miRNAs 4 and 6 hours post-infection.

KEGG Pathway	FDR 4 hpi	FDR 6 hpi
MAPK signalling pathway	4.15E-33	2.79E−21
Cytokine-cytokine receptor interaction	4.71E-08	1.96E−32
Axon guidance	ns	3.35E−11
Calcium signalling pathway	ns	1.15E−06
MTOR signalling pathway	ns	1.36E−06
Colorectal cancer	ns	1.81E−06
Insulin signalling pathway	ns	3.69E−06
Jak-STAT signalling pathway	ns	3.56E−05
Fatty acid biosynthesis	ns	4.86E−05

*FDR = false discovery rate. *hpi = hours post-infection.

Taken together, these analyses strongly suggest that miRNAs that are differentially expressed during infection of BMEs with *S. uberis* are key regulators of the host response to this pathogen.

MicroRNA isomiRs

MicroRNA isomiRs are heterogeneous variants of canonical miRNA species, which are, increasingly, being suggested to be of functional importance (Cloonan et al. 2011). It has been suggested that these miRNA variants can be cell type specific, have functional differences, and vary in their response to biological stimuli. Evidence suggests that although isomiRs show similar expression patterns to their equivalent canonical miRNA, their targets can vary (Peng et al. 2012). Deletions at both the 5′ and 3′ end of isomiRs may change the specificity of the seed binding region effecting miRNA function.

We have found that the expression of isomiRs was common for the majority of BME expressed miRNAs (Table 5-4). 100 known miRNAs were found to have at least one isomiR expressed at a level of >100 reads and more than 1,000 different isomiRs were identified. Notably, in 40% of cases at least one isomiR was more highly expressed than the miRbase consensus sequence, suggesting that the isomiR should in fact be annotated as the consensus.

On further examination, we found that isomiR 'nibbling' was 1.4 more times likely than post-transcriptional additions and isomiR editing was 2.3 times more likely to be 3′ modified than 5′ modified, agreeing with current literature (Neilsen, Goodall, and Bracken 2012). That said, almost 40% of isomiRs were 5′ modified, potentially impacting on which targets they regulate, though the majority of 5′ modified isomiRs were expressed at low levels. Further work is required to determine whether isomiRs have a functional role in response to infection.

Table 5-4 Analysis of isomiR heterogeneity across 24 miRNAseq samples.

Sample	# miRs with isomiRs*	# isomiRs	# longer than consensus	# shorter than consensus	**# 5'** modified	**# 3'** modified	# cases where isomiR expressed more highly than consensus
1 hour control repl.1	78	592	185	276	213	494	30
1 hour control repl.2	108	1159	354	520	437	948	42
1 hour control repl.3	95	818	242	372	260	694	36
1 hour infected repl.1	103	1148	355	514	442	936	39
1 hour infected repl.2	94	810	233	378	269	682	34
1 hour infected repl.3	52	341	111	167	97	309	21
2 hour control repl.1	104	1114	341	515	426	922	41
2 hour control repl.2	94	999	307	460	390	828	40
2 hour control repl.3	92	876	256	416	280	753	37
2 hour infected repl.1	98	1075	321	513	437	881	37

2 hour infected repl.2	91	870	259	418	346	713	40
2 hour infected repl.3	91	829	249	366	297	676	37
1 hour infected repl.1	114	1165	362	526	478	945	41
1 hour infected repl.2	101	1086	310	482	411	879	39
1 hour infected repl.3	141	1629	572	651	549	1341	59
2 hour control repl.1	98	1048	314	446	343	867	37
2 hour control repl.2	100	1008	313	467	410	824	32
2 hour control repl.3	91	899	269	406	338	738	43
2 hour infected repl.1	96	1056	321	495	419	881	40
2 hour infected repl.2	97	1002	307	459	349	851	39
2 hour infected repl.3	102	1062	330	504	441	881	38
4 hour control repl.1	120	1245	448	533	397	1070	51
4 hour control repl.2	119	1165	408	510	371	998	52
4 hour control repl.3	137	1796	633	723	597	1500	63

*Only isomiRs present at >100 reads are shown.

Novel miRNA Discovery

In addition to profiling the expression of known miRNAs, miRNAseq data can also be used to identify the expression of potentially novel miRNAs. To do this, miRNAseq data from this study was analysed using the software package, miRDeep2 (Mackowiak 2011). We identified 21 high-confidence, putatively novel, bovine miRNAs that were independently predicted in multiple BME miRNAseq datasets (Table 5-5). Homology searching of the miRBase database (v 19) (Kozomara and Griffiths-Jones 2010) using BLAST (Altschul et al. 1990) identified that 2 of the novel miRNAs had 100% identity to known miRNAs in other species, ssc-miR-664-3p (pig) and hsa-miR-219-1 (human).

Additionally 5 of the novel bovine miRNAs had significant homology with the bta-mir-2285 family. The bta-mir-2285 family has over 40 members spanning the entire bovine genome (Guduric-Fuchs et al. 2012). Two additional novel miRNAs showed homology to the bta-mir-2284 family. The remaining miRNAs did not show significant homology to other known miRNAs in other species. However, given the very high read counts observed for several of these predicted miRNAs, and the fact that were independently predicted in multiple different samples, it would suggest that many of these predictions represent true novel bovine miRNAs.

Table 5-5 Putative novel bovine miRNAs discovered through miRDeep2 analysis of miRNAseq data from 24 bovine primary mammary epithelial cell samples.

Name *	Mature Sequence Best miRBase BLAST Hit (e-value <1)	# Samples miRNAis Predicted in	Mature Tag Count **	Predicted Mature Sequence
bta-mir-6537	N/A	7	272,924	Gugggacgcgugcguuuu
bta-mir-6538	N/A	22	22,094	Auagccaguuggggaagaaugc
bta-mir-6539	N/A	20	9,687	Acgcaauucuucaaaaucuuagc
bta-mir-6540	N/A	16	2,840	Aaaaacuggcagcuucauguaa
bta-mir-2285i-1	bta-miR-2285i	13	2,241	Aaaacuggaacgaacuuuugggc
bta-mir-2285f-3	bta-miR-2285f	18	2,202	Aaaaccugaaugaacuucuugg
bta-mir-2284z-8	bta-miR-2284z	15	2,013	Uaaaaguuugguuggguuuuu
bta-mir-664b	ssc-miR-664-3p	20	1,736	Uauucauuuaucucccagccuac

bta-mir-6541	N/A	8	1,501	Uggagcggcugcacagagcgu
bta-mir-2285c-1	bta-miR-2285c	14	694	Aaaaccugaagagacuuuuugg
bta-mir-6542	N/A	13	693	Ugcuaccuagucugagugaguga
bta-mir-6544	N/A	18	332	Uggugcucccuggagcugagc
bta-mir-6516	gga-miR-6516-5p	7	242	Uuugcaguaacaggugugaac
bta-mir-219-1	hsa-miR-219-1-3p	3	173	Agaguugagucuggacgucccg
bta-mir-6545	N/A	3	121	Auggacugucaccugaggagc
bta-mir-2285m-6	bta-miR-2285m	2	107	Aaaacccaaaugaacuuuuugg
bta-mir-2284b-1	bta-miR-2284b	5	86	Aaauguucgcuuggcuuuuucc
bta-mir-2285f-4	bta-miR-2285f	2	64	Agaaaguucauuuagguuuuuc
bta-mir-6546	N/A	4	52	Cuuccucuuccgguuggcaga
bta-mir-6547	N/A	3	29	Auucuccauuggauauaauagu
bta-mir-6643	gga-miR-6643-5p	2	21	Cagggagggcaggggaggg

5.3.4. Discussion

In this study, we have used a next generation sequencing approach to profile the expression of bovine miRNAs at multiple time-points in primary bovine mammary epithelial cells (BMEs) infected *in vitro* with *S. uberis*, a causative agent of bovine mastitis. In comparison to previous NGS studies investigating the host miRNA response to infection, we have sequenced un-pooled miRNA libraries to a previously unprecedented sequencing depth from multiple replicates and controls across multiple time-points, allowing us to explore statistically significant temporal changes in miRNA expression in response to infection. Analysing over 450 million sequenng reads, we found that approximately 20% of known bovine miRNAs are expressed in BMEs. A similar diversity of miRNA expression has also been recently reported in other tissues, including bovine retinal microvascular endothelial cells (RMECs) and in testicular and ovarian tissues (J. Huang 2011). As has also been reported in other studies, there is a significant dynamic range in the expression of known miRNAs in BMEs. A few miRNAs are expressed at very high levels, with the majority being expressed at low levels. The top 10 most highly expressed miRNAs account for >80% of all aligned reads and are highly conserved across species. Whether or not the other more lowly expressed miRNAs play a significant biological role remains an open question.

We have also found that the expression of isomiRs was common for many of the BME expressed miRNAs. Most significantly, in 40% of cases, at least one isomiR was more highly expressed than the miRbase consensus sequence, indicating that studies such as ours can also be used to improve miRNA annotation. In particular, changes to the 5′ of the consensus sequence could lead to dramatically different target genes being computationally predicted. We identified a large number of 5′ isomiRs although they were mostly expressed at low levels.

In addition to profiling known miRNAs, we have also analysed the sequencing data to identify potentially novel bovine miRNAs. Twenty-one high-confidence, putatively novel, bovine miRNAs were identified independently across multiple different samples. The mature sequences of two of the novel miRNAs were 100% identical to known miRNAs in other species. Seven of the other predicted miRNAs exhibited significant homology to two bovine miRNA families, bta-mir-2284 and bta-mir-2285.

Few studies have previously investigated temporal changes in global miRNA expression using an NGS approach (Cui et al. 2010), although several have used microarray technology (P. Mukhopadhyay et al. 2010; Reinsbach et al. 2012; Li et al. 2010; F.-Z. Wang et al. 2008). Thus far, many immune-relevant miRNAs have been identified as part of the host response to LPS stimulation (J. Qi et al. 2012; L. A. O'Neill, Sheedy, and McCoy 2011). We completed a literature survey and identified more than 145 miRNAs across multiple different species and tissues that have been shown to be LPS responsive. In our study, the miRNA response to the Gram-positive *S. uberis* was markedly different to the reported LPS miRNA response. Over the 6 hour time-course, we identified 21 known bovine miRNAs as significantly differentially expressed in response to the *S. uberis* challenge. Only 9 of these miRNAs have also been reported to be differentially expressed in response to LPS. For those in common, an inverse pattern of expression was observed in 5 of the 9 cases suggesting that the miRNA response to Gram-positive bacteria may be markedly different to Gram-negative. Further global studies of the miRNA response to Gram-positive and Gram-negative bacteria in the same tissue at the same time-points will be required to confirm this.

It is notable that we found most miRNAs were differentially expressed at different time-points post-infection, suggesting that miRNAs exhibit rapid dynamics in their

temporal expression. It is also notable that the majority of miRNAs that we report as being differentially expressed exhibit relatively subtle changes in gene expression in response to infection. This subtle change in expression is in line with existing literature and strengthens the hypothesis that miRNAs are fine-tuners of gene expression (L. A. O'Neill, Sheedy, and McCoy 2011; Bartel 2009). For example, miR-let-7d, miR-652 and miR-494 demonstrated similar levels of differential expression 6 hours post LTA stimulation in mouse tissues (Hsieh et al. 2012).

Computational analysis revealed that the predicted target genes of *S. uberis* down-regulated miRNAs were statistically enriched for roles in innate immunity. This would suggest that these miRNAs may significantly regulate the sentinel capacity of mammary epithelial cells to mobilise the innate immune system (Swamy et al. 2010). Pathway analysis of the predicted targets of up-regulated miRNAs has also identified the statistical over-representation of several pathways previously implicated in the host response to mastitis, such as the MAPK, JAK-STAT and other cytokine signaling pathways. Furthermore, several of the differentially expressed miRNAs have been shown to have roles in the immune systems of other species. For example, bta-let-7 miRNAs were up-regulated at both 4 and 6 hours post-infection with *S. uberis*. The let-7 family has been extensively described in the literature for having a role in immunity. The down-regulation of let-7 family members, for example, was shown to promote expression of IL-10 and IL-6 in HeLa cells infected with *Salmonella enterica* serovar Typhimurium (Schulte et al. 2011). The observed up-regulation of let-7 miRNAs in our study may lead to the repression of anti-inflammatory cytokines to promote innate immunity.

We also report the down-regulation of two other miRNAs, bta-miR-29b-2 and bta-miR-130a, both of which have known roles in immunity and infection in other species. miR-29a/miR-29b down-regulation has been demonstrated to facilitate IFN-γ up-regulation in NK cells and T_H1 cells (F. Ma et al. 2011; K. M. Smith et al. 2012). IFN- γ is well known as an innate inflammatory mediator and its secretion promotes host resistance against viral and intracellular bacteria. Furthermore, IFN-γ mRNA expression has been demonstrated in human mammary epithelial cells (Khalkhali-Ellis et al. 2008), suggesting that this may be a relevant target in our model. LPS induced TNF-α expression in neonatal and adult monocytes has been shown to be greatly suppressed by the induction of miR-130a (H.-C. Huang et al. 2012).

Taken together, the evidence suggests that the differentially expressed miRNAs identified in this study are likely regulators of the innate immune response to *S. uberis* and thus might represent potential therapeutic targets or novel biomarkers of infection and inflammation.

Chapter 6. Profiling microRNA expression in bovine alveolar macrophages using RNA-seq

This chapter is based on a modified version of the article "Profiling microRNA expression in bovine alveolar macrophages using RNA-seq.", co-authored by Peter Vegh, Amir B.K. Foroushani, David A. Magee, Matthew S. McCabe, John A. Brown, Nicolas C. Nalpas, Kevin M. Conlon, Stephen V. Gordon, Daniel G. Bradley, David E. MacHugh and David J. Lynn in Vet Immunology Immunopathology *© 2013 Published by Elsevier B.V. My contribution has been the prediction of putative targets for the detected miRNAs, the analysis of overrepresentation of innate immunity related genes among these targets, pathway analysis of the targets, and the discussion of these results.*

6.1. Abstract

MicroRNAs (miRNAs) are important regulators of gene expression and are known to play a key role in regulating both adaptive and innate immunity. Bovine alveolar macrophages (BAMs) help maintain lung homeostasis and constitute the front line of host defense against several infectious respiratory diseases, such as bovine tuberculosis. Little is known, however, about the role miRNAs play in these cells. In this study, we used a high-throughput sequencing approach, RNA-seq, to determine the expression levels of known and novel miRNAs in unchallenged BAMs isolated from lung lavages of eight different healthy Holstein–Friesian male calves. Approximately 80 million sequence reads were generated from eight BAM miRNA Illumina sequencing libraries, and 80 miRNAs were identified as being expressed in BAMs at a threshold of at least 100 reads per million (RPM). The expression levels of miRNAs varied over a large dynamic range, with a few miRNAs expressed at very high levels (up to 800,000 RPM), and the majority lowly expressed. Notably, many of the most highly expressed miRNAs in BAMs have known roles in regulating immunity in other species (e.g. bta-let-7i, bta-miR-21, bta-miR-27, bta-miR-99b, bta-miR-146, bta-miR-147, bta-miR-155 and bta-miR-223). The most highly expressed miRNA in BAMs was miR-21,

which has been shown to regulate the expression of antimicrobial peptides in *Mycobacterium leprae*-infected human monocytes. Furthermore, the predicted target genes of BAM-expressed miRNAs were found to be statistically enriched for roles in innate immunity. In addition to profiling the expression of known miRNAs, the RNA-seq data was also analysed to identify potentially novel bovine miRNAs. One putatively novel bovine miRNA was identified. To the best of our knowledge, this is the first RNA-seq study to profile miRNA expression in BAMs and provides an important reference dataset for investigating the regulatory roles miRNAs play in this important immune cell type.

6.2. Introduction

MicroRNAs (miRNAs) are an approximately 22 nucleotide (nt) long subset of non-coding RNAs, which post-transcriptionally regulate gene expression by base-pairing with target messenger RNAs (mRNAs). miRNAs are transcribed as pri-miRNAs in the nucleus and are then processed into pre-miRNAs. After export to the cytoplasm, a mature 22 nt duplex is formed. One miRNA strand is then incorporated into the RNA-induced silencing complex (RISC), and interacts with its target mRNA via base-pairing at binding sites usually located within 3′ untranslated regions (UTRs), meanwhile the other strand is usually degraded (Holley and Topkara 2011). Depending on the level of miRNA-mRNA complementarity, the target mRNA can be degraded or its translation repressed (Bartel 2009). Several diseases and conditions have been linked to abnormal expression of miRNAs (Alvarez-Garcia and Miska 2005; Bushati and Cohen 2007), as they have a regulatory role in most biological processes, such as differentiation, apoptosis, and development (Ivey and Srivastava 2010; O'Connell et al. 2010; Xiao and Rajewsky 2009).

It is also becoming increasingly clear that both adaptive and innate immunity are finely regulated by miRNAs. In the adaptive immune system, the differentiation of B cells, antibody generation, and T cell development and function, are all influenced by miRNAs (Belver, Papavasiliou, and Ramiro 2011). Innate immune cell activation is also regulated by miRNAs, including miR-155, miR-146a, miR-21, and miR-9 (Gantier 2010). For example, miR-155 is a positive regulator of Toll-like receptor (TLR) signalling, and is

induced upon stimulation of murine macrophages with interferon beta (IFN-β) or TLR ligands (Liston, Linterman, and Lu 2010; O'Connell et al. 2007).

Furthermore, tumour necrosis factor (TNF) biosynthesis has been shown to be inhibited by *Mycobacterium tuberculosis*, an intracellular mycobacterial pathogen that infects alveolar macrophages, by regulating levels of a human macrophage miRNA, miR-125b (Rajaram et al. 2011). Alveolar macrophages have important roles in lung homeostasis and in many respiratory diseases, such as asthma in humans (Peters-Golden 2004), and are the first cells to encounter several respiratory pathogens during the early stages of infection (Lambrecht 2006; Marriott and Dockrell 2007). In cattle, bovine alveolar macrophages (BAMs) are the major target cell type infected by *Mycobacterium bovis*, the causative agent of bovine tuberculosis (BTB) (Pollock et al. 2006), which results in losses of approximately US$3 billion to global agriculture annually (Garnier et al. 2003).

Currently, 755 bovine miRNAs are annotated in miRBase (version 19, http://www.mirbase.org) (Kozomara and Griffiths-Jones 2010), of which only 22 have been shown to be expressed in BAMs (G. Xu et al. 2009). In this study, we present the first next-generation sequencing approach to profile miRNA expression in unchallenged BAMs, providing an important reference atlas for further elucidating the role miRNAs play in regulating immune networks in this important immune cell type.

6.3. Materials and methods

6.3.1. Ethics statement

All animal procedures were performed according to the provisions of the Irish Cruelty to Animals Act, and ethical approval for the study was obtained from the University College Dublin (UCD) Animal Ethics Committee (protocol number AREC-13-14-Gordon).

6.4. Animals

Eight unrelated Holstein–Friesian male calves (aged between 7 and 12 weeks old) were used in this study. All animals were maintained under uniform housing conditions and nutritional regimens at the UCD Lyons Research Farm (Newcastle, County Kildare, Ireland). The animals were selected from a herd without a recent history of bovine tuberculosis infection.

6.4.1. Lung lavages, alveolar cell preparation and storage

BAMs were harvested by pulmonary lavage of lungs obtained post-mortem from eight animals. Lungs were washed using a total of 3 L of calcium- and magnesium-free sterile Hank's Balanced Salt Solution (HBSS, Invitrogen, Life Technologies Ltd., Paisley, UK); HBSS was infused into the lungs (500 ml per infusion) via the trachea. After each 500 ml infusion of HBSS, the lungs were gently massaged and resulting HBSS-cell suspension was collected into sterile beakers. All lung washes were performed in a laminar flow hood.

50 ml of HBSS-cell suspension collected from the first 500 ml wash was centrifuged (200 × *g* for 10 min at room temperature) and the resulting cell pellet was resuspended in 10 ml HBSS and screened for microbial contamination by incubation on agar plates using the following conditions: Columbia blood agar with 5% defibrinated sheep blood (aerobic and CO_2-enriched atmosphere, 37 °C, 36 h); Chocolate agar (CO_2-enriched atmosphere, 37 °C, 36 h); Columbia-colistin-nalidixic acid agar (aerobic, 37 °C, 36 h); MacConkey agar number 2 (aerobic, 37 °C 36 h); Sabouraud dextrose agar (aerobic, 37 °C, 5 days); and Mycoplasma agar (CO_2-enriched atmosphere, 14 days). All media was obtained from Oxoid Ltd. (Basingstoke, Hampshire, UK). All animals were negative for microbial contamination.

The remaining HBSS-cell suspension (~2 L) was transferred to 50 ml sterile tubes and centrifuged (200 × g for 10 min at room temperature). After centrifugation, the supernatants were discarded and the resulting cell pellets were pooled and resuspended in 50 ml cold R10 media (RPMI 1640 medium [Invitrogen], supplemented with 10% foetal bovine serum [FBS; Sigma–Aldrich, Dublin, Ireland], 2.5 µg/ml amphotericin B

[Sigma–Aldrich], 2 mM l-glutamine [Sigma–Aldrich]). The cell suspension was repelleted (200 × *g* for 10 min at room temperature) and resuspended in 10 ml $R10^+$ media (RPMI 1640 medium supplemented with 10% FBS, 2.5 µg/ml amphotericin B, 2 mM l-glutamine, 100 µg/ml ampicillin [Sigma–Aldrich] and 25 µg/ml gentamycin [Sigma–Aldrich]). Cells were then counted using a haemocytometer, recentrifuged at 200 × *g* for 10 min and re-suspended in freezing solution (10% DMSO [Sigma–Aldrich], 90% FCS) at a density of 2.5×10^7 cells/ml. 1 ml cell aliquots were made in 2 ml sterile cryovials (Sarstedt Ltd., Wexford, Ireland) and placed into Mr. Frosty® Cryo 1 °C Freezing Containers (Nalgene®, Thermo Fisher Scientific, Waltham, MA, USA) containing 100% isopropanol. Cryovials were stored at −80 °C for a period of 18 h after which they were removed from the freezing containers and transferred to −140 °C storage conditions until required for further use.

6.4.2. Alveolar cell culture

Cells were thawed and incubated in a 175 cm^2 flask (Cellstar, Greiner Bio-One Ltd., Stonehouse, UK) in $R10^+$ media for 24 h at 37 C, 5% CO_2. After incubation, media was removed together with nonadherent cells, and replaced with HBSS. After removing HBSS, adherent cells were dissociated with 15 ml 1× non-enzymatic cell dissociation solution (Sigma–Aldrich). Collected cells were pelleted at 400×*g* for 5 min, then resuspended in 10 ml sterile PBS (Invitrogen). Cells were counted with a haemocytometer (BRAND, Wertheim, Germany), and then 4×10^6cells were pelleted and lysed for each RNA extraction, except for sample BAM-8 which had a cell count of 0.8×10^6.

6.4.3. Small RNA

In total, eight RNA-seq libraries were prepared for sequencing. Eight small and total RNA fractions were prepared from the pelleted cells, using the RNeasy Plus Mini kit and RNeasy MinElute Cleanup Kit (Qiagen Ltd., Manchester, UK), according to the appendix E part of the manufacturer's protocol. The quality and quantity of the prepared total and small RNA were assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies Ireland Ltd., Cork, Ireland) with the 6000 Nano and small RNA LabChip kits (Agilent Technologies). Total RNA had RIN values of 9.4–9.8 and 28S/18S rRNA

ratios of 1.5–1.9. Small RNA concentrations were 10,000–25,000 pg/µl, and miRNA concentrations were ≈660–1900 pg/µl (Appendix N.). Samples were stored at −80 °C until further use.

6.4.4. RNA-seq libraries

Illumina RNA-seq libraries were prepared with the Epicentre Scriptminer multiplex kit (Epicentre Biotechnologies, Illumina Inc., Madison, WI, USA), according to the manufacturer's protocol, using 8 µl of the prepared small RNA. The Epicentre FailSafe™ Enzyme mix was used in the PCR amplification step, and Epicentre RNA-Seq barcode primers (Epicentre Biotechnologies) were used for indexing. Libraries were sequenced (50 bp single-end) on one lane with an Illumina HiSeq 2000 machine (Norwegian Sequencing Centre, Oslo, Norway).

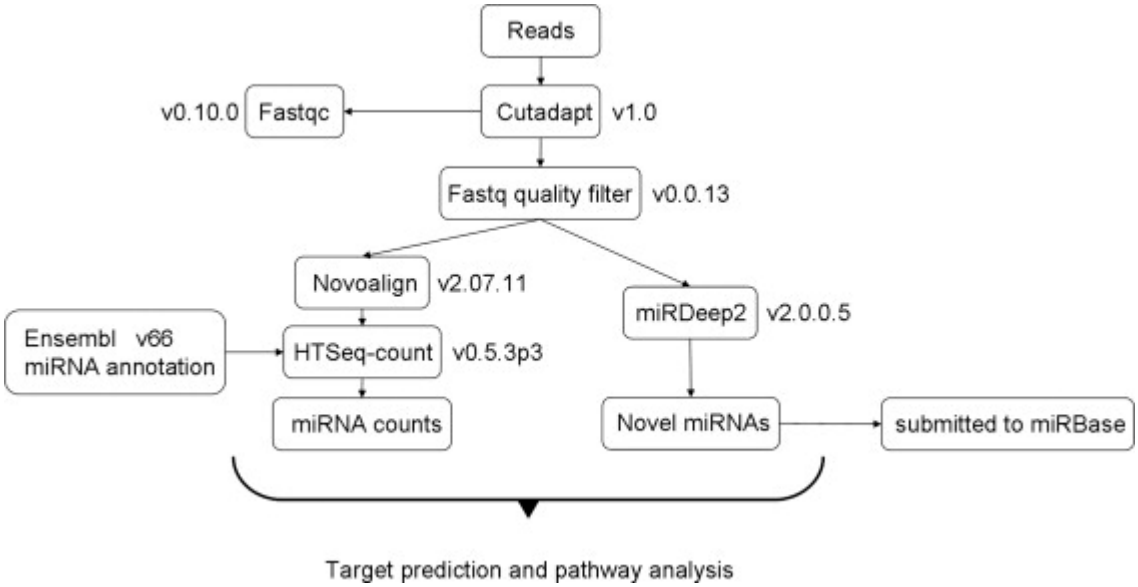
6.4.5. Analysis of RNA-seq data

The quality and number of the reads for each sample were assessed using FASTQC v0.10.0 (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). Cutadapt v1.0 (http://code.google.com/p/cutadapt/) was used to trim 3′ Illumina adapter sequences from reads. Reads which were less than 18 nt after trimming and all untrimmed reads were discarded. The remaining reads were then further filtered using FASTQ Quality Filter (FASTX Toolkit v0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/). Reads where at least 50% of the bases had a Phred score <20 were removed. Finally, reads passing all the above filters were also trimmed at their ends using FASTQ Quality Trimmer (FASTX Toolkit v0.0.13) to remove low quality bases (Phred score <20), and reads less than 18 nt after the trimming were discarded.

Reads which passed all quality control steps were aligned to the bovine genome (UMD3.1 assembly (Zimin et al. 2009)) using Novoalign (Novocraft Technologies, version 2.07.11) in ‘miRNA’ mode. HTSeq-count (part of the HTSeq framework, version 0.5.3p3) in ‘union’ mode was then used to count aligned reads that overlapped with known miRNA gene annotation from Ensembl version 66 (www.ensembl.org). To investigate the proportion of reads sequenced from non-miRNA genes, HTSeq-count was also separately used to count aligned reads that overlapped with all bovine gene

annotations from Ensembl. A graphical depiction of the analysis pipeline is shown in Figure 6-1.

6-1 Bioinformatics analysis pipeline overview



miRDeep2 (version 2.0.0.5) (Friedländer et al. 2008) was used to identify potentially novel miRNAs from the data. The predicted novel miRNAs that were independently predicted in at least three samples and fulfilled all of the following criteria were submitted to miRBase: both the mature and star strand had a minimum five reads each; the predicted miRNA had a high (>90%) probability of being a true miRNA by miRDeep2; the hairpin structure had a RandFold *P* value <0.05.

Customised Perl scripts were also written in house to examine the miRDeep2 output for the presence of miRNA isomiRs as described recently by us (Lawless et al. 2013). The Perl code is available upon request from the authors.

Target genes that are potentially regulated by expressed miRNAs were predicted using the consensus of two computational approaches, miRanda v3.3a (Betel et al. 2007) and TargetScan v6.2 (Friedman et al. 2008; Grimson et al. 2007; Lewis, Burge, and Bartel 2005). Given the high false positive rates for miRNA target prediction, we identified only those potential target genes that were predicted by both methods as described in detail previously (Lawless et al. 2013). The InnateDB (www.innatedb.com)

(Lynn et al. 2008) pathway analysis tool was used to identify potential pathways that were statistically overrepresented among predicted targets. Predicted target genes with a role in innate immunity were also identified using InnateDB and a hypergeometric test was used to investigate whether innate immune genes were statistically overrepresented.

R (v2.14.0, http://cran.r-project.org) was used to create the boxplot showing the miRNA expression levels, and the scatter plots showing the correlation between samples. The RNA-seq fastq files have been uploaded to the NCBI Gene Expression Omnibus (GEO) database (Barrett et al. 2010) with experiment series accession number GSE41138.

6.4.6. RT-qPCR validation

The relative expression of three miRNAs (bta-miR-21, bta-miR-148a, and bta-miR-708) in BAM-5 and BAM-7 was also determined by quantitative reverse transcription PCR (RT-qPCR), using the Taqman MicroRNA Reverse Transcription Kit (Applied Biosystems, Life Technologies Ltd.), according to the manufacturer's protocol, and using 5µl of 1:30 dilution of the prepared small RNA fraction, in 15µl reverse transcription reaction (30 min 16 C, 30 min 42°C, 5 min at 85°C, and then maintained at 4 °C). Taqman MicroRNA Assays (Applied Biosystems) with TaqMan Universal Master Mix II (no UNG) (Applied Biosystems), were used according to the manufacturer's protocol. A 4-point 1:10 serial dilution of miR-21 and a 1:10 dilution of miR-148a was also measured. Additionally, the expression of miR-21 and miR-148a were also measured in BAM-3 and BAM-4.

RT-qPCR was performed on a 7500 Fast Real-Time PCR System (Applied Biosystems) with 7500 Software (version 2.0.6), using MicroAmp Fast optical 96-well plate and Optical Adhesive Film (Applied Biosystems). 20 µl reaction volumes were used for each well and amplification was performed using the following thermal cycle: 95°C 10min, and 40 cycles: 95°C for 15s, 60° C for 1 min. Technical replication was performed in triplicate. Non-template controls had no amplification.

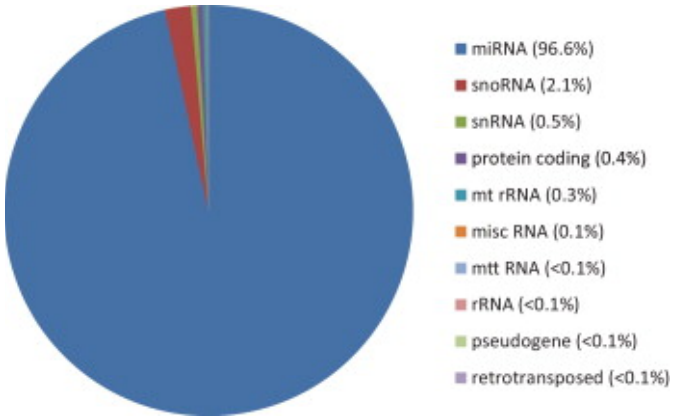
6.5. Results and discussion

6.5.1. BAM-expressed miRNAs

The aim of this study was to identify, catalogue and quantify the expression of all known and novel miRNAs in non-activated BAMs using RNA-seq. In total, 86 million reads were generated by sequencing the eight libraries. After the sequence processing steps of quality control and adapter removal, 62.5 million reads remained for further analysis; 86% of these reads were 19–24 nt long, thus validating the miRNA extraction and library preparation procedure.

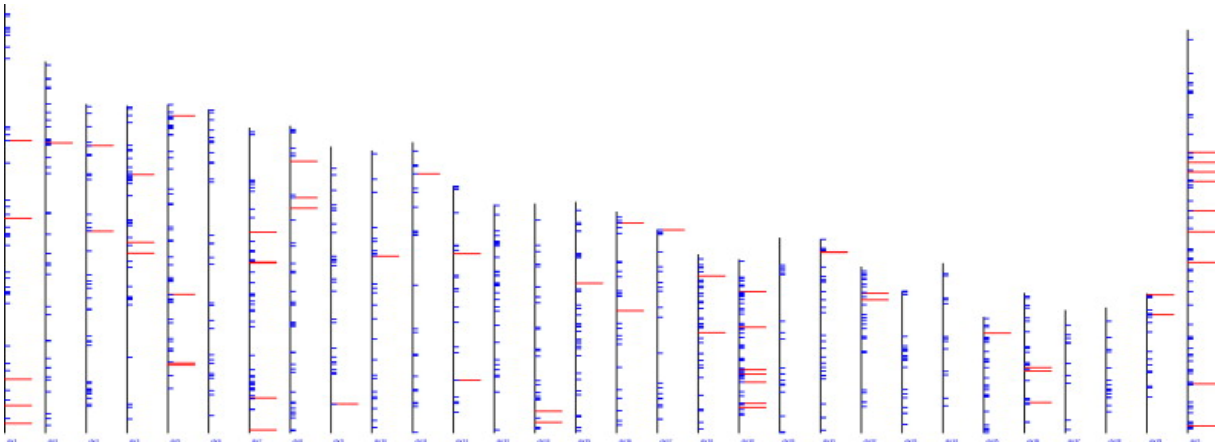
42.5 million reads aligned uniquely to the reference genome. 96.6% of reads aligning to known gene annotations from Ensembl (v66) aligned to miRNAs (Figure 6-2), 2.1% to small nucleolar RNAs (snoRNAs), 0.5% to small nuclear RNAs (snRNAs), and the rest to other RNA species including messenger RNAs (mRNAs), transfer RNAs (tRNAs), ribosomal RNAs (rRNAs) and mitochondrial RNAs (mtRNAs). Approximately one million reads aligned to unannotated regions of the genome; these possibly include novel miRNAs, other novel non-coding RNAs (ncRNAs), or mRNA degradation products. Highly expressed miRNAs were distributed relatively evenly across the genome (Figure 6-3).

6-2 The proportion of reads aligning uniquely to annotated bovine genes (averaged across eight samples).



96.6% percent of annotated reads aligned to known miRNAs. snoRNA, small nucleolar RNA; snRNA, small nuclear RNA; mt, mitochondrial; rRNA, ribosomal RNA; tRNA, transfer RNA.

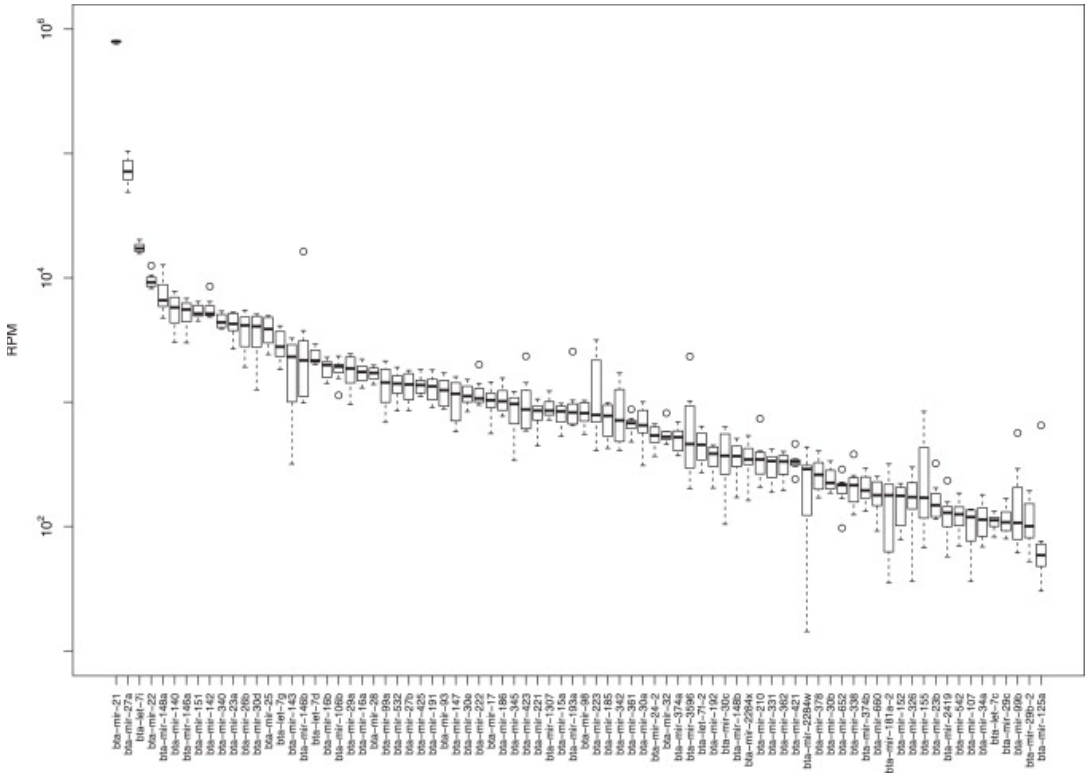
6-3 Distribution of miRNAs in the bovine genome.



Highly expressed miRNAs were relatively evenly distributed across the genome. Vertical bars represent chromosomes, blue horizontal bars known miRNA locations. Red horizontal bars show highly expressed (>100 RPM) miRNA locations.

Recently it has been reported that miRNAs expressed below 100 reads per million (RPM) are unlikely to be functional (Mullokandov et al. 2012). Using 100 RPM as a threshold for functional expression, we found that more than 10% (80/755) of known bovine miRNAs were expressed in BAMs. As has been previously reported in the RAW 264.7 mouse macrophage cell line (Garmire and Subramaniam 2012), there is a large dynamic range in the expression of known miRNAs – a few miRNAs (bta-mir-21, miR-27a and bta-let-7i) are expressed at very high levels (up to 800,000 RPM), with the majority being expressed at low levels (around 1000 RPM). Figure 6-4 shows all miRNAs expressed in BAMs above 100 RPM (the read counts for all known bovine miRNAs can be found in Appendix O). Read counts were highly correlated between biological replicates (Figure 6-4).

6-4 Box plot of miRNAs expressed in BAMs above a threshold of 100 RPM.



The expression of Ensembl annotated bovine miRNAs in BAMs. Read counts shown are the number of reads aligning uniquely to that miRNA.

Several of the miRNAs which we identified as expressed in BAMs (>100 RPM) have been shown to have regulatory roles in monocytes and macrophages of other species. The most highly expressed BAM miRNA, miR-21, for example, regulates the expression of antimicrobial peptides and was found to be up-regulated in *M. leprae* infected human monocytes (P. T. Liu et al. 2012). miR-27, which was also found to be highly expressed in BAMs, has been shown to be involved in the activation of human macrophages (Cheng et al. 2012; Graff et al. 2012). The third mostly highly BAM-expressed miRNA, bta-let-7i, belongs to the let-7 family, a family of miRNAs that has several well-documented roles in immunity (Androulidaki et al. 2009; Satoh et al. 2012). For example, several let-7 family members, including let-7i, were observed to be down-regulated in murine macrophages during *Salmonella* infection (Schulte et al. 2011). Other BAM-expressed miRNAs identified in our study, including miR-21, miR-27b, miR-146, miR-147, miR-155, and miR-223, have all been found to be up-regulated by TLR

signalling in other studies (Ghorpade et al. 2012; G. Liu et al. 2009; L. A. O'Neill, Sheedy, and McCoy 2011). In addition to regulation by the host's immune system, miRNAs may also be regulated by pathogens to modulate the immune response in a way that is advantageous to the pathogen. For example, miR-99b, another BAM-expressed miRNA, is up-regulated by *M. tuberculosis* in infected murine dendritic cells, and this inhibits the production of pro-inflammatory cytokines (Singh et al. 2013).

6.5.2. Relative expression of miR-21, miR-148a and miR-708

To assess reliability of the RNA-seq data presented in this study and to confirm that the high relative expression of miR-21 in BAMs is not an artefact of the sequencing technology, we measured the relative expression of three miRNAs (miR-21, miR-148a, and miR-708) using RT-qPCR, in two samples (BAM-5 and BAM-7). We chose these miRNAs for the following reasons: according to our RNA-seq data, miR-21 was the highest expressed miRNA in BAMs (Appendix O); miR-148a was expressed at a level approximately 100-fold less than miR-21; and miR-708 was expressed at a very low level (only a few reads mapped to it). RT-qPCR confirmed the same profile of expression of these miRNAs. miR-148a was found to be expressed two orders of magnitude less than miR-21 (C_q (quantification cycle) values for all four samples are shown in Appendix P); while miR-708 was observed to be very lowly expressed compared to miR-21 ($\Delta C_q \approx 16$), which corresponds to the count numbers. These data were further confirmed using serial dilutions of the miRNA cDNA. We have found that 100× dilution of the miR-21 RT product cDNA used in the qPCR resulted in approximately the same C_q values as undiluted miR-148a ($\Delta C_q < 0.35$) in both samples. We had similar results for 1000× dilution of miR-21 and 10× dilution of miR-148a cDNA ($\Delta C_{q\text{BAM-5}} = -1$ and $\Delta C_{q\text{BAM-7}} = 0.1$).

6.5.3. Analysis of predicted miRNA target genes

To identify the genes and pathways that may be under miRNA control in BAMs, we performed target prediction on the miRNAs which are expressed above 100 RPM. The targets that were predicted by both miRanda and TargetScan, are listed in Appendix Q. Using InnateDB, we found that innate immunity related genes were (slightly) overrepresented ($P = 0.049$) among predicted targets of miRNAs. Pathway analysis of the predicted genes showed no pathways to be significantly overrepresented.

Figure 6-5 Combined targeting frequencies.

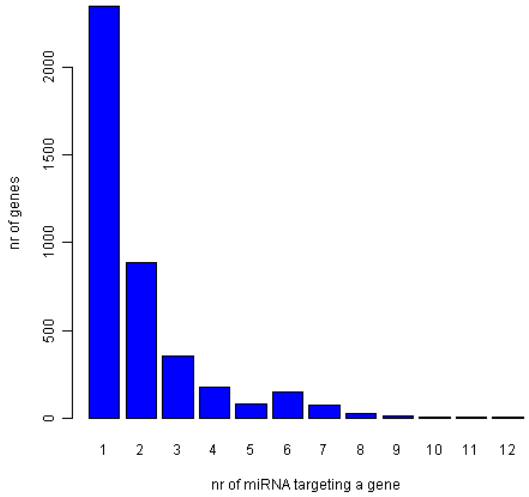
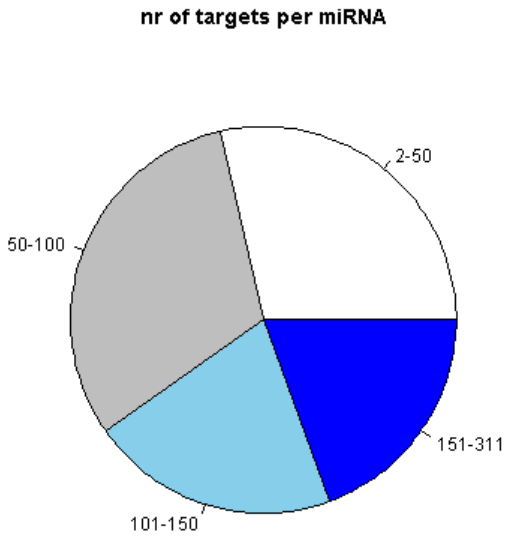


Figure 6-6 Distribution of the number of predicted targets per microRNA.



For two of the eighty microRNAs, no targets could be predicted. These two microRNAs happened to be among the top five most abundant microRNAs in the samples, at position 2 and 4. The remaining 78 microRNAs had collectively 4,119 unique predicted targets. Examining the predicted targets showed a wide range in number of microRNAs predicted to act on each target (Figure 6-5) as well as a wild

range in the number of predicted targets per microRNA (Figure 6-6). For instance, 2,354 of the 4,119 genes were predicted to be targeted by just one miRNA, were as 8 genes (GATM , MFSD8 , ARRDC4 , HIF3A , SNX16 , TMOD2 , PPARGC1B,UBN2) were predicted targets of 10-12 (out of 78) microRNAs. Notably, the members of the *let* family (bta-let-7b, bta-let-7i, bta-let-7c, bta-let-7d, bta-let-7e, bta-let-7g) had 280-311 predicted targets each, followed by bta-mir-98, bta-mir-326, bta-mir-186 (232-274 predicted targets each). On the other end of the spectrum, there were only 2-6 predicted targets for bta-mir-99b, bta-mir-99a and bta-mir-2419 and 15-26 predicted targets for bta-mir-147,bta-mir-191 and bta-mir-147.

In contrast to the study in chapter 5, the aim of the present study was to determine 'which microRNAs are present in unchallenged alveolar macrophages', without any information on 'how these microRNA profiles change over time after exposure to a particular pathogen and which -if any- subsets of these micoRNAs change most significantly'. Hence, any pathway analysis results would only suggest that in resting condition, certain pathways are more likely to be partially maintained by the totality of the encountered microRNA repertoire (including those at very low expression levels) than others.

Using InnateDB, there were no statistically significant pathways for the predicted targets. Depending on the number of most abundant microRNAs that were considered, analysis of the predicted targets by SIGORA highlighted different sets of pathways. Regardless of how many of the top abundant microRNAs were taken into consideration, endocytosis, MAPK-signaling and JAK-Stat signaling always among SIGORA's top 5 results. These pathways are known to be closely intertwined with microRNA networks -(Johnson et al. 2005; Paroo et al. 2009; Bode, Ehlting, and Häussinger 2012; Collins et al. 2013)- and have a known function in macrophages (e.g. sampling of the environment by endocytosis).

6.5.4. IsomiR expression in BAMs

IsomiRs are isoforms of miRNAs that differ in a few nucleotides from the canonical sequence, and may have functional importance. Compared to the canonical sequence, isomiRs can be modified at either ends, with modifications at the 3′ end being

more common (Neilsen, Goodall, and Bracken 2012). We have found that isomiRs are commonly expressed in BAMs. More than 100 miRNAs had isomiRs that were expressed at >100 reads. Furthermore, over half of these miRNAs had an isomiR more highly expressed than the miRBase consensus sequence.

IsomiRs were generally a few nucleotides shorter than the consensus miRNAs; in agreement with earlier findings (L. W. Lee et al. 2010; M. A. Newman, Mani, and Hammond 2011) modifications at the 3′ end were twice as common as at the 5′ end (Appendix R). This is in concordance with observations that usually a short seed sequence at the 5′ end is responsible for most of the miRNA target specificity (Brennecke et al. 2005).

6.5.5. Prediction of novel bovine miRNAs expressed in BAMs

In comparison with miRBase version 19, we also identified five putatively novel bovine miRNAs in the RNA-seq data. Four of these were also recently discovered by us in bovine mammary epithelial cells (Lawless et al. 2013). One putatively novel miRNA, bta-miR-8550, found on chromosome 7, with a predicted mature sequence ‘caggcucuggaacacgggagc’, is entirely novel and showed no homology to miRNAs in miRBase.

6.6. Conclusion

Here we have provided the first atlas of miRNA expression in unchallenged BAMs, which will serve as a reference point for future functional studies or challenge experiments directed to uncover the role of miRNAs in these critical immune cells.

Chapter 7. Concluding Remarks

A better understanding of the processes that govern the immune response is crucial to reduce the direct and indirect burden of infectious disease and to limit the unintended damage caused by dysregulated inflammatory processes. The immense complexity and the intricate nature of immunological processes necessitate the use of the tools of systems biology. The goal of this thesis was the development and application of systems biology tools to help further this understanding. Acknowledging the importance of agricultural model organisms, several specific aims of this thesis focused on utilizing the relatively recently completed draft of bovine genome and analyzing high throughput bovine 'omics' data, but -as computational approaches- they are relatively easily transferable to other mammals.

The first goal of this thesis was to computationally reconstruct bovine interaction networks and pathways and to make these available to the bovine research community. I tackled this task by orthology based transfer of gene functions and protein interactions and subsequently incorporating these into the InnateDB knowledge discovery and analysis platform. Overall, about three quarter of experimentally validated human protein interactions –as present in InnateDB- and about 80% of human pathways could be reconstructed in cow. A subsequent analysis of the global interaction partners showed that human genes with a predicted ortholog and innate immunity related genes have distinct topological properties in the network. The quality of these predictions could be further improved, e.g. for interactions classified as direct physical interactions, one could examine the conservation of interaction sites. Examining the inferred pathways showed that although some immunity related pathways do not seem to be very well conserved, this issue is more prevalent among basic, metabolism related pathways. This observation was rather surprising, but can be partly explained by several factors; including the gene-centric view of pathways in the repository (and the selected high-throughput transfer methodology), which neglects the question of conservation of small molecules and metabolites, the less than optimal quality of the current draft of the bovine

genome (and the confidence in bovine gene calling methods), but also real biological divergence, including duplications of bovine interferon families (Walker and Roberts 2009) and new insights into slower metabolic rates in primates than other mammals (Pontzer et al, 2014). To be sure, even after further improvements, the inferred bovine pathways and networks would remain a temporary surrogate for the real networks, and would have to be eventually replaced or supplemented by experimentally validated data. This is particularly true for bovine specific genes without human orthologs such as members of duplicated interferon families. InnateDB's submission platform (Lynn et al. 2010) will be useful in integrating such data, if and when they become available. In the meantime, InnateDB remains one of the first analysis platforms for systems level analysis of bovine datasets in a pathway- and network oriented context.

Two observations in course of completion of this first thesis-goal motivated the development of two new thesis goals.

The first observation was that genes with known pathway annotations tend to be simultaneously annotated in multiple, at times seemingly unrelated pathways. This observation led me to examine the implications of this issue for pathway analysis methods that treat pathways as collections of individual genes and treat each gene in a pathway as equally informative. Microarray and next generation sequencing (NGS) investigations of pathological conditions (such as infection and cancer) often report dramatic changes in expression levels of hundreds or thousands of genes. Pathway analysis statistically links these molecular changes to higher level cellular or organismal processes and thereby facilitates the biological interpretation of the experimental results. It is, however, not uncommon for pathway analysis methods to replace the long list of differentially expressed genes with an only slightly less confusing long list of potentially perturbed pathways. In addition to the inherent complexity of pathological conditions, several methodological factors contribute to this situation: first, traditional pathway analysis methods treat different members of a pathway as interchangeable markers of that pathway-- regardless of differences in number of pathways associated with each gene. Second, every pathway associated with a multifunctional gene is commonly treated as an equally likely candidate for the biological role assumed by that gene in the examined condition-- despite the fact that the biological role of a gene is highly context-sensitive and dependent on the set of available interaction partners. Third, as the

statistical tests for significance of a pathway are affected the size of the pathway (measured in number of genes in the repository that are annotated in the pathway), the differential expression of just a few multifunctional genes can push several irrelevant but relatively small pathways towards the top of the list of significant pathways. Once I realized that “single-gene multi-functionality” does result in identification of potentially misleading, spurious pathways as statistically significant in a range of situations, I tried to improve pathway analysis results by developing a new method that focuses on the statistical over-representation of pathway-specific gene combinations in a user defined list of genes of interest (e.g. list of genes that show differential expression behaviour in a case-control study). The shift from single genes to specific gene-pairs as markers of a pathway offers a more context-oriented perspective and the resulting method (Signature over-representation analysis, SIGORA) improves upon the performance of many existing, popular methods, according to a range of different evaluation criteria. Like all pathway analysis methods, SIGORA’s results ultimately depend on the extent and quality of the underlying repositories.

This leads to the second observation: in higher organisms, the biological role of most genes is simply unknown. Currently, only about a third of all human protein coding genes have a pathway affiliation in any pathway repository. In the case of inferred bovine pathways, this fraction becomes even smaller. Gene ontology biological process (GO-BP) annotations are often used as an alternative means of functional analysis, but here too, a substantial fraction of genes are not associated with any biological processes. As a minor step towards ameliorating this issue, I applied a “guilt by association” approach to identify highly co-expressed, potentially functionally related clusters of genes in a large scale bovine tissue expression dataset extracted from a broad and diverse set of tissue samples. The identified clusters were then subjected to functional analysis, and under strict pre-conditions, the un-annotated genes in each cluster were assigned a statistically significant function of the respective cluster. Despite the ‘proof of concept’ nature of this approach, about 20% of obtained functional predictions showed literature support, but more work is needed to improve the results.

Finally, I also had the opportunity to contribute to two high throughput studies on the role of microRNA in bovine infectious disease, one of which (a mastitis study) was a time-series, post-infection study. Bovine mastitis (an inflammatory disease of the

mammary gland in response to infection or physical damage) is the most prevalent health disorder in dairy farms that can be caused by a range of different pathogens (Staphylococcus *aureus*, Streptococcus spp., coliforms, gram-positive bacilli, Corynebacterium *bovis*, Staphylococcus spp., among others) (Sargeant et al. 1998) and results in significant losses in milk yield and quality. The frequency of infection and the type of causative agents vary strongly depending on geographical location, dairy farm management practices, usage of growth hormones and implemented mastitis control programs (Dohoo et al. 2003; Keane et al. 2013). Estimations of Incidence rates are additionally affected by variations in perception, definition and detection method of mastitis (somatic cell counts above an arbitrary threshold, abnormalities in milk and/or udder) as well as farmers' reporting habits. Conservative estimates report that -in average- annually 20% of dairy cows experience clinical mastitis during lactation.

An interesting possibility arising from the study presented in chapter 5 (which examined the differential expression of microRNAs at several time points after exposure of bovine mammary epithelial cells to *S. uberis*) is a notable difference in miRNA response to Gram-positive bacteria vs. Gram-negative ones (e.g. E. *coli*). If confirmed by future studies (the same tissue at the same time-points), one might speculate that – similar to recent developments in oncology (Ajit 2012)- specific combinations of microRNAs could one day be used as biomarkers that distinguish different causative agents of infection and possibly inform the most appropriate treatment course. For the time being, however, the observation of dynamic changes in the microRNA profile in course of the same study (chapter 5) points to additional confounding factors for translational potential of such markers in infectious disease studies.

A more basic question posed by the differentially expressed microRNA is: what are they doing? Here, I used a combination of three target prediction methods to arrive at a target list for the differentially expressed microRNAs. Subsequent statistical analysis showed that predicted target genes of *S. uberis* down-regulated miRNAs were statistically enriched for roles in innate immunity, and that the results of pathway analysis of the predicted targets (by SIGORA) was largely in line with the known biology of bovine mastitis. Furthermore, in contrast to previously reported results, this combination of methods seemed to deliver a more plausible biological picture than any of the methods

taken individually. Whether this is due to specific properties of this dataset or a possibly more generalizable approach, remains to be seen.

It has been noted that due to biases in the bio-medical publishing system -where reports on positive results are preferred- most published results findings are wrong (Ioannidis 2005). It will be interesting for me to revisit some of the results presented here in a few years, when more mature technologies are likely to be available at a smaller cost for detecting, quantifying, modifying and functionally analyzing cellular and organismal constituents, to learn from the extent of systematic biases in my approaches. Examples of such technologies include HITS-CLIP (Thomson, Bracken, and Goodall 2011) and RNA immunoprecipitation (RIP) experiments (Zhang et al., 2007), which could be used to identify the actual targets of microRNAs in vivo and a context specific manner, as well as clustered regularly interspaced short palindromic repeats (CRISPR), which –by providing powerful and precise genome editing tools capable of simultaneously targeting several genes- is poised to strongly accelerate the experimental investigation of biological functions of genes and genomic regions, including those coding for miRNA (L. S. Qi et al. 2013; Y. Zhao et al. 2014; Cong et al. 2013).

References

Agrawal, Anshu. 2013. “Mechanisms and Implications of Age-Associated Impaired Innate Interferon Secretion by Dendritic Cells: A Mini-Review.” *Gerontology* (April 18). doi:10.1159/000350536. http://www.ncbi.nlm.nih.gov/pubmed/23615484.

Ajit, Seena K. 2012. “Circulating microRNAs as Biomarkers, Therapeutic Targets, and Signaling Molecules.” *Sensors (Basel, Switzerland)* 12 (3) (January 8): 3359–69. doi:10.3390/s120303359. http://www.mdpi.com/1424-8220/12/3/3359.

Albert, I, J Thakar, S Li, R Zhang, and R Albert. 2008. “Boolean Network Simulations for Life Scientists.” *Source Code for Biology and Medicine* 3 (November): 16.

Alexa, Adrian, Jörg Rahnenführer, and Thomas Lengauer. 2006. “Improved Scoring of Functional Groups from Gene Expression Data by Decorrelating GO Graph Structure.” *Bioinformatics (Oxford, England)* 22 (13) (July 1): 1600–7. doi:10.1093/bioinformatics/btl140. http://www.ncbi.nlm.nih.gov/pubmed/16606683.

Alexiou, P, M Maragkakis, G L Papadopoulos, M Reczko, and A G Hatzigeorgiou. 2009. “Lost in Translation: An Assessment and Perspective for Computational microRNA Target Identification.” *Bioinformatics (Oxford, England)* 25 (23) (December): 3049–3055.

Allali, Oussama, Clémence Magnien, and Matthieu Latapy. 2013. “Internal Link Prediction : a New Approach for Predicting Links in Bipartite Graphs.” *Dynamic Networks and Knowledge Discovery, Special Issue of Intelligent Data Analysis* 7 (1): 5–25.

Alto, Neal M, and Kim Orth. 2012. “Subversion of Cell Signaling by Pathogens.” *Cold Spring Harbor Perspectives in Biology* 4 (9) (September 1): a006114. doi:10.1101/cshperspect.a006114. http://cshperspectives.cshlp.org/content/4/9/a006114.full.

Altschul, Stephen F, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3) (October): 403–410. doi:10.1016/S0022-2836(05)80360-2. http://linkinghub.elsevier.com/retrieve/pii/S0022283605803602.

Alvarez-Garcia, Ines, and Eric A Miska. 2005. “MicroRNA Functions in Animal Development and Human Disease.” *Development (Cambridge, England)* 132 (21) (November): 4653–62. doi:10.1242/dev.02073. http://www.ncbi.nlm.nih.gov/pubmed/16224045.

Androulidaki, Ariadne, Dimitrios Iliopoulos, Alicia Arranz, Christina Doxaki, Steffen Schworer, Vassiliki Zacharioudaki, Andrew N Margioris, Philip N Tsichlis, and Christos Tsatsanis. 2009. “The Kinase Akt1 Controls Macrophage Response to Lipopolysaccharide by Regulating microRNAs.” *Immunity* 31 (2) (August 21): 220–31. doi:10.1016/j.immuni.2009.06.024. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2865583&tool=pmcentrez&rendertype=abstract.

Antonov, Alexey V, Thorsten Schmidt, Yu Wang, and Hans W Mewes. 2008. “ProfCom: a Web Tool for Profiling the Complex Functionality of Gene Groups Identified from High-Throughput Data.” *Nucleic Acids Research* 36 (Web Server issue) (July 1): W347–51. doi:10.1093/nar/gkn239. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2447768&tool=pmcentrez&rendertype=abstract.

Arpaia, Nicholas, Jernej Godec, Laura Lau, Kelsey E Sivick, Laura M McLaughlin, Marcus B Jones, Tatiana Dracheva, Scott N Peterson, Denise M Monack, and Gregory M Barton. 2011. “TLR Signaling Is Required for Salmonella Typhimurium Virulence.” *Cell* 144 (5) (March 4): 675–88. doi:10.1016/j.cell.2011.01.031. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3063366&tool=pmcentrez&rendertype=abstract.

Ashburner, M, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.” *Nature Genetics* 25 (1) (May): 25–9. doi:10.1038/75556. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3037419&tool=pmcentrez&rendertype=abstract.

Asmann, Yan W, Eric W Klee, E Aubrey Thompson, Edith A Perez, Sumit Middha, Ann L Oberg, Terry M Therneau, et al. 2009. “3’ Tag Digital Gene Expression Profiling of Human Brain and Universal Reference RNA Using Illumina Genome Analyzer.” *BMC Genomics* 10 (1) (January): 531. doi:10.1186/1471-2164-10-531. http://www.biomedcentral.com/1471-2164/10/531.

Assenov, Yassen, Fidel Ramírez, Sven-Eric Schelhorn, Thomas Lengauer, and Mario Albrecht. 2008. “Computing Topological Parameters of Biological Networks.” *Bioinformatics (Oxford, England)* 24 (2) (January 15): 282–4. doi:10.1093/bioinformatics/btm554. http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/2/282.

Ayele, W Y, S D Neill, J Zinsstag, M G Weiss, and I Pavlik. 2004. “Bovine Tuberculosis: An Old Disease but a New Threat to Africa.” *The International Journal of Tuberculosis and Lung Disease : the Official Journal of the International Union Against Tuberculosis and Lung Disease* 8 (8) (August): 924–37. http://www.ncbi.nlm.nih.gov/pubmed/15305473.

Baek, Daehyun, Judit Villén, Chanseok Shin, Fernando D Camargo, Steven P Gygi, and David P Bartel. 2008. “The Impact of microRNAs on Protein Output.” *Nature* 455 (7209) (September 4): 64–71. doi:10.1038/nature07242.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2745094&tool=pmcentrez&rendertype=abstract.

Balachandar, S, and A Katyal. 2011. “Peroxisome Proliferator Activating Receptor (PPAR) in Cerebral Malaria (CM): a Novel Target for an Additional Therapy.” *European Journal of Clinical Microbiology & Infectious Diseases : Official Publication of the European Society of Clinical Microbiology* 30 (4) (April): 483–98. doi:10.1007/s10096-010-1122-9. http://www.ncbi.nlm.nih.gov/pubmed/21140187.

Barabási, A.L. 2007. “Network Medicine — From Obesity to the ‘ Diseasome ’.” *New England Journal of Medicine* 357 (4): 404–407. http://www.nejm.org/doi/full/10.1056/NEJMe078114.

Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo. 2011. “Network Medicine: a Network-Based Approach to Human Disease.” *Nature Reviews. Genetics* 12 (1) (January): 56–68. doi:10.1038/nrg2918. http://www.ncbi.nlm.nih.gov/pubmed/21164525.

Barabási, Albert-László, and Zoltán N Oltvai. 2004. “Network Biology: Understanding the Cell’s Functional Organization.” *Nature Reviews Genetics* 5 (2): 101–113. http://www.ncbi.nlm.nih.gov/pubmed/14735121.

Barrett, T, D B Troup, S E Wilhite, P Ledoux, C Evangelista, I F Kim, M Tomashevsky, et al. 2010. “NCBI GEO: Archive for Functional Genomics Data Sets--10 Years On.” *Nucleic Acids Research* 39 (Database) (November): D1005–D1010. doi:10.1093/nar/gkq1184. http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkq1184.

Barsky, Aaron, Jennifer L Gardy, Robert E W Hancock, and Tamara Munzner. 2007. “Cerebral: a Cytoscape Plugin for Layout of and Interaction with Biological Networks Using Subcellular Localization Annotation.” *Bioinformatics (Oxford, England)* 23 (8) (April 15): 1040–2. doi:10.1093/bioinformatics/btm057. http://www.ncbi.nlm.nih.gov/pubmed/17309895.

Bartel, David P. 2009. “MicroRNAs: Target Recognition and Regulatory Functions.” *Cell* 136 (2) (January): 215–233. doi:10.1016/j.cell.2009.01.002. http://linkinghub.elsevier.com/retrieve/pii/S0092867409000087.

Batz, Michael B, Evan Henke, and Barbara Kowalcyk. 2013. “Long-Term Consequences of Foodborne Infections.” *Infectious Disease Clinics of North America* 27 (3) (September): 599–616. doi:10.1016/j.idc.2013.05.003. http://www.ncbi.nlm.nih.gov/pubmed/24011832.

Bell, M D, D D Taub, and V H Perry. 1996. “Overriding the Brain’s Intrinsic Resistance to Leukocyte Recruitment with Intraparenchymal Injections of Recombinant Chemokines.” *Neuroscience* 74 (1) (September): 283–92. http://www.ncbi.nlm.nih.gov/pubmed/8843093.

Belver, Laura, F Nina Papavasiliou, and Almudena R Ramiro. 2011. “MicroRNA Control of Lymphocyte Differentiation and Function.” *Current Opinion in Immunology* 23 (3)

(June): 368–73. doi:10.1016/j.coi.2011.02.001. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3109091&tool=pmcentrez&rendertype=abstract.

Benjamini, Y, and Y Hochberg. 1995. “Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B Methodological* 57 (1): 289–300. doi:10.2307/2346101. http://www.jstor.org/stable/2346101.

Benoit, Marie, Benoît Desnues, and Jean-Louis Mege. 2008. “Macrophage Polarization in Bacterial Infections.” *Journal of Immunology (Baltimore, Md.: 1950)* 181 (6) (September 15): 3733–9. http://www.jimmunol.org/content/181/6/3733.full.

Betel, D, M Wilson, A Gabow, D S Marks, and C Sander. 2007. “The microRNA.org Resource: Targets and Expression.” *Nucleic Acids Research* 36 (Database) (December): D149–D153. doi:10.1093/nar/gkm995. http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkm995.

Bi, Yujing, Guangwei Liu, and Ruifu Yang. 2009. “MicroRNAs: Novel Regulators During the Immune Response.” *Journal of Cellular Physiology* 218 (3) (March): 467–72. doi:10.1002/jcp.21639. http://www.ncbi.nlm.nih.gov/pubmed/19034913.

Blanchet, Xavier, Marcella Langer, Christian Weber, Rory R Koenen, and Philipp von Hundelshausen. 2012. “Touch of Chemokines.” *Frontiers in Immunology* 3 (January): 175. doi:10.3389/fimmu.2012.00175. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3394994&tool=pmcentrez&rendertype=abstract.

Blatchford, David R, Lynda H Quarrie, Elizabeth Tonner, Corinna McCarthy, David J Flint, and Colin J Wilde. 1999. “Influence of Microenvironment on Mammary Epithelial Cell Survival in Primary Culture.” *Journal of Cellular Physiology* 181 (2) (November): 304–311. doi:10.1002/(SICI)1097-4652(199911)181:2<304::AID-JCP12>3.0.CO;2-5. http://doi.wiley.com/10.1002/(SICI)1097-4652(199911)181:2<304::AID-JCP12>3.0.CO;2-5.

Bode, Johannes G, Christian Ehlting, and Dieter Häussinger. 2012. “The Macrophage Response Towards LPS and Its Control through the p38(MAPK)-STAT3 Axis.” *Cellular Signalling* 24 (6) (June): 1185–94. doi:10.1016/j.cellsig.2012.01.018. http://www.sciencedirect.com/science/article/pii/S0898656812000411.

Boone, David L, Emre E Turer, Eric G Lee, Regina-Celeste Ahmad, Matthew T Wheeler, Colleen Tsui, Paula Hurley, et al. 2004. “The Ubiquitin-Modifying Enzyme A20 Is Required for Termination of Toll-Like Receptor Responses.” *Nature Immunology* 5 (10) (October 29): 1052–60. doi:10.1038/ni1110. http://www.nature.com.proxy.lib.sfu.ca/ni/journal/v5/n10/full/ni1110.html.

Bopp, Selina E R, Vandana Ramachandran, Kerstin Henson, Angelina Luzader, Merle Lindstrom, Muriel Spooner, Brian M Steffy, et al. 2010. “Genome Wide Analysis of Inbred Mouse Lines Identifies a Locus Containing Ppar-Gamma as Contributing to Enhanced Malaria Survival.” Edited by Denise L. Doolan. *PloS One* 5 (5) (January):

e10903. doi:10.1371/journal.pone.0010903. http://dx.plos.org/10.1371/journal.pone.0010903.

Brawand, David, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, et al. 2011. “The Evolution of Gene Expression Levels in Mammalian Organs.” *Nature* 478 (7369) (October 20): 343–8. doi:10.1038/nature10532. http://www.ncbi.nlm.nih.gov/pubmed/22012392.

Brennecke, Julius, Alexander Stark, Robert B Russell, and Stephen M Cohen. 2005. “Principles of microRNA-Target Recognition.” *PLoS Biology* 3 (3) (March): e85. doi:10.1371/journal.pbio.0030085. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1043860&tool=pmcentrez&rendertype=abstract.

Breuer, Karin, Amir K Foroushani, Matthew R Laird, Carol Chen, Anastasia Sribnaia, Raymond Lo, Geoffrey L Winsor, Robert E W Hancock, Fiona S L Brinkman, and David J Lynn. 2013. “InnateDB: Systems Biology of Innate Immunity and Beyond--Recent Updates and Continuing Curation.” *Nucleic Acids Research* 41 (D1) (January 1): D1228–33. doi:10.1093/nar/gks1147. http://nar.oxfordjournals.org/content/41/D1/D1228.full.

Buermans, Henk P J, Yavuz Ariyurek, Gertjan van Ommen, Johan T den Dunnen, and Peter A C ’t Hoen. 2010. “New Methods for Next Generation Sequencing Based microRNA Expression Profiling.” *BMC Genomics* 11 (1): 716. doi:10.1186/1471-2164-11-716. http://www.biomedcentral.com/1471-2164/11/716.

Bullard, James H, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. 2010. “Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments.” *BMC Bioinformatics* 11 (1): 94. doi:10.1186/1471-2105-11-94. http://www.biomedcentral.com/1471-2105/11/94.

Bushati, Natascha, and Stephen M Cohen. 2007. “microRNA Functions.” *Annual Review of Cell and Developmental Biology* 23 (January): 175–205. doi:10.1146/annurev.cellbio.23.090506.123406. http://www.ncbi.nlm.nih.gov/pubmed/17506695.

Chavali, Arvind K, Erwin P Gianchandani, Kenneth S Tung, Michael B Lawrence, Shayn M Peirce, and Jason a Papin. 2008. “Characterizing Emergent Properties of Immunological Systems with Multi-Cellular Rule-Based Computational Modeling.” *Trends in Immunology* 29 (12) (December): 589–99. doi:10.1016/j.it.2008.08.006. http://www.ncbi.nlm.nih.gov/pubmed/18964301.

Chen, Nansheng, and Lincoln D Stein. 2006. “Conservation and Functional Significance of Gene Topology in the Genome of Caenorhabditis Elegans.” *Genome Research* 16 (5) (May): 606–17. doi:10.1101/gr.4515306. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1457050&tool=pmcentrez&rendertype=abstract.

Chen, Xi, Chao Gao, Haijin Li, Lei Huang, Qi Sun, Yanye Dong, Chunliang Tian, et al. 2010. “Identification and Characterization of microRNAs in Raw Milk During

Different Periods of Lactation, Commercial Fluid, and Powdered Milk Products." *Cell Research* 20 (10) (June): 1128–1137. doi:10.1038/cr.2010.80. http://www.nature.com/doifinder/10.1038/cr.2010.80.

Cheng, Ying, Wenhua Kuang, Yongchang Hao, Donglin Zhang, Ming Lei, Li Du, Hanwei Jiao, Xiaoru Zhang, and Fengyang Wang. 2012. "Downregulation of miR-27a* and miR-532-5p and Upregulation of miR-146a and miR-155 in LPS-Induced RAW264.7 Macrophage Cells." *Inflammation* 35 (4) (August): 1308–13. doi:10.1007/s10753-012-9443-8. http://www.ncbi.nlm.nih.gov/pubmed/22415194.

Cho, Sooyoung, Insu Jang, Yukyung Jun, Suhyeon Yoon, Minjeong Ko, Yeajee Kwon, Ikjung Choi, et al. 2013. "MiRGator V3.0: a microRNA Portal for Deep Sequencing, Expression Profiling and mRNA Targeting." *Nucleic Acids Research* 41 (Database issue) (January 1): D252–7. doi:10.1093/nar/gks1168. http://nar.oxfordjournals.org/content/41/D1/D252.full.

Chua, Hon Nian, Wing-Kin Sung, and Limsoon Wong. 2006. "Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein-Protein Interactions." *Bioinformatics (Oxford, England)* 22 (13) (July 1): 1623–30. doi:10.1093/bioinformatics/btl145. http://bioinformatics.oxfordjournals.org/content/22/13/1623.

Cichocki, Frank, Martin Felices, Valarie McCullar, Steven R Presnell, Ahmad Al-Attar, Charles T Lutz, and Jeffrey S Miller. 2011. "Cutting Edge: microRNA-181 Promotes Human NK Cell Development by Regulating Notch Signaling." *Journal of Immunology (Baltimore, Md.: 1950)* 187 (12) (December 15): 6171–5. doi:10.4049/jimmunol.1100835. http://www.jimmunol.org/content/187/12/6171.full.

Clarkson, Benjamin D, Erika Héninger, Melissa G Harris, JangEun Lee, Matyas Sandor, and Zsuzsanna Fabry. 2012. "Innate-Adaptive Crosstalk: How Dendritic Cells Shape Immune Responses in the CNS." *Advances in Experimental Medicine and Biology* 946 (January): 309–33. doi:10.1007/978-1-4614-0106-3_18. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3666851&tool=pmcentrez&rendertype=abstract.

Cloonan, Nicole, Shivangi Wani, Qinying Xu, Jian Gu, Kristi Lea, Sheila Heater, Catalin Barbacioru, et al. 2011. "MicroRNAs and Their isomiRs Function Cooperatively to Target Common Biological Pathways." *Genome Biology* 12 (12): R126. doi:10.1186/gb-2011-12-12-r126. http://genomebiology.com/2012/12/12/R126.

Cock, P J A, C J Fields, N Goto, M L Heuer, and P M Rice. 2009. "The Sanger FASTQ File Format for Sequences with Quality Scores, and the Solexa/Illumina FASTQ Variants." *Nucleic Acids Research* 38 (6) (December): 1767–1771. doi:10.1093/nar/gkp1137. http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkp1137.

Collins, Aideen S, Claire E McCoy, Andrew T Lloyd, Cliona O'Farrelly, and Nigel J Stevenson. 2013. "miR-19a: An Effective Regulator of SOCS3 and Enhancer of JAK-STAT Signalling." *PloS One* 8 (7) (January): e69090. doi:10.1371/journal.pone.0069090.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3718810&tool=pmcentrez&rendertype=abstract.

Colobran, R, R Pujol-Borrell, M P Armengol, and M Juan. 2007. “The Chemokine Network. II. On How Polymorphisms and Alternative Splicing Increase the Number of Molecular Species and Configure Intricate Patterns of Disease Susceptibility.” *Clinical and Experimental Immunology* 150 (1) (October): 1–12.

Cong, Le, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, et al. 2013. “Multiplex Genome Engineering Using CRISPR/Cas Systems.” *Science (New York, N.Y.)* 339 (6121) (February 15): 819–23. doi:10.1126/science.1231143. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3795411&tool=pmcentrez&rendertype=abstract.

Copeland, N.G., and N.A. Jenkins. 2009. “Deciphering the Genetic Landscape of Cancer-from Genes to Pathways.” *Trends in Genetics* 25 (10): 455–462. doi:10.1016/j.tig.2009.08.004. http://www.sciencedirect.com/science/article/pii/S0168952509001735.

Costantini, Claudio, and Marco A Cassatella. 2011. “The Defensive Alliance Between Neutrophils and NK Cells as a Novel Arm of Innate Immunity.” *Journal of Leukocyte Biology* 89 (2) (February): 221–33. doi:10.1189/jlb.0510250. http://www.ncbi.nlm.nih.gov/pubmed/20682626.

Coutinho, L L, L K Matukumalli, T S Sonstegard, C P Van Tassell, L C Gasbarre, A V Capuco, and T P L Smith. 2006. “Discovery and Profiling of Bovine microRNAs from Immune-Related and Embryonic Tissues.” *Physiological Genomics* 29 (1) (December): 35–43. doi:10.1152/physiolgenomics.00081.2006. http://physiolgenomics.physiology.org/cgi/doi/10.1152/physiolgenomics.00081.2006.

Crozat, Karine, Eric Vivier, and Marc Dalod. 2009. “Crosstalk Between Components of the Innate Immune System: Promoting Anti-Microbial Defenses and Avoiding Immunopathologies.” *Immunological Reviews* 227 (1) (January): 129–49. doi:10.1111/j.1600-065X.2008.00736.x. http://www.ncbi.nlm.nih.gov/pubmed/19120481.

Csardi, Gabor, and Tamas Nepusz. 2006. “The Igraph Software Package for Complex Network Research.” *InterJournal* Complex Sy: 1695. http://igraph.sf.net.

Cui, Lunbiao, Xiling Guo, Yuhua Qi, Xian Qi, Yiyue Ge, Zhiyang Shi, Tao Wu, et al. 2010. “Identification of microRNAs Involved in the Host Response to Enterovirus 71 Infection by a Deep Sequencing Approach.” *Journal of Biomedicine and Biotechnology* 2010: 1–8. doi:10.1155/2010/425939. http://www.hindawi.com/journals/bmri/2010/425939/.

Daszak, Peter, Carlos Zambrana-Torrelio, Tiffany L Bogich, Miguel Fernandez, Jonathan H Epstein, Kris A Murray, and Healy Hamilton. 2013. “Interdisciplinary Approaches to Understanding Disease Emergence: The Past, Present, and Future Drivers of

Nipah Virus Emergence." *Proceedings of the National Academy of Sciences of the United States of America* 110 Suppl (Supplement_1) (February 26): 3681–8. doi:10.1073/pnas.1201243109. http://www.pnas.org/content/110/Supplement_1/3681.

De Martel, Catherine, and Silvia Franceschi. 2009. "Infections and Cancer: Established Associations and New Hypotheses." *Critical Reviews in Oncology/hematology* 70 (3) (June): 183–94. doi:10.1016/j.critrevonc.2008.07.021. http://www.ncbi.nlm.nih.gov/pubmed/18805702.

De Smet, Riet, and Kathleen Marchal. 2010. "Advantages and Limitations of Current Network Inference Methods." *Nature Reviews. Microbiology* 8 (10) (October): 717–29. doi:10.1038/nrmicro2419. http://www.ncbi.nlm.nih.gov/pubmed/20805835.

De Swart, Rik L, W Paul Duprex, and Albert D M E Osterhaus. 2012. "Rinderpest Eradication: Lessons for Measles Eradication?" *Current Opinion in Virology* 2 (3) (June): 330–4. doi:10.1016/j.coviro.2012.02.010. http://www.ncbi.nlm.nih.gov/pubmed/22709518.

Delano, Matthew J, Terri Thayer, Sonia Gabrilovich, Kindra M Kelly-Scumpia, Robert D Winfield, Philip O Scumpia, Alex G Cuenca, et al. 2011. "Sepsis Induces Early Alterations in Innate Immunity That Impact Mortality to Secondary Infection." *Journal of Immunology (Baltimore, Md.: 1950)* 186 (1) (January 1): 195–202. doi:10.4049/jimmunol.1002104. http://www.ncbi.nlm.nih.gov/pubmed/21106855.

Dezso, Zoltán, Yuri Nikolsky, Evgeny Sviridov, Weiwei Shi, Tatiana Serebriyskaya, Damir Dosymbekov, Andrej Bugrim, et al. 2008. "A Comprehensive Functional Analysis of Tissue Specificity of Human Gene Expression." *BMC Biology* 6 (1) (January): 49. doi:10.1186/1741-7007-6-49. http://www.biomedcentral.com/1741-7007/6/49.

Diacovich, Lautaro, and J.P. Gorvel. 2010. "Bacterial Manipulation of Innate Immunity to Promote Infection." *Nature Reviews Microbiology* 8 (2): 117–128. doi:10.1038/nrmicro2295. http://dx.doi.org/10.1038/nrmicro2295.

Didiano, Dominic, and Oliver Hobert. 2006. "Perfect Seed Pairing Is Not a Generally Reliable Predictor for miRNA-Target Interactions." *Nature Structural & Molecular Biology* 13 (9) (September): 849–51. doi:10.1038/nsmb1138. http://www.ncbi.nlm.nih.gov/pubmed/16921378.

Dilda, Francesca, Gloria Gioia, Laura Pisani, Laura Restelli, Cristina Lecchi, Francesca Albonico, Valerio Bronzo, Michele Mortarino, and Fabrizio Ceciliani. 2012. "Escherichia Coli Lipopolysaccharides and Staphylococcus Aureus Enterotoxin B Differentially Modulate Inflammatory microRNAs in Bovine Monocytes." *The Veterinary Journal* 192 (3) (June): 514–516. doi:10.1016/j.tvjl.2011.08.018. http://linkinghub.elsevier.com/retrieve/pii/S1090023311003005.

Dinu, Irina, John D Potter, Thomas Mueller, Qi Liu, Adeniyi J Adewale, Gian S Jhangri, Gunilla Einecke, Konrad S Famulski, Philip Halloran, and Yutaka Yasui. 2009.

"Gene-Set Analysis and Reduction." *Briefings in Bioinformatics* 10 (1): 24–34. http://www.ncbi.nlm.nih.gov/pubmed/18836208.

Dohoo, I R, L DesCôteaux, K Leslie, A Fredeen, W Shewfelt, A Preston, and P Dowling. 2003. "A Meta-Analysis Review of the Effects of Recombinant Bovine Somatotropin. 2. Effects on Animal Health, Reproductive Performance, and Culling." *Canadian Journal of Veterinary Research = Revue Canadienne de Recherche Vétérinaire* 67 (4) (October): 252–64. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=280709&tool=pmcentrez&rendertype=abstract.

Draghici, Sorin, Purvesh Khatri, A.L. Tarca, Kashyap Amin, Arina Done, Calin Voichita, Constantin Georgescu, and Roberto Romero. 2007. "A Systems Biology Approach for Pathway Level Analysis." *Genome Research* 17 (10): 1537–1545. doi:10.1101/gr.6202607. http://genome.cshlp.org/content/17/10/1537.short.

Dunne, Aisling, and Luke A J O'Neill. 2005. "Adaptor Usage and Toll-Like Receptor Signaling Specificity." *FEBS Letters* 579 (15) (June 13): 3330–5. doi:10.1016/j.febslet.2005.04.024. http://dx.doi.org/10.1016/j.febslet.2005.04.024.

Eisen, M B, P T Spellman, P O Brown, and D Botstein. 1998. "Cluster Analysis and Display of Genome-Wide Expression Patterns." *Proceedings of the National Academy of Sciences of the United States of America* 95 (25) (December 8): 14863–8. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=24541&tool=pmcentrez&rendertype=abstract.

Emmert-Streib, Frank, and Galina V Glazko. 2011. "Pathway Analysis of Expression Data: Deciphering Functional Building Blocks of Complex Diseases." Edited by Fran Lewitter. *PLoS Computational Biology* 7 (5) (May): e1002053. doi:10.1371/journal.pcbi.1002053. http://dx.plos.org/10.1371/journal.pcbi.1002053.

Enright, Anton J, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S Marks. 2003. "MicroRNA Targets in Drosophila." *Genome Biology* 5 (1) (January): R1. doi:10.1186/gb-2003-5-1-r1. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=395733&tool=pmcentrez&rendertype=abstract.

Ernst, J D. 1998. "Macrophage Receptors for Mycobacterium Tuberculosis." *Infection and Immunity* 66 (4) (April 26): 1277–81. http://pubmedcentralcanada.ca/pmcc/articles/PMC108049/.

Essbauer, Sandra, Martin Pfeffer, and Hermann Meyer. 2010. "Zoonotic Poxviruses." *Veterinary Microbiology* 140 (3-4) (January 27): 229–36. doi:10.1016/j.vetmic.2009.08.026. http://www.ncbi.nlm.nih.gov/pubmed/19828265.

Flicek, P, M R Amode, D Barrell, K Beal, S Brent, D Carvalho-Silva, P Clapham, et al. 2011. "Ensembl 2012." *Nucleic Acids Research* 40 (D1) (November): D84–D90. doi:10.1093/nar/gkr991. http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkr991.

Foroushani, Amir B K, Fiona S L Brinkman, and David J Lynn. 2013. “Pathway-GPS and SIGORA: Identifying Relevant Pathways Based on the over-Representation of Their Gene-Pair Signatures.” *PeerJ* 1 (December 19): e229. doi:10.7717/peerj.229. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3883547&tool=pmcentrez&rendertype=abstract.

Francischetti, Ivo M B, Karl B Seydel, and Robson Q Monteiro. 2008. “Blood Coagulation, Inflammation, and Malaria.” *Microcirculation (New York, N.Y.: 1994)* 15 (2) (February): 81–107. doi:10.1080/10739680701451516. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2892216&tool=pmcentrez&rendertype=abstract.

Friedländer, Marc R, Wei Chen, Catherine Adamidi, Jonas Maaskola, Ralf Einspanier, Signe Knespel, and Nikolaus Rajewsky. 2008. “Discovering microRNAs from Deep Sequencing Data Using miRDeep.” *Nature Biotechnology* 26 (4) (April): 407–15. doi:10.1038/nbt1394. http://dx.doi.org/10.1038/nbt1394.

Friedman, R C, K K.-H. Farh, C B Burge, and D P Bartel. 2008. “Most Mammalian mRNAs Are Conserved Targets of microRNAs.” *Genome Research* 19 (1) (October): 92–105. doi:10.1101/gr.082701.108. http://genome.cshlp.org/cgi/doi/10.1101/gr.082701.108.

Friend, S H. 2010. “The Need for Precompetitive Integrative Bionetwork Disease Model Building.” *Clinical Pharmacology and Therapeutics* 87 (5) (May): 536–539.

Frohlich, H, T Beissbarth, A Tresch, D Kostka, J Jacob, R Spang, and F Markowetz. 2008. “Analyzing Gene Perturbation Screens with Nested Effects Models in R and Bioconductor.” *Bioinformatics (Oxford, England)* 24 (21) (November): 2549–2550.

Fulton, D L, Y Y Li, M R Laird, B G Horsman, F M Roche, and F S Brinkman. 2006. “Improving the Specificity of High-Throughput Ortholog Prediction.” *BMC Bioinformatics* 7 (May): 270.

Furuse, Yuki, Akira Suzuki, and Hitoshi Oshitani. 2010. “Origin of Measles Virus: Divergence from Rinderpest Virus Between the 11th and 12th Centuries.” *Virology Journal* 7 (January): 52. doi:10.1186/1743-422X-7-52. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2838858&tool=pmcentrez&rendertype=abstract.

Gantier, Michael P. 2010. “New Perspectives in MicroRNA Regulation of Innate Immunity.” *Journal of Interferon & Cytokine Research : the Official Journal of the International Society for Interferon and Cytokine Research* 30 (5) (May): 283–9. doi:10.1089/jir.2010.0037. http://www.ncbi.nlm.nih.gov/pubmed/20477549.

Garcia, David M, Daehyun Baek, Chanseok Shin, George W Bell, Andrew Grimson, and David P Bartel. 2011. “Weak Seed-Pairing Stability and High Target-Site Abundance Decrease the Proficiency of Lsy-6 and Other microRNAs.” *Nature Structural & Molecular Biology* 18 (10) (October): 1139–46. doi:10.1038/nsmb.2115. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3190056&tool=pmcentrez&rendertype=abstract.

Gardy, Jennifer L, David J Lynn, Fiona S L Brinkman, and Robert E W Hancock. 2009. “Enabling a Systems Biology Approach to Immunology: Focus on Innate Immunity.” *Trends in Immunology* 30 (6) (June): 249–62. doi:10.1016/j.it.2009.03.009. http://www.ncbi.nlm.nih.gov/pubmed/19428301.

Garmire, L X, and S Subramaniam. 2012. “Evaluation of Normalization Methods in Mammalian microRNA-Seq Data.” *RNA* 18 (6) (April): 1279–1288. doi:10.1261/rna.030916.111. http://rnajournal.cshlp.org/cgi/doi/10.1261/rna.030916.111.

Garnier, Thierry, Karin Eiglmeier, Jean-Christophe Camus, Nadine Medina, Huma Mansoor, Melinda Pryor, Stephanie Duthoy, et al. 2003. “The Complete Genome Sequence of Mycobacterium Bovis.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (13) (June 24): 7877–82. doi:10.1073/pnas.1130426100. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=164681&tool=pmcentrez&rendertype=abstract.

Gasch, Audrey P, and Michael B Eisen. 2002. “Exploring the Conditional Coregulation of Yeast Gene Expression through Fuzzy k-Means Clustering.” *Genome Biology* 3 (11) (October 10): RESEARCH0059. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=133443&tool=pmcentrez&rendertype=abstract.

Gatherer, Derek. 2010. “So What Do We Really Mean When We Say That Systems Biology Is Holistic?” *BMC Systems Biology* 4 (January): 22. doi:10.1186/1752-0509-4-22. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2850881&tool=pmcentrez&rendertype=abstract.

Getoor, Lise, and Christopher P. Diehl. 2005. “Link Mining.” *ACM SIGKDD Explorations Newsletter* 7 (2) (December 1): 3–12. doi:10.1145/1117454.1117456. http://dl.acm.org/citation.cfm?id=1117454.1117456.

Ghorpade, Devram Sampat, Rebecca Leyland, Mariola Kurowska-Stolarska, Shripad A Patil, and Kithiganahalli Narayanaswamy Balaji. 2012. “MicroRNA-155 Is Required for Mycobacterium Bovis BCG-Mediated Apoptosis of Macrophages.” *Molecular and Cellular Biology* 32 (12) (June): 2239–53. doi:10.1128/MCB.06597-11. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3372268&tool=pmcentrez&rendertype=abstract.

Gillis, Jesse, and Paul Pavlidis. 2011. “The Impact of Multifunctional Genes on ‘Guilt by Association’ Analysis.” Edited by Joel Bader. *PloS One* 6 (2) (January): e17258. doi:10.1371/journal.pone.0017258. http://dx.plos.org/10.1371/journal.pone.0017258.

———. 2012. “‘Guilt by Association’ Is the Exception Rather Than the Rule in Gene Networks.” *PLoS Computational Biology* 8 (3) (January): e1002444. doi:10.1371/journal.pcbi.1002444.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3315453&tool=pmcentrez&rendertype=abstract.

Glazov, Evgeny A, Kritaya Kongsuwan, Wanchai Assavalapsakul, Paul F Horwood, Neena Mitter, and Timothy J Mahony. 2009. “Repertoire of Bovine miRNA and miRNA-Like Small Regulatory RNAs Expressed Upon Viral Infection.” Edited by Lennart Randau. *PLoS ONE* 4 (7) (July): e6349. doi:10.1371/journal.pone.0006349. http://dx.plos.org/10.1371/journal.pone.0006349.

Goeman, Jelle J, and Peter Bühlmann. 2007. “Analyzing Gene Expression Data in Terms of Gene Sets: Methodological Issues.” *Bioinformatics (Oxford, England)* 23 (8) (April 15): 980–7. doi:10.1093/bioinformatics/btm051. http://www.ncbi.nlm.nih.gov/pubmed/17303618.

Gonose, Thandubuhle, Anthony M Smith, Karen H Keddy, Arvinda Sooka, Victoria Howell, Charlene Ann Jacobs, Sumayya Haffejee, and Premi Govender. 2012. “Human Infections Due to Salmonella Blockley, a Rare Serotype in South Africa: a Case Report.” *BMC Research Notes* 5 (1) (January): 562. doi:10.1186/1756-0500-5-562. http://www.biomedcentral.com/1756-0500/5/562.

Gouwy, Mieke, Milena Schiraldi, Sofie Struyf, Jo Van Damme, and Mariagrazia Uguccioni. 2012. “Possible Mechanisms Involved in Chemokine Synergy Fine Tuning the Inflammatory Response.” *Immunology Letters* 145 (1-2) (July 30): 10–4. doi:10.1016/j.imlet.2012.04.005. http://www.ncbi.nlm.nih.gov/pubmed/22698178.

Graff, Joel W, Anne M Dickson, Gwendolyn Clay, Anton P McCaffrey, and Mary E Wilson. 2012. “Identifying Functional microRNAs in Macrophages with Polarized Phenotypes.” *The Journal of Biological Chemistry* 287 (26) (June 22): 21816–25. doi:10.1074/jbc.M111.327031. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3381144&tool=pmcentrez&rendertype=abstract.

Gray, Kristian A, Louise C Daugherty, Susan M Gordon, Ruth L Seal, Mathew W Wright, and Elspeth A Bruford. 2013. “Genenames.org: The HGNC Resources in 2013.” *Nucleic Acids Research* 41 (Database issue) (January): D545–52. doi:10.1093/nar/gks1066. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531211&tool=pmcentrez&rendertype=abstract.

Greger, Michael. 2008. “The Human/Animal Interface: Emergence and Resurgence of Zoonotic Infectious Diseases” (October 11). http://informahealthcare.com/doi/abs/10.1080/10408410701647594.

Grimson, Andrew, Kyle Kai-How Farh, Wendy K Johnston, Philip Garrett-Engele, Lee P Lim, and David P Bartel. 2007. “MicroRNA Targeting Specificity in Mammals: Determinants Beyond Seed Pairing.” *Molecular Cell* 27 (1) (July): 91–105. doi:10.1016/j.molcel.2007.06.017. http://linkinghub.elsevier.com/retrieve/pii/S1097276507004078.

Grossmann, Steffen, Sebastian Bauer, Peter N Robinson, and Martin Vingron. 2007. “Improved Detection of Overrepresentation of Gene-Ontology Annotations with Parent Child Analysis.” *Bioinformatics (Oxford, England)* 23 (22) (November 15): 3024–31. doi:10.1093/bioinformatics/btm440. http://www.ncbi.nlm.nih.gov/pubmed/17848398.

Guduric-Fuchs, Jasenka, Anna O’Connor, Angela Cullen, Laura Harwood, Reinhold J Medina, Christina L O’Neill, Alan W Stitt, Tim M Curtis, and David A Simpson. 2012. “Deep Sequencing Reveals Predominant Expression of miR-21 Amongst the Small Non-Coding RNAs in Retinal Microvascular Endothelial Cells.” *Journal of Cellular Biochemistry* 113 (6) (June): 2098–2111. doi:10.1002/jcb.24084. http://doi.wiley.com/10.1002/jcb.24084.

Hajishengallis, George, and John D Lambris. 2011. “Microbial Manipulation of Receptor Crosstalk in Innate Immunity.” *Nature Reviews. Immunology* 11 (3) (March 25): 187–200. doi:10.1038/nri2918. http://www.nature.com.proxy.lib.sfu.ca/nri/journal/v11/n3/full/nri2918.html.

Hammell, Molly. 2010. “Computational Methods to Identify miRNA Targets.” *Seminars in Cell & Developmental Biology* 21 (7) (September): 738–44. doi:10.1016/j.semcdb.2010.01.004. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2891825&tool=pmcentrez&rendertype=abstract.

Hansen, Diana S, Mary-Anne Siomos, Lynn Buckingham, Anthony A Scalzo, and Louis Schofield. 2003. “Regulation of Murine Cerebral Malaria Pathogenesis by CD1d-Restricted NKT Cells and the Natural Killer Complex.” *Immunity* 18 (3) (March): 391–402. http://www.ncbi.nlm.nih.gov/pubmed/12648456.

Harhay, Gregory P, Timothy Pl Smith, Leeson J Alexander, Christian D Haudenschild, John W Keele, Lakshmi K Matukumalli, Steven G Schroeder, et al. 2010. “An Atlas of Bovine Gene Expression Reveals Novel Distinctive Tissue Characteristics and Evidence for Improving Genome Annotation.” *Genome Biology* 11 (10) (January): R102. doi:10.1186/gb-2010-11-10-r102. http://www.ncbi.nlm.nih.gov/pubmed/20961407.

Harizi, Hedi. 2013. “Reciprocal Crosstalk Between Dendritic Cells and Natural Killer Cells Under the Effects of PGE2 in Immunity and Immunopathology.” *Cellular & Molecular Immunology* 10 (3) (May): 213–21. doi:10.1038/cmi.2013.1. http://www.ncbi.nlm.nih.gov/pubmed/23524652.

Heaton, Nicholas S, and Glenn Randall. 2010. “Dengue Virus-Induced Autophagy Regulates Lipid Metabolism.” *Cell Host & Microbe* 8 (5) (November 18): 422–32. doi:10.1016/j.chom.2010.10.006. http://www.cell.com/cell-host-microbe/fulltext/S1931-3128(10)00343-4.

Hirata, Noriyuki, Yoshiki Yanagawa, Takashi Ebihara, Tsukasa Seya, Satoshi Uematsu, Shizuo Akira, Fumie Hayashi, Kazuya Iwabuchi, and Kazunori Onoé. 2008. “Selective Synergy in Anti-Inflammatory Cytokine Production Upon Cooperated Signaling via TLR4 and TLR2 in Murine Conventional Dendritic Cells.” *Molecular*

Immunology 45 (10) (May): 2734–42. doi:10.1016/j.molimm.2008.02.010. http://www.ncbi.nlm.nih.gov/pubmed/18372043.

Hoang, Long Truong, David J Lynn, Matt Henn, Bruce W Birren, Niall J Lennon, Phuong Thi Le, Kien Thi Hue Duong, et al. 2010. “The Early Whole-Blood Transcriptional Signature of Dengue and Features Associated with Progression to Dengue Shock Syndrome in Vietnamese Children and Young Adults.” *Journal of Virology* (October) (October 13). doi:10.1128/JVI.01224-10. http://www.ncbi.nlm.nih.gov/pubmed/20943967.

Hoffmann, Robert. 2008. “A Wiki for the Life Sciences Where Authorship Matters.” *Nature Genetics* 40 (9) (September): 1047–51. doi:10.1038/ng.f.217. http://www.ncbi.nlm.nih.gov/pubmed/18728691.

Holley, Christopher L, and Veli K Topkara. 2011. “An Introduction to Small Non-Coding RNAs: miRNA and snoRNA.” *Cardiovascular Drugs and Therapy / Sponsored by the International Society of Cardiovascular Pharmacotherapy* 25 (2) (April): 151–9. doi:10.1007/s10557-011-6290-z. http://www.ncbi.nlm.nih.gov/pubmed/21573765.

Horan, Kevin, Charles Jang, Julia Bailey-Serres, Ron Mittler, Christian Shelton, Jeff F Harper, Jian-Kang Zhu, John C Cushman, Martin Gollery, and Thomas Girke. 2008. “Annotating Genes of Known and Unknown Function by Large-Scale Coexpression Analysis.” *Plant Physiology* 147 (1) (May): 41–57. doi:10.1104/pp.108.117366. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2330292&tool=pmcentrez&rendertype=abstract.

Hou, Qinlei, Jinming Huang, Zhihua Ju, Qiuling Li, Liming Li, Changfa Wang, Tao Sun, et al. 2012. “Identification of Splice Variants, Targeted MicroRNAs and Functional Single Nucleotide Polymorphisms of the BOLA-DQA2 Gene in Dairy Cattle.” *DNA and Cell Biology* 31 (5) (May): 739–744. doi:10.1089/dna.2011.1402. http://online.liebertpub.com/doi/abs/10.1089/dna.2011.1402.

Hsieh, Ching-Hua, Cheng-Shyuan Rau, Jonathan Jeng, Yi-Chun Chen, Tsu-Hsiang Lu, Chia-Jung Wu, Yi-Chan Wu, Siou-Ling Tzeng, and Johnson Yang. 2012. “Whole Blood-Derived microRNA Signatures in Mice Exposed to Lipopolysaccharides.” *Journal of Biomedical Science* 19 (1): 69. doi:10.1186/1423-0127-19-69. http://www.jbiomedsci.com/content/19/1/69.

Hsueh, R C, M Natarajan, I Fraser, B Pond, J Liu, S Mumby, H Han, et al. 2009. “Deciphering Signaling Outcomes from a System of Complex Networks.” *Science Signaling* 2 (71) (May): ra22.

Hu, Wei, and Chandrashekhar Pasare. 2013. “Location, Location, Location: Tissue-Specific Regulation of Immune Responses.” *Journal of Leukocyte Biology* 94 (3) (September): 409–21. doi:10.1189/jlb.0413207. http://www.ncbi.nlm.nih.gov/pubmed/23825388.

Huang, Da Wei, Brad T Sherman, and Richard a Lempicki. 2009. “Bioinformatics Enrichment Tools: Paths Toward the Comprehensive Functional Analysis of Large

Gene Lists." *Nucleic Acids Research* 37 (1) (January): 1–13. doi:10.1093/nar/gkn923. http://www.ncbi.nlm.nih.gov/pubmed/19033363.

Huang, H.-C., H.-R. Yu, L.-T. Huang, R.-F. Chen, I.-C. Lin, C.-Y. Ou, T.-Y. Hsu, and K D Yang. 2012. "miRNA-125b Regulates TNF- Production in CD14+ Neonatal Monocytes via Post-Transcriptional Regulation." *Journal of Leukocyte Biology* 92 (1) (May): 171–182. doi:10.1189/jlb.1211593. http://www.jleukbio.org/cgi/doi/10.1189/jlb.1211593.

Huang, J C, Q D Morris, and B J Frey. 2007. "Bayesian Inference of MicroRNA Targets from Sequence and Expression Data." *Journal of Computational Biology : a Journal of Computational Molecular Cell Biology* 14 (5) (June): 550–563.

Huang, Jinming. 2011. "Solexa Sequencing of Novel and Differentially Expressed MicroRNAs in Testicular and Ovarian Tissues in Holstein Cattle." *International Journal of Biological Sciences*: 1016–1026. doi:10.7150/ijbs.7.1016. http://www.ijbs.com/v07p1016.htm.

Huang, Yu, Haifeng Li, Haiyan Hu, Xifeng Yan, Michael S Waterman, Haiyan Huang, and Xianghong Jasmine Zhou. 2007. "Systematic Discovery of Functional Modules and Context-Specific Functional Annotation of Human Genome." *Bioinformatics (Oxford, England)* 23 (13) (July 1): i222–9. doi:10.1093/bioinformatics/btm222. http://www.ncbi.nlm.nih.gov/pubmed/17646300.

Hubbard, T J P, B L Aken, S Ayling, B Ballester, K Beal, E Bragin, S Brent, et al. 2009. "Ensembl 2009." *Nucleic Acids Research* 37 (Database) (January): D690–D697. doi:10.1093/nar/gkn828. http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkn828.

Hussein, H S. 2007. "Prevalence and Pathogenicity of Shiga Toxin-Producing Escherichia Coli in Beef Cattle and Their Products." *Journal of Animal Science* 85 (13 Suppl) (March): E63–72. doi:10.2527/jas.2006-421. http://www.ncbi.nlm.nih.gov/pubmed/17060419.

Ioannidis, John P A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8). doi:10.1371/journal.pmed.0020124.

Isci, S., C. Ozturk, J. Jones, and H. H. Otu. 2011. "Pathway Analysis of High Throughput Biological Data Within a Bayesian Network Framework." *Bioinformatics* 27 (12) (May 5): 1667–1674. doi:10.1093/bioinformatics/btr269. http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btr269.

Ivey, Kathryn N, and Deepak Srivastava. 2010. "MicroRNAs as Regulators of Differentiation and Cell Fate Decisions." *Cell Stem Cell* 7 (1) (July 2): 36–41. doi:10.1016/j.stem.2010.06.012. http://www.ncbi.nlm.nih.gov/pubmed/20621048.

Janeway, C A. 1989. "Approaching the Asymptote? Evolution and Revolution in Immunology." *Cold Spring Harbor Symposia on Quantitative Biology* 54 Pt 1 (January): 1–13. http://www.ncbi.nlm.nih.gov/pubmed/2700931.

Jennewein, C, A von Knethen, T Schmid, and B Brune. 2010. “MicroRNA-27b Contributes to Lipopolysaccharide-Mediated Peroxisome Proliferator-Activated Receptor (PPAR) mRNA Destabilization.” *Journal of Biological Chemistry* 285 (16) (February): 11846–11853. doi:10.1074/jbc.M109.066399. http://www.jbc.org/cgi/doi/10.1074/jbc.M109.066399.

Jiao, Xiaoli, Brad T Sherman, Da Wei Huang, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard A Lempicki. 2012. “DAVID-WS: a Stateful Web Service to Facilitate Gene/protein List Analysis.” *Bioinformatics (Oxford, England)* 28 (13) (July 1): 1805–6. doi:10.1093/bioinformatics/bts251. http://bioinformatics.oxfordjournals.org/content/28/13/1805.

Johnson, Steven M, Helge Grosshans, Jaclyn Shingara, Mike Byrom, Rich Jarvis, Angie Cheng, Emmanuel Labourier, Kristy L Reinert, David Brown, and Frank J Slack. 2005. “RAS Is Regulated by the Let-7 microRNA Family.” *Cell* 120 (5) (March 11): 635–47. doi:10.1016/j.cell.2005.01.014. http://www.ncbi.nlm.nih.gov/pubmed/15766527.

Jones, Kate E, Nikkita G Patel, Marc A Levy, Adam Storeygard, Deborah Balk, John L Gittleman, and Peter Daszak. 2008. “Global Trends in Emerging Infectious Diseases.” *Nature* 451 (7181) (February 21): 990–3. doi:10.1038/nature06536. http://www.nature.com.proxy.lib.sfu.ca/nature/journal/v451/n7181/full/nature06536.html#B5.

Jopling, Catherine. 2012. “Liver-Specific microRNA-122: Biogenesis and Function.” *RNA Biology* 9 (2) (February): 137–42. doi:10.4161/rna.18827. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3346312&tool=pmcentrez&rendertype=abstract.

Jupiter, D., J. Sahutoglu, and V. VanBuren. 2009. “TreeHugger: A New Test for Enrichment of Gene Ontology Terms.” *INFORMS Journal on Computing* 22 (2) (October 2): 210–221. doi:10.1287/ijoc.1090.0356. http://joc.journal.informs.org/cgi/doi/10.1287/ijoc.1090.0356.

Kanehisa, M, and S Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28 (1) (January 1): 27–30. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcentrez&rendertype=abstract.

Kanehisa, Minoru, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. 2011a. “KEGG for Integration and Interpretation of Large-Scale Molecular Data Sets.” *Nucleic Acids Research*: 1–6. doi:10.1093/nar/gkr988.

———. 2011b. “KEGG for Integration and Interpretation of Large-Scale Molecular Data Sets.” *Nucleic Acids Research* 40 (Database issue) (November 10): D109–14. doi:10.1093/nar/gkr988. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245020&tool=pmcentrez&rendertype=abstract.

———. 2012. “KEGG for Integration and Interpretation of Large-Scale Molecular Data Sets.” *Nucleic Acids Research* 40 (D1): D109–D114. doi:10.1093/nar/gkr988. http://nar.oxfordjournals.org/content/40/D1/D109.abstract.

Karin, Michael, Toby Lawrence, and Victor Nizet. 2006. “Innate Immunity Gone Awry: Linking Microbial Infections to Chronic Inflammation and Cancer.” *Cell* 124 (4) (February 24): 823–35. doi:10.1016/j.cell.2006.02.016. http://www.ncbi.nlm.nih.gov/pubmed/16497591.

Kawai, Taro, and Shizuo Akira. 2011. “Toll-Like Receptors and Their Crosstalk with Other Innate Receptors in Infection and Immunity.” *Immunity* 34 (5) (May 27): 637–50. doi:10.1016/j.immuni.2011.05.006. http://www.ncbi.nlm.nih.gov/pubmed/21616434.

Keane, O M, K E Budd, J Flynn, and F McCoy. 2013. “Pathogen Profile of Clinical Mastitis in Irish Milk-Recording Herds Reveals a Complex Aetiology.” *The Veterinary Record* 173 (1) (July 6): 17. doi:10.1136/vr.101308. http://www.ncbi.nlm.nih.gov/pubmed/23694921.

Keestra, a Marijke, and Andreas J Bäumler. 2011. “Host Defenses Trigger Salmonella’s Arsenal.” *Cell Host & Microbe* 9 (3) (March 17): 167–8. doi:10.1016/j.chom.2011.03.003. http://www.ncbi.nlm.nih.gov/pubmed/21402352.

Kerrien, Samuel, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, et al. 2012. “The IntAct Molecular Interaction Database in 2012.” *Nucleic Acids Research* 40 (Database issue) (January): D841–6. doi:10.1093/nar/gkr1088. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245075&tool=pmcentrez&rendertype=abstract.

Kerrien, Samuel, Sandra Orchard, Luisa Montecchi-Palazzi, Bruno Aranda, Antony F Quinn, Nisha Vinod, Gary D Bader, et al. 2007. “Broadening the Horizon--Level 2.5 of the HUPO-PSI Format for Molecular Interactions.” *BMC Biology* 5 (January): 44. doi:10.1186/1741-7007-5-44. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2189715&tool=pmcentrez&rendertype=abstract.

Khakpoor, Atefeh, Mingkwan Panyasrivanit, Nitwara Wikan, and Duncan R Smith. 2009. “A Role for Autophagolysosomes in Dengue Virus 3 Production in HepG2 Cells.” *The Journal of General Virology* 90 (Pt 5) (May 1): 1093–103. doi:10.1099/vir.0.007914-0. http://vir.sgmjournals.org/content/90/5/1093.short.

Khalkhali-Ellis, Zhila, Daniel E Abbott, Caleb M Bailey, William Goossens, Naira V Margaryan, Stephen L Gluck, Moshe Reuveni, and Mary J C Hendrix. 2008. “IFN-γ Regulation of Vacuolar pH, Cathepsin D Processing and Autophagy in Mammary Epithelial Cells.” *Journal of Cellular Biochemistry* 105 (1) (September): 208–218. doi:10.1002/jcb.21814. http://doi.wiley.com/10.1002/jcb.21814.

Khatri, Purvesh, and Sorin Drăghici. 2005. “Ontological Analysis of Gene Expression Data: Current Tools, Limitations, and Open Problems.” *Bioinformatics (Oxford,*

England) 21 (18) (September 15): 3587–95. doi:10.1093/bioinformatics/bti565. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2435250&tool=pmcentrez&rendertype=abstract.

Khatri, Purvesh, Marina Sirota, and Atul J Butte. 2012. “Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges.” *PLoS Computational Biology* 8 (2) (January): e1002375. doi:10.1371/journal.pcbi.1002375. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3285573&tool=pmcentrez&rendertype=abstract.

Kim, S K, J Lund, M Kiraly, K Duke, M Jiang, J M Stuart, A Eizinger, B N Wylie, and G S Davidson. 2001. “A Gene Expression Map for Caenorhabditis Elegans.” *Science (New York, N.Y.)* 293 (5537) (September 14): 2087–92. doi:10.1126/science.1061603. http://www.ncbi.nlm.nih.gov/pubmed/11557892.

Kim, V Narry, Jinju Han, and Mikiko C Siomi. 2009. “Biogenesis of Small RNAs in Animals.” *Nature Reviews Molecular Cell Biology* 10 (2) (February): 126–139. doi:10.1038/nrm2632. http://www.nature.com/doifinder/10.1038/nrm2632.

Kistemann, Thomas, Sonja Zimmer, Ivar Vågsholm, and Yvonne Andersson. 2004. “GIS-Supported Investigation of Human EHEC and Cattle VTEC O157 Infections in Sweden: Geographical Distribution, Spatial Variation and Possible Risk Factors.” *Epidemiology and Infection* 132 (3) (June): 495–505. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2870128&tool=pmcentrez&rendertype=abstract.

Kitano, Hiroaki. 2002. “Systems Biology: a Brief Overview.” *Science (New York, N.Y.)* 295 (5560) (March 1): 1662–4. doi:10.1126/science.1069492. http://www.ncbi.nlm.nih.gov/pubmed/11872829.

Kitano, Hiroaki, and Kanae Oda. 2006. “Robustness Trade-Offs and Host-Microbial Symbiosis in the Immune System.” *Molecular Systems Biology* 2 (January): 2006.0022. doi:10.1038/msb4100039. http://www.ncbi.nlm.nih.gov/pubmed/16738567.

Kluger, Yuval, Ronen Basri, Joseph T Chang, and Mark Gerstein. 2003. “Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions.” *Genome Research* 13 (4) (April 1): 703–16. doi:10.1101/gr.648603. http://genome.cshlp.org/content/13/4/703.long.

Kowarsch, A, C Marr, D Schmidl, A Ruepp, and F J Theis. 2010. “Tissue-Specific Target Analysis of Disease-Associated microRNAs in Human Signaling Pathways.” *PloS One* 5 (6) (June): e11154.

Kozomara, A, and S Griffiths-Jones. 2010. “miRBase: Integrating microRNA Annotation and Deep-Sequencing Data.” *Nucleic Acids Research* 39 (Database) (October): D152–D157. doi:10.1093/nar/gkq1027. http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkq1027.

Krek, Azra, Dominic Grün, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J Epstein, Philip MacMenamin, et al. 2005. “Combinatorial microRNA Target Predictions.” *Nature Genetics* 37 (5) (May): 495–500. doi:10.1038/ng1536. http://www.ncbi.nlm.nih.gov/pubmed/15806104.

Kumar, Himanshu, Taro Kawai, and Shizuo Akira. 2011. “Pathogen Recognition by the Innate Immune System.” *International Reviews of Immunology* 30 (1) (February): 16–34. doi:10.3109/08830185.2010.529976. http://www.ncbi.nlm.nih.gov/pubmed/21235323.

Lambrecht, Bart N. 2006. “Alveolar Macrophage in the Driver’s Seat.” *Immunity* 24 (4) (April): 366–8. doi:10.1016/j.immuni.2006.03.008. http://www.ncbi.nlm.nih.gov/pubmed/16618595.

Lander, A D. 2010. “The Edges of Understanding.” *BMC Biology* 8 (April): 40.

Laubenbacher, R, V Hower, A Jarrah, S V Torti, V Shulaev, P Mendes, F M Torti, and S Akman. 2009. “A Systems Biology View of Cancer.” *Biochimica et Biophysica Acta* 1796 (2) (December): 129–139.

Lawless, Nathan, Amir B K Foroushani, Matthew S McCabe, Cliona O’Farrelly, and David J Lynn. 2013. “Next Generation Sequencing Reveals the Expression of a Unique miRNA Profile in Response to a Gram-Positive Bacterial Infection.” *PloS One* 8 (3) (January): e57543. doi:10.1371/journal.pone.0057543. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3589390&tool=pmcentrez&rendertype=abstract.

Lee, Homin K, Amy K Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis. 2004. “Coexpression Analysis of Human Genes Across Many Microarray Data Sets.” *Genome Research* 14 (6) (June): 1085–94. doi:10.1101/gr.1910904. http://genome.cshlp.org/cgi/content/abstract/14/6/1085.

Lee, Insuk, Bindu Ambaru, Pranjali Thakkar, Edward M Marcotte, and Seung Yon Rhee. 2010. “Rational Association of Genes with Traits Using a Genome-Scale Gene Network for Arabidopsis Thaliana.” *Nature …* (June 2009). doi:10.1038/nbt.1603. http://dx.doi.org/10.1038/nbt.1603.

Lee, Insuk, Ben Lehner, Catriona Crombie, Wendy Wong, Andrew G Fraser, and Edward M Marcotte. 2008. “A Single Gene Network Accurately Predicts Phenotypic Effects of Gene Perturbation in Caenorhabditis Elegans.” *Nature Genetics* 40 (2) (March): 181–8. doi:10.1038/ng.2007.70. http://dx.doi.org/10.1038/ng.2007.70.

Lee, L W, S Zhang, A Etheridge, L Ma, D Martin, D Galas, and K Wang. 2010. “Complexity of the microRNA Repertoire Revealed by Next-Generation Sequencing.” *RNA* 16 (11) (September): 2170–2180. doi:10.1261/rna.2225110. http://rnajournal.cshlp.org/cgi/doi/10.1261/rna.2225110.

Lee, Rosalind C, Rhonda L Feinbaum, and Victor Ambros. 1993. “The C. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to

Lin-14." *Cell* 75 (5) (December): 843–854. doi:10.1016/0092-8674(93)90529-Y. http://linkinghub.elsevier.com/retrieve/pii/009286749390529Y.

Lehuen, Agnès, Julien Diana, Paola Zaccone, and Anne Cooke. 2010. "Immune Cell Crosstalk in Type 1 Diabetes." *Nature Reviews. Immunology* 10 (7) (July): 501–13. doi:10.1038/nri2787. http://www.ncbi.nlm.nih.gov/pubmed/20577267.

Lewis, Benjamin P, Christopher B Burge, and David P Bartel. 2005. "Conserved Seed Pairing, Often Flanked by Adenosines, Indicates That Thousands of Human Genes Are MicroRNA Targets." *Cell* 120 (1) (January): 15–20. doi:10.1016/j.cell.2004.12.035. http://linkinghub.elsevier.com/retrieve/pii/S0092867404012607.

Lewis, Benjamin P., I-hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. 2003. "Prediction of Mammalian MicroRNA Targets." *Cell* 115 (7): 787–798. http://www.sciencedirect.com/science/article/pii/S0092867403010183.

Li, Zhiguang, William S Branham, Stacey L Dial, Yexun Wang, Lei Guo, Leming Shi, and Tao Chen. 2010. "Genomic Analysis of microRNA Time-Course Expression in Liver of Mice Treated with Genotoxic Carcinogen N-Ethyl-N-Nitrosourea." *BMC Genomics* 11 (1): 609. doi:10.1186/1471-2164-11-609. http://www.biomedcentral.com/1471-2164/11/609.

Lin, Wan-Wan, and Michael Karin. 2007. "A Cytokine-Mediated Link Between Innate Immunity, Inflammation, and Cancer." *The Journal of Clinical Investigation* 117 (5) (May): 1175–83. doi:10.1172/JCI31537. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1857251&tool=pmcentrez&rendertype=abstract.

Liston, Adrian, Michelle Linterman, and Li-Fan Lu. 2010. "MicroRNA in the Adaptive Immune System, in Sickness and in Health." *Journal of Clinical Immunology* 30 (3) (May): 339–46. doi:10.1007/s10875-010-9378-5. http://www.ncbi.nlm.nih.gov/pubmed/20191314.

Liu, Gang, Arnaud Friggeri, Yanping Yang, Young-Jun Park, Yuko Tsuruta, and Edward Abraham. 2009. "miR-147, a microRNA That Is Induced Upon Toll-Like Receptor Stimulation, Regulates Murine Macrophage Inflammatory Responses." *Proceedings of the National Academy of Sciences of the United States of America* 106 (37) (September 15): 15819–24. doi:10.1073/pnas.0901216106. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2747202&tool=pmcentrez&rendertype=abstract.

Liu, Philip T, Matthew Wheelwright, Rosane Teles, Evangelia Komisopoulou, Kristina Edfeldt, Benjamin Ferguson, Manali D Mehta, et al. 2012. "MicroRNA-21 Targets the Vitamin D–dependent Antimicrobial Pathway in Leprosy." *Nature Medicine* 18 (2) (January): 267–273. doi:10.1038/nm.2584. http://www.nature.com/doifinder/10.1038/nm.2584.

Lloyd-Smith, James O, Dylan George, Kim M Pepin, Virginia E Pitzer, Juliet R C Pulliam, Andrew P Dobson, Peter J Hudson, and Bryan T Grenfell. 2009. “Epidemic Dynamics at the Human-Animal Interface.” *Science (New York, N.Y.)* 326 (5958) (December 4): 1362–7. doi:10.1126/science.1177345. http://www.sciencemag.org/content/326/5958/1362.abstract.

Loo, Yueh-Ming, and Michael Gale. 2011. “Immune Signaling by RIG-I-Like Receptors.” *Immunity* 34 (5) (May 27): 680–92. doi:10.1016/j.immuni.2011.05.003. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3177755&tool=pmcentrez&rendertype=abstract.

Lovegrove, Fiona E, Sina A Gharib, Samir N Patel, Cheryl A Hawkes, Kevin C Kain, and W Conrad Liles. 2007. “Expression Microarray Analysis Implicates Apoptosis and Interferon-Responsive Mechanisms in Susceptibility to Experimental Cerebral Malaria.” *The American Journal of Pathology* 171 (6) (December): 1894–903. doi:10.2353/ajpath.2007.070630. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2111112&tool=pmcentrez&rendertype=abstract.

Lu, Tzu-Pin, Chien-Yueh Lee, Mong-Hsun Tsai, Yu-Chiao Chiu, Chuhsing Kate Hsiao, Liang-Chuan Lai, and Eric Y Chuang. 2012. “miRSystem: An Integrated System for Characterizing Enriched Functions and Pathways of microRNA Targets.” *PloS One* 7 (8) (January): e42390. doi:10.1371/journal.pone.0042390. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3411648&tool=pmcentrez&rendertype=abstract.

Lynn, David J, Calvin Chan, Misbah Naseer, Melissa Yau, Raymond Lo, Anastasia Sribnaia, Giselle Ring, et al. 2010. “Curating the Innate Immunity Interactome.” *BMC Systems Biology* 4 (January): 117. doi:10.1186/1752-0509-4-117. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2936296&tool=pmcentrez&rendertype=abstract.

Lynn, David J, Geoffrey L Winsor, Calvin Chan, Nicolas Richard, Matthew R Laird, Aaron Barsky, Jennifer L Gardy, et al. 2008. “InnateDB: Facilitating Systems-Level Analyses of the Mammalian Innate Immune Response.” *Molecular Systems Biology* 4 (218) (January): 218. doi:10.1038/msb.2008.55. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2564732&tool=pmcentrez&rendertype=abstract.

Ma, Feng, Sheng Xu, Xingguang Liu, Qian Zhang, Xiongfei Xu, Mofang Liu, Minmin Hua, Nan Li, Hangping Yao, and Xuetao Cao. 2011. “The microRNA miR-29 Controls Innate and Adaptive Immune Responses to Intracellular Bacterial Infection by Targeting Interferon-Γ.” *Nature Immunology* 12 (9) (September): 861–9. doi:10.1038/ni.2073. http://dx.doi.org/10.1038/ni.2073.

Ma, Jun, Maureen a Sartor, and H.v. Jagadish. 2011. “Appearance Frequency Modulated Gene Set Enrichment Testing.” *BMC Bioinformatics* 12 (1): 81. doi:10.1186/1471-2105-12-81. http://www.biomedcentral.com/1471-2105/12/81.

Mackowiak, Sebastian D. 2011. “Identification of Novel and Known miRNAs in Deep-Sequencing Data with miRDeep2.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* Chapter 12 (December): Unit 12.10. doi:10.1002/0471250953.bi1210s36. http://www.ncbi.nlm.nih.gov/pubmed/22161567.

Maere, Steven, Karel Heymans, and Martin Kuiper. 2005. “BiNGO: a Cytoscape Plugin to Assess Overrepresentation of Gene Ontology Categories in Biological Networks.” *Bioinformatics (Oxford, England)* 21 (16) (August 15): 3448–9. doi:10.1093/bioinformatics/bti551. http://www.ncbi.nlm.nih.gov/pubmed/15972284.

Maertzdorf, J, D Repsilber, S K Parida, K Stanley, T Roberts, G Black, G Walzl, and S H E Kaufmann. 2011. “Human Gene Expression Profiles of Susceptibility and Resistance in Tuberculosis.” *Genes and Immunity* 12 (1) (January): 15–22. doi:10.1038/gene.2010.51. http://www.ncbi.nlm.nih.gov/pubmed/20861863.

Maglione, Paul J, Jiayong Xu, Arturo Casadevall, and John Chan. 2008. “Fc Gamma Receptors Regulate Immune Activation and Susceptibility During Mycobacterium Tuberculosis Infection.” *Journal of Immunology (Baltimore, Md.: 1950)* 180 (5) (March 1): 3329–38. http://www.ncbi.nlm.nih.gov/pubmed/18292558.

Marcenaro, Emanuela, Simona Carlomagno, Silvia Pesce, Alessandro Moretta, and Simona Sivori. 2012. “NK/DC Crosstalk in Anti-Viral Response.” *Advances in Experimental Medicine and Biology* 946 (January): 295–308. doi:10.1007/978-1-4614-0106-3_17. http://www.ncbi.nlm.nih.gov/pubmed/21948375.

Marcinowski, Lisa, Mélanie Tanguy, Astrid Krmpotic, Bernd Rädle, Vanda J Lisnić, Lee Tuddenham, Béatrice Chane-Woon-Ming, et al. 2012. “Degradation of Cellular miR-27 by a Novel, Highly Abundant Viral Transcript Is Important for Efficient Virus Replication In Vivo.” Edited by Bryan R Cullen. *PLoS Pathogens* 8 (2) (February): e1002510. doi:10.1371/journal.ppat.1002510. http://dx.plos.org/10.1371/journal.ppat.1002510.

Marcotte, E M, M Pellegrini, M J Thompson, T O Yeates, and D Eisenberg. 1999. “A Combined Algorithm for Genome-Wide Prediction of Protein Function.” *Nature* 402 (6757) (November 4): 83–6. doi:10.1038/47048. http://www.ncbi.nlm.nih.gov/pubmed/10573421.

Markowetz, Florian, and Rainer Spang. 2007. “Inferring Cellular Networks – a Review.” *BMC Bioinformatics* 17. doi:10.1186/1471-2105-8-S6-S5.

Marriott, Helen M, and David H Dockrell. 2007. “The Role of the Macrophage in Lung Disease Mediated by Bacteria.” *Experimental Lung Research* 33 (10) (December): 493–505. doi:10.1080/01902140701756562. http://www.ncbi.nlm.nih.gov/pubmed/18075824.

Marshak-Rothstein, Ann. 2006. “Toll-Like Receptors in Systemic Autoimmune Disease.” *Nature Reviews. Immunology* 6 (11) (November): 823–35. doi:10.1038/nri1957. http://dx.doi.org/10.1038/nri1957.

Mason, O., and M. Verwoerd. 2007. “Graph Theory and Networks in Biology.” *IET Systems Biology* 1 (2) (March 1): 89–119. doi:10.1049/iet-syb:20060038. http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=4140672.

Matthews, Lisa, Gopal Gopinath, Marc Gillespie, Michael Caudy, David Croft, Bernard de Bono, Phani Garapati, et al. 2009. “Reactome Knowledgebase of Human Biological Pathways and Processes.” *Nucleic Acids Research* 37 (Database issue) (January): D619–22. doi:10.1093/nar/gkn863. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686536&tool=pmcentrez&rendertype=abstract.

Mazumder, B, X Li, and S Barik. 2010. “Translation Control: a Multifaceted Regulator of Inflammatory Response.” *Journal of Immunology (Baltimore, Md.: 1950)* 184 (7) (April): 3311–3319.

McCartney, Stephen A, Larissa B Thackray, Leonid Gitlin, Susan Gilfillan, Herbert W Virgin, Herbert W Virgin Iv, and Marco Colonna. 2008. “MDA-5 Recognition of a Murine Norovirus.” *PLoS Pathogens* 4 (7) (July): e1000108. doi:10.1371/journal.ppat.1000108. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2443291&tool=pmcentrez&rendertype=abstract.

McGary, Kriston L, Tae Joo Park, John O Woods, Hye Ji Cha, John B Wallingford, and Edward M Marcotte. 2010. “Systematic Discovery of Nonobvious Human Disease Models through Orthologous Phenotypes.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (14) (April 6): 6544–9. doi:10.1073/pnas.0910200107. http://www.ncbi.nlm.nih.gov/pubmed/20308572.

McKenna, Maryn. 2012. “Food Poisoning’s Hidden Legacy.” *Scientific American* 306 (4) (April): 26–7. http://www.ncbi.nlm.nih.gov/pubmed/22486112.

Medzhitov, Ruslan. 2009. “Approaching the Asymptote: 20 Years Later.” *Immunity* 30 (6) (June 19): 766–75. doi:10.1016/j.immuni.2009.06.004. http://dx.doi.org/10.1016/j.immuni.2009.06.004.

Miller, Lance D, Johanna Smeds, Joshy George, Vinsensius B Vega, Liza Vergara, Alexander Ploner, Yudi Pawitan, et al. 2005. “An Expression Signature for P53 Status in Human Breast Cancer Predicts Mutation Status, Transcriptional Effects, and Patient Survival.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (38) (September 20): 13550–5. doi:10.1073/pnas.0506230102. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1197273&tool=pmcentrez&rendertype=abstract.

Mitchell, Daniel, and Colleen Olive. 2010. “Regulation of Toll-Like Receptor-Induced Chemokine Production in Murine Dendritic Cells by Mitogen-Activated Protein Kinases.” *Molecular Immunology* 47 (11-12) (July): 2065–73. doi:10.1016/j.molimm.2010.04.004. http://www.ncbi.nlm.nih.gov/pubmed/20451253.

Mittelbrunn, María, Cristina Gutiérrez-Vázquez, Carolina Villarroya-Beltri, Susana González, Fátima Sánchez-Cabo, Manuel Ángel González, Antonio Bernad, and Francisco Sánchez-Madrid. 2011. “Unidirectional Transfer of microRNA-Loaded Exosomes from T Cells to Antigen-Presenting Cells.” *Nature Communications* 2 (January): 282. doi:10.1038/ncomms1285. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3104548&tool=pmcentrez&rendertype=abstract.

Mocellin, Simone, Monica C Panelli, Ena Wang, Dirk Nagorsen, and Francesco M Marincola. 2003. “The Dual Role of IL-10.” *Trends in Immunology* 24 (1) (January): 36–43. http://www.ncbi.nlm.nih.gov/pubmed/12495723.

Moelants, Eva Av, Anneleen Mortier, Jo Van Damme, and Paul Proost. 2013. “In Vivo Regulation of Chemokine Activity by Post-Translational Modification.” *Immunology and Cell Biology* (April 30). doi:10.1038/icb.2013.16. http://www.ncbi.nlm.nih.gov/pubmed/23628804.

Mogensen, Trine H. 2009. “Pathogen Recognition and Inflammatory Signaling in Innate Immune Defenses.” *Clinical Microbiology Reviews* 22 (2) (April 1): 240–73, Table of Contents. doi:10.1128/CMR.00046-08. http://cmr.asm.org/content/22/2/240.full.

Mootha, Vamsi K, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, et al. 2003. “PGC-1alpha-Responsive Genes Involved in Oxidative Phosphorylation Are Coordinately Downregulated in Human Diabetes.” *Nature Genetics* 34 (3) (July): 267–73. doi:10.1038/ng1180. http://www.ncbi.nlm.nih.gov/pubmed/12808457.

Morens, David M, Gregory K Folkers, and Anthony S Fauci. 2004. “The Challenge of Emerging and Re-Emerging Infectious Diseases.” *Nature* 430 (6996) (July 8): 242–9. doi:10.1038/nature02759. http://www.ncbi.nlm.nih.gov/pubmed/15241422.

Morens, David M, Edward C Holmes, A Sally Davis, and Jeffery K Taubenberger. 2011. “Global Rinderpest Eradication: Lessons Learned and Why Humans Should Celebrate Too.” *The Journal of Infectious Diseases* 204 (4) (August 15): 502–5. doi:10.1093/infdis/jir327. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3144172&tool=pmcentrez&rendertype=abstract.

Moschos, Sterghios A, Andrew E Williams, Mark M Perry, Mark A Birrell, Maria G Belvisi, and Mark A Lindsay. 2007. “Expression Profiling in Vivo Demonstrates Rapid Changes in Lung microRNA Levels Following Lipopolysaccharide-Induced Inflammation but Not in the Anti-Inflammatory Action of Glucocorticoids.” *BMC Genomics* 8 (1): 240. doi:10.1186/1471-2164-8-240. http://www.biomedcentral.com/1471-2164/8/240.

Mostafavi, Sara, Debajyoti Ray, David Warde-Farley, Chris Grouios, and Quaid Morris. 2008. “GeneMANIA: a Real-Time Multiple Association Network Integration Algorithm for Predicting Gene Function.” *Genome Biology* 9 Suppl 1 (January): S4. doi:10.1186/gb-2008-9-s1-s4.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2447538&tool=pmcentrez&rendertype=abstract.

Moyes, Kasey M, James K Drackley, Dawn E Morin, Massimo Bionaz, Sandra L Rodriguez-Zas, Robin E Everts, Harris A Lewin, and Juan J Loor. 2009. “Gene Network and Pathway Analysis of Bovine Mammary Tissue Challenged with Streptococcus Uberis Reveals Induction of Cell Proliferation and Inhibition of {PPARγ} Signaling as Potential Mechanism for the Negative Relationships Between Immune Response and l.” *BMC Genomics* 10 (1): 542. doi:10.1186/1471-2164-10-542. http://www.biomedcentral.com/1471-2164/10/542.

Mukhopadhyay, Partha, Guy Brock, Vasyl Pihur, Cynthia Webb, M Michele Pisano, and Robert M Greene. 2010. “Developmental microRNA Expression Profiling of Murine Embryonic Orofacial Tissue.” *Birth Defects Research Part A: Clinical and Molecular Teratology* 88 (7) (June): 511–534. doi:10.1002/bdra.20684. http://doi.wiley.com/10.1002/bdra.20684.

Mukhopadhyay, Rupak, Partho Sarothi Ray, Abul Arif, Anna K Brady, Michael Kinter, and Paul L Fox. 2008. “DAPK-ZIPK-L13a Axis Constitutes a Negative-Feedback Module Regulating Inflammatory Gene Expression.” *Molecular Cell* 32 (3) (November 7): 371–82. doi:10.1016/j.molcel.2008.09.019. http://www.cell.com/molecular-cell/fulltext/S1097-2765(08)00699-0.

Mullokandov, Gavriel, Alessia Baccarini, Albert Ruzo, Anitha D Jayaprakash, Navpreet Tung, Benjamin Israelow, Matthew J Evans, Ravi Sachidanandam, and Brian D Brown. 2012. “High-Throughput Assessment of microRNA Activity and Function Using microRNA Sensor and Decoy Libraries.” *Nature Methods* 9 (8) (August): 840–6. doi:10.1038/nmeth.2078. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3518396&tool=pmcentrez&rendertype=abstract.

Naeem, A, K Zhong, S J Moisá, J K Drackley, K M Moyes, and J J Loor. 2012. “Bioinformatics Analysis of microRNA and Putative Target Genes in Bovine Mammary Tissue Infected with Streptococcus Uberis.” *Journal of Dairy Science* 95 (11) (November): 6397–408. doi:10.3168/jds.2011-5173. http://www.ncbi.nlm.nih.gov/pubmed/22959936.

Nahid, Md A, Bing Yao, Paul R Dominguez-Gutierrez, Lakshmyya Kesavalu, Minoru Satoh, and Edward K L Chan. 2013. “Regulation of TLR2-Mediated Tolerance and Cross-Tolerance through IRAK4 Modulation by miR-132 and miR-212.” *Journal of Immunology (Baltimore, Md.: 1950)* 190 (3) (February 1): 1250–63. doi:10.4049/jimmunol.1103060. http://www.jimmunol.org/content/190/3/1250.full.

Neilsen, Corine T, Gregory J Goodall, and Cameron P Bracken. 2012. “IsomiRs – the Overlooked Repertoire in the Dynamic microRNAome.” *Trends in Genetics* 28 (11) (November): 544–549. doi:10.1016/j.tig.2012.07.005. http://linkinghub.elsevier.com/retrieve/pii/S0168952512001126.

Newman, M E. 2001. “The Structure of Scientific Collaboration Networks.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (2)

(January 16): 404–9. doi:10.1073/pnas.021544898. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=14598&tool=pmcentrez&rendertype=abstract.

Newman, Martin A, Vidya Mani, and Scott M Hammond. 2011. “Deep Sequencing of microRNA Precursors Reveals Extensive 3’ End Modification.” *RNA (New York, N.Y.)* 17 (10) (October): 1795–803. doi:10.1261/rna.2713611. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3185913&tool=pmcentrez&rendertype=abstract.

Newton, Kim, and Vishva M Dixit. 2012. “Signaling in Innate Immunity and Inflammation.” *Cold Spring Harbor Perspectives in Biology* 4 (3) (March 1). doi:10.1101/cshperspect.a006049. http://cshperspectives.cshlp.org/content/4/3/a006049.full.

Nicolae, Marius, and Ion Măndoiu. 2011. “Accurate Estimation of Gene Expression Levels from DGE Sequencing Data” (May 27): 392–403. http://dl.acm.org/citation.cfm?id=2009164.2009201.

O’Connell, Ryan M, Dinesh S Rao, Aadel A Chaudhuri, and David Baltimore. 2010. “Physiological and Pathological Roles for microRNAs in the Immune System.” *Nature Reviews. Immunology* 10 (2) (March): 111–22. doi:10.1038/nri2708. http://www.ncbi.nlm.nih.gov/pubmed/20098459.

O’Connell, Ryan M, Konstantin D Taganov, Mark P Boldin, Genhong Cheng, and David Baltimore. 2007. “MicroRNA-155 Is Induced During the Macrophage Inflammatory Response.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (5) (January 30): 1604–9. doi:10.1073/pnas.0610731104. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1780072&tool=pmcentrez&rendertype=abstract.

O’Neill, L.A., F.J. Sheedy, and C.E. McCoy. 2011. “MicroRNAs: The Fine-Tuners of Toll-Like Receptor Signalling.” *Nature Reviews Immunology* 11 (3): 163–175. doi:10.1038/nri2957. http://www.nature.com/nri/journal/v11/n3/abs/nri2957.html.

O’Neill, Luke A J, and D Grahame Hardie. 2013. “Metabolism of Inflammation Limited by AMPK and Pseudo-Starvation.” *Nature* 493 (7432) (January 17): 346–55. doi:10.1038/nature11862. http://dx.doi.org/10.1038/nature11862.

Obayashi, Takeshi, and Kengo Kinoshita. 2011. “COXPRESdb: a Database to Compare Gene Coexpression in Seven Model Animals.” *Nucleic Acids Research* 39 (Database issue) (January): D1016–22. doi:10.1093/nar/gkq1147. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013720&tool=pmcentrez&rendertype=abstract.

Oghabian, Ali, Sami Kilpinen, Sampsa Hautaniemi, and Elena Czeizler. 2014. “Biclustering Methods: Biological Relevance and Application in Gene Expression Analysis.” *PloS One* 9 (3) (January): e90801. doi:10.1371/journal.pone.0090801. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3961251&tool=pmcentrez&rendertype=abstract.

Oh, Ji Eun, and Heung Kyu Lee. 2013. “Autophagy as an Innate Immune Modulator.” *Immune Network* 13 (1) (March): 1–9. doi:10.4110/in.2013.13.1.1. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3607704&tool=pmcentrez&rendertype=abstract.

Orchard, Sandra, Samuel Kerrien, Sara Abbani, Bruno Aranda, Jignesh Bhate, Shelby Bidwell, Alan Bridge, et al. 2012. “Protein Interaction Data Curation: The International Molecular Exchange (IMEx) Consortium.” *Nature Methods* 9 (4) (April): 345–50. doi:10.1038/nmeth.1931. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3703241&tool=pmcentrez&rendertype=abstract.

Orchard, Sandra, Samuel Kerrien, Philip Jones, Arnaud Ceol, Andrew Chatr-Aryamontri, Lukasz Salwinski, Jason Nerothin, and Henning Hermjakob. 2007. “Submit Your Interaction Data the IMEx Way: a Step by Step Guide to Trouble-Free Deposition.” *Proteomics* 7 Suppl 1 (September): 28–34. doi:10.1002/pmic.200700286. http://www.ncbi.nlm.nih.gov/pubmed/17893861.

Orchard, Sandra, Lukasz Salwinski, Samuel Kerrien, Luisa Montecchi-Palazzi, Matthias Oesterheld, Volker Stümpflen, Arnaud Ceol, et al. 2007. “The Minimum Information Required for Reporting a Molecular Interaction Experiment (MIMIx).” *Nature Biotechnology* 25 (8) (August): 894–8. doi:10.1038/nbt1324. http://www.ncbi.nlm.nih.gov/pubmed/17687370.

Ozinsky, A, D M Underhill, J D Fontenot, A M Hajjar, K D Smith, C B Wilson, L Schroeder, and A Aderem. 2000. “The Repertoire for Pattern Recognition of Pathogens by the Innate Immune System Is Defined by Cooperation Between Toll-Like Receptors.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (25) (December 5): 13766–71. doi:10.1073/pnas.250476497. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=17650&tool=pmcentrez&rendertype=abstract.

Padrón, Benigno, Manuel Nogales, and Anna Traveset. 2011. “Alternative Approaches of Transforming Bimodal into Unimodal Mutualistic Networks. The Usefulness of Preserving Weighted Information.” *Basic and Applied Ecology* 12 (8) (December): 713–721. doi:10.1016/j.baae.2011.09.004. http://dx.doi.org/10.1016/j.baae.2011.09.004.

Paroo, Zain, Xuecheng Ye, She Chen, and Qinghua Liu. 2009. “Phosphorylation of the Human microRNA-Generating Complex Mediates MAPK/Erk Signaling.” *Cell* 139 (1) (October 2): 112–22. doi:10.1016/j.cell.2009.06.044. http://www.sciencedirect.com/science/article/pii/S0092867409007910.

Parrish, Colin R, Edward C Holmes, David M Morens, Eun-Chung Park, Donald S Burke, Charles H Calisher, Catherine A Laughlin, Linda J Saif, and Peter Daszak. 2008. “Cross-Species Virus Transmission and the Emergence of New Epidemic Diseases.” *Microbiology and Molecular Biology Reviews : MMBR* 72 (3) (September 1): 457–70. doi:10.1128/MMBR.00004-08. http://mmbr.asm.org/content/72/3/457.

Peña-Castillo, Lourdes, Murat Tasan, Chad L Myers, Hyunju Lee, Trupti Joshi, Chao Zhang, Yuanfang Guan, et al. 2008. “A Critical Assessment of Mus Musculus Gene Function Prediction Using Integrated Genomic Evidence.” *Genome Biology* 9 Suppl 1 (Suppl 1) (January): S2. doi:10.1186/gb-2008-9-s1-s2. http://genomebiology.com/2008/9/S1/S2.

Peng, Zhiyu, Yanbing Cheng, Bertrand Chin-Ming Tan, Lin Kang, Zhijian Tian, Yuankun Zhu, Wenwei Zhang, et al. 2012. “Comprehensive Analysis of RNA-Seq Data Reveals Extensive RNA Editing in a Human Transcriptome.” *Nature Biotechnology* 30 (3) (February): 253–260. doi:10.1038/nbt.2122. http://www.nature.com/doifinder/10.1038/nbt.2122.

Peters-Golden, Marc. 2004. “The Alveolar Macrophage: The Forgotten Cell in Asthma.” *American Journal of Respiratory Cell and Molecular Biology* 31 (1) (July): 3–7. doi:10.1165/rcmb.f279. http://www.ncbi.nlm.nih.gov/pubmed/15208096.

Pfeffer, Sébastien, Mihaela Zavolan, Friedrich A Grässer, Minchen Chien, James J Russo, Jingyue Ju, Bino John, et al. 2004. “Identification of Virus-Encoded microRNAs.” *Science (New York, N.Y.)* 304 (5671) (April 30): 734–6. doi:10.1126/science.1096781. http://www.ncbi.nlm.nih.gov/pubmed/15118162.

Pichlmair, Andreas, Kumaran Kandasamy, Gualtiero Alvisi, Orla Mulhern, Roberto Sacco, Matthias Habjan, Marco Binder, et al. 2012. “Viral Immune Modulators Perturb the Human Molecular Network by Common and Unique Strategies.” *Nature* 487 (7408) (July 26): 486–90. doi:10.1038/nature11289. http://www.ncbi.nlm.nih.gov/pubmed/22810585.

Pollock, J M, J D Rodgers, M D Welsh, and J McNair. 2006. “Pathogenesis of Bovine Tuberculosis: The Role of Experimental Models of Infection.” *Veterinary Microbiology* 112 (2-4) (February 25): 141–50. doi:10.1016/j.vetmic.2005.11.032. http://www.ncbi.nlm.nih.gov/pubmed/16384665.

Pothlichet, Julien, and Lluis Quintana-Murci. 2013. “The Genetics of Innate Immunity Sensors and Human Disease.” *International Reviews of Immunology* 32 (2) (April): 157–208. doi:10.3109/08830185.2013.777064. http://www.ncbi.nlm.nih.gov/pubmed/23570315.

Pulliam, Juliet R C, and Jonathan Dushoff. 2009. “Ability to Replicate in the Cytoplasm Predicts Zoonotic Transmission of Livestock Viruses.” *The Journal of Infectious Diseases* 199 (4) (February 15): 565–8. doi:10.1086/596510. http://jid.oxfordjournals.org/content/199/4/565.abstract?ijkey=d28ac9589e1452c2b0c21f24da840e3e02708119&keytype2=tf_ipsecsha.

Qi, Jianni, Yu Qiao, Peng Wang, Shuqing Li, Wei Zhao, and Chengjiang Gao. 2012. “microRNA-210 Negatively Regulates LPS-Induced Production of Proinflammatory Cytokines by Targeting NF-κB1 in Murine Macrophages.” *FEBS Letters* 586 (8) (April): 1201–1207. doi:10.1016/j.febslet.2012.03.011. http://linkinghub.elsevier.com/retrieve/pii/S0014579312001986.

Qi, Lei S, Matthew H Larson, Luke A Gilbert, Jennifer A Doudna, Jonathan S Weissman, Adam P Arkin, and Wendell A Lim. 2013. “Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression.” *Cell* 152 (5) (February 28): 1173–83. doi:10.1016/j.cell.2013.02.022. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3664290&tool=pmcentrez&rendertype=abstract.

Quinn, Susan R, and Luke A O’Neill. 2011. “A Trio of microRNAs That Control Toll-Like Receptor Signalling.” *International Immunology* 23 (7) (July 1): 421–5. doi:10.1093/intimm/dxr034. http://intimm.oxfordjournals.org/content/23/7/421.full.

R Development Core Team. 2008. “R: A Language and Environment for Statistical Computing”. Vienna, Austria. http://www.r-project.org.

Rajaram, Murugesan V S, Bin Ni, Jessica D Morris, Michelle N Brooks, Tracy K Carlson, Baskar Bakthavachalu, Daniel R Schoenberg, Jordi B Torrelles, and Larry S Schlesinger. 2011. “Mycobacterium Tuberculosis Lipomannan Blocks TNF Biosynthesis by Regulating Macrophage MAPK-Activated Protein Kinase 2 (MK2) and microRNA miR-125b.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (42) (October 18): 17408–13. doi:10.1073/pnas.1112660108. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3198317&tool=pmcentrez&rendertype=abstract.

Rajewsky, Nikolaus. 2006. “microRNA Target Predictions in Animals.” *Nature Genetics* 38 Suppl (June): S8–13. doi:10.1038/ng1798. http://www.ncbi.nlm.nih.gov/pubmed/16736023.

Ramos, Theresa N, Meghan M Darley, Sebastian Weckbach, Philip F Stahel, Stephen Tomlinson, and Scott R Barnum. 2012. “The C5 Convertase Is Not Required for Activation of the Terminal Complement Pathway in Murine Experimental Cerebral Malaria.” *The Journal of Biological Chemistry* 287 (29) (July 13): 24734–8. doi:10.1074/jbc.C112.378364. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3397900&tool=pmcentrez&rendertype=abstract.

Ramsey, S A, E S Gold, and A Aderem. 2010. “A Systems Biology Approach to Understanding Atherosclerosis.” *EMBO Molecular Medicine* 2 (3) (March): 79–89.

Ramsey, Stephen A, Sandy L Klemm, Daniel E Zak, Kathleen A Kennedy, Vesteinn Thorsson, Bin Li, Mark Gilchrist, et al. 2008. “Uncovering a Macrophage Transcriptional Program by Integrating Evidence from Motif Scanning and Expression Dynamics.” *PLoS Computational Biology* 4 (3). doi:10.1371/journal.pcbi.1000021.

Rapin, Nicolas, Ole Lund, Massimo Bernaschi, and Filippo Castiglione. 2010. “Computational Immunology Meets Bioinformatics : The Use of Prediction Tools for Molecular Binding in the Simulation of the Immune System.” *Bioinformatics* 5 (4). doi:10.1371/journal.pone.0009862.

Reimand, Jüri, Meelis Kull, Hedi Peterson, Jaanus Hansen, and Jaak Vilo. 2007. “g:Profiler--a Web-Based Toolset for Functional Profiling of Gene Lists from Large-Scale Experiments.” *Nucleic Acids Research* 35 (Web Server issue) (July): W193–200. doi:10.1093/nar/gkm226. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1933153&tool=pmcentrez&rendertype=abstract.

Reinsbach, Susanne, Petr V Nazarov, Demetra Philippidou, Martina Schmitt, Anke Wienecke-Baldacchino, Arnaud Muller, Laurent Vallar, Iris Behrmann, and Stephanie Kreis. 2012. “Dynamic Regulation of microRNA Expression Following Interferon-γ-Induced Gene Transcription.” *RNA Biology* 9 (7) (July): 978–989. doi:10.4161/rna.20494. http://www.landesbioscience.com/journals/rnabiology/article/20494/.

Reynolds, Craig W. 1987. “Flocks, Herds and Schools: A Distributed Behavioral Model.” *ACM SIGGRAPH Computer Graphics* 21 (4) (August 1): 25–34. doi:10.1145/37402.37406. http://dl.acm.org/citation.cfm?id=37402.37406.

Rhee, Seung Yon, Valerie Wood, Kara Dolinski, and Sorin Draghici. 2008. “Use and Misuse of the Gene Ontology Annotations.” *Nature Reviews. Genetics* 9 (7) (July): 509–15. doi:10.1038/nrg2363. http://www.ncbi.nlm.nih.gov/pubmed/18475267.

Rhrissorrakrai, Kahn, and Kristin C Gunsalus. 2011. “MINE: Module Identification in Networks.” *BMC Bioinformatics* 12 (1) (January): 192. doi:10.1186/1471-2105-12-192. http://www.biomedcentral.com/1471-2105/12/192.

Robinson, M D, D J McCarthy, and G K Smyth. 2009. “edgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1) (November): 139–140. doi:10.1093/bioinformatics/btp616. http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp616.

Rossato, Marzia, Graziella Curtale, Nicola Tamassia, Monica Castellucci, Laura Mori, Sara Gasperini, Barbara Mariotti, et al. 2012. “IL-10-Induced microRNA-187 Negatively Regulates TNF-A, IL-6, and IL-12p40 Production in TLR4-Stimulated Monocytes.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (45) (November 6): E3101–10. doi:10.1073/pnas.1209100109. http://www.pnas.org/content/109/45/E3101.full.

Royer, Loïc, Matthias Reimann, Bill Andreopoulos, and Michael Schroeder. 2008. “Unraveling Protein Networks with Power Graph Analysis.” *PLoS Computational Biology* 4 (7) (January): e1000108. doi:10.1371/journal.pcbi.1000108. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2424176&tool=pmcentrez&rendertype=abstract.

Sargeant, J M, H M Scott, K E Leslie, M J Ireland, and A Bashiri. 1998. “Clinical Mastitis in Dairy Cattle in Ontario: Frequency of Occurrence and Bacteriological Isolates.” *The Canadian Veterinary Journal. La Revue Vétérinaire Canadienne* 39 (1) (January): 33–8. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1539829&tool=pmcentrez&rendertype=abstract.

Satoh, Mamoru, Tsuyoshi Tabuchi, Yoshitaka Minami, Yuji Takahashi, Tomonori Itoh, and Motoyuki Nakamura. 2012. “Expression of Let-7i Is Associated with Toll-Like Receptor 4 Signal in Coronary Artery Disease: Effect of Statins on Let-7i and Toll-Like Receptor 4 Signal.” *Immunobiology* 217 (5) (May): 533–9. doi:10.1016/j.imbio.2011.08.005. http://www.ncbi.nlm.nih.gov/pubmed/21899916.

Scallan, Elaine, Patricia M Griffin, Frederick J Angulo, Robert V Tauxe, and Robert M Hoekstra. 2011. “Foodborne Illness Acquired in the United States--Unspecified Agents.” *Emerging Infectious Diseases* 17 (1) (January): 16–22. doi:10.3201/eid1701.091101p2. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3204615&tool=pmcentrez&rendertype=abstract.

Schaefer, Carl F, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. 2009. “PID: The Pathway Interaction Database.” *Nucleic Acids Research* 37 (Database issue) (January): D674–9. doi:10.1093/nar/gkn653. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686461&tool=pmcentrez&rendertype=abstract.

Schaefer, Martin H, Jae-Seong Yang, Luis Serrano, and Christina Kiel. 2014. “Protein Conservation and Variation Suggest Mechanisms of Cell Type-Specific Modulation of Signaling Pathways.” Edited by Anand R. Asthagiri. *PLoS Computational Biology* 10 (6) (June): e1003659. doi:10.1371/journal.pcbi.1003659. http://dx.plos.org/10.1371/journal.pcbi.1003659.

Schulte, Leon N, Ana Eulalio, Hans-Joachim Mollenkopf, Richard Reinhardt, and Jörg Vogel. 2011. “Analysis of the Host microRNA Response to Salmonella Uncovers the Control of Major Cytokines by the Let-7 Family: MicroRNA and Bacterial Infection.” *The EMBO Journal* 30 (10) (May): 1977–1989. doi:10.1038/emboj.2011.94. http://emboj.embopress.org/cgi/doi/10.1038/emboj.2011.94.

Schulte, Leon N, Alexander J Westermann, and Jörg Vogel. 2013. “Differential Activation and Functional Specialization of miR-146 and miR-155 in Innate Immune Sensing.” *Nucleic Acids Research* 41 (1) (January 7): 542–53. doi:10.1093/nar/gks1030. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3592429&tool=pmcentrez&rendertype=abstract.

Schwartz, E A, W Y Zhang, S K Karnik, S Borwege, V R Anand, P S Laine, Y Su, and P D Reaven. 2010. “Nutrient Modification of the Innate Immune Response: a Novel Mechanism by Which Saturated Fatty Acids Greatly Amplify Monocyte Inflammation.” *Arteriosclerosis, Thrombosis, and Vascular Biology* 30 (4) (April): 802–808.

Schwarz, Dianne S, György Hutvágner, Tingting Du, Zuoshang Xu, Neil Aronin, and Phillip D Zamore. 2003. “Asymmetry in the Assembly of the RNAi Enzyme Complex.” *Cell* 115 (2) (October 17): 199–208. http://www.ncbi.nlm.nih.gov/pubmed/14567917.

Senaldi, G, C Vesin, R Chang, G E Grau, and P F Piguet. 1994. “Role of Polymorphonuclear Neutrophil Leukocytes and Their Integrin CD11a (LFA-1) in the Pathogenesis of Severe Murine Malaria.” *Infection and Immunity* 62 (4) (April): 1144–9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=186241&tool=pmcentrez&rendertype=abstract.

Serghides, Lena. 2012. “The Case for the Use of PPARγ Agonists as an Adjunctive Therapy for Cerebral Malaria.” *PPAR Research* 2012 (January): 513865. doi:10.1155/2012/513865. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3135089&tool=pmcentrez&rendertype=abstract.

Serghides, Lena, Samir N Patel, Kodjo Ayi, Ziyue Lu, D Channe Gowda, W Conrad Liles, and Kevin C Kain. 2009. “Rosiglitazone Modulates the Innate Immune Response to Plasmodium Falciparum Infection and Improves Outcome in Experimental Cerebral Malaria.” *The Journal of Infectious Diseases* 199 (10) (May 15): 1536–45. doi:10.1086/598222. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2854576&tool=pmcentrez&rendertype=abstract.

Sethupathy, Praveen, Molly Megraw, and Artemis G Hatzigeorgiou. 2006. “A Guide through Present Computational Approaches for the Identification of Mammalian microRNA Targets.” *Nature Methods* 3 (11) (November): 881–6. doi:10.1038/nmeth954. http://www.ncbi.nlm.nih.gov/pubmed/17060911.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Research* 13 (11) (November): 2498–504. doi:10.1101/gr.1239303. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=403769&tool=pmcentrez&rendertype=abstract.

Sharan, Roded, Igor Ulitsky, and Ron Shamir. 2007. “Network-Based Prediction of Protein Function.” *Molecular Systems Biology* 3 (January 13): 88. doi:10.1038/msb4100129. http://dx.doi.org/10.1038/msb4100129.

Shi, F D, H G Ljunggren, and N Sarvetnick. 2001. “Innate Immunity and Autoimmunity: From Self-Protection to Self-Destruction.” *Trends in Immunology* 22 (2) (February): 97–101. http://www.ncbi.nlm.nih.gov/pubmed/11286711.

Shinde, Kaustubh, Mukta Phatak, Freudenberg M Johannes, Jing Chen, Qian Li, Joshi K Vineet, Zhen Hu, Krishnendu Ghosh, Jaroslaw Meller, and Mario Medvedovic. 2010. “Genomics Portals: Integrative Web-Platform for Mining Genomics Data.” *BMC Genomics* 11 (1) (January): 27. doi:10.1186/1471-2164-11-27. http://www.biomedcentral.com/1471-2164/11/27.

Shresta, S. 2012. “Role of Complement in Dengue Virus Infection: Protection or Pathogenesis?” *mBio* 3 (1) (February 7): e00003–12–e00003–12. doi:10.1128/mBio.00003-12. /pmc/articles/PMC3280461/?report=abstract.

Singh, Ram Pyare, Israel Massachi, Sudhir Manickavel, Satendra Singh, Nagesh P Rao, Sascha Hasan, Deborah K Mc Curdy, et al. 2013. “The Role of miRNA in Inflammation and Autoimmunity.” *Autoimmunity Reviews* 12 (12) (October): 1160–5. doi:10.1016/j.autrev.2013.07.003. http://www.ncbi.nlm.nih.gov/pubmed/23860189.

Smith, K M, M Guerau-de-Arellano, S Costinean, J L Williams, A Bottoni, G Mavrikis Cox, A R Satoskar, et al. 2012. “miR-29ab1 Deficiency Identifies a Negative Feedback Loop Controlling Th1 Bias That Is Dysregulated in Multiple Sclerosis.” *The Journal of Immunology* 189 (4) (July): 1567–1576. doi:10.4049/jimmunol.1103171. http://www.jimmunol.org/cgi/doi/10.4049/jimmunol.1103171.

Smith, Kelly D, and Hamid Bolouri. 2005. “Dissecting Innate Immune Responses with the Tools of Systems Biology.” *Current Opinion in Immunology* 17 (1): 49–54. http://www.sciencedirect.com/science/article/pii/S0952791504001876.

Smoot, Michael E, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. 2011. “Cytoscape 2.8: New Features for Data Integration and Network Visualization.” *Bioinformatics (Oxford, England)* 27 (3) (February 1): 431–2. doi:10.1093/bioinformatics/btq675. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3031041&tool=pmcentrez&rendertype=abstract.

Sorathiya, A, A Bracciali, and P Lio. 2010. “Formal Reasoning on Qualitative Models of Coinfection of HIV and Tuberculosis and HAART Therapy.” *BMC Bioinformatics* 11 Suppl 1 (January): S67.

Sotiriou, Christos, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren, et al. 2006. “Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade to Improve Prognosis.” *Journal of the National Cancer Institute* 98 (4) (February 15): 262–72. doi:10.1093/jnci/djj052. http://www.ncbi.nlm.nih.gov/pubmed/16478745.

Staedel, Cathy, and Fabien Darfeuille. 2013. “MicroRNAs and Bacterial Infection.” *Cellular Microbiology* 15 (9) (September): 1496–507. doi:10.1111/cmi.12159. http://www.ncbi.nlm.nih.gov/pubmed/23795564.

Stark, Alexander, Michael F Lin, Pouya Kheradpour, Jakob S Pedersen, Leopold Parts, Joseph W Carlson, Madeline A Crosby, et al. 2007. “Discovery of Functional Elements in 12 Drosophila Genomes Using Evolutionary Signatures.” *Nature* 450 (7167) (November 8): 219–32. doi:10.1038/nature06340. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2474711&tool=pmcentrez&rendertype=abstract.

Stelzl, Ulrich, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H Brembeck, Heike Goehler, Martin Stroedicke, et al. 2005. “A Human Protein-Protein Interaction Network : A Resource for Annotating the Proteome Max-Planck-Institute for Molecular Genetics” 122: 957–968. doi:10.1016/j.cell.2005.08.029.

Sturm, Martin, Michael Hackenberg, David Langenberger, and Dmitrij Frishman. 2010. “TargetSpy : a Supervised Machine Learning Approach for microRNA Target Prediction.” *BMC Bioinformatics*.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–15550. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1239896&tool=pmcentrez&rendertype=abstract.

Sutherst, Robert W. 2004. “Global Change and Human Vulnerability to Vector-Borne Diseases.” *Clinical Microbiology Reviews* 17 (1) (January): 136–73. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=321469&tool=pmcentrez&rendertype=abstract.

Swamy, Mahima, Colin Jamora, Wendy Havran, and Adrian Hayday. 2010. “Epithelial Decision Makers: In Search of the ‘Epimmunome’.” *Nature Immunology* 11 (8) (July): 656–665. doi:10.1038/ni.1905. http://www.nature.com/doifinder/10.1038/ni.1905.

Tamayo, Pablo, George Steinhardt, Arthur Liberzon, and Jill P Mesirov. “Gene Set Enrichment Analysis Made Right.” *Pancreas*.

Tarca, Adi Laurentiu, Sorin Draghici, Gaurav Bhatti, and Roberto Romero. 2012. “Down-Weighting Overlapping Genes Improves Gene Set Analysis.” *BMC Bioinformatics* 13 (1) (January): 136. doi:10.1186/1471-2105-13-136. http://www.biomedcentral.com/1471-2105/13/136.

Tegnér, Jesper, Roland Nilsson, Vladimir B Bajic, Johan Björkegren, and Timothy Ravasi. 2006. “Systems Biology of Innate Immunity.” *Cellular Immunology* 244 (2) (December): 105–9. doi:10.1016/j.cellimm.2007.01.010. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1947944&tool=pmcentrez&rendertype=abstract.

Theocharidis, Athanasios, Stjin van Dongen, Anton J Enright, and Tom C Freeman. 2009. “Network Visualization and Analysis of Gene Expression Data Using BioLayout Express(3D).” *Nature Protocols* 4 (10) (January): 1535–50. doi:10.1038/nprot.2009.177. http://www.ncbi.nlm.nih.gov/pubmed/19798086.

Thomson, Daniel W, Cameron P Bracken, and Gregory J Goodall. 2011. “Experimental Strategies for microRNA Target Identification.” *Nucleic Acids Research* 39 (16) (September 1): 6845–53. doi:10.1093/nar/gkr330. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3167600&tool=pmcentrez&rendertype=abstract.

Thuong, Nguyen Thuy Thuong, Sarah J Dunstan, Tran Thi Hong Chau, Vesteinn Thorsson, Cameron P Simmons, Nguyen Than Ha Quyen, Guy E Thwaites, et al. 2008. “Identification of Tuberculosis Susceptibility Genes with Human Macrophage

Gene Expression Profiles." Edited by Jeffery S. Cox. *PLoS Pathogens* 4 (12) (December): e1000229. doi:10.1371/journal.ppat.1000229. http://dx.plos.org/10.1371/journal.ppat.1000229.

Tomley, Fiona M, and Martin W Shirley. 2009. "Livestock Infectious Diseases and Zoonoses." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 364 (1530) (September 27): 2637–42. doi:10.1098/rstb.2009.0133. http://rstb.royalsocietypublishing.org/content/364/1530/2637.full.

Tripathi, Shailesh, Galina V Glazko, and Frank Emmert-Streib. 2013. "Ensuring the Statistical Soundness of Competitive Gene Set Approaches: Gene Filtering and Genome-Scale Coverage Are Essential." *Nucleic Acids Research* (February 6). doi:10.1093/nar/gkt054. http://www.ncbi.nlm.nih.gov/pubmed/23389952.

Usadel, Björn, Takeshi Obayashi, Marek Mutwil, Federico M Giorgi, George W Bassel, Mimi Tanimoto, Amanda Chow, Dirk Steinhauser, Staffan Persson, and Nicholas J Provart. 2009. "Co-Expression Tools for Plant Biology: Opportunities for Hypothesis Generation and Caveats." *Plant, Cell & Environment* 32 (12) (December): 1633–51. doi:10.1111/j.1365-3040.2009.02040.x. http://www.ncbi.nlm.nih.gov/pubmed/19712066.

Valadi, Hadi, Karin Ekström, Apostolos Bossios, Margareta Sjöstrand, James J Lee, and Jan O Lötvall. 2007. "Exosome-Mediated Transfer of mRNAs and microRNAs Is a Novel Mechanism of Genetic Exchange Between Cells." *Nature Cell Biology* 9 (6) (June): 654–9. doi:10.1038/ncb1596. http://www.ncbi.nlm.nih.gov/pubmed/17486113.

Valcour, James E, Pascal Michel, Scott A McEwen, and Jeffrey B Wilson. 2002. "Associations Between Indicators of Livestock Farming Intensity and Incidence of Human Shiga Toxin-Producing Escherichia Coli Infection." *Emerging Infectious Diseases* 8 (3) (March): 252–7. doi:10.3201/eid0803.010159. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2732476&tool=pmcentrez&rendertype=abstract.

Van der Heyde, Henri C, John Nolan, Valéry Combes, Irene Gramaglia, and Georges E Grau. 2006. "A Unified Hypothesis for the Genesis of Cerebral Malaria: Sequestration, Inflammation and Hemostasis Leading to Microcirculatory Dysfunction." *Trends in Parasitology* 22 (11) (November): 503–8. doi:10.1016/j.pt.2006.09.002. http://www.ncbi.nlm.nih.gov/pubmed/16979941.

Vasudevan, Shobha, Yingchun Tong, and Joan A Steitz. 2007. "Switching from Repression to Activation: microRNAs Can up-Regulate Translation." *Science (New York, N.Y.)* 318 (5858) (December 21): 1931–4. doi:10.1126/science.1149460. http://www.sciencemag.org/content/318/5858/1931.abstract.

Vegh, Peter, Amir B K Foroushani, David A Magee, Matthew S McCabe, John A Browne, Nicolas C Nalpas, Kevin M Conlon, et al. 2013. "Profiling microRNA Expression in Bovine Alveolar Macrophages Using RNA-Seq." *Veterinary*

Immunology and Immunopathology 155 (4) (October 1): 238–44. doi:10.1016/j.vetimm.2013.08.004. http://www.ncbi.nlm.nih.gov/pubmed/24021155.

Vergoulis, Thanasis, Ioannis S Vlachos, Panagiotis Alexiou, George Georgakilas, Manolis Maragkakis, Martin Reczko, Stefanos Gerangelos, Nectarios Koziris, Theodore Dalamagas, and Artemis G Hatzigeorgiou. 2012. "TarBase 6.0: Capturing the Exponential Growth of miRNA Targets with Experimental Support." *Nucleic Acids Research* 40 (Database issue) (January): D222–9. doi:10.1093/nar/gkr1161. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245116&tool=pmcentrez&rendertype=abstract.

Vidal, Marc, Michael E Cusick, and Albert-László Barabási. 2011. "Interactome Networks and Human Disease." *Cell* 144 (6) (March 18): 986–98. doi:10.1016/j.cell.2011.02.016. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3102045&tool=pmcentrez&rendertype=abstract.

Wain-Hobson, S, and A Meyerhans. 1999. "On Viral Epidemics, Zoonoses and Memory." *Trends in Microbiology* 7 (10) (October): 389–91. http://www.ncbi.nlm.nih.gov/pubmed/10498942.

Walker, Angela M, and R Michael Roberts. 2009. "Characterization of the Bovine Type I IFN Locus: Rearrangements, Expansions, and Novel Subfamilies." *BMC Genomics* 10 (January): 187. doi:10.1186/1471-2164-10-187. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2680415&tool=pmcentrez&rendertype=abstract.

Wang, F.-Z., F Weber, C Croce, C.-G. Liu, X Liao, and P E Pellett. 2008. "Human Cytomegalovirus Infection Alters the Expression of Cellular {MicroRNA} Species That Affect Its Replication." *Journal of Virology* 82 (18) (July): 9065–9074. doi:10.1128/JVI.00961-08. http://jvi.asm.org/cgi/doi/10.1128/JVI.00961-08.

Wang, Fei, Ping Li, Arnd Christian König, and Muting Wan. 2011. "Improving Clustering by Learning a Bi-Stochastic Data Similarity Matrix." *Knowledge and Information Systems* 32 (2) (July 14): 351–382. doi:10.1007/s10115-011-0433-1. http://link.springer.com/10.1007/s10115-011-0433-1.

Wang, Hui-Juan, Jie Deng, Jing-Yang Wang, Peng-Jun Zhang, Zhang Xin, Kun Xiao, Dan Feng, Yan-Hong Jia, You-Ning Liu, and Li-Xin Xie. 2014. "Serum miR-122 Levels Are Related to Coagulation Disorders in Sepsis Patients." *Clinical Chemistry and Laboratory Medicine : CCLM / FESCC* 52 (6) (June): 927–33. doi:10.1515/cclm-2013-0899. http://www.ncbi.nlm.nih.gov/pubmed/24421215.

Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. "Analysing Biological Pathways in Genome-Wide Association Studies." *Nature Reviews Genetics* 11 (12) (December): 843–854. doi:10.1038/nrg2884. http://www.ncbi.nlm.nih.gov/pubmed/21085203.

Wang, Peggy I, Sohyun Hwang, Rodney P Kincaid, Christopher S Sullivan, Insuk Lee, and Edward M Marcotte. 2012. "RIDDLE: Reflective Diffusion and Local Extension

Reveal Functional Associations for Unannotated Gene Sets via Proximity in a Gene Network." *Genome Biology* 13 (12) (December 26): R125. doi:10.1186/gb-2012-13-12-r125. http://genomebiology.com/2012/13/12/R125.

Wen, Li, Ruth E Ley, Pavel Yu Volchkov, Peter B Stranges, Lia Avanesyan, Austin C Stonebraker, Changyun Hu, et al. 2008. "Innate Immunity and Intestinal Microbiota in the Development of Type 1 Diabetes." *Nature* 455 (7216) (October 23): 1109–13. doi:10.1038/nature07336. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2574766&tool=pmcentrez&rendertype=abstract.

Westholm, Jakub O, and Eric C Lai. 2011. "Mirtrons: microRNA Biogenesis via Splicing." *Biochimie* 93 (11) (November): 1897–904. doi:10.1016/j.biochi.2011.06.017. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3185189&tool=pmcentrez&rendertype=abstract.

Wherry, E John, Sang-Jun Ha, Susan M Kaech, W Nicholas Haining, Surojit Sarkar, Vandana Kalia, Shruti Subramaniam, Joseph N Blattman, Daniel L Barber, and Rafi Ahmed. 2007. "Molecular Signature of CD8+ T Cell Exhaustion During Chronic Viral Infection." *Immunity* 27 (4): 670–684. doi:10.1016/j.immuni.2007.09.006. http://www.sciencedirect.com/science/article/pii/S1074761307004542.

Williams, Andrew E, Mark M Perry, Sterghios A Moschos, Hanna M Larner-Svensson, and Mark A Lindsay. 2008. "Role of miRNA-146a in the Regulation of the Innate Immune Response and Cancer." *Biochemical Society Transactions* 36 (Pt 6) (December): 1211–5. doi:10.1042/BST0361211. http://www.ncbi.nlm.nih.gov/pubmed/19021527.

Wilson, Anthony J, and Philip S Mellor. 2009. "Bluetongue in Europe: Past, Present and Future." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 364 (1530) (September 27): 2669–81. doi:10.1098/rstb.2009.0091. http://rstb.royalsocietypublishing.org/content/364/1530/2669.long.

Wilson, Ian G, and Esme Whitehead. 2006. "Emergence of Salmonella Blockley, Possible Association with Long-Term Reactive Arthritis, and Antimicrobial Resistance." *FEMS Immunology and Medical Microbiology* 46 (1) (February): 3–7. doi:10.1111/j.1574-695X.2005.00010.x. http://www.ncbi.nlm.nih.gov/pubmed/16420591.

Winter, Julia, and Sven Diederichs. 2013. "Argonaute-3 Activates the Let-7a Passenger Strand microRNA." *RNA Biology* 10 (10) (October 2). http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3866245&tool=pmcentrez&rendertype=abstract.

Wolfe, Cecily J, Isaac S Kohane, and Atul J Butte. 2005. "Systematic Survey Reveals General Applicability of 'Guilt-by-Association' Within Gene Coexpression Networks." *BMC Bioinformatics* 6 (1) (January): 227. doi:10.1186/1471-2105-6-227. http://www.biomedcentral.com/1471-2105/6/227.

Woolhouse, Mark E J, and Sonya Gowtage-Sequeria. 2005. “Host Range and Emerging and Reemerging Pathogens.” *Emerging Infectious Diseases* 11 (12) (December): 1842–7. doi:10.3201/eid1112.050997. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3367654&tool=pmcentrez&rendertype=abstract.

Wu, Hailong, Shoumin Zhu, and Yin-Yuan Mo. 2009. “Suppression of Cell Growth and Invasion by miR-205 in Breast Cancer.” *Cell Research* 19 (4) (April): 439–448. doi:10.1038/cr.2009.18. http://www.nature.com/doifinder/10.1038/cr.2009.18.

Xiao, Changchun, Dinis Pedro Calado, Gunther Galler, To-Ha Thai, Heide Christine Patterson, Jing Wang, Nikolaus Rajewsky, Timothy P Bender, and Klaus Rajewsky. 2007. “MiR-150 Controls B Cell Differentiation by Targeting the Transcription Factor c-Myb.” *Cell* 131 (1) (October): 146–159. doi:10.1016/j.cell.2007.07.021. http://linkinghub.elsevier.com/retrieve/pii/S0092867407009580.

Xiao, Changchun, and Klaus Rajewsky. 2009. “MicroRNA Control in the Immune System: Basic Principles.” *Cell* 136 (1): 26–36. http://www.sciencedirect.com/science/article/pii/S0092867408016334.

Xu, Guangxian, Yan Zhang, Hao Jia, Juan Li, Xiaoming Liu, John F Engelhardt, and Yujiong Wang. 2009. “Cloning and Identification of microRNAs in Bovine Alveolar Macrophages.” *Molecular and Cellular Biochemistry* 332 (1-2) (June): 9–16. doi:10.1007/s11010-009-0168-4. http://link.springer.com/10.1007/s11010-009-0168-4.

Xu, Xiequn. 2007. “Same Computational Analysis, Different miRNA Target Predictions.” *Nature Methods* 4 (3) (March 1): 191; author reply 191. doi:10.1038/nmeth0307-191a. http://www.nature.com.proxy.lib.sfu.ca/nmeth/journal/v4/n3/full/nmeth0307-191a.html.

Xue, Jia, Susanne V. Schmidt, Jil Sander, Astrid Draffehn, Wolfgang Krebs, Inga Quester, Dominic De Nardo, et al. 2014. “Transcriptome-Based Network Analysis Reveals a Spectrum Model of Human Macrophage Activation.” *Immunity* (February 11). doi:10.1016/j.immuni.2014.01.006. http://www.ncbi.nlm.nih.gov/pubmed/24530056.

Yamamoto, Satoko, Noriko Sakai, Hiromi Nakamura, Hiroshi Fukagawa, Ken Fukuda, and Toshihisa Takagi. 2011. “INOH: Ontology-Based Highly Structured Database of Signal Transduction Pathways.” *Database: The Journal of Biological Databases and Curation* 2011 (January): bar052. doi:10.1093/database/bar052. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3225078&tool=pmcentrez&rendertype=abstract.

Yang, Jr-Shiuan, Michael D Phillips, Doron Betel, Ping Mu, Andrea Ventura, Adam C Siepel, Kevin C Chen, and Eric C Lai. 2011. “Widespread Regulatory Activity of Vertebrate microRNA* Species.” *RNA (New York, N.Y.)* 17 (2) (March): 312–26. doi:10.1261/rna.2537911. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3022280&tool=pmcentrez&rendertype=abstract.

Yang, Rongxi, Bettina Schlehe, Kari Hemminki, Christian Sutter, Peter Bugert, Barbara Wappenschmidt, Juliane Volkmann, et al. 2009. “A Genetic Variant in the Pre-miR-27a Oncogene Is Associated with a Reduced Familial Breast Cancer Risk.” *Breast Cancer Research and Treatment* 121 (3) (November): 693–702. doi:10.1007/s10549-009-0633-5. http://link.springer.com/10.1007/s10549-009-0633-5.

Yousef, M, S Jung, A V Kossenkov, L C Showe, and M K Showe. 2007. “Naive Bayes for microRNA Target Predictions--Machine Learning for microRNA Targets.” *Bioinformatics (Oxford, England)* 23 (22) (November): 2987–2992.

Yu, J, D G Ryan, S Getsios, M Oliveira-Fernandes, A Fatima, and R M Lavker. 2008. “MicroRNA-184 Antagonizes microRNA-205 to Maintain SHIP2 Levels in Epithelia.” *Proceedings of the National Academy of Sciences* 105 (49) (November): 19300–19305. doi:10.1073/pnas.0803992105. http://www.pnas.org/cgi/doi/10.1073/pnas.0803992105.

Yu, Y, S S Kanwar, B B Patel, P.-S. Oh, J Nautiyal, F H Sarkar, and A P N Majumdar. 2011. “MicroRNA-21 Induces Stemness by Downregulating Transforming Growth Factor Beta Receptor 2 (TGFR2) in Colon Cancer Cells.” *Carcinogenesis* 33 (1) (November): 68–76. doi:10.1093/carcin/bgr246. http://www.carcin.oxfordjournals.org/cgi/doi/10.1093/carcin/bgr246.

Yue, Xiao, Peiguo Wang, Jun Xu, Yufang Zhu, Guan Sun, Qi Pang, and Rongjie Tao. 2012. “MicroRNA-205 Functions as a Tumor Suppressor in Human Glioblastoma Cells by Targeting VEGF-A.” *Oncology Reports* 27 (4) (April): 1200–6. doi:10.3892/or.2011.1588. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3583473&tool=pmcentrez&rendertype=abstract.

Zak, Daniel E, and Alan Aderem. 2009. “Systems Biology of Innate Immunity.” *Immunological Reviews* 227 (1) (January): 264–82. doi:10.1111/j.1600-065X.2008.00721.x. http://www.ncbi.nlm.nih.gov/pubmed/21111589.

Zaslavsky, Elena, Uri Hershberg, Jeremy Seto, Alissa M, Susanna Marquez, Jamie L Duke, James G, Benjamin R, Stuart C Sealfon, and Steven H Kleinstein. 2010. “Antiviral Response Dictated by Choreographed Cascade of Transcription Factors.” *The Journal of Immunology*. doi:10.4049/jimmunol.0903453.

Zhao, Yicheng, Zhen Dai, Yang Liang, Ming Yin, Kuiying Ma, Mei He, Hongsheng Ouyang, and Chun-Bo Teng. 2014. “Sequence-Specific Inhibition of microRNA via CRISPR/CRISPRi System.” *Scientific Reports* 4 (January): 3943. doi:10.1038/srep03943. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3909901&tool=pmcentrez&rendertype=abstract.

Zhao, Zhongying, Li Fang, Nansheng Chen, Robert C Johnsen, Lincoln Stein, and David L Baillie. 2005. “Distinct Regulatory Elements Mediate Similar Expression Patterns in the Excretory Cell of Caenorhabditis Elegans.” *The Journal of Biological*

Chemistry 280 (46) (November 18): 38787–94. doi:10.1074/jbc.M505701200. http://www.ncbi.nlm.nih.gov/pubmed/16159881.

Zhu, Shu, Wen Pan, and Youcun Qian. 2013. “MicroRNA in Immunity and Autoimmunity.” *Journal of Molecular Medicine (Berlin, Germany)* (May 1). doi:10.1007/s00109-013-1043-z. http://www.ncbi.nlm.nih.gov/pubmed/23636510.

Zimin, Aleksey V, Arthur L Delcher, Liliana Florea, David R Kelley, Michael C Schatz, Daniela Puiu, Finnian Hanrahan, et al. 2009. “A Whole-Genome Assembly of the Domestic Cow, Bos Taurus.” *Genome Biology* 10 (4) (January): R42. doi:10.1186/gb-2009-10-4-r42. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2688933&tool=pmcentrez&rendertype=abstract.

Zoubarev, Anton, Kelsey M Hamer, Kiran D Keshav, E Luke McCarthy, Joseph Roy C Santos, Thea Van Rossum, Cameron McDonald, et al. 2012. “Gemma: a Resource for the Reuse, Sharing and Meta-Analysis of Expression Profiling Data.” *Bioinformatics (Oxford, England)* 28 (17) (September 1): 2272–3. doi:10.1093/bioinformatics/bts430. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3426847&tool=pmcentrez&rendertype=abstract.

Appendices

Appendix A.

Human pathways with a relatively low conservation rate in cow

Pathway name	Source	Number of bovine genes (predicted)	Number of human genes	Conserved genes ratio (%)
Defensins	REACTOME	3	37	8
RNA Polymerase I Promoter Opening	REACTOME	3	29	10
Classical antibody-mediated complement activation	REACTOME	3	28	11
IFN alpha signaling pathway(JAK1 TYK2 STAT1) (IFN alpha signaling(JAK1 TYK2 STAT1 STAT2 STAT3))	INOH	2	16	13
Beta defensins	REACTOME	5	35	14
Xenobiotics	REACTOME	2	13	15
Sema3A PAK dependent Axon repulsion	REACTOME	2	11	18
IFN alpha signaling pathway(JAK1 TYK2 STAT1 STAT3) (IFN alpha signaling(JAK1 TYK2 STAT1 STAT2 STAT3))	INOH	3	17	18
IFN alpha signaling pathway(JAK1 TYK2 STAT3) (IFN alpha signaling(JAK1 TYK2 STAT1 STAT2 STAT3))	INOH	3	16	19
Packaging Of Telomere Ends	REACTOME	6	30	20
Alpha-defensins	REACTOME	2	9	22
Glutathione conjugation	REACTOME	2	9	22
IFN alpha signaling pathway((JAK1 TYK2 STAT1 STAT2) (IFN alpha signaling(JAK1 TYK2 STAT1 STAT2 STAT3))	INOH	4	18	22
Generic Transcription Pathway	REACTOME	65	214	30
Ascorbate and aldarate metabolism	KEGG	8	25	32
Metabolism of xenobiotics by cytochrome P450	KEGG	22	69	32
Caffeine metabolism	KEGG	2	6	33
IFN gamma signaling pathway(JAK1 JAK2 STAT1) (IFN gamma signaling(JAK1 JAK2 STAT1))	INOH	2	6	33
Olfactory Signaling Pathway	REACTOME	116	351	33

Pathway name	Source	Number of bovine genes (predicted)	Number of human genes	Conserved genes ratio (%)
Drug metabolism - cytochrome P450	KEGG	24	70	34
Graft-versus-host disease	KEGG	13	37	35
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	REACTOME	32	90	36
Olfactory transduction	KEGG	133	364	37
Autoimmune thyroid disease	KEGG	19	50	38
Eicosanoids	REACTOME	2	5	40
Ifn gamma signaling pathway	PID BIOCARTA	2	5	40
Pentose and glucuronate interconversions	KEGG	12	30	40
Retinol metabolism	KEGG	25	62	40
Steroid hormone biosynthesis	KEGG	22	54	41
Endosomal/Vacuolar pathway	REACTOME	3	7	43
IL-3 signaling pathway(JAK1 JAK2 STAT5) (IL-3 signaling(JAK1 JAK2 STAT5))	INOH	3	7	43
RNA Polymerase I Chain Elongation	REACTOME	20	46	43
Ifn alpha signaling pathway	PID BIOCARTA	4	9	44
Systemic lupus erythematosus	KEGG	39	88	44
Drug metabolism - other enzymes	KEGG	23	50	46
Deposition of New CENPA-containing Nucleosomes at the Centromere	REACTOME	21	44	48
Allograft rejection	KEGG	17	35	49
Meiotic Recombination	REACTOME	26	53	49
Amine ligand-binding receptors	REACTOME	3	6	50
Calcitonin-like ligand receptors	REACTOME	4	8	50
Ras-independent pathway in nk cell-mediated cytotoxicity	PID BIOCARTA	11	22	50
Asthma	KEGG	14	28	50
Antigen processing and presentation	KEGG	34	66	52
Regulation of autophagy	KEGG	18	34	53
Natural killer cell mediated cytotoxicity	KEGG	70	133	53
NCAM signaling for neurite out-growth	REACTOME	7	13	54
Linoleic acid metabolism	KEGG	15	28	54
Type I diabetes mellitus	KEGG	22	41	54
Chemokine receptors bind chemokines	REACTOME	25	46	54
Interferon gamma signaling	REACTOME	25	46	54
Meiotic Synapsis	REACTOME	30	56	54
Interferon alpha/beta signaling	REACTOME	24	44	55

Pathway name	Source	Number of bovine genes (predicted)	Number of human genes	Conserved genes ratio (%)
Digestion of dietary carbohydrate	REACTOME	4	7	57
NAD+ + 3 4-Dihydroxy-phenyl-ethyleneglycol = NADH + 3 4-Dihydroxy-mandelaldehyde (Tyrosine metabolism)	INOH	4	7	57
Starch and sucrose metabolism	KEGG	28	49	57
Staphylococcus aureus infection	KEGG	31	53	58
Smooth Muscle Contraction	REACTOME	13	22	59
Porphyrin and chlorophyll metabolism	KEGG	24	41	59
Intestinal immune network for IgA production	KEGG	27	46	59
APOBEC3G mediated resistance to HIV-1 infection	REACTOME	3	5	60
JNK (c-Jun kinases) phosphorylation and activation mediated by activated human TAK1	REACTOME	3	5	60
LDL endocytosis	REACTOME	3	5	60
Miscellaneous substrates	REACTOME	3	5	60
Post-chaperonin tubulin folding pathway	REACTOME	11	18	61
Termination of O-glycan biosynthesis	REACTOME	14	23	61
Activation of DNA fragmentation factor	REACTOME	8	13	62
Other types of O-glycan biosynthesis	KEGG	28	45	62
Ca-calmodulin-dependent protein kinase activation	PID BIOCARTA	5	8	63
Gene expression of SOCS1 by STAT dimer (JAK-STAT pathway and regulation pathway Diagram)	INOH	5	8	63
Gene expression of SOCS3 by STAT dimer (JAK-STAT pathway and regulation pathway Diagram)	INOH	5	8	63
Signal dependent regulation of myogenesis by corepressor mitr	PID BIOCARTA	5	8	63
Cytosolic DNA-sensing pathway	KEGG	34	54	63
Recycling of bile acids and salts	REACTOME	7	11	64
Translocation of ZAP-70 to Immunological synapse	REACTOME	9	14	64
Stathmin and breast cancer resistance to antimicrotubule agents	PID BIOCARTA	14	22	64
Cytokine-cytokine receptor interaction	KEGG	174	270	64
Taste transduction	KEGG	31	48	65
Ribosome	KEGG	57	88	65
Activation of the AP-1 family of transcription factors	REACTOME	4	6	67

Pathway name	Source	Number of bovine genes (predicted)	Number of human genes	Conserved genes ratio (%)
Basic mechanisms of sumoylation	PID BIOCARTA	4	6	67
Cytosolic sulfonation of small molecules	REACTOME	4	6	67
ERK cascade (EGF signaling pathway Diagram)	INOH	4	6	67
ERK cascade (HGF signaling pathway)	INOH	4	6	67
ERK cascade (Integrin signaling pathway)	INOH	4	6	67
ERK cascade (PDGF signaling pathway)	INOH	4	6	67
ERK cascade (VEGF signaling pathway)	INOH	4	6	67
Gene expression of smad7 by R-smad:smad4 (TGF-beta super family signaling pathway(canonical))	INOH	4	6	67
JNK cascade (TGF-beta signaling(through TAK1))	INOH	4	6	67
Methylation	REACTOME	4	6	67
Synthesis of bile acids and bile salts via 27-hydroxycholesterol	REACTOME	4	6	67
TRAIL signaling	REACTOME	4	6	67
p75NTR negatively regulates cell cycle via SC1	REACTOME	4	6	67
Inactivation of Cdc42 and Rac	REACTOME	6	9	67
Opsins	REACTOME	6	9	67
Antigen processing and presentation	PID BIOCARTA	8	12	67
Il 3 signaling pathway	PID BIOCARTA	8	12	67
Antigen Presentation: Folding assembly and peptide loading of class I MHC	REACTOME	14	21	67
Biosynthesis of unsaturated fatty acids	KEGG	14	21	67
Formation of tubulin folding intermediates by CCT/TriC	REACTOME	14	21	67
IL3-mediated signaling events	PID NCI	14	21	67
TSLP	NETPATH	16	24	67
Eukaryotic Translation Termination	REACTOME	56	84	67
Toll-like receptor signaling pathway	KEGG	68	102	67
Arachidonic acid metabolism	KEGG	39	57	68
Viral myocarditis	KEGG	46	68	68
Peptide chain elongation	REACTOME	57	84	68
Viral mRNA Translation	REACTOME	57	84	68

Appendix B.

Detailed results for the pathway analysis of a TB expression dataset (GSE11199) by 6 different methods

The attached worksheets file (Appendix_B.xlsx) forms part of this work. The file can be opened with Microsoft Excel.

The worksheet contains 9 tabs: the list of the input genes, the full results of each of the six compared analysis tools for this data-set, a summarizing 6way-comparison sheet of the ranks of pathways that were identified as significant by at least one method, and an additional tab for DAVID's functional clusters.

Appendix C.

Detailed results for the pathway analysis of a mouse experimental cerebral malaria (ECM) expression dataset (GSE7814) by 6 different methods

The attached worksheets file (Appendix_C.xlsx) forms part of this work. The file can be opened with Microsoft Excel.

The worksheet contains 9 tabs: the list of the input genes (for the ORA-based methods), the full results of each of the six compared analysis tools for this data-set, a summarizing 6way-comparison sheet of the ranks of pathways that were identified as significant by at least one method, and an additional tab for DAVID's functional clusters.

Appendix D.

Detailed results for the pathway analysis of a Dengue fever dataset (GSE25001) by 6 different methods.

The attached worksheets file (Appendix_D.xlsx) forms part of this work. The file can be opened with Microsoft Excel.

The worksheet contains 8 tabs: the list of the input genes, the full results of each of the six compared analysis tools for this data-set, a summarizing 6 way-comparison sheet of the ranks of pathways that were identified as significant by at least one method.

Appendix E.

Clusters of highly co-expressed genes in BGA and predictions based on GBA.

The attached worksheets file (Appendix_E.xlsx) forms part of this work. The file can be opened with Microsoft Excel.

The worksheet contains three tabs: 'Clusters identified by MINE', 'Predictions part 1' and 'Predictions part 2'.

In both prediction tabs, genes listed in column L are predicted to be associated with functions in column J based on GBA.

In the tab labeled *'Predictions part 1'*, the functions in column J **are** the top scoring functions for the clusters in column B, and the function passes the proportional filter, i.e. at least 40% of all annotated genes in the cluster are associated with the function in J.

In the '*Predictions part 2*' tab, the functions in column J **are not** the top scoring functions for the clusters in column B, but the first significant function that fulfills the fractional test, i.e. at least 40% all annotated genes in the cluster are associated with the function in J.

Appendix F.

Sample descriptions for the miRNA study

The attached spreadsheets file (Appendix_F.xls) can be opened with Microsoft Excel.

Appendix G.

RNA integrity values and miRNA concentrations for the samples in the miRNA study

The attached spreadsheets file (Appendix_G.xls) can be opened with Microsoft Excel.

Appendix H.

Average transcript counts per miRNA

The attached spreadsheets file (Appendix_H.xls) can be opened with Microsoft Excel.

Appendix I.

miRNAs involved in response to S uberis in the literature

The attached spreadsheets file (Appendix_I.xls) can be opened with Microsoft Excel.

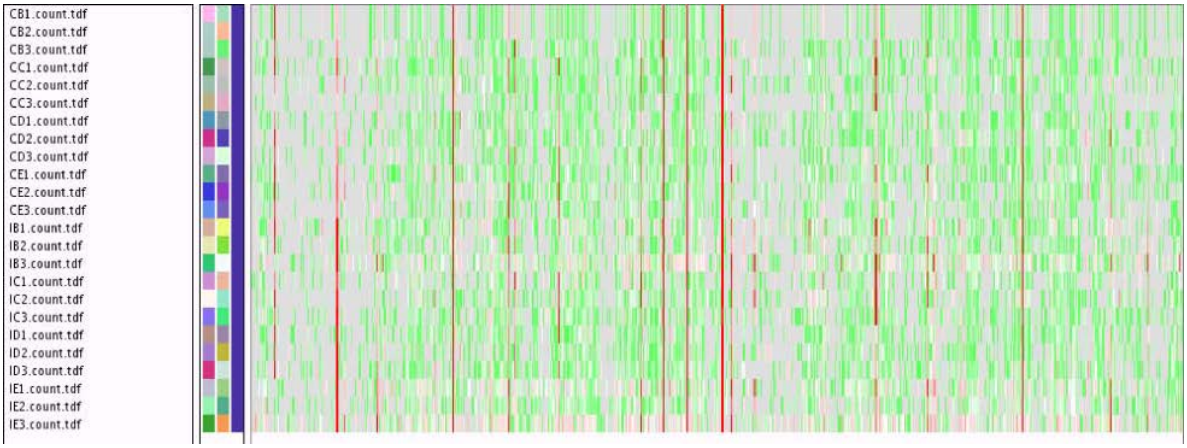
Appendix J.

Genes predicted to be targeted by differentially expressed miRNAs

The attached worksheets file (Appendix_J.xlsx) forms part of this work. The file can be opened with Microsoft Excel.

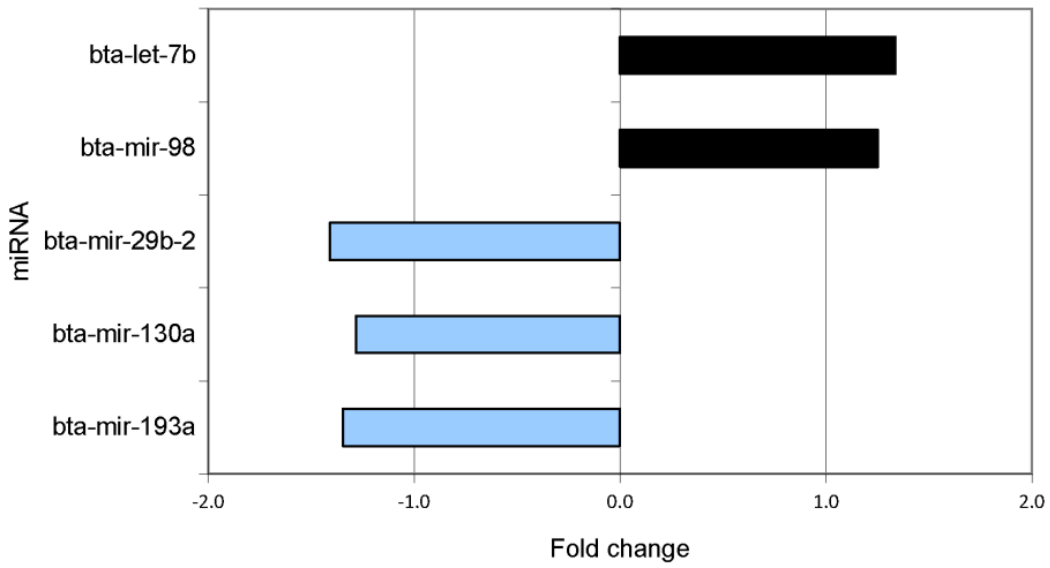
Appendix K.

Alignment of reads to bovine ncRNAs



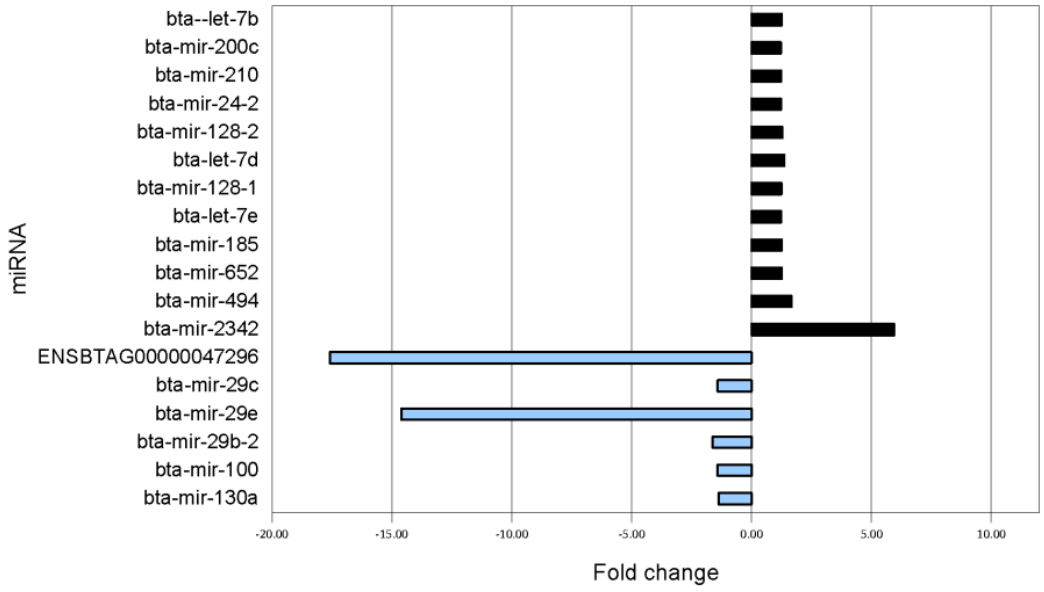
Appendix L.

Fold changes in expression of differentially expressed miRNAs at 4 hpi



Appendix M.

Fold changes in expression of differentially expressed miRNAs at 6hpi



Appendix N.

RNA quality and read numbers.

Sample ID	RNA conc. [ng/uL]	rRNA ratio (28s/18s)	RIN	Small RNA conc. [pg/µL]	miRNA conc. [pg/µL]	Total # of reads	#of reads after quality control steps	#of uniquely aligned reads	Uniquely aligned as % of total
BAM-1	383	1.6	9.4	15,674	746	11,268,106	7,767,806	5,370,197	48%
BAM-2	530	1.6	9.4	19,602	1,144	10,359,205	6,486,994	4,657,271	45%
BAM-3	483	1.7	9.4	18,441	1,404	8,379,883	5,628,363	3,867,637	46%
BAM-4	446	1.6	9.6	13,306	1,005	11,322,893	8,759,869	5,928,652	52%
BAM-5	372	1.7	9.7	17,380	1,875	9,957,402	8,195,175	5,641,224	57%
BAM-6	470	1.7	9.6	11,565	1,830	13,108,422	9,174,465	6,270,685	48%
BAM-7	656	1.5	9.4	9,854	1,820	10,561,796	8,550,114	5,941,310	56%
BAM-8	80	1.9	9.8	9,869	660	11,380,422	8,018,972	4,916,793	43%

RNA Integrity Number (RIN) values were above 9 for all samples.

Appendix O.

The expression of Ensembl annotated bovine miRNAs in BAMs.

The attached worksheets file (Appendix_O.xlsx) forms part of this work. The file can be opened with Microsoft Excel. Read counts shown are the number of reads aligning uniquely to that miRNA.

Appendix P.

The RT-qPCR results for miR-21, miR-148a and miR-708 in four samples.

	bta-miR-21 (RPM)	bta-miR-21 (Cq)	bta-miR-21 (Cq) 1:100 dilution	bta-miR-148a (RPM)	bta-miR-148a (Cq)	bta-miR-708 (RPM)	bta-miR-708 (Cq)
BAM-5	796,997	19.23	26.92	8,045	26.98	0.00	35.78
BAM-7	792,299	18.73	26.10	12,734	25.78	0.72	34.38
BAM-3	798,224	17.56	N/A	6,909	25.06	0.27	N/A
BAM-4	759,241	19.03	N/A	5,541	26.80	0.72	N/A

Means of Cq values of three wells are shown. (* Cq Threshold = 0.15)

Appendix Q.

List of target genes that are computationally predicted to be regulated by miRNAs expressed above a threshold of 100 RPM in BAMs.

The attached worksheets file (Appendix_Q.xlsx) forms part of this work. The file can be opened with Microsoft Excel. Only genes that were predicted by both of the computational approaches, miRanda v3.3a and TargetScan v6.2, are listed.

Appendix R.

Summary of isomiR expression across samples.

Sample	Strand	# miRs with isomiRs	# of isomiRs >100 reads	# isomiRs longer than consensus	# isomiRs shorter than consensus	# isomiRs 5' modified (first 5nt)	# isomiRs 3' modified (last 5nt)	# cases where isomiR expressed more highly than consensus
BAM-1	5p	54	456	65	344	183	418	31
BAM-1	3p	39	352	100	136	167	268	27
BAM-2	5p	53	432	63	321	176	393	28
BAM-2	3p	42	308	95	132	135	245	28
BAM-3	5p	49	435	59	327	180	395	26
BAM-3	3p	41	303	99	117	132	241	25
BAM-4	5p	55	512	76	378	218	463	28
BAM-4	3p	42	423	125	175	202	336	28
BAM-5	5p	56	463	74	333	198	409	29
BAM-5	3p	39	377	119	136	177	290	25
BAM-6	5p	57	498	76	361	220	445	30
BAM-6	3p	43	436	125	173	215	338	28
BAM-7	5p	55	479	63	362	183	435	32
BAM-7	3p	42	389	104	164	181	298	27
BAM-8	5p	49	462	64	340	191	416	25
BAM-8	3p	39	295	101	115	119	240	26

IsomiRs were found to be generally shorter than consensus sequences. Modifications at 3′ ends were more common than at 5′ ends. Several isomiRs were more highly expressed than their consensus sequences.