2011

Prediction and verification of NF-κB targets in the porcine MHC through the use of sequence similarity and pathway inhibition

Oliver Couture
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd
Part of the Animal Sciences Commons

Recommended Citation
Couture, Oliver, "Prediction and verification of NF-κB targets in the porcine MHC through the use of sequence similarity and pathway inhibition" (2011). Graduate Theses and Dissertations. 11980.
http://lib.dr.iastate.edu/etd/11980

This Dissertation is brought to you for free and open access by the Graduate College at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Prediction and verification of NF-κB targets in the porcine MHC through the use of sequence similarity and pathway inhibition

by

Oliver Paul Couture

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Christopher K. Tuggle, Major Professor
Vasant Honavar
Jack Dekkers
Dan Nettleton
James Reecy

Iowa State University

Ames, Iowa

2011

Copyright © Oliver Paul Couture, 2011. All rights reserved.
TABLE OF CONTENTS

ABSTRACT iv

CHAPTER 1. GENERAL INTRODUCTION 1
   Introduction 1
   Research Objectives 2
   Dissertation Organization 2
   References 4

CHAPTER 2. LITERATURE REVIEW 5
   The Importance and Usage of Swine 5
   Impact and Spread of Salmonella on Swine 6
   Current State of Porcine Sequence Annotation 9
   Clustering Expression Data 11
   Predicting Targets of Transcription Factors 13
   Multiple Dimensional Clustering 14
   Biological Verification of Predictions 16
   Macrophages and their Activation by Endotoxin 18
   Known Effects of NF-κB in the Inflammatory Response 21
   Chemical Inhibition of NF-κB 23
   Use of High Throughput Expression Analysis Methods in Pigs 25
   References 27
   Figure Legends 34

CHAPTER 3. ANEXDB: AN INTEGRATED ANIMAL ANNOTATION AND MICROARRAY EXPRESSION DATABASE 38
   Abstract 38
   Introduction 39
   Materials and Methods 41
   Results 47
   Discussion 50
   Acknowledgements 56
   References 57
   Figure Legends 64

CHAPTER 4. ANALYSIS OF PORCINE TRANSCRIPTIONAL RESPONSE TO SALMONELLA ENTERICA SEROVAR CHOLERAESUIS SUGGESTS NOVEL TARGETS OF NFKAPPAB ARE ACTIVATED IN THE MESENTERIC LYMPH NODE 69
   Abstract 69
   Introduction 70
   Materials and Methods 72
   Results 78
   Discussion 87
CHAPTER 5. CONFIRMATION OF BIOINFORMATICALLY PREDICTED
NOVEL TARGETS OF NF-κB: H2-EB1 AND TRIM26 IN MOUSE AND C2
AND UBD IN PIG

Abstract 117
Background 118
Materials and Methods 121
Results 131
Discussion 139
Conclusion 149
Authors’ Contributions 150
Acknowledgements 150
References 152
Figure Legends 163

CHAPTER 6. GENERAL CONCLUSIONS
Discussion 181
Maturation of ANEXdb since publication 183
Future Research 184
Conclusions 186
References 187

APPENDIX A. STRUCTURE OF THE ANEXDB WEB PAGES
AND UNDERLYING MYSQL DATABASES
User Submission to ANEXdb 189
Administrator Finalization of Submissions 192
Other Administrator Duties for Managing Expression Data 194
Figure Legends 196

APPENDIX B. CONSTRUCTION AND ANNOTATION OF THE
ASSEMBLY BY USING TGICL
General Assembly Procedure 201
Annotating, Parsing, and Loading Data into annotDB 204
Web Pages 209
Figure Legends 211

ACKNOWLEDGMENTS 213
ABSTRACT

With the advent of high throughput technologies for both the sequencing of genomic DNA and the measure of the expression of RNA a tremendous amount of information has been generated and deposited into public databases. This large amount of data has led to the better understanding of how a genome is organized, the number of regions encoding information for transcripts, as well as how the amount of these transcripts change due to various perturbations a cell or organism encounters, whether it be an outside stimuli, such as bacteria or viruses, or internal, such as a mutation within the genome.

Some species, such as the human and mouse, have had a significant amount of sequencing completed, leading to excellent reference genome sequences, as well as these sequences being well understood at the function and structure level, termed gene annotation. However, for most vertebrate species, their genomes are in various states of completion; from being nearly completed with partial annotation, like the pig, to having only portions of their genomes completed, such as *Alatina moseri*, a species of Hawaiian jellyfish. For these species the amount of direct annotation is greatly lacking compared to that of other species, such as human and mouse. When annotation is lacking for one species, it is possible to leverage the information already obtained for closely related but better-studied species by comparing sequences across species and identifying similar regions between them, allowing the annotations of these regions to be inferred across species.

Once a species has sufficient sequence annotation, high throughput expression data, such as that from microarrays, can be better understood. One area of research that is under development, which can utilize high throughput expression measures, is understanding how a set of transcripts changing together in response to perturbations in the environment is
controlled by specific proteins, called transcription factors, such as NF-κB. NF-κB is an important transcription factor, having a role in a variety of cellular functions, such as mounting a response to infection and preventing cell death by inhibiting apoptosis. While some transcription factors, like NF-κB, have been well studied and many of its target genes identified, this identification is typically done one or a few genes at a time. However, as more genomes are sequenced, better algorithms developed for identification of possible targets, and new biological techniques optimized, the ability to predict and verify targets is also moving toward high throughput. In order to create more reliable gene annotation for the pig, raw porcine sequences were assembled into more full length sequences to create an accurate base for which to compare to other species, as well as identify possible sequence variation within the assembled sequences. This annotation was then used in a high throughput experiment to look for genes changing expression due to an inoculation of *Salmonella choleraesuis* in pigs, and to determine which genes are potential NF-κB targets. Then, potential target genes found in an immune related region of the genome were tested in response of bacterial endotoxin either in the presence or absence of an NF-κB inhibitor. The ability of NF-κB to bind to their promoters was also tested using a labeled EMSA probe. Using these two methods, we show the murine H2-Eb1 and Trim26 and porcine C2 and UBD are novel targets of NF-κB and that such bioinformatic predictions can be confirmed using molecular assays.
CHAPTER 1. GENERAL INTRODUCTION

INTRODUCTION

The use of the domestic pig is varied and important: ranging from being a viable, edible protein source to being a source of replacement organs for humans (xenotransplantation, Lai and Prather 2002) and as a model for human disease, such as cystic fibrosis (Rogers et al. 2008) and tissue scarring (Zhu et al. 2008; Gallant-Behm et al. 2007). As a source of food, pigs also bring along some health risks through contamination of food, mainly of a family of bacteria called Gram negative bacteria. This group of bacteria consists of both Salmonella and Escherichia coli, which are two of the main sources of food contamination in the United States, and along with Campylobacter (found in birds) and the calicivirus group of viruses, commonly known as Norwalk and Norwalk-like viruses in birds or other species, comprise the majority of foodborne illness each year (Center for Disease Control, CDC, 2010). The CDC estimates that about 1 in 10 people get hospitalized each year, and approximately 5,000 deaths occur in the United States due to foodborne illness. A better understanding of how animals, like the pig, control the response to infectious agents through the regulation of transcripts by specific transcription factors could allow for the selection of animals with a specific version of a gene (allele) to breed into a population (Griesbeck-Zilch et al. 2009).

One of the central transcription factors for the immune response, and a key regulator of gram negative bacteria recognition, is NF-κB (Sharif et al., 2007). Since NF-κB is a well studied transcription factor, previously known targets can be utilized to help identify novel targets. This can be done by identifying genes with similar patterns of expression due to the same stimuli to known targets, and having similar regions of sequence within their promoters
to the known binding sites of NF-κB. To do this, however, reliable annotation of transcripts and genomic regions are needed.

RESEARCH OBJECTIVES

The increase in the amount of both sequence and expression data available to the public has led to the development of new algorithms to utilize and analyze this data. One such area is in the development of the discovery of which transcription factors are responsible for the changes in RNA due to perturbations to a cell or organism. The objective of Chapter 3 is to develop annotation for transcripts of any poorly annotated species, specifically in this case swine. The objective of Chapter 4 is to use the annotation created in Chapter 3 to improve the annotation of the Affymetrix porcine GeneChip®, and use that platform to measure the changes in transcript expression responding to an inoculation of *Salmonella* in swine within the mesenteric lymph nodes of swine. Chapter 5’s objective was to verify the predicted targets of NF-κB identified from the work in Chapter 4 by measuring the response to endotoxin in mouse and pig macrophage-like cells.

DISSECTATION STRUCTURE

The following chapter of this thesis is a literature review to provide background information of the research done for this thesis. The next three chapters of this thesis are comprised of individual papers that have either been published or are in preparation for publication. Chapter 3 is a published manuscript in Mammalian Genome 20, 768-777 titled “ANEXdb: An Integrated Animal ANnotation and Microarray EXpression Database.” It was written primarily by Oliver Couture under the supervision of Christopher Tuggle. It outlines and describes a publicly available database which houses microarray expression data as well as expressed transcript sequence assemblies, and the annotations of the resulting consensus
sequences. Keith Callenberg helped in the design and implementation of the expression
database, Neeraj Koul, Sushain Pandit, and Remy Younes created the object model to handle
dataflow between the submission and static database, as well as the output of files that can be
submitted to NCBI’s expression database GEO. Zhi-Liang Hu, James Reecy, Jack Dekkers,
and Vasant Honavar provided suggestions on analysis and manuscript preparations. Chapter
4 is a paper published in BMC Genomics 9, 437-456 entitled “Analysis of porcine
transcriptional response to Salmonella enterica serovar Choleraesuis suggest novel targets of
NFkappaB are activated in the mesenteric lymph node” and was written primarily by
Yanfang Wang with significant input from Oliver Couture, both under the supervision of
Christopher Tuggle. It describes the work done to identify genes responding to Salmonella
enterica serovar Choleraesuis in the mesenteric lymph node of pigs, as well as identify a
subset of genes that could be potential targets of NF-κB. Yanfang Wang also did most of the
post-statistical microarray analysis and contributed to the RT-Q-PCR work. Oliver Couture
did the NF-κB promoter target searching, as well as creating the GO-slim. Long Qu, Dan
Nettleton, and Jack Dekkers performed the statistical analysis, and Shawn Bearson and Jolita
Uthe helped with sample collection and preparation. Daniel Kuhar and Joan Lunney
performed the RT-Q-PCR. Chapter 5 is a manuscript in preparation to be submitted to BMC
Genomics entitled “Verification of bioinformatically predicted targets of NF-κB in mouse
and pig reveal novel targets including UBD and TRIM26” that was written primarily by
Oliver Couture under the supervision of Christopher Tuggle. Jolita Uthe and Nowlan Freese
contributed to the lab work, Mike Wannemuehler provided suggestions about the cell culture
experiments, and Beth Kenkel and Amanda Ramer-Tait provided help with lab work, as well
as, suggestions on the analysis and manuscript preparations.
REFERENCES


CHAPTER 2. LITERATURE REVIEW

THE IMPORTANCE AND USAGE OF SWINE

Swine are used for a variety of purposes in this country, from consumption to medical models; swine are even a potential candidate for organ transplant into humans. In 2004, the last year summarized by the Economic Research Service (ERS) section of the United States Department of Agriculture (USDA), pork accounted for nearly half of red meat (45.2%; ERS, USDA, 2006) and almost a quarter of all meat (24.0%; ERS, USDA, 2006) produced in the United States. In 2007 the pork industries retail value in the United States was worth $48.5 billion (ERS, USDA, 2008). Outside of just the United States, pork consists of about 39% of the global meat production, the second being poultry at 30%. The rate of meat consumption is also increasing throughout the world (Halweil, 2003). While swine are commonly used as food source, its importance to the medical field is also increasing.

Swine are being used as a medical model for such things as tissue scarring (Zhu et al., 2008 and Gallant-Behm et al., 2007) and specific diseases such as cystic fibrosis (Rogers et al., 2008) and Huntington’s disease (Uchida et al., 2001). In addition to being used as a disease model, transgenic pigs are also being created for use as a potential source of organs for xenotransplantation due to their similar size and physiology to humans (see for example van der Windt et al., 2009, and Lai and Prather, 2002). Even without creating a transgenic pig to model a specific disease, pigs are sometimes a better model than the mouse or rat (the more commonly used models) due to the similarity in human and pig physiology (Butler, 2009). There have also been several differences identified between human and mouse pertaining to the immune system (Mestas and Hughes, 2004).
One example of these differences is how humans and mice respond to *Salmonella enterica* serovar Typhimurium (*S.* Typhimurium). In humans it creates a less severe infection, usually resulting in gastroenteritis, while in mouse it can cause a systemic reaction more similar to what is seen in a human infected with *Salmonella enterica* serovar Typhi (Mittrücker and Kaufmann, 2000). Pigs also can become infected with *S.* Typhimurium, having a more similar infection to the milder case seen by humans than the severe case seen in mouse (Boyen et al., 2008). This wide zoonotic ability of *S.* Typhimurium can result in food borne infection of *S.* Typhimurium in humans from the consumption of eating infected pork, as well as eggs (Muir et al., 1998 and Calloway, 2010). Since swine and chicken are the most commonly consumed sources of protein worldwide (accounting for 69% of meat consumption globally), it is important to understand how *Salmonella* and its hosts interact to reduce its negative impact on both meat production and human health.

**IMPACT AND SPREAD OF *SALMONELLA* ON SWINE**

While *S.* Typhimurium infection usually results in enterocolitis in the pigs, it has been shown that an acute infection can significantly affect a pig’s average daily gain and reduce their daily lean growth rate when compared to uninfected pigs within the first three days of the infection, even when being fed *ad libitum*. However, by day 17 no adverse effects were observed. This effect is mostly due to a reduction in appetite and food take, but a significant portion, about 33%, is due to the energy spent on the immune system to fight off the infection (Loughmiller et al, 2007). Though an acute infection does not seem to have long-term effects on the growth rate of pigs, *S.* Typhimurium can create a carrier state in pigs that can last for more than six months (Charles, 2000). During this carrier state, the pig can start shedding bacteria again when stressed, such as trucking between farms, which could explain
why approximately 58% of farms across five sampled states in the United States tested positive for *Salmonella* in 2005 (CAHFSE, 2005). This can lead to spreading of infection not only within a barn, but from farm to farm. Rao et al. (2009) observed that the spread of *S. Typhimurium* was localized around the pens that contained infected pigs, this was followed by genotyping of bacteria in multiple pens on multiple visits over a three-month period. While Rao et al. (2009) did not measure it, they also speculated that in the barn they used, which was a farrow-to-finish barn and replacement pigs being breed on the same farm, the spread of bacteria from pen to pen seen in the barn was most likely due to moving infected pigs from pen to pen. They also stated that all-in, all-out operations could minimize the impact of carrier-state pigs being transferred from barn to barn, due to the localized spread of *Salmonella* within a barn, as long as the infected pig was identified and could be kept separate from the rest of the herd. The spread and impact of *S. Typhimurium* has been estimated to costs millions of dollars to the pork industry (Fedorka-Cray et al., 1995). The pig is also a reservoir for food-borne pathogens for humans (Hoffmann et al., 2007).

*S. Typhimurium* has been the most common *Salmonella* spp. to be isolated from humans, unfortunately many of these are resistant to multiple drugs (CDC, 2008). The incidence of *S. Typhimurium* in human infection is also increasing throughout the world. In 2006 it tied with *Salmonella enteriditis* for the most number of outbreaks. This is unusual in that *Salmonella enteriditis* typically has more outbreaks, while *S. Typhimurium* typically has more sporadic cases in the United States (CDC, 2008). In Europe, an estimated 15-23% of reported cases of salmonellosis are associated with pork consumption (Boyen et al., 2007). Since pork is the most commonly consumed meat worldwide, this proportion could be higher in countries outside of Europe and the United States. Also, while pork consumption accounts
for only about 3.1% of the overall food-borne illness in the United States (with non-typhoidal 
*Salmonella* being the third most common infection from pigs, Norwalk-like viruses and 
*Campylobacter* spp. being first and second), non-typhoidal *Salmonella* illness from pigs 
ranks 15th in food-borne related hospitalizations (out of 121) and food-borne related 17th in 
fatalities due to infection (Hoffmann et al., 2007). In addition to contributing directly to 
human infections, if manure from infected animals is used as fertilizer for produce, infection 
can spread that way (Guan and Holley, 2003) and non-typhoidal *Salmonella* from produce 
ranks 7th for food-borne related hospitalizations and 9th for food-borne related fatalities 
(Hoffmann et al., 2007).

While there are porcine immunizations and vaccines for *Salmonella* (Roesler et al., 
2006 and 2004, and Fagan et al., 2001), there has not been a drop in the rate of *Salmonella* 
infection in humans in the past 20 years in the United States (Hall-Baker et al., 2010). This, 
coupled with the increase in drug resistance of *Salmonella*, indicates that alternate routes to 
prevent food-borne diseases may be a better than drug based methods. One such method is 
to select animals for better innate immune systems. Targeting the improvement of the innate 
immune system can lead to broad pathogen resistance, including not only different kinds of 
bacteria, but extending to viruses, protazoans, and fungi, as it is the part of the immune 
system that recognizes a broad range of pathogens (the other part, the adaptive immune 
system, contains strategies for dealing with specific pathogens, Janeway et al., 2001). 
Selection for better innate immune systems can also lead to generational stability of 
improved immunity, unlike immunizations, which would have to be continuously given.

Using high throughput expression methods, such as microarrays and new sequencing 
methods, to identify differences in genetic expression within a given population in response
to stimuli, in addition to identifying single nucleotide polymorphisms (SNPs) or other sequence variations, it is possible to look at entire functional and structural variations in genomes at once for potential targets for herd improvement. However, to look at targets at the genomic scale, a well-annotated genome and transcripts are essential. Without this high level of annotation, the utility of high throughput measures are limited as important information, such as protein domains or transcription start sites, will only be known for a subset of the genes being measured.

**CURRENT STATE OF PORCINE SEQUENCE ANNOTATION**

The annotation of porcine sequence is improving, but not as good as what is available for other organisms. For example, as of November 15th, 2010, at NCBI’s UniGene database, where similar sequences are clustered together regardless of alternative splicing, human and mouse each have more than 24,000 clusters with genes names other than “Transcribed locus” and humans have over 123,000 total clusters and mouse with over 79,000 clusters; while pig only has 6,791 clusters with names other than “Transcribed locus” with a total of 53,600 clusters. Similarly, for the RefSeq database, where individual sequences are curated and alternative splice variants are filtered into individual RefSeqs, human and mouse have about 40,000 sequences, and the pig has 7,979. Also, while human and mouse have well annotated full genome sequences, the pig’s genome sequence is currently being assembled and then will be annotated, but public release of the annotated genome will not be until 2011. Human and mice also have a greater amount of functional annotation by way of GO terms: human having 12,335 and mouse having 16,370 annotated proteins as part of the GOA project (current to October 19th, 2010). Currently, there is no GOA project for porcine genes. Even looking at the largest source of sequence data, expressed sequence tags (ESTs; short reads
representing full-length transcribed sequences, Tuggle et al., 2003), humans have over 8,000,000, mouse has almost 5,000,000, while the pig has slightly more than 1,600,000 sequences, about three times less that the mouse and six times less than human, at NCBI’s EST database.

For sequences that do not have good annotation, they must rely on sequence similarity, whether full gene sequences or functional subunits, for insight into the possible function (Nagaraj et al., 2006). One of the most commonly used methods to measure sequence similarity is to search for short regions, or local regions, of sequences that match, using variants of the BLAST algorithm (Altschul et al., 2000). This method gives a score based on how similar regions of sequences are (bit score) and gives an estimate of how often you would find a region of random sequences of the size of the region of similarity within a target database of a given size (E-value). While this method is fast and provides a good foundation for annotation based on sequence similarity, it can lead to incorrect annotation if only partial sequence is available. If the sequences are incomplete or are created incorrectly during an assembly, this can lead to lacking important discriminating regions of the gene, like protein domains or cellular localization signals, leading to incomplete or misleading annotations (Sjölander, 2004).

To counter this, short, but numerous, Expressed Sequence Tags (ESTs) can be assembled along with the more rare full-length clones in an attempt to group them nearly identical sequences, called sequence clustering, then to combine the sequences belonging to a cluster into longer consensus sequences, sometimes labeled contigs, called sequence assembly. Currently, there are three sources of porcine assemblies: the Dana Farber Cancer Institute Porcine Gene Index (DFGI PGI; Quackenbush et al., 2000), the Sino-Danish Pig
Genome Project (SD; Gorodkin et al., 2007) and ANEXdb (Couture et al., 2009; see the following chapter). The DFGI PGI was recently updated (March 11th, 2010) and uses sequences downloaded from NCBI’s dbEST and dbCore databases. It is comprised of approximately 1.4 million individual sequences and includes sequences submitted from the Sino-Danish project, which is made of approximately 1 million sequences. ANEXdb used not only all the information from the Sino-Danish project, which has been deposited in both dbEST and TRACE Archive, but also utilized dbEST, dbCore, and TRACE Archive from NCBI, resulting in approximately 2.5 million sequences, with about 800,000 sequences that overlap between TRACE Archive and dbEST. This was done to obtain the greatest number of sequences possible, of both EST and full-length clones. Through sequence assembly, the number of sequences is reduced, for instance ANEXdb has 243,975 unique sequences and DFGI PGI has 240,420 unique sequences, but increases the average length of the consensus sequences: in ANEXdb, the average length of the contributing sequences is 562 bases, while the consensus sequences have an average length of 896 bases. Once a sequence assembly is annotated, expression probes or sequence reads can be mapped to it and can utilize the annotation provided by the assembly.

CLUSTERING EXPRESSION DATA

Similar to clustering sequences together, expression based results can also be clustered. This is done to group genes that show similar responses to the same perturbations, either from internal signals like cell cycle, or external signals like the endotoxin from gram-negative bacteria. Clustering can be done independent of other statistical analysis, such as ANOVAs, when analyzing microarray data, especially when looking across different datasets generated by different technologies. Jenner and Young (2005) used this approach to identify
genes that responded to either bacteria or viruses or both by compiling 32 studies of 77 different host-pathogen interactions, for a total of 785 experiments. They were able to identify 511 common host response genes, i.e. genes with similar expression patterns across many cell types and pathogens, of 5042 unique genes, as well as virus and bacteria and cell specific responses across the experiments.

There are multiple methods for clustering expression data, each using different strategies to group genes together. Some of the most commonly used methods start with the expression values of a gene across different conditions, then join the two most similar gene expression patterns across the same conditions, then the next two most similar until all sets are joined, such as in agglomerative hierarchical clustering (Eisen et al., 1998, Dougherty et al., 2002, Shannon et al., 2003, de Hoon et al., 2004). Others start with all expression sets in one group and start dividing the group by finding the two most dissimilar sets, then the next two most dissimilar until genes belong to their own sets, as in divisive hierarchical clustering and the deterministic-annealing algorithm (Alon et al., 1999, de Hoon et al., 2004). Other methods, like k-means (Dougherty et al., 2002), fuzzy k-means (Gasch and Eisen, 2002, Dougherty et al., 2002), and self-organizing maps (Tamayo et al., 1999, Dougherty et al., 2002, de Hoon et al., 2004), use a random point in the expression sets to group similar sets together, updating the spot when a new member is added to look for other members. Similarly, adaptive resonance theory (Lonosky et al., 2004) and cluster affinity search techniques (Ben-Dor et al., 1999) use the data itself to create the clusters by first making the assumption that each set of values is its own cluster, only adding to it if another set is more similar than a given threshold, if not, a new cluster is formed. Other methods create distributions based on the data to create assumptions about the genes to create the clusters on
(such as Baysian infinite mixture model, Medvedovic and Sivaganesan, 2003, Yeung et al., 2003), and still others map to high dimensional space, then look for cuts in the data to separate out the different clusters, like spectral clustering (Verma and Melia, 2005). Once the genes are grouped together, they can be analyzed for similar function, such as sharing GO (Gene Ontology Consortium, 2000) terms or being within the same KEGG (Kanehisa and Goto, 2000) pathways, as well as identifying their control mechanisms.

PREDICTING TARGETS OF TRANSCRIPTION FACTORS

Genes that cluster together generally share common functions and act in pathways in the cell. This commonality is believed to be due to evolutionarily conserved control over their expression through the sharing of common transcription factor binding regions in their promoters (Alon et al., 1999 and Moreau et al., 2002). There are two general ways of searching for these binding regions: using Position Specific Scoring Matrices (PSSM; sometimes referred to as Position Weighted Matrix, PWM) and *de novo* methods. Position Specific Scoring Matrices (PSSMs) are created from known binding sequences for specific transcription factors, such as what is found at TRANSFAC (Matys et al., 2003) and JASPAR (Sandelin et al., 2004). The other, *de novo*, method requires multiple sequences to search for conserved regions of similar sequence within a larger sequence, such as MEME (Bailey and Elkan, 1994). Two of the benefits of using the PSSM over the *de novo* method are the search time and the ability to use it on individual (Matys et al., 2003) or multiple sequences as input (Frith et al., 2004, Defrance and Touzet, 2006). Search time is reduced due to the matrices already being created and can be quickly scored, whereas in the *de novo* method the equivalent of the matrix has to be discovered within the sequences before scoring. Another benefit for using PSSMs is that they can be used on individual sequences and do not require
multiple sequences for use. The *de novo* method has its advantage in identifying putative binding sequences that do not have a matrix previously created, or for regions of the sequences that may contain multiple transcription factor binding sites (TFBS) in a conserved order to each other (Bailey and Elkan, 1994).

While PSSMs can be used on single sequences, there are several methods that also use them to search across multiple sequences to look for conservation of motifs within a given set. Two of these methods were used in Wang et al. (2008) to look for novel targets of NF-κB, an important transcription factor in innate immunity (see below). One of these methods, CLOVER (Frith et al., 2004) uses random sequences of similar GC composition and length as the sequences of interest to see if TFBS appear more often in the sequences of interest than in the random group. The other, TFM-Explorer (Defrance and Touzet, 2006), uses a distribution generated from the sequences of interest to create an estimate of how often the TFBS of interest would appear at random in that composition of nucleotides, and then compares it to how often it is observed. In addition, it adds location restraints and gives higher scores to TFBS within a certain physical distance of each other using the assertion that transcription factors are more likely to bind in regulatory modules. This requires the user to know where some sort of similar anchor point for the gene is, typically the transcription start site (TSS).

**MULTIPLE DIMENSIONAL CLUSTERING**

While in Wang et al. (2008, Chapter Four) the expression clustering and promoter prediction were done in separate steps, several algorithms have been designed to do this simultaneously. One such program, INCLUSive (Coessens et al., 2003) initially clusters microarray data on the expression of the transcripts across conditions. Following this, the
INCLUSive algorithm then retrieves the upstream promoter sequences for the probe sets to re-cluster the genes. This re-clustering is based on conserved subsequences within the promoter sequences using both a *de novo* and a PSSM method to identify common transcription factors, allowing the creation of subclusters within the expression clusters.

By leveraging both the expression and sequence data, clustering methods have been developed to create regulatory modules. Using microarray expression and a list of known regulatory molecules (both transcription factors and signal transduction molecules), Segal et al. (2003) created regulatory modules based on the changes in expression of the regulators and the rest of the genes on microarrays. While this method, and methods like it, require prior knowledge of regulatory genes, places like KEGG and GO provide easy access to such information. In addition to helping identify regulators, they also used KEGG and GO to provide annotation for the modules by looking for enrichment of terms within the genes comprising the different modules. Segal et al. (2003) also looked at TFBS by *de novo* and PSSM verification in genes in the transcription factors module, providing better evidence of that transcription factor being able to control the expression of that gene. They were also able to look at combinations of regulator genes using this method, and found that by pairing regulators, previously non-significant regulators showed a significant effect on gene expression. Similarly, Seok et al. (2010) showed that adding information to expression data increased the similarity of predicted transcription factors effects and what is seen with chromatin immuno-precipitation (ChIP; pulling down sequence that is bound by a specific transcription factor under a specific condition). However, unlike Segal et al. (2003), their method is limited to only transcription factors and not other regulators, such as signal transduction molecules.
In Hudson et al. (2009), the authors reported that a method was developed to look for transcription factors responsible for difference in phenotypes based on microarray expression and verified it through correctly predicting genes known to cause the differences in phenotypes: the same group later verified the method on four other previously created datasets with known regulatory controls by correctly predicting transcription factors that previously had biological evidence for causing the differences in phenotypes (Reverter et al., 2010).

**BIOLOGICAL VERIFICATION OF PREDICTIONS**

The methods mentioned above generally rely on previously created datasets, not creating their own. Nor were they followed up experimentally to see if the predicted novel genes expression of the transcription factor was effected through knocking down, or out, the ability of the transcription factors regulatory ability. This is most likely due in part to the limit of available techniques or chemicals to target specific transcription factors. However, experimental verification of the bioinformatics predictions through expression or other quantitative methods is becoming more common.

For instance, Lu et al. (2008) used a set of rules to select two genes from a quantitative trait locus (QTL) that affected the weight of the mouse forebrain and then used archived expression data and their own quantitative real-time polymerase chain reaction (q-RT-PCR) to verify the difference in the amount of RNA between two mouse strains that showed the difference. The Reverter et al. (2010) method does appear to be robust, working across many different phenotypes, however, the algorithm has not yet been used to first predict transcription factors responsible for phenotypic variation, and then have the predictions verified through experimental methods.
While few papers have started by prediction from expression data, followed by biological verification, the Institute for Systems Biology (ISB) have begun to use ChIP-on-chip (ChIP followed by hybridizing the pulled down chromatin on a microarray) to verify predicted targets of transcription factors. Gilchrist et al. (2006), showed that ATF3 is responsible for down-regulation of the TLR4, one of the the main receptors for Gram-negative bacterial recognition (see below), response in RAW 264.7 murine macrophage (MΦ) like cells using clustered microarray expression data and motif searching to predict the effect of ATF3, then used ChIP-on-chip and ATF3+/− mice to verify the predicted effect. Then, Ramsey et al. (2008) extended the study to look at how different TLR mutant mice responded to different TLR agonists to predict transcription factors that were responsible for various stimuli and then verified using ChIP-on-chip. Litvak et al. (2009) looked specifically at interplay between three transcription factors, two with known effects, NF-κB initiating stimulation of genes and ATF3 inhibiting stimulated genes, and predicting the effect of C/EBPδ to further increase the expression of NF-κB targets, and then verified their prediction using of various knockout mice and ChIP-on-chip. Similarly to the ISB, Quatrini et al. (2007) used EMSA and other methods to verify targets of the Fur gene in Acidithiobacillus ferrooxidans that were predicted by identifying possible binding sites close the TSS through bioinformatic methods.

Except for one of the verification tests in Reverter et al. (2010), all the others were done on species other than pigs. The typical datasets to test microarray analysis methods on are the yeast dataset created by Gash et al. (2000) and Spellman et al. (1998), which combined have nearly 5,500 citations. The mouse, due to their technical advantages, such as ease of making transgenic or knockdown animals, stable cell lines, well annotated genome
and transcripts, availability of antibodies for their proteins, is another commonly used organism for studies, being the model organism used by the ISB in their studies. Application of transcription factor target prediction or regulatory motif prediction directly to the pig in a high-throughput manner has not been done, and even in Wang et al. (2008) we had to rely on human sequence as a surrogate for pig genes due to the lack of a fully sequenced and assembled genome.

MACROPHAGES AND THEIR ACTIVATION BY ENDOTOXIN

These methods, due to their ability to predict targets of specific transcription factors, would help in immunology because of the complexity involved in recognition of foreign compounds in the body and the careful regulation of the immune system itself. Also, due to the pigs potential as a good model for human physiology, applying these methods to look for regulatory pathways could help in the understanding of human disease, as well as, to find candidate immunity related genes to improve herd health and to prevent the spread of pathogens from pigs to humans, especially *S. Typhimurium*. *S. Typhimurium* is an intracellular parasitic bacterium that invades and uses cells of the immune system, such as macrophages (Mφs), to live and divide. Macrophages are one of the first cells that recognize infection in the innate immune system through the recognition of pathogen-associated molecular patterns (PAMPs) of Toll-like receptors (TLRs) on their surface. There are different TLRs for different PAMPs, and some TLRs even have accessory proteins to increase their sensitivity to specific PAMPs (Ramsey et al., 2008 and Rallabhandi et al., 2006).

One of the TLR responsible for recognition of *S. Typhimurium*, and other Gram-negative bacteria, is TLR4 through the recognition of the lipopolysaccharide (LPS) portion of
the bacterial cell wall. The binding to TLR4 by LPS is currently thought to happen in vivo through another high affinity receptor, CD14. LPS is bound by, and then transferred by the LPS-binding protein (LBP) in vivo to CD14. The CD14/LPS complex then interacts with the TLR4 molecule bound by MD-2. It appears all three proteins, CD14, TLR4, and MD-2, are necessary for full response to LPS (Rallabhandi et al., 2006).

The TLR4 protein, which spans the cellular membrane, is then activated and has its intracellular portion recruit the MyD88/TIRAP (Mal in mice; appears to only aid in recruitment of MyD88 to TLR4) complex and TRAM initiating several pathways. MyD88 has two pathways it contributes to. In one, MyD88 interacts with FADD, which activates CASP8, which can result in apoptosis during inflammation if left unchecked (Aliprantis et al., 2000). Activation and continuation of apoptosis, however, does have several known repressors: activation of p53 by Bcl-2 (Deng et al., 2006), initiation through FADD by FLIP (Haag et al., 2011), the continuation of the caspase cascade by XIAP, cIAP1, and cIAP2 (Vince et al., 2007), and the continuation due to the release of cytochrome C from the mitochondria by Bcl-2 and Bcl-XL (Deng et al., 2006).

The other MyD88 pathway recruits IRAK1/4 and TRAF6 to TLR4, and IRAK4 undergoes autophosphorylation, causing IRAK1 to phosphorylate then IRAK1 hyperphosphorylates, most likely through another round of autophosphorylation, causing it to disassociate from MyD88 but not TRAF6, while IRAK4 remains behind. The IRAK1/TRAF6 complex then interacts with a membrane bound complex of proteins: MAP3K7/MAP3K7IP1/MAP3K7IP2. IRAK1 remains behind to become poly-ubiquitinated (either being degraded or remaining to form new complexes), while TRAF6/MAP3K7/MAP3K7IP1/MAP3K7IP2 leaves the membrane and translocates to the
cytosol. This complex then interacts with an E2 ubiquitin-conjugative enzyme complex (Ubc13/Uev1A), causing TRAF6 to become poly-ubiquitinated and associating with MALT1, BCL10, and Pellino2. Following this, MAP3K7 autophosphorylates and phosphorylates MAP3K7IP1, causing MAP3K7 to become an active kinase. MAP3K7 targets IKKβ, causing the IKK complex (comprised of IKKα, IKKβ, and IKKγ) to active and phosphorylating IκBα. IκBα is bound to NF-κB, keeping it from activating and translocating into the nucleus to bind DNA. The phosphorylation of NFKBIA causes it to be marked for degradation, releasing NF-κB and exposing its nuclear import signal (Verstrepen et al., 2008, Figure 1). In addition to phosphorylating IκBα, IKKβ can also phosphorylate on of the subunits of NF-κB, p65, at a specific amino acid, which can affect its transcriptional activity. This phosphorylation appears to be specific to IKKβ, while other kinases phosphorylate p65 at other amino acids (Kishore et al., 2003). Activated MAP3K7 complex can also activate the MAPK signaling pathway, resulting in the phosphorylation of both p38 and JNK, both of which activate AP-1 (Brown et al., 2010, Figure 2).

In addition to the MyD88 dependent pathway, TLR4 also has a MyD88 independent pathway to activate NF-κB through TRAM and TRIF. This pathway becomes activated during the endocytosis of TLR4 and after the MyD88 dependent pathway. TRIF associates with TRAF6 and RIPK1, activating TRAF6 and ultimately resulting in NF-κB activation (Verstrepen et al., 2008). In addition to NF-κB, IRF3 and IRF7 are activated through this pathway. TRIF associates with TRAF3, causing the activation of TBK1 and its association with IKKe. The TBK1/IKe dimer then phosphorylates both IRF7 and IRF3, causing interferon expression (Oganesyan et al., 2006).
KNOWN EFFECTS OF NF-κB IN THE INFLAMMATORY RESPONSE

The activation and translocation of NF-κB, AP-1, IRF3, and IRF7 into the nucleus of MΦs activate multiple intracellular signaling pathways. NF-κB is considered a master regulator in LPS induced inflammation (Sharif et al., 2007). NF-κB is known to have more than 150 target genes that it is at least partially responsible for their increase in expression due to some sort of stimulus. These target genes have functions such as being cytokines/chemokines, immune receptors, antigen presenting and preprocessing, cell adhesion, complement and acute inflammation phase responders, oxidative stress responders, cell-surface receptors, enzymes, other transcription factors, growth factors, apoptosis regulation, and others (Pahl, 1999, www.nf-kb.org, bioinfo.lifl.fr/NF-KB/). However, NF-κB, as pointed out by Litvak et al. (2009) and in the above section of the TLR4 activation of MΦ, does not act on its own, i.e. NF-κB induced C/EBPδ for long term, full activation, yet also induces ATF3, which in turn turns off NF-κB targets in a predictable network, so while it may be responsible for a given amount of changes in expression, it may not directly account for all the changes in expression seen given a stimulus.

In fact, after the initial recognition of LPS by TLR4, other cascades initiate through other receptors, like TNFR1. TNFR1 is one of the classic death receptors as it typically signals apoptosis, with its ligand being one of the targets of NF-κB: TNF. TNFR1 generally starts an intracellular cascade that initiates apoptosis through FADD, the same protein MyD88 can bind to and initiate the caspase cascade through CASP8. In addition, TNFR1 can also initiate the formation of reactive oxygen species (ROS) through TRADD and RIPK1. ROS creation in MΦs is one of the mechanisms of killing bacteria; hence ROS
production is an important step of innate immunity. However, reactive oxygen is a dangerous molecule to have around as it can cause damage to macromolecules, such as proteins and DNA, so a careful balance has to be maintained with ROS production to keep the cell from apoptosis. If the level gets too high, reactive oxygen can signal through the activation of JNK and the inhibition of the MAPK pathway through oxidation. JNK activates ITCH, which causes the ubiquitzaion and degradation of CFLAR, a target of NF-κB and a caspase pathway inhibitor, allowing the cleavage of pro-caspase molecules into their active forms, thereby initiating apoptosis. JNK also causes the downstream cleavage of BID, another signal for apoptosis at the mitochondrial level. While ROS, though currently unknown methods, also appears to be able to down regulate NF-κB, it is thought that ROS can directly cause the oxidation of NF-κB, preventing its binding to DNA. It has also been shown that ROS is required for full caspase pathway activation (Bubici et al., 2006 and Simon et al., 2000, Figure 3).

There is evidence that NF-κB can inhibit the effect of ROS and JNK within the cell, which increases a cell’s likelihood of surviving inflammation as these two pathways are pro-apoptotic. Several targets of NF-κB inhibit different stages of ROS formation as well as effect the activation of JNK. For example: GADD45B can bind to and inhibit the MAP2K7, which is the main kinase responsible for activation of JNK downstream from TNFR1. Other kinds of inhibition are not well understood yet, however, XIAP and A20 also seem to exert an effect on the JNK pathway and both are inducible by NF-κB. FHC will form complexes with other proteins to create ferritin and bind iron in cells, with iron being the primary metal ion used to microchondrial ROS (mainly ·O₂−). SOD2, one of the superoxide dismutases, directly interacts in the ROS pathway and converts ·O₂− into H₂O₂, which, without iron (the
function of FHC), is disposed of by other enzymes: peroxidases and catalases. NF-κB also targets the apoptosis inhibitory genes Bcl-2 and Bcl-XL, which prevent the release of cytochrome c from the mitochondria and preventing the activation of CASP9 (Bubici et al., 2006, Simon et al., 2000, and Pahl 1999). It is mostly likely a balance between these four systems, JNK singaling, ROS formation, caspase activation (which creates a positive feedback loop for apoptosis), and NF-κB that determines if a cell will survive during an inflammation.

CHEMICAL INHIBITION OF NF-κB

Due to the wide range of pathways affected by the known NF-κB regulated genes; NF-κB is a potential target for drug development in order to control inflammation and cancers, both in which apoptosis plays a key part. As detailed above, the regulation of NF-κB can be at several points: preventing NF-κB from translocating into the nucleus, oxidation of important amino acids on NF-κB, preventing its disassociation from NFKBIA, preventing IκB from degradation directly, or preventing IκB degradation indirectly by impairing the functionality of the IKK complex, specifically IKKβ. Going any higher in the regulatory network could also impact other pathways, which may lead to confounded results due to the ability of MAP3K7 to target other proteins than IKBK.

One small molecule, SC-514, selectively targets and inhibits IKKβ, while not inhibiting at least 30 other tested kinases, including other kinases in the TLR4 pathways. It has been shown that SC-514 reduces the phosphorylation of NFKBIA. It is thought that SC-514 causes the IKBKB deactivation by preventing IKKβ autophosphorylating its C terminus, allowing it to provide phosphate during the MAP2K pathway, without effecting the final
inactivation of the IKKβ by phosphatase(s). However, this has not yet been confirmed.

Unlike dominant negatives of IKKκ and IκBα, in which high concentrations are normally used, to completely remove p65 translocation, SC-514 fails to do so, instead causing a dose-dependent delay and significant inhibition in NFKBIA degradation. Using a small molecule, however, does not disrupt the IKBK complex, whereas a dominant negative version could, especially in the dosages normally used for their expression. It has also been shown that SC-514 increases the nuclear export of p65 by NFKBIA, as well as preventing the IKKβ phosphorylation of p65, further reducing the effectiveness of p65 on gene expression in an IKKβ specific way (Kishore et al., 2003).

SC-514 is also effective in multiple tissues, cell lines, and in vivo by reducing the effect of NF-κB due to various stimuli. More than 20 different tissues, including three different MΦ cell lines, have had SC-514 used to inhibit the effect of NF-κB induction by at least 15 different treatments, including LPS and RAW 264.7 cells (Clavel et al., 2010, Hwang et al., 2010, Koo et al., 2010, Kuntz et al., 2010, Moreno et al., 2010, Yarmo et al., 2010, Zhou et al., 2010, Gagnon et al., 2009, Lu et al., 2009, Motagally et al., 2009, Thompson and Van Eldik, 2009, Antunes et al., 2008, Choo et al., 2008, de Oliverira et al., 2008, Jeong et al., 2008, Voronov et al., 2008, Hwang et al., 2007, Killeen et al., 2006, Gomez et al., 2006, Jeong et al., 2005, Peng et al., 2005). Some studies have used not only SC-514, but also other inhibitors specific to the other TLR4 pathways to start to sort out which genes are responding to which pathway (Moreno et al., 2010 and de Oliverira et al., 2008). It has also been used to guarantee the effect of NF-κB is not being activated to study other pathways (Clavel et al., 2010). However, it has never been used in pigs, nor has it ever
been used for assessing regulatory pathway predictions from microarray studies, typically only for ChIP and PCR experiments.

USE OF HIGH THROUGHPUT EXPRESSION ANALYSIS METHODS IN PIGS

There are currently two datasets, but over 130 data series for pigs stored in NCBI’s Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/). Among these are immune related datasets to various infections, both viral and bacterial, and in different tissues. The response to PRRSV in alveolar macrophages (GSE10346), lung and lymph node tissue (GSE12194), two different types of \textit{Salmonella enteritidis} serovars (Typhimurium [GSE3713] in mesenteric lymph node and Choleraesuis in mesenteric lymph node [GSE7314] and lung tissue [GSE2339]), \textit{Streptococcus suis} in brain, lungs, and PBMCs (GSE24889), \textit{Haemophilus parasuis} in lung (GSE19126) and spleen (GSE11787), \textit{Actinobacillus pleuropneumoniae} in lung (GSE13134, GSE4577), liver (GSE11404, GSE4577), and lung lymph nodes (GSE4577), porcine circovirus type 2 in mediastinal lymph nodes (GSE14758), \textit{Brucella suis} (GSE17492), and psudorabies in a porcine cell line (GSE8676 and GSE9259) have all been looked at for changes in gene expression using various microarrays. But while these datasets and lists of differentially expressed genes and pathways exist, there are few analyses that directly investigate predictions of regulation, or even potential regulation, of these genes and pathways.

Through the understanding of regulatory mechanisms, whether they are a transcription factor like NF-κB or another member of an intracellular signaling cascade like IKBKB, improvements can be made to porcine populations due to the discovery of potential for drug targets or improvement through genetic selection. However, the ability to fully use
the above methods in pigs is currently limited due to the little direct annotation of porcine
genes and the lack of well-annotated genome, but through the leveraging of other species it is
possible to gain insights in how pigs work.
REFERENCES


Moreno, L., McMaster, S.K., Gatheral, T., Bailey, L.K., Harrington, L.S., Cartwright, N.,
oligomerization domain 1 is a dominant pathway for NOS2 induction in vascular
smooth muscle cells: comparison with Toll-like receptor 4 responses in macrophages.

necrosis factor α activates nuclear factor κB signaling to reduce N-type voltage-gated


Immunization in Reducing Salmonella typhimurium Infection in Chickens. Poult Sci.
77, 1874-1883.

Nagaraj, S.H., Gasser, R.B., Ranganathan, S., 2006. A hitchhiker’s guide to expressed
sequence tag (EST) analysis. Brief Bioinform. 8, 6-21.

Oganesyan G., Saha, S.K., Guo, B., He, J.Q., Shahangian, A., Zarnegar, B., Perry, A., Cheng,
independent antiviral response. Nature. 439, 208-211.

Oncogene. 18, 6853-6866.

IgE-induced TNF production by mast cells: a role for the IKK-IκB-NF-κB pathway
in IgE-dependent mast cell activation. J Leucok Biol. 77, 975-984.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea,
G., Sultan, R., White, J., 2000. The TIGR Gene Indices: reconstruction and
representation of expressed gene sequences. Nucleic Acid Res. 28, 141-145.

Bioinformatic prediction and experimental verification of Fur-regulated genes in the
extreme acidophile Acidithiobacillus ferrooxidans. Nucleic Acids Res. 35, 2153-
2166.

Rallabhandi, P., Bell, J., Boukhalvalova, M.S., Medvedev, A., Lorenz, E., Arditi, M.,
Polymorphic Variants: New Insights into TLR4/MD-2/CD14 Stoichiometry,

Ramsey, S.A., Klemm, S.L., Zak, D.E., Kennedy, K.A., Thorsson, V., Li, B., Gilchrist, M.,
Gold, E.S., Johnson, C.D., Litvak, V., Navarro, G., Roach, J.C., Rosenberger, C.M.,
Rust, A.G., Yudkovsky, N., Aderem, A., Shmulevich, I., 2009. Uncovering a
Macrophage Transcriptional Program by Integrating Evidence from Motif Scanning
and Expression Dynamics. PLOS Comput Biol. 4, e1000021.

Regulatory impact factors: unraveling the transcriptional regulation of complex traits
from expression data. Bioinformatics. 76, 896-904.

Rogers, C.S., Stoltz, D.A., Meyerholz, D.K., Ostedgaard, L.S., Rokhlina, T., Taft, P.J.,
Rogan, M.P., Pezzulo, A.A., Karp, P.H., Itani, O.A., Kabel, A.C., Wohlford-Lenane,
C.L., Davis, G.J., Hanfland, R.A., Smith, T.L., Samuel, M., Wax, D., Murphy, C.N.,


FIGURE LEGENDS

**Figure 1.** NF-κB signaling due to LPS recognition at TLR4 extracellular complex. Taken from Vestrepén et al. (2008, Figure 2).

**Figure 2.** TLR4 signaling of NF-κB, p38, and JNK. Taken from Brown et al. (2010, Figure 2).

**Figure 3.** TNFα, upregulated due to LPS, recognition activating reactive oxygen species. Taken from Bubici et al. (2006, Figure 1).
Figure 1. Vestrepen et al., 2008.
Figure 2. Brown et al., 2010.
Figure 3. Bubici et al., 2006.
CHAPTER 3: ANEXDB: AN INTEGRATED ANIMAL ANNOTATION AND MICROARRAY EXPRESSION DATABASE

Modified from a paper published in *Mammalian Genome*¹

Oliver Couture², Keith Callenberg, Neeraj Koul, Sushain Pandit, Remy Younes, Zhi-Liang Hu, Jack Dekkers, James M. Reecy, Vasant Honavar, Christopher Tuggle³

ABSTRACT

To determine annotations of the sequence elements on microarrays used for transcriptional profiling experiments in livestock species, currently researchers must either use the sparse direct annotations available for these species, or create their own annotations. ANEXdb (http://www.anexdb.org) is an open-source web application that supports integrated access of two databases that house microarray expression (ExpressDB) and EST annotation (AnnotDB) data. The expression database currently supports storage and querying of Affymetrix-based expression data, as well as retrieval of experiments in a form ready for NCBI-GEO submission; these services are available online. AnnotDB currently houses a novel assembly of ~2.5 million porcine expressed sequences called the Iowa Porcine Assembly (IPA), which consists of 140,087 consequences sequences, the Iowa Tentative Consensus (ITC) sequences, and 103,888 singletons. The IPA has been annotated via transfer of information from homologs identified through sequence alignment to NCBI RefSeq. These annotated sequences have been mapped to the Affymetrix porcine array elements, providing annotation for 22,569 of the 23,937 (94%) of the porcine specific probe sets, of which 19,253 (80%) are linked to an NCBI RefSeq entry. The ITC has also been

1 Reprinted with permissions of *Mammalian Genome*, 2009, 20, 768-777.
2 Primary author and researcher
3 To whom correspondence should be addressed: cktuggle@iastate.edu
mined for sequence variation: providing evidence for up to 267,968 SNPs, 54,915 deletions, and 576 insertions in porcine expressed sequence. These results create a single location to obtain porcine annotation of, and sequence variation in, differently expressed genes in expression experiments, thus permitting possible identification of causal variants in such genes of interest. The ANEXdb application is open source and available from SourceForge.net.

Keywords: porcine, RNA expression, database, microarray, annotation, EST assembly, SNP

INTRODUCTION

Microarray and other high throughput expression platforms can provide a vast amount of information about transcriptional products within a biological sample under a variety of conditions. It is essential that functions be assigned to these transcripts so that transcriptome data can be fully utilized to explore important biological questions. Many human and mouse transcripts have functional annotations (for instance, currently there are 11,616 manually annotated human proteins and 14,182 manually annotated mouse proteins with GO terms through the GOA project; see www.ebi.ac.uk/GOA/human_release.html and http://www.ebi.ac.uk/GOA/mouse_release.html for results on both manual and automated GO annotations, released July 2009); most of these are indirect annotation predicted from structural similarity to genes and proteins in prokaryotes and lower eukaryotes. However, for species such as the pig, relatively little direct or indirect annotation is available for most expressed sequences (ESTs). One approach that can be used to address this problem is to leverage sequence annotations from the better-annotated species, e.g. mouse, to infer annotations of their homologs in other species, e.g. pig (Nagaraj et al. 2006).
Such an approach to porcine EST annotation requires that as many porcine ESTs as possible be assembled into putative consensus transcripts. There are currently two sources of porcine EST assemblies: the Dana Farber Cancer Institute Porcine Gene Index (PGI) (originally housed at The Institute for Genome Research (TIGR); Quackenbush et al. 2000) and the Sino-Danish Pig Genome Project (SD) (Gorodkin et al. 2007). These two resources, however, do not fully exploit all available sequence data: PGI uses the public data within NCBI’s dbEST (Boguski et al. 1993) and dbCore, but not TRACE (Wheeler et al. 2008). The TRACE database is a repository for the SD raw sequence data, which, as the largest single source of pig cDNA sequence data to date, is a significant resource for porcine. Conversely, the SD assembly utilized less than half of the public porcine sequences. Other sequences databases, e.g. PEDE (Uenishi et al. 2007) and PiGenome (Lim et al. 2007), while useful for other purposes, do not provide comprehensive coverage of porcine ESTs.

Once the sequences are assembled, sequence similarity is improved due to longer sequences containing more structural information, such as multiple protein domains, than a single EST, which may only contain one, or part of a protein domain. Hence, an assembled sequence can be more reliably used to leverage the better-annotated species for the less well-annotated species (Sjölander 2004). Not only can homologies be used to transfer annotation from one species to another, they can be used in cross-species comparisons to test if the same structural homolog has similar characteristics such as expression pattern or protein function in different species, or if its character is unique to a given species.

The domestic pig, which has significant EST data and a developing genome sequence, has been used as a model for a number of investigations, including cystic fibrosis (Rogers et al. 2008) and tissue scarring (Zhu et al. 2008; Gallant-Behm et al., 2007). The pig
genome has been modified for possible use in xenotransplantation (Lai and Prather 2002), but does not have well established annotations for most genes due to lack of a draft genome sequences. This makes it challenging to analyze the underlying pathways in such investigations. To fully utilize these models in large-scale transcriptomic studies, it is essential to create the best annotation of available ESTs.

Against this background, ANEXdb, an open sourced web application that supports integrated access of two databases, ExpressDB and AnnotDB, that house a) microarray expression and EST assembly, and b) annotation data, respectively, was created.

MATERIALS AND METHODS

ANEXdb Application Implementation

ANEXdb was implemented using a LAMP (Linux, Apache, MySQL, PHP/Perl) system: utilizing Red Hat Enterprise Linux (RHEL) 4, Apache 2.0.53, MySQL5, and PHP5. The administrator uses Perl scripts to upload data into AnnotDB, to download the sequences via integration of NCBI’s eUtils (http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html), and to control the assembly through the TGI Clustering tools (TGICL, Quackenbush et al. 2000, available from http://compbio.dfci.harvard.edu/tgi/software/). PHP scripts are used to control the data flow and analysis in ExpressDB via a Flash applet, a Java application for data migration, to upload data and run R procedures from the Bioconductor library for MAS/RMA calculations. A web based query interface was created with PHP5, which allows users to query AnnotDB with a NCBI sequence ID, consensus ID, and Affymetrix probe ID; experiment, hybridization, or sample information in the case of ExpressDB (Figure 1).
Currently, ExpressDB houses microarray expression data only from the Affymetrix platform. Support for Affymetrix was chosen since such data is available for many species, including porcine, bovine, chicken, human, and mouse, and is also inherently comparable across experiments (Mongan et al. 2008). A submission interface implemented with PHP allows users to submit data from their microarray experiments to a temporary submission database. Such data includes information about individual hybridizations, such as sample information, i.e. organism and sample source, and technical information, i.e. hybridization and treatment protocols, and is MIAME compliant. An administrator (via an admin interface) verifies and approves the submission (a process called finalization). Once a submitted experiment is finalized, the system uses Bioconductor packages (Gentleman et al. 2004) to calculate MAS/RMA data for the experiment and transfers this and the previously submitted data to a final database that only an administrator can alter. The system can also output the data from a finalized experiment in the SOFT format for submission to GEO (Wheeler et al. 2008) by using a Java application that is invoked from the PHP interface. This application constructs an object model (an in-memory data representation) and uses the Apache Velocity template engine (http://velocity.apache.org/) to generate the SOFT output from the object model (Wheeler et al. 2008). The SOFT output, along with the .cel and .chp files are then zipped together into a single file that can be downloaded and submitted to GEO.

AnnotDB is designed to house individual or assemblies of sequences in a form that facilitates annotation of transcripts in the target species. Perl scripts were written to parse relevant data and to upload that data into AnnotDB. Such data includes mappings of NCBI Gene database to GO terms, KEGG pathways, and the RefSeq database, as well as Pfam
mappings to GO terms to help annotation of the sequences. To provide a basis for annotation AnnotDB also includes several analyses done on the sequences from the assembly: sequence alignments (using BLAST and Exonerate), ORF prediction, and SNP analysis. To query AnnotDB, users use the same PHP interface that is used for ExpressDB. AnnotDB also has a MySQL guest account, which allows users access to the complete database so that they can perform database queries using Perl or another scripting language. This was not created for ExpressDB due to the potential private nature of some of the data.

**Sequence Assembly**

Currently within AnnotDB, 2,529,315 public porcine ESTs have been assembled into consensus sequences (contigs). This assembly was compiled from NCBI’s dbEST 1,475,958 sequences, 18,157 sequences from dbCore, and 1,035,200 sequences from the TRACE Archive (February 2008). All sequences were downloaded in FASTA format and cleaned using the included SeqClean program (available from [http://compbio.dfci.harvard.edu/tgi/software/](http://compbio.dfci.harvard.edu/tgi/software/)) and two additional sequence files. One file was the sequences found in UniVec (Wheeler et al. 2008), while the other contained porcine specific long and short repeated sequences created by querying Nucleotide at NCBI (available from [http://www.anexdb.org/download.php](http://www.anexdb.org/download.php)). Using these two files, SeqClean functions removed or trimmed sequences that were vector sequence, regions of low complexity, repeated sequences, and any sequence less than 100 bases long. The removal of short sequences is performed twice: once prior to trimming the sequences, then after trimming poor sequences off of good sequences. This resulted in 2,369,608 clean sequences, of which 1,144,310 were trimmed, that were available for assembly.
The original FASTA annotation lines for each sequence were then mined for evidence that the sequence represented a full-length sequence by searching for “complete CDS” or “full length mRNA” keywords. Entries with such annotations were given the TGICL based full-length “et|” designation in their annotation line and were then used for seeded clustering (see below).

Briefly, TGICL runs an alignment algorithm similar to megablast (Zhang et al. 2000) to calculate the overlap between each sequence using a given cutoff, and stores the alignment data in tab-delimited files. These files are then used to cluster the sequences into groups via a process called transitive clustering (tclust). Transitive clustering begins with the highest aligning pair and then groups other sequences that also align to either member of the pair. This is repeated with all sequences until none are left that align to any member in the cluster above the given threshold. An optional method, which requires sequences to be assigned the “et|” in their annotation line, is called seeded clustering (sclust). This starts with the “et|” sequences as a basis for a full-length and complete transcript and uses a stricter cutoff for sequences that overhang either end of the full-length sequence. We used sclust only on clusters with more than 20,000 individual sequence members. After the clustering, TGICL then uses CAP3 (Huang and Madan, 1993) to assemble the clusters into contigs, and also performs a more stringent alignment that may cause a single cluster to create multiple contigs. Because TGICL calculates fewer overall alignments, this method is faster than the alignment and clustering procedure used by CAP3, which performs both at the same time.

**Assembly Annotation**

The IPA was aligned against the RefSeq RNA database (Pruitt et al. 2007) using BLASTN (cutoff E-value ≤ 1e-10, downloaded Dec. 2008, Altschul et al, 2000), and against
RefSeq Protein (cutoff E-value \( \leq 1e-5 \), downloaded Feb. 2009) and Pfam (Bateman et al., 2002) databases (cutoff E-value \( \leq 1e-5 \), downloaded Dec. 2008) using BLASTX. They were also aligned with human chromosomes with Exonerate (est2genome model, requiring at least 60% of the sequence to be aligned, with at least one HSP with a minimum score of 100, Slater and Birney 2005). Using the resulting alignments to RefSeq, the RefSeq accession numbers were mapped to their NCBI Gene IDs (Wheeler et al. 2008) using the gene2refseq flat file (downloaded Dec. 2008) from NCBI, which allowed the GO (The Gene Ontology Consortium 2000) and KEGG (Kanehisa and Goto 2000) annotations to be transferred to the IPA from their homologues using gene2go (available from NCBI, downloaded Dec. 2008). Likewise, the GO terms associated with each Pfam were transferred to the IPA using the pfam2go (downloaded Dec. 2008) file from GO and the resulting alignments.

Similarly, to map the Exonerate alignments to the corresponding human gene, the gene2refseq file was used. This provided the chromosomal interval covered for each RefSeq ID and its corresponding Gene ID. These reported locations were then compared to the Exonerate alignments of the porcine sequences to the human chromosome. To allow for evolutionary variation in gene size, the human chromosomal interval for each RefSeq was extended 250 base pairs in each direction along the chromosome (an additional 500 base pairs). This overall interval was then compared with the interval along the human chromosome, which was aligned with the porcine sequence by using Exonerate, requiring the interval aligned with the porcine sequence to be within the extended human RefSeq interval. To validate the annotations, the Gene IDs of the top human hits transferred the two RefSeq BLASTs (BLASTN and BLASTX) were compared with each other as well as the Gene IDs retrieved from Exonerate for each of the top hits.
Open Reading Frame (ORF) Prediction

Using a Perl script, each sequence was submitted one at a time to the ORF Finder program (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) using the Standard genetic code. Perl scripting was used to extract all predicted reading frames in each of the six frames (if available) from the resulting output and inserted into the database. Open Reading Frames were predicted for all sequences in the entire IPA.

Array Mapping

BLASTN was used to align both the Affymetrix porcine target and the individual 25-mer probe sequences to the entire IPA. A cutoff of E-value $\leq 1e^{-5}$ was used for the target sequences, while, due to the shortness of the 25-mer, a cutoff of E-value $\leq 0.05$ was used for these sequences. For direct sequence comparisons between Affymetrix elements and IPA entities, the Affymetrix target sequences were blasted against the same RefSeq database at the same cutoff.

Assembly Sequence Variation and Comparison to Porcine dbSNP

To detect sequence variation from the assembly results, a Perl script was used to mine the ACE output file from CAP3 for differences between the consensus sequence and the sequences that contributed to each location along the consensus sequence. Variant frequency was calculated using only the number of contributing sequences at the specific nucleotide location, not the overall number of contributing sequences for the entire consensus sequence.

To compare the SNPs identified from this data set to the porcine SNPs available in dbSNP at NCBI, the quarterly updated FASTA files were downloaded from dbSNP, joined, and then aligned using BLASTN with a match cutoff E-value $\leq 1e^{-5}$ (due to their short length and further criteria used, see below) to the ITCs. This created a subset of SNP-containing
sequences from dbSNP that could be found in our assembly. This subset of sequences from dbSNP were then mined to compare the exact location of the SNP within the submitted dbEST sequence to the corresponding base in the ITC. A percentage cutoff of the length of the dbSNP sequence in the alignment from 20 to 100%, as well as the minimum number of minor allele reads, were both utilized as further criteria in the analysis. The ITC accession number was then used to query the sequence variation table within the database to determine if that specific base was reported to contain a SNP or not. A percentage cutoff of the length of the dbSNP sequence in the alignment from 20 to 100%, as well as the minimum number of minor allele reads, were both utilized as further criteria in the analysis. Also, since the ITCs are expressed sequences, but as dbSNP contains SNPs from both cDNA and genomic DNA sources, the FASTA annotation lines in dbEST were mined for the term “cDNA” to label SNPs specifically in expressed sequences. However, both genomic and cDNA clones were aligned and analyzed.

RESULTS

**Assembly and Annotation of Porcine Expressed Sequences**

Our assembly of all publicly available porcine expressed sequences, called the Iowa Porcine Assembly (IPA), consists of 140,087 consensus sequences (contigs), called the Iowa Tentative Consensus (ITC), and 103,888 singletons. In comparison, the PGI contains 104,293 contigs and 133,455 singletons. A comparative analysis of the coverage of human RefSeqs by the IPA using BLAST showed that the IPA matched as many of the human sequences as did mouse RefSeqs: 72% vs. 73% respectively (see Table 1 and Supplemental Figure 1). For this BLAST analysis and all others in the following descriptions, we count a
BLAST alignment for those results showing a BLASTN E-value no larger than 1e-10 and a
BLASTX E-value no larger than 1e-05.

Of the 243,975 total sequences in the IPA, 191,602 (79%) have a BLASTN alignment
to the RefSeq RNA database. The number of alignments to the RefSeq Protein and Pfam had
a lower hit count: 71,332 (29%) and 76,386 (31%) respectively. However, the transfer of
GO terms (from the two RefSeq databases and Pfam) and of KEGG terms (from the RefSeq
databases) provided 166,119 and 92,263 sequences (68% and 38%), respectively, with
additional functional information (see Table 2 for a summary of annotations).

The RefSeq annotations of the IPA sequences rely on the validity of the BLAST
results between the IPA and RefSeq. To provide further evidence of these relationships, we
compared the results of BLASTN and BLASTX hits to RefSeq with Exonerate alignments of
the pig to human genome sequence. We found that the majority of the top scoring
alignments overlapped: 80.1% of the top Exonerate alignments agreed with the top scoring
hits returned by at least one of the two BLAST algorithms. As well, 82% of the top
BLASTX hits agreed with the top BLASTN hit, regardless of Exonerate results (see Figure
2).

A BLAST analysis of Affymetrix porcine GeneChip® target sequences to the IPA
provided alignments for 22,569 of the 23,937 (94%) of the probe sets, of which 19,253
(80%) had an IPA to NCBI RefSeq alignment. This resulted in 1,293 (5.4%) more probe sets
that aligned to RefSeq RNA than a direct BLASTN alignment of the Affymetrix target
sequences to RefSeq RNA, which resulted in 17,960 alignments. The IPA-RefSeq
alignments were also longer on average, which creates a higher score; with the IPA having an
average BLASTN score of 1,244 versus an average score of 392 for the direct Affymetrix target sequence BLASTN against RefSeq.

**Sequence Variation Inferred from Overlapping Sequences in AnnotDB**

The ACE files containing the CAP3 assembly output for the ITCs were also analyzed for sequence variation. This analysis yielded a maximum of 2,004,432 candidate SNPs, 14,536 candidate insertions, and 428,637 candidate deletions in 63,995 ITCs with putative variation using the minimum requirement of one minor allele variant. Although these numbers decreased as the number of minor allele reads required for declaration of variation increased, we were able to provide evidence for many SNPs; for example, nearly 40,000 SNPs were predicted in 4,441 different ITCs, even when 10 sequences were required to show the minor allele variant (Table 3).

It is important to evaluate the accuracy of these predictions. We approached this by comparing our predictions to available information in dbSNP, for sequences in common with ITC sequences. First, we identified those sequences entries in dbSNP that match any SNP-containing assembled contigs using BLAST. Then we determined whether the dbSNP entry has the exact SNP location as predicted in our assembly. Through the comparison of dbSNP to our assembly (requiring a minimum of a 60% overlap length of the sequence from the dbSNP, and at least two minor allele reads) we were able to confirm the majority of SNPs that were reported in regions in common between our predicted SNPs and those reported in dbSNP. Specifically, of the total number of porcine SNPs in dbSNP (34,508 including both genomic and cDNA entries), 5,010 had an alignment to an ITC, of which 4,244 (85%) were found to match putative sequence variation present in an ITC. When considering SNPs annotated as found only in cDNA sequences, this percentage is 79% (161 out of 203
alignments) using the same criteria. When the criteria for declaring a putative SNP as matched to the public SNP data is reduced to the least stringent (only a single read of the minor allele is required, with any amount of sequence alignment), the percent of matches decreases only slightly to 80% of the total porcine dbSNPs (7,695 out of 9,583), and 75% (310 out of 414) for the cDNA SNPs. Increasing the minimum number of minor alleles to three or increasing the minimum alignment length from 60% to 100% made little impact on these percentages. (see and Table 4 and Supplemental Figure 2 for full results).

We also compared our SNP predictions to the available SD SNP data by using the SNP data reported on the SD website which requires at least three reads of the minor allele. We found a slightly higher number of ITC sequences with SNPs: 16,253 (ITC) vs. 14,120 (SD). However, there are more SNPs within the ITCs as compared to those of the SD consortium: 267,968 (ITC) versus 54,828 (SD) (Table 3).

DISCUSSION

Assembly and Annotation of Porcine Expressed Sequences

When our IPA is compared to the SD assembly (version 1, Gorodkin et al. 2007), which contains 48,629 contigs and 73,171 singletons, and the DFCI Pig Gene Index assembly (PGI, version 13, assembled July 2008), with 104,293 contigs and 133,455 singletons, the IPA has a higher number of contigs, while having fewer singletons. This is most likely due to our inclusion of more initial sequences, which increased the ability to cluster more of the starting sequences together into sequence clusters and then into contigs. In addition, unlike the SD assembly, we did not filter out ribosomal or mitochondrial mRNAs from the starting sequences. While we produced more contigs than what is currently thought to be the number of genes, this is expected due to the software distinguishing alternative
splice products into distinct contigs or singletons. This has been documented for other assemblies; the human Gene Index at Dana Farber, for example, starts with more than seven million individual ESTs to create a Human Gene Index with over one million sequences. Similar results are seen with the mouse assembly. It is also likely many of these currently independent sequences will be merged once additional porcine sequence data becomes available.

The Iowa Porcine Assembly covers as much of the human RefSeq database as the mouse RefSeq database (see Table 1 and Supplemental Figure 1). This indicates that, relative to human sequences, the IPA is as complete as the mouse RefSeq database. Also, due to the large percentage of the sequences in the IPA aligning to the entire RefSeq database (80%), putative gene annotation is provided to the majority of the IPA. By using Exonerate to map IPA sequences to the human chromosomes, it was also possible to get putative exon boundaries for 33,016 of the sequences in the IPA. When the porcine genome is complete and available, this result will be updated. In addition to the direct sequence comparison to the RefSeq and Pfam databases, a large number of the sequences were annotated at varying levels by linking them to GO terms or KEGG pathways through using data from the Gene and Pfam databases. Thus we believe this annotation to be as complete as possible using available comparative information, and these data should be useful to researchers interested in annotation porcine sequences such as those obtained in transcriptional profiling experiments.

Array Mapping to Annotate Probeset Elements on the Affymetrix Porcine GeneChip®

An important goal of our work was to update the functional annotations of the porcine Affymetrix GeneChip® probesets. Through the use of homolog mapping to annotate
Affymetrix GeneChip® elements, it is also possible to obtain a better understanding of cross-species similarities and differences in transcriptional profiling experiments as the effects of structurally similar genes can be compared under similar experimental conditions. By aligning the Affymetrix target sequence to the IPA, then the IPA to RefSeq, we obtained a higher average BLAST score than by aligning the Affymetrix target sequences directly to RefSeq. We also found more overall alignments.

Our method is similar to that reported by Tsai et al. (2006), but differs in a few ways. One is that we only use our consensus sequences to map to RefSeq entries instead of a mix of the Affymetrix target sequences and the PGI assembly while it was housed at TIGR (Quackenbush et al., 2000). We also used RefSeq instead of Ensembl (cDNAs are now a part of EMBL-Bank), as RefSeq is a more specific database than EMBL-Bank, containing well-annotated, non-redundant sequences (the NCBI equivalent of EMBL-Bank is GenBank). We also used a stricter cutoff: we set a cutoff of an E-value $\leq 1e^{-10}$, which translates into a minimum score of 78 in this situation, while Tsai and co-workers used a minimum score of 50. Finally, we only mapped the non-control porcine specific sequences of the Affymetrix platform to the IPA, so we started with fewer probe sets (24,123 vs. 23,937). Due to our more stringent cut-off score, we did find slightly fewer alignments to RefSeq with the IPA sequences that have an Affymetrix sequence alignment than did their method: 19,253 (80%, IPA) vs. 19,675 (82%, Tsai et al.) at the comparative nucleotide level. Further, 22,569 (94%) of the probe sets have an alignment to sequences in the IPA, and thus we can map annotations to Affymetrix probesets that were recognized only when the longer and more complete ITC sequences were aligned with RefSeq. Additionally, we provide online at AnnotDB a list of all the sequences within the IPA that each of the 25-mers hit to
help flag possible cross-hybridization of close gene family members or alternate transcripts. Hence, by using RefSeq and a higher cutoff, we have higher confidence in the homologies being correct, and by leveraging all the additional information of newer sequences, we provide the most complete annotation of the Affymetrix platform for biological interpretation.

**Sequence Variation Inferred from Overlapping Sequences in AnnotDB**

Alignment of cDNA sequences from different individuals can provide evidence of structural variation, provided the depth of sequence data is sufficiently high. We found 39,943 to 2,004,432 putative SNP variants, depending on the threshold for number of minor allele sequences required. A comparison between our assembly-based predictions and those SNPs reported within dbSNP entries, which were matched to our SNP-containing sequences, found a high frequency of dbSNP reported variation in our SNP predictions-. These results provide evidence that many of our SNP predictions are accurate. However, this analysis could be biased because we could only compare a subset of our predicted SNPs to the SNPs reported in dbSNP to our predicted SNPs; those with similarity to sequences that are already known to be polymorphic due to their inclusion in dbSNP. This could increase our rate of agreement; however, it is difficult to estimate the level of bias. Furthermore, we do identify in many cases the exact same polymorphic position as that reported in dbSNP for these matched sequences. Sequencing errors could be a source of false predictions. However, it would be very difficult to directly determine the false positive rate of our predictions through resequences the source material; the public expressed sequence dataset has more than two million sequences from animal samples from many contributing laboratories.
Another group has used sequence assembly to predict SNPs in porcine expressed sequences (Panitz et al. 2007, see SD assembly information at http://pigest.ku.dk/download/pigest_kvl_1.0.1/snp_candidates.html). We found approximately five-fold more SNPs in the IPC than were found in the SD assembly at the same minor allele count. This large increase is most likely due to the inclusion of more than twice the number of sequences in the IPC compared to the SD assembly, and that additional sequences from breeds outside of those used in the SD project were included in the IPA. While the inclusion of additional sequence also brings the possibility of increased incorrect pseudogene or close gene family member integration during the assembly process, artifacts created by assembling ESTs (Picoult-Newberg et al. 1999), as well as bringing in the inherent sequencing error rate from EST sequencing, full validation of the predicted SNPs in AnnotDB is beyond the scope of this paper, as such validation would require a draft porcine genome sequence to provide information on ITCs that may have been assembled incorrectly. However, it has been shown by others that approximately 40% to 74% of SNPs predicted by EST assembly were confirmed by resequencing or through Genetic Bit Analysis (see for example Picoult-Newberg et al. 1999; Brett et al. 2000; Huntley et al. 2006; Panitz et al. 2007; Gorbach et al. 2009) Combining these results with the analysis above that shows substantial agreement with the known variations reported in dbSNP, we predict that the majority of predicted SNPs with at least three minor alleles (265,968; Table 3) are likely to be accurate.

The porcine-specific sequence variation derived from these sequence assemblies can also be used in genome mapping or be combined with expression data to integrate changes in expression with the genome determinants of those changes (Cookson et al. 2009; Schadt
2006). For example, genome-wide SNP mapping using large-scale SNP panels are being used in many species, including the pig. Our SNP data is largely outside the randomly selected, genome sequencing based SNPs in these panels, and thus is complementary to these genotyping tools.

**Availability, Extensibility and Maintenance of Annotation**

By making ANEXdb open source, we provide a local installation option available to users in addition to the web-based tool. While ANEXdb can be a local installation for microarray data storage, it is GEO compatible due to its ability to output expression data in SOFT format. Although ANEXdb is currently set up to only output in SOFT, the velocity template system will allow users to create custom outputs, such as XML-based MINiML, for submission to other databases, such as ArrayExpress. Likewise, ANEXdb as an open source application will also allow users to create novel plugins to directly access both expression and annotation data contained within the database. These user-created plugins can also be shared with other ANEXdb users.

We plan to re-assemble the IPA and revise annotations approximately every six months, and will incorporate the draft porcine genome sequence when it becomes available. Additional sequences will continue to help refine assemblies and the genome sequence will provide an anchor to help determine alternative transcripts and mis-assembled highly similar gene family members. In addition to providing the new assembly and annotation data for online download, we will also store archived assembly and annotation data.

Although ANEXdb currently houses porcine-specific data, it has been designed to be species independent and ANEXdb can be easily customized for other species by populating the databases with the relevant annotation and expression data from a variety of platforms,
such as Affymetrix GeneChips® in other species or custom arrays including the new porcine 20k spotted oligonucleotide array (Ernst et al. 2008). Finally, because ANEXdb holds both expression and annotations in a single location, its use will allow easier and faster analysis of the large amounts of data generated through high throughput expression experiments.

ACKNOWLEDGEMENTS

We would like to thank the USDA CSREES-NRI-2005-3560415618 and the ISU Center for Integrated Animal Genomics for funding this project. A USDA MGET 2001-52100-11506 Fellowship to O.C. is gratefully acknowledged.
REFERENCES


Uenishi H, Eguchi T, Suzuki K, Sawazaki T, Toki D, Shinkai H, Okumura N, Hamasima N, Awata T (2007) PEDE (Pig EST Data Explorer) has been expanded into Pig


TABLE 1

Overlap of the Iowa Porcine Assembly with human and mouse RNA RefSeq shows IPA has similar complexity to these datasets.

A cutoff of an E-value <= 1e-10 for BLAST match was used to compare the distinct numbers of matches to human or mouse RNA RefSeqs matches within and across species. See Supplemental Figure 1 for coverage.

<table>
<thead>
<tr>
<th>Query Source</th>
<th>Subject Source</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Porcine Assembly</td>
<td>Human RefSeq</td>
<td>33,341</td>
</tr>
<tr>
<td>Porcine Assembly</td>
<td>Mouse RefSeq</td>
<td>27,109</td>
</tr>
<tr>
<td>Human RefSeq</td>
<td>Human RefSeq</td>
<td>46,049</td>
</tr>
<tr>
<td>Human RefSeq</td>
<td>Mouse RefSeq</td>
<td>29,510</td>
</tr>
<tr>
<td>Mouse RefSeq</td>
<td>Human RefSeq</td>
<td>33,846</td>
</tr>
<tr>
<td>Mouse RefSeq</td>
<td>Mouse RefSeq</td>
<td>40,158</td>
</tr>
</tbody>
</table>
TABLE 2

Number of Iowa Porcine Assembly sequences with various types of annotation.

Total represent the total number of rows within the appropriate MySQL table, while distinct is the number of individual sequences contributing to the total; for example 191,602 sequences contribute to the 7,643,037 BLASTN hits to RefSeq RNA database entries.

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Total Number</th>
<th>Total Consensus</th>
<th>DISTINCT NUMBER</th>
<th>Distinct Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLASTN to RefSeq RNA</td>
<td>7,643,037</td>
<td>4,114,965</td>
<td>191,602</td>
<td>105,514</td>
</tr>
<tr>
<td>BLASTX to RefSeq Protein</td>
<td>7,624,127</td>
<td>3,372,389</td>
<td>71,332</td>
<td>31,195</td>
</tr>
<tr>
<td>BLASTX to Pfam</td>
<td>6,716,663</td>
<td>3,364,108</td>
<td>76,385</td>
<td>40,813</td>
</tr>
<tr>
<td>Exonerate</td>
<td>71,116</td>
<td>33,016</td>
<td>34,573</td>
<td>22,083</td>
</tr>
<tr>
<td>Associated GO Terms</td>
<td>20,436,544</td>
<td>10,326,109</td>
<td>166,119</td>
<td>87371</td>
</tr>
<tr>
<td>Associated KEGG Pathways</td>
<td>6,695,077</td>
<td>3,438,014</td>
<td>92,263</td>
<td>48,763</td>
</tr>
<tr>
<td>Putative SNP</td>
<td>2,528,653</td>
<td>2,528,653</td>
<td>63,995</td>
<td>63,995</td>
</tr>
<tr>
<td>ORF</td>
<td>1,200,483</td>
<td>723,047</td>
<td>227,954</td>
<td>127,978</td>
</tr>
</tbody>
</table>
TABLE 3

SNPs predicted by automated inspection of expressed sequence alignments.

Sequence variation by type and count of minor allele decreases as minimum count of the minor allele is increased.

<table>
<thead>
<tr>
<th>Number of Minor alleles required</th>
<th>Variation Type</th>
<th>ITC Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNP</td>
<td>Deletion</td>
</tr>
<tr>
<td>1</td>
<td>2,004,432</td>
<td>428,637</td>
</tr>
<tr>
<td>2</td>
<td>628,944</td>
<td>116,653</td>
</tr>
<tr>
<td>3</td>
<td>267,968</td>
<td>54,915</td>
</tr>
<tr>
<td>4</td>
<td>171,034</td>
<td>34,093</td>
</tr>
<tr>
<td>5</td>
<td>117,496</td>
<td>23,303</td>
</tr>
<tr>
<td>6</td>
<td>88,594</td>
<td>17,114</td>
</tr>
<tr>
<td>7</td>
<td>69,598</td>
<td>13,144</td>
</tr>
<tr>
<td>8</td>
<td>56,640</td>
<td>10,407</td>
</tr>
<tr>
<td>9</td>
<td>47,084</td>
<td>8,475</td>
</tr>
<tr>
<td>10</td>
<td>39,943</td>
<td>7,053</td>
</tr>
</tbody>
</table>
### TABLE 4

**Number of ITC-predicted SNPs found reported by NCBI dbSNP**

Within each SNP source (cDNA, genomic or total) are reported the comparison results when there was a requirement of 1, 2 or 3 minor alleles for the predicted SNP within the ITC. The upper number is the count of SNP agreements between dbSNP and ITC, while the lower number is number of sequences with variations as reported in dbSNP that align to an ITC entry. Number under the SNP source category indicates the total number of each type of SNP within dbSNP.

<table>
<thead>
<tr>
<th>Minimum Percent of dbEST Sequence in Alignment</th>
<th>&gt;0</th>
<th>20</th>
<th>40</th>
<th>60</th>
<th>80</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SNP agreements/total aligned sequences</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Number of Minor alleles required</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cDNA (645)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>310/414</td>
<td>310/414</td>
<td>310/414</td>
<td>303/407</td>
<td>291/391</td>
<td>267/355</td>
</tr>
<tr>
<td>2</td>
<td>166/208</td>
<td>166/208</td>
<td>166/208</td>
<td>161/203</td>
<td>155/195</td>
<td>142/178</td>
</tr>
<tr>
<td>3</td>
<td>94/114</td>
<td>94/114</td>
<td>94/114</td>
<td>92/112</td>
<td>87/106</td>
<td>79/94</td>
</tr>
<tr>
<td>Genomic (33,863)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7385/9169</td>
<td>7341/9122</td>
<td>7187/8955</td>
<td>6894/8617</td>
<td>6245/7882</td>
<td>4788/6121</td>
</tr>
<tr>
<td>2</td>
<td>4320/5082</td>
<td>4301/5063</td>
<td>4231/4986</td>
<td>4083/4807</td>
<td>3757/4439</td>
<td>2925/3463</td>
</tr>
<tr>
<td>3</td>
<td>2513/2977</td>
<td>2510/2974</td>
<td>2478/2938</td>
<td>2411/2857</td>
<td>2248/2673</td>
<td>1717/2038</td>
</tr>
<tr>
<td>Total (34,508)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7695/9583</td>
<td>7651/9536</td>
<td>7497/9369</td>
<td>7197/9024</td>
<td>6536/8273</td>
<td>5055/6476</td>
</tr>
<tr>
<td>2</td>
<td>4486/5290</td>
<td>4467/5271</td>
<td>4397/5194</td>
<td>4244/5010</td>
<td>3912/4634</td>
<td>3067/3641</td>
</tr>
<tr>
<td>3</td>
<td>2607/3091</td>
<td>2604/3088</td>
<td>2572/3052</td>
<td>2503/2969</td>
<td>2335/2779</td>
<td>1796/2132</td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

**Figure 1.** Schema of the ANEXdb application. Arrows indicate data flow within ANEXdb; levels show which language or host is being used by ANEXdb to control data flow. A user has access to the submission and querying aspects of ANEXdb, while an administrator controls the finalization and migration of submission from a temporary database to the storage database; the administrator also controls the information going into AnnotDB.

**Figure 2.** Highly consistent agreement of the top hits from the different alignment algorithms indicates IPA annotations are accurate. Chromosomal location of each Gene ID match from BLAST results was obtained using NCBI’s gene2refseq assignments and the top scoring RefSeq (with E-value ≤ 1e-10 for BLASTN and E-value ≤ 1e-5 for BLASTX alignments). The Gene ID location was then compared to the human genome location of the Exonerate alignment to the porcine sequence query; a match required the Exonerate-based alignment to be within 250 bases on either side of the BLAST Gene assignment. Note: numbers are different from Table 2 due to E-value cutoffs and because not all RefSeq IDs can be mapped to a Gene ID.

**Supplemental Figure 1.** The coverage by Iowa Porcine Assembly of human and mouse RefSeq databases is similar to coverage of human by mouse, and of mouse by human. Black shows the coverage of subject database by the query database, as measured by BLAST matches, while gray shows the number of sequences unique to the subject database.

**Supplemental Figure 2.** Alignment to dbSNP provides verification of predicted SNPs. Total number of porcine SNPs in dbEST is 34,508 (as of May, 2008) which is comprised of 645 cDNA annotated SNPs and 33,863 genomic SNPs.
Figure 1. Couture et al., 2009.
Figure 2. Couture et al., 2009.

<table>
<thead>
<tr>
<th>BLASTN</th>
<th>Total: 128,659</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>69,351</td>
<td></td>
</tr>
<tr>
<td>32,990</td>
<td></td>
</tr>
<tr>
<td>10,531</td>
<td></td>
</tr>
<tr>
<td>9,629</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td></td>
</tr>
<tr>
<td>15,787</td>
<td></td>
</tr>
<tr>
<td>6,581</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BLASTX</th>
<th>Total: 53,278</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exonerate</th>
<th>Total: 33,027</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Supplemental Figure 1. Couture et al., 2009.
Supplemental Figure 2. Couture et al., 2009.

[Diagram of bar chart showing percent agreement of SNPs from dbSNP aligning to ITCs for different minimum number of minor allele counts, with categories for cDNA, Genomic, and Total, and alignment length criteria for 0%, 20%, 40%, 60%, and 80% of dbSNP sequence length.]
CHAPTER 4: ANALYSIS OF PORCINE TRANSCRIPTIONAL RESPONSE TO SALMONELLA ENTERICA SEROVAR CHOLERAESUIS SUGGESTS NOVEL TARGETS OF NFKAPPAB ARE ACTIVATED IN THE MESENTERIC LYMPH NODE

Modified from a manuscript published in BMC Genomics


ABSTRACT

The Affymetrix GeneChip® porcine genome array was used to identify differentially expressed genes in pig mesenteric lymph nodes (MLN) responding to infection with Salmonella enterica serovar Choleraesuis (S. Choleraesuis) at acute (8 hours (h), 24h and 48h post-inoculation (pi)) and chronic stages (21 days (d) pi). Analysis of variance with false discovery rate control showed that 1,853 genes exhibited significant changes in expression level with analysis of variance p-value<0.01 and estimated fold change >2 (q<0.26). Hierarchical clustering analysis on gene expression revealed several specific features of host response to infection. Down-regulation of translation-related genes at 8 hpi and 24 hpi implied that S. Choleraesuis repressed several steps in host protein translation. Genes involved in the Th1, innate immune/inflammation response and apoptosis pathways were induced significantly. However, the antigen presentation/dendritic cell (DC) function pathways were not affected significantly during infection. We suspect that a lack of strong DC-mediated antigen presentation in the MLN might be the reason that S. Choleraesuis

5 Primary author and researcher
6 Contributed significantly to the analysis and writing
7 To whom correspondence should be addressed to: cktuggle@iastate.edu
infected pigs can develop a systemic infection. Furthermore, Quantitative-PCR analyses of 22 genes confirmed a strong NFκB pathway transcriptional response detected by the microarray results, as 59 known NFκB target genes were induced significantly at 8, 24 and/or 48 hpi. These target genes can be classified as an “Early group” (genes induced at either 8 or 24 hpi) and a “Late group” (genes only induced at 48 hpi) based on their expression patterns. Expanded GO-slim annotation revealed that functions for cytokine activity or chemokine activity were enriched within the “Early group” genes (Fisher exact test $p=0.053$), while the “Late group” was predominantly composed of signal transduction and cell metabolism annotated genes. In addition, our analysis suggests a number of putative novel NFκB targets, due to these genes having 1) similar expression pattern as known NFκB target genes and 2) computationally-identified NFκB binding motifs within the promoter region of the orthologous human gene. Our study not only provides novel genome-wide transcriptional profiling data on the porcine response to $S$. Choleraesuis, but also expands the understanding of NFκB signaling in response to Salmonella infection.

INTRODUCTION

Salmonella enterica serovar Choleraesuis ($S$. Choleraesuis) is a narrow host range $Salmonella$ serovar that predominately colonizes swine, although this serovar can also cause serious infections in human patients (Chiu et al., 2004, 2006). With clinical manifestations of enterocolitis, pneumonia, septicemia and hepatitis, $S$. Choleraesuis infections result in a higher mortality rate in pigs than $S$. Typhimurium, a broad host range $Salmonella$ serovar that typically causes only enterocolitis in swine. This highly invasive serovar is of particular concern to the swine industry since pig salmonellosis results in about $100$ million in annual production losses nationwide (Fedorka-Cray et al., 1995). $S$. Choleraesuis can result in higher
mortality than S. Typhimurium infection in pigs. Moreover, S. Choleraesuis infected pigs can develop a carrier state lasting as long as 12 weeks (Gray et al., 1996); resulting in shedding of the bacteria by carrier pigs and Salmonella contamination the environment. Researchers have speculated that human systemic infections caused by S. Choleraesuis were acquired from pigs, as affirmed by DNA fingerprints of the S. Choleraesuis isolated from humans and swine (Chiu et al., 2004, 2005, 2006). Although an oral S. Choleraesuis vaccine (ENTERISOL® SC-54) has been shown to enhance humoral and cellular immune responses (Shiau et al., 2005), S. Choleraesuis infections can be particularly difficult to treat due to its resistance to multiple antimicrobial agents. Thus S. Choleraesuis is both a swine industry and public health problem.

While several serovars such as S. Typhimurium have been extensively studied, a relatively limited number of experiments on the S. Choleraesuis serovar have been published using a swine infection model (Chiu et al., 2004; Utue et al., 2005, 2007; Zhao et al., 2006; Hyland et al., 2006a,b). The transcriptional response to Salmonella infection is a predominantly a Th1 immune response as observed in porcine lung during S. Choleraesuis infection (Zhao et al., 2006) and in porcine mesenteric lymph nodes (MLN) during S. Typhimurium and S. Choleraesuis infection (Utue et al., 2007; Wang et al., 2007). The antigen processing- and apoptosis-related pathways were strongly induced at 24 and 48 hpi in porcine lung during S. Choleraesuis infection based on oligonucleotide microarray data (Zhao et al., 2006). Two important early proinflammatory cytokines, IL8 and IL1B, were rapidly induced in porcine jejunal and distal ileal Peyer’s patches during S. Choleraesuis infection (Hyland et al., 2006a,b). Utue et al. (2005) used suppression subtractive hybridization (SSH) to identify genes which were up- or down- regulated at 24 hpi in porcine
MLN during *S. Choleraesuis* infection compared to non-infected animals. Most of these genes have been annotated as being involved in host cellular functions including innate immunity and cytoskeleton regulation (Uthe et al., 2005). Skjolaas et al. (2006) tested the effect of *S. Choleraesuis* infection on gene expression in pig jejunal epithelial cells; IL8 and CCL20 were significantly induced at 1.5 h and 3 h post exposure, whereas CCL20 reached peak response at 6 h post exposure.

In this study, a complex genome-wide investigation of the host’s response to *S. Choleraesuis* infection was conducted using the Affymetrix GeneChip® Porcine Genome Array containing oligonucleotides representing approximately 23,256 transcripts from 20,201 *S. scrofa* genes. We report the global transcriptional profile of the porcine MLN, the largest lymph nodes in the human and animal body and one of the components of gut-associated lymphoid tissues (GALT), to *S. Choleraesuis* infection at the acute and chronic stages. We characterize the host immune responses, targeting the understanding of the host’s innate immunity, and in the process detecting previously unidentified potential NFκB target genes.

**MATERIALS AND METHODS**

**Experimental design**

Fifteen piglets from *Salmonella* spp.-free sows were weaned at 10 days (d) of age, shipped to the National Animal Disease Center, Ames, IA, and raised in isolation facilities. To confirm that all piglets were fecal-negative for *Salmonella* spp. prior to challenge, bacteriological cultures were performed twice on rectal swabs before the experiments. At seven weeks of age, 3 and 12 pigs were randomly allocated to the non-infected group or to the infected group, respectively. The three non-infected control pigs were necropsied 3 days prior to experimental infection. On day 0, pigs in the infected groups were intranasally
challenged with $1 \times 10^9$ CFU of *S. Choleraesuis* χ3246. Three randomly chosen infected pigs were necropsied at each time point of 8 hpi, 24 hpi, 48 hpi and 21 dpi, respectively. Tissue samples from the MLN were collected and immediately frozen in liquid nitrogen. Total RNA was isolated from ~200 mg of these samples by using the RNeasy Midi kit with on-column RNase-free DNase digestion (Qiagen, Valencia, CA) based on the manufacturer’s protocol. RNA integrity, quality and quantity were assessed using the Agilent Bioanalyser 2100 and RNA Nano 6000 Labchip kit (Agilent technologies, Palo Alto, CA).

**Microarray hybridizations and data analysis**

Five $\mu$g total RNA was used for first and second strand cDNA synthesis according to the manufacturer instructions (Affymetrix, Inc. Santa Clara, CA). The double stranded cDNA was purified, tested on an Agilent Bioanalyser 2100, and served as the template for the subsequent *in vitro* transcription (IVT) reaction for cRNA amplification. Labeling cRNA with biotin was performed by the GeneChip® One-Cycle target labeling kit (Affymetrix; Expression Analysis Technical Manual). Quality of the labeled cRNA was tested on an Agilent Bioanalyser 2100. Subsequently, labeled cRNA was fractionated and hybridized with the GeneChip® Porcine Genome Array according to the standard procedures provided by the manufacturer. Chips were washed and stained with a GeneChip Fluidics Station 450 (Affymetrix, Inc. Santa Clara, CA) using the standard fluidics protocol. Chips were then scanned with an Affymetrix GeneChip Scanner 3000 (Affymetrix, Inc. Santa Clara, CA).

MAS 5.0 (Microarray Analysis System 5.0, Affymetrix, Inc. Santa Clara, CA) default normalization methods were used to obtain the expression measure for each probeset. Base-2 logarithms were then taken on these expression measures. The median of the log expression measures for each chip was then subtracted from all log expression measures on
the same chip. Differentially expressed genes were identified by analyzing these normalized data with a linear model using SAS Proc GLM (SAS Insititue, Cary, NC) on a gene by gene basis. The statistical model for gene $g$ was as follows: $y_{ijg} = \mu_g + T_{ig} + \epsilon_{ijg}$, where $y_{ijg}$ is the log of the normalized signal for gene $g$ on pig $j$ necropsied at time $i$, $\mu_g$ is an intercept term for gene $g$, $T_{ig}$ is the fixed effect of the $i^{th}$ time-point on expression of gene $g$, and the $\epsilon_{ijg}$ values are independently normally distributed random errors with mean 0 and gene-specific variances $\sigma^2_g > 0$, assumed to be independent for each $g$. An $F$ test for differences in expression across all time points during infection and $t$-tests for all ten pair-wise comparisons among the five treatment groups (non-infected, 8 hpi, 24 hpi, 48 hpi, and 21 dpi) were conducted as part of the analysis for each gene. This yielded eleven sets of $p$-values for the effect of infection. Each set of $p$-values was converted to a set of $q$-values using the method of Storey and Tibshirani (2003). The largest $q$-value in a list of genes declared to be differentially expressed provides an estimate of the upper bound of the positive False Discovery Rate (pFDR) associated with the list. Fold changes (FC) for each of the 10 pairwise comparisons were estimated as $2^{\hat{d}_{kg}}$, where $\hat{d}_{kg}$ is the estimated difference in means for gene $g$ on the log scale for comparison $k, k=1,2,\cdots,10$. As the $p$-values tend to be stochastically smaller for genes with larger estimated fold changes in our data, we still use the largest $q$-value in a list of genes as a conservative estimate of the upper bound of pFDR even though we have excluded the genes with estimated FC<2 (or estimated FC<10 in some cases as noted in the Results section). Our microarray data have been submitted to the NCBI GEO database and the accession number is GSE7314.
Transcriptome determination

The transcriptome of normal and S. Choleraesuis infected MLN were determined as described previously (Wang et al., 2007). Briefly, transcripts which showed a Present call for all three non-infected animals were counted in the transcriptome of normal porcine MLN tissue, while transcripts which showed a Present call for all three replicates in at least one time point during infection were counted as the transcriptome of infected porcine MLN tissue.

Hierarchical Cluster analysis

After removing duplicate probe sets, a total of 1,853 genes showed \( p \)-value<0.01 and estimated FC>2 \( (q\)-value<0.26) in at least one of the 10 possible time point pair-wise comparisons \( (8h-C, 24h-C, 48h-C, 21d-C, 24h-8h, 48h-8h, 21d-8h, 48h-24h, 21d-24h \text{ and } 21d-48h) \) during S. Choleraesuis infection and were designated as the DE (differentially expressed) genes. This list was used to perform a hierarchical cluster analysis and to construct a heat map using the Gene Cluster 3.0 and tree view software (Stanford University, 2002).

GO-slim creation and GO annotation of Affymetrix probesets

A set of high level GO terms which represent host response categories in biological process was selected by using OBO-Edit, which is part of the go-dev software provided by GO at Sourceforge (https://sourceforge.net/project/showfiles.php?group_id=36855&package_id=33201). An expanded GO slim was created and GO analysis was performed for transcriptome and DE genes as described before (Wang et al., 2007). Fisher’s exact test was used for gene
enrichment analysis.

Quantitative PCR (QPCR) RNA analysis

Quantitative PCR technology was used to verify the differential expression of 21 genes at early response stages (8 hpi, 24 hpi and 48 hpi), as identified by the microarray. The TGM3 gene, which has not yet been annotated on the microarray, was also analyzed. The RPL32 gene, a reference gene for high abundance gene transcripts, was used as a positive control. All probes and primers for real time TaqMan PCR were designed as previously described (Royaee et al., 2004; Dawson et al., 2005). Total RNA was isolated from the MLN of the 3 non-infected pigs and the 3 infected pigs at each time point of 8 hpi, 24 hpi and 48 hpi, and QPCR were performed as previously described (Wang et al, 2007). The interpolated number \( (C) \) of cycles to reach a fixed threshold above background noise was used to quantify amplification. The fold change in expression of the target gene was estimated as \( 2^{ΔCt} \), where \( ΔCt \) is the difference between average \( Ct \) values for the control and infected pigs. Resulting Q-PCR data were analyzed by one-way ANOVA on a gene by gene basis, as done in analyzing microarray data, but using JMP 5.0 Software (SAS Inc, Cary, NC). Fisher’s LSD post-hoc test was applied to assess differences between groups of pigs at different time points post infection. A value of \( p≤0.05 \) was considered statistically significant.

NFkB motif searching

TFM-Explorer identifies windows conserved amongst a group of sequences sharing a common transcription factor binding site. To do this, TFM-Explorer compares the input sequences to a set of previously derived position specific scoring matrices (PSSM) for the binding sites of interest to obtain a score for each sequence. It then compares this to the
probability of the binding site appearing at random in the genomic background sequences. It
then sees how many of the sequences have the same matrix within a window from 300 to
1500 bases long. The closer the sequence matches the PSSM, and the higher the percentage
of sequences with the binding site within a window, the more significant that window
becomes. The default parameters of TFM-Explorer, including the limit of 500 input
sequences, were used for all runs.

The 1,500 bp 5’ and 500 3’ bp (relative to the annotated transcription start site) of the
human orthologs for all annotated and upregulated DE porcine genes were obtained using
PERL scripts. These scripts use human RNA RefSeq accession number, obtained from
BLAST results using the Affymetrix porcine consensus sequence (Couture et al.,
unpublished results), to identify and download human chromosome GenBank and FASTA
files from NCBI RefSeq. Information on transcription start sites from the human
chromosomal GenBank files, strand information, and sub-sequences for each transcript were
extracted from the FASTA files. Human orthologs of the porcine DE genes were then
separated into different groups: two early sets (one of 83 sequences with known and
unknown NFκB targets [E83] and one of 61 sequences without these known targets [E61]);
two late sets (one of 319 sequences with known NFκB targets [L319] and one of 283
sequences without these known targets [L283]); five different randomly selected sets of 500
(A500) from the 560 up-regulated and annotated genes across all time points (due to the
limitation of TFM-Explorer of allowing a maximum of 500 input sequences); and a set of
475 genes from this 560 gene set with the known targets of NFκB removed (U475). A list of
genes known to be direct targets of NFκB was obtained at two websites: http://www.nf-
kb.org and http://bioinfo.lifl.fr/NF-KB/. The promoter sequences in these different groups
were then searched for windows of sequence with over-representation of any of the available known NFκB motifs defined by position-specific sequence matrices deposited in public TRANSFAC 7.0 (http://www.gene-regulation.com/pub/databases.html) or JASPAR (http://jaspar.genereg.net/) using the default parameters of TFM-Explorer (Defrance and Touzet, 2006).

In addition to TFM-Explorer, Clover (Martin et al., 2004) was also used to find putative binding sites in the above promoter sequences for E83, L319, and for all 560 DE genes. Clover can use a permutation test to see how often a given matrix is found in a group of sequences when compared to how often it is found in a randomly generated sequence set selected from background sequences that maintain the same number of sequences, the same length, and same G/C content as the input set. Since human sequences were used as input, sets of human sequence was used as background sequence as well; human cpg islands, the entire chromosome 20, and the 2,000 upstream bases from genes. Only the default number of permutation tests was changed: using 5,000 instead of the default of 1,000. Also, since Clover does not have an input size limit all 560 sequences were run as a single group. After running both programs on the same set of sequences, the intercept of the sequences was defined to be putative NFκB targets.

RESULTS

Transcriptome

The normal porcine MLN tissue transcriptome was represented by all probesets (14,348) which showed a Present call for all three non-infected animals. All probesets that showed a Present call for all three replicates in at least one time point during infection
(15,935) were counted as the transcriptome of infected porcine MLN tissue. Combining these lists gave 16,046 probesets (70 % of the whole chip) that detected expression in MLN (Supplementary Table 1).

**Differentially expressed gene analysis during S. choleraesuis infection**

Swine infected with S. Choleraesuis exhibit high bacterial loads as early as 24 hours post infection and reach maximal levels at 2-7 dpi (Uthe et al., 2007). Thus we have targeted our analyses at the early time points of infection where unique control mechanisms might be revealed. Genes with a p-value<0.01 and an estimated FC >2 in at least one of the 10 time point pair-wise comparisons were declared to be differentially expressed (DE). A total of 1,853 genes satisfied these criteria (Supplementary Table 2). The largest q- value associated with this list of genes was 0.26, which is an estimated upper bound of the False Discovery Rate (FDR). Of these 1,853 DE genes, 1,189 genes (64%) have sequence similarity to a human Refseq based on our BLAST analysis.

Also, compared to non-infected pigs, the transcriptional response to S. Choleraesuis infection identified 85, 160, 954, and 111 transcripts differentially expressed (p-value<0.01 and estimated FC>2) at 8 hpi, 24 hpi, 48 hpi and 21 dpi, respectively (Fig 1). Clearly, a strong transcriptional response was observed at 48 hpi, with a similar number of up- and down-regulated genes. Table 1 shows 32 up-regulated and 9 down-regulated genes with a fold change greater than 10 at 48 hpi in infected MLNs, relative to non-infected MLNs (p-value<0.01). Of the 30 annotated up-regulated genes in Table 1, 23 genes (77 %) are involved in immune response, inflammation and apoptosis. The other 7 up-regulated genes, STEAP4, TNFAIP6, PSTPIP2, LIPG, HK3, STXBP1 and ITPR1, have not been previously associated with bacterial infection. The S100 calcium binding protein A9 (S100A9)
exhibited the highest fold change (226.8) at 48 hpi, compared to non-infected pigs. A gene from the same family, S100A12, also showed strong up-regulation (38.9 estimated FC). Compared to the 32 up-regulated genes, only 9 transcripts were down-regulated using the same criteria of $p$-value<0.01 and estimated FC>10. Of these, only 3 genes (CA3, KRT17 and PLN) have been annotated, and none of these previously have been shown to be involved in the response to bacterial infection.

**Cluster analysis**

To get a broad overview of the changes in gene expression during infection, a heat map was created using Gene cluster 3.0 software for the 1,853 DE genes. Two distinct clusters, an induced gene cluster and a repressed gene cluster, were identified and several interesting gene expression patterns were exhibited (Fig 2). Group A includes 52 genes that were repressed at 8 hpi and 24 hpi, after which RNA levels returned to those seen in the non-infected animals. Many ribosomal protein genes, as well as eukaryotic translation elongation factor 1 alpha 1 (EEF1A1) were included in this group. A large number of genes that presented a comparatively lower expression level at 48 hpi during infection were found in group B. The annotated genes in this group had diverse biological functions, without an obvious overrepresentation of any specific pathway. The common feature of genes in group C, D and E is that they had a peak RNA response at 48 hpi during infection. However, genes in each group had different peak response times: 47 genes in group C, such as INFG, PSMB9, PSMB10 and C2, were only up-regulated at 48 hpi, while 141 genes in group D, such as TNF, IL6, IL8, IL1A and IL1B, increased their expression level at 24 hpi and expression peaked at 48 hpi. Genes in group E were slightly up-regulated as early as 8 hpi during infection; some heat shock proteins, such as HSPB1, HSPA6, HSPA1B were found in
this group. An additional group F contained genes that were up-regulated at both the acute and chronic stages of infection, as compared to non-infected animals. This group included 38 annotated genes, including two complement related genes, C1QA and CFH.

**Gene Ontology annotation**

Additional GO annotation of the 16,046 transcripts expressed in MLN were performed using our laboratory-designed GO-slim, built with terms in the biological process and molecular function categories of the GO database relevant for immune biology (Fig. 3A, gray bars). About 4,730 transcripts were assigned specific GO terms. Approximately 39% of these transcripts were annotated as being involved in cellular metabolic processes, and a significant number of transcripts were assigned known functions in signal transduction (15.6%), cell differentiation (7.2%), cell cycle (6.3%), calcium ion binding (5.6%), apoptosis (5.3%), cell adhesion (3.8%) and cell proliferation (3.7%). An additional 13% of the transcripts were assigned GO terms related to protein folding, immune response, cell migration, inflammatory response, defense response, antigen processing, and antigen presentation. The immune-oriented GO-slim annotation of the 1,853 DE genes was also performed (Figure 3A, white bars) and compared to the global transcriptome GO term assignment. The proportion of genes assigned GO terms associated with cell adhesion, apoptosis, immune response, inflammatory response, cellular metabolic process, calcium ion binding activity and acute-phase response was significantly enriched in our DE gene list ($p<0.01$), compared to the transcriptome GO assignment.

An over-representation of a specific biological process does not indicate whether the process in question is being stimulated or repressed overall. To investigate over- or under-represented functional activities specifically within the up-regulated and down-regulated
genes, GO annotations were also assigned to transcripts from the induced and repressed clusters. Statistical analysis revealed that genes annotated with GO terms of immune response, innate immune response, defense response, apoptosis and cellular metabolic process were significantly enriched in the induced cluster, while the repressed cluster had a significantly higher percentage of genes related to cell adhesion \( (p<0.05) \) (Figure 3B).

**Pathway analysis**

During bacterial infection, T cells will migrate into the T-cell zone of the MLN and scan the surface of the antigen-presenting cells, primarily macrophages and dendritic cells, for specific peptide:MHC complexes (Janeway et al., 2005). The levels of any RNA transcript present in the MLN can be changed by cells migrating into or out of the lymph node, and these RNA changes might be erroneously interpreted as transcriptional response to infection within an immune cell type. To verify that the observed differences in gene expression are true transcriptional differences, Affymetrix data-based expression levels of specific markers for T cells, macrophages, dendritic cells and granulocytes were checked. No evidence of major changes in cell migration was observed since most RNA levels for these marker genes did not change significantly (Fig 4A). As we have further concentrated our analyses and interpretations on genes with large differences in expression, these data indicate the RNA differences we observed are likely representative of specific transcriptional responses within cells, and not due to significant changes in the abundance of specific cell types in the MLN.

To investigate immune-related pathways stimulated during infection, the Affymetrix-based expression patterns of genes that are known to be involved in specific immune pathways were collected (Fig 4B). Results showed that 6 out of 11 selected Th1-related
genes, such as INFG, IRF1, SOCS1, STAT1, TNF and WARS, were significantly up-regulated at 48 hpi, while genes known to be predominately associated with Th2 response (IL4, IL13, MAPK14 and STAT6), were down-regulated or unchanged respectively at all time points during infection (Fig 4B). These data suggest that S. Choleraesuis elicited primarily a Th1-associated response within the MLN during infection.

Most genes known to be involved in innate/inflammatory pathways in Fig 4B, such as IL-8, IL6, SLC11A1, TLR4 and IL1B, increased their RNA expression level significantly at 24hpi and/or 48hpi during infection. Strong induction of apoptosis pathways was also observed in response to S. Choleraesuis infection. Seven apoptosis related genes (CASP1, CASP3, CASP4, GZMB, PTDSR, TGM1 and TGM2) were up-regulated significantly at 24 hpi and/or 48 hpi.

Two antigen-processing related genes, PSMB10 and TAP1, were induced at 48 hpi, which indicates that the antigen processing pathway was activated. However, S. Choleraesuis infection did not appear to significantly influence the antigen presentation/DC function pathway, because two genes known to be involved in dendritic cell activation, CD80 and CD86, did not show elevated expression level at either 24 or 48 hpi. In addition, CD209, which is also named DC-SIGN, showed non-significant changes during infection ($p>0.05$).

We observed a significant induction of genes at 48 hpi that have GO annotation of calcium binding activity, including S100A9, S100A12, several ANXA family genes, and CALU. Therefore, we speculate that calcium pathways were strongly affected by infection at this time (Fig. 4B). Further, annotation showed that many of the genes that were up-regulated at 24 and 48 hpi are known direct NFκB targets (Schreiber et al, 2006; http://bioinfo.lifl.fr/NF-KB/; http://www.nf-kb.org/). These known NFκB genes are listed in
Supplementary Table 3. To determine whether NFκB dependent genes formed a central part of the host transcriptional response to S. Choleraesuis infection, GO analysis was performed to compare the functional activities of these differentially expressed known NFκB target genes from an “Early” group, defined as genes that were up-regulated significantly at 8 or 24 hpi (84 genes; \( p \)-value<0.1 and estimated FC>2), and a “Late” group, defined as genes that were only up-regulated at 48 hpi (324 genes; \( p \)-value<0.01 and estimated FC>2). Figure 5 demonstrates that cytokine or chemokine activity GO terms found in the known NFκB target genes (such as CXCL1, CXCL2, CXCL6, IL1A, IL1B, IL1RN and CSF3), were enriched in the Early group as compared to the Late group (Fisher’s exact test \( p=0.053 \)). The Late group NFκB target genes were predominantly annotated in the signal transduction and cell metabolism categories.

**Q-PCR analysis of differentially-expressed porcine genes**

To confirm the NFκB genes that were declared differentially expressed in our microarray analysis, a panel of 22 genes was selected for real-time PCR analysis and validation of the expression patterns at 8 hpi, 24 hpi and 48 hpi. The RNA levels of 18 known NFκB target genes (IL1A, IL15, CCL2, CCL3, CXCL5, PPBP, GBP1, GBP2, PTX3, IKBA, JUNB, NFKBIZ, CD14, ICAM1, TLR2, GZMB, TAP1 and CCR5) was measured. We also analyzed the expression for a T cell marker gene (CD4) and a macrophage marker gene CD163. Although an oligonucleotide set representing the TGM3 gene could not be found on the Affymetrix microarray, Q-PCR was performed for this gene since our earlier work (Zhao et al., 2006) showed a strong up-regulation of TGM3 in porcine lung during acute infection with S. Choleraesuis, and the temporal expression pattern is similar to known NFκB target genes. Finally, TREM1 was also selected for Q-PCR validation because our
microarray data showed its expression pattern to be similar to known NFκB target genes, although no reports have shown that TREM1 is directly regulated by NFκB. Comparison of the QPCR results with the microarray data demonstrated that expression differences for 20 of the 21 genes were statistically significant by QPCR, confirming the Affymetrix-based results for these genes and RNA samples (Table 3 and Fig. 6). Results showed that, except for CD14, all known NFκB-regulated genes were significantly up-regulated at 48 hpi during infection, and many of the genes experienced a strong induction from 24 hpi to 48 hpi. We also confirmed that TGM3, TREM1 and CD163 exhibited a similar transcriptional profile with many known NFκB target genes.

**Identification of putative NFκB target genes within the DE Gene lists through bioinformatic analysis of the human orthologous promoter sequences**

To further explore the possibility that the response observed might have identified novel NF-κB target genes, we collected *in silico* comparative evidence that NFκB may bind to flanking sequences of these genes. We used human genomic sequence data for the DE genes in the Early group (83 human orthologs found for 84 Early genes; E83), the Late group (319 orthologs found for 324 Late genes; L319), and all DE up-regulated genes to identify those genes containing statistically significant NFκB binding sequence motifs near their promoters. The latter group contained 560 genes total, with 544 orthologs found, of which we used only 500 human orthologs (A500) in the TFM-Explorer analyses, as this was the maximum allowed by TFM-Explorer. However, all 544 were used in Clover as a single group; A544. We also investigated the effect on finding putative NFκB targets in these groups when removing the known NFκB target genes from these groups (Methods; Supplementary Table 3), to explore the underlying regulatory “signal” in these promoters.
Thus we generated an early group without 22 known targets (E61), a Late group with 36 known NFκB targets removed (L283), and a complete group of all upregulated DE genes with the 69 known NFκB targets removed (U475).

TFM-Explorer analysis of E61, representing all genes upregulated in the first 8-24 hpi but with known NFκB targets removed, found 51 promoters (84%) with at least one NFκB binding site (Fig 7). However, when the full E83 set, including all known NFκB targets, was analyzed, 73 sequences (88%) were identified with one or more NFκB binding sites (Fig 7). The software identified almost all of the known targets (21 of 22) (Supplementary Table 4). Similarly, TFM-Explorer analysis of L283 (Late genes with known NFκB target genes removed) identified 178 sequences with target sites, while analysis of L319 identified 202 sequences with NFκB sites, including 24 out of 37 known NFκB targets (Fig 7). Similar analysis of U475 identified 300 gene promoters with NFκB motifs, and an average of the results for testing five A500 groups (iterated sampling of the complete 560 up-regulated and annotated complete gene set) shows 310 sequences were identified (totaling 427 putative targets), which on average found 45 of the 70 known NFκB target genes in this group (totaling 58 when summed across all five replicates).

Clover analysis shows similar results using 5,000 permutations and p < 0.01. In the E83 group 80 of the 83 input sequences were found to have a NFκB binding sequence, with 21 of the 22 known targets being found. For L319, Clover identified 239 putative targets including 30 out of the 37 known targets. When all DE genes are grouped together, (A544), 528 promoter regions were identified to contain a putative binding site; finding 70 out of 70 known targets. Taking the intersection of TFM-Explorer and Clover there are 72 promoter
sequences with a potential NFκB binding site in the Early group. Within these 72 promoters, 21 are known to contain a binding site, leaving 51 potential novel NFκB targets. Similarly, for the Late group 169 promoters were found by both programs, and with 24 of the promoters being known targets of NFκB, this identifies 145 potential new direct targets of NFκB. As for the complete set, using the sum of found targets by TFM-Explorer in the five subsets and the single complete set ran in Clover, 419 promoter sequences were found by both programs, with 58 being previously known targets of NFκB, yielding 358 potential new NFκB targets.

DISCUSSION

Transcriptome, DE genes and GO annotation

In this study, the Affymetrix GeneChip® Porcine Genome Array and Q-PCR were used to monitor whole-genome expression profiling of porcine MLN in response to *S. Choleraesuis* infection. Our results show that the MLN transcriptomes from non-infected and infected pigs were significantly interrogated by this approach. Expression of more than 16,000 transcripts were reproducibly detected and of these, 11,020 (68.7 %) transcripts had a significant hit to human RefSeq. Thus, we believe the transcriptional response detected in our study represents a high proportion of the porcine genomic response to *S. Choleraesuis* infection within the MLN.

Statistical analysis of differential expression revealed 1,853 genes changed their expression level within at least one of 10 possible pair wise comparisons during infection, and about 63 % of these DE genes were annotated using BLAST analysis. Because of limited availability of full-length porcine cDNA, as well as the fact that many human/mouse
genes do not have functional annotation, the 37% non-annotated genes are currently not very helpful in further understanding gene functions and pathways responding to infection. However, the discovery of these non-annotated genes as part of the transcriptional response to bacterial infection is novel, and should contribute to an improved understanding of the pathways of host response to bacteria infection across host species as better annotation becomes available.

A strong host transcriptional response at 48 hpi with S. Choleraesuis was observed, as 954 genes were differentially expressed at 48 hpi compared to non-infected pigs (Fig 1). As expected, we found many genes with large gene expression changes, such as cytokines, chemokines and heat shock proteins; these generally overlap with sets of genes that have been implicated in host response to infection by others (Eckmann et al., 2001; Tukel et al., 2006), indicating our data can be integrated with similar results in other species. However, we also identified a number of genes that had not previously been shown to be involved in host’s response to bacterial infection, such as STEAP4, PSTPIP2, LIPG, HK3, STXBP1, ITPR1, CA3, KRT17 and PLN. Even though the biological function of these genes during infection remains unclear, these data more fully describe the transcriptional response to S. Choleraesuis and suggest additional functional roles for these genes. Furthermore, the DE genes with known immune functions and those with unknown functions add to the list of candidate genes to investigate for associations between immune related traits and DNA-level variation; polymorphisms at these candidate genes might result in valuable markers for enhancing disease resistance, pig health, and food safety through molecular breeding methods.

The GO consortium provides a defined vocabulary of gene functions in cells. GO
terms are now widely accepted as a useful means to annotate gene array elements. Our laboratory-designed GO-slim, which emphasized immune response annotations, was helpful in further defining the biological significance of the observed transcriptional response. Compared to the transcriptome GO term totals, genes which were assigned GO terms associated with cell adhesion, apoptosis, immune response, inflammatory response and acute-phase response were significantly enriched in our DE gene list. We also analyzed up-regulated and down-regulated groups of genes identified by cluster analysis, and found that a significant number of up-regulated genes were annotated as immune response, inflammation and defense response related genes, indicating that induction of gene expression (rather than repression) is the main result of immune response during Salmonella infection.

**Cluster analysis**

Co-expression of genes of known function with poorly characterized or novel genes may provide a simple means of gaining leads to the functions of genes for which information is not currently available. In our study, hierarchical cluster analysis was performed on 1,853 genes that were differentially expressed during infection. As we observed in porcine MLN response to S. Typhimurium (Wang et al., 2007), some ribosome- and/or translation-related genes were repressed at 8 hpi and 24 hpi. This effect is also similar to the response to LPS in skeletal muscle of neonatal and adult pigs (Lang et al., 2000; Orellana et al., 2004) and to the response to endotoxin in human blood leukocytes (Calvano et al., 2005), where a large number of genes involved in translation were repressed. Thus, we speculate that an early pathogenic effect of multiple serovars of Salmonella on the host is suppression of translation.

Three gene groups (Fig. 2, group C, D and E) in the induced cluster exhibited different expression patterns but shared a common feature, a peak expression response at 48
hpi. As large numbers of NFκB target genes were found in groups C and D, we discuss their expression features in the context of NFκB signaling pathways in Section 4.4.

**Pathway analysis**

A predominant Th1 immune response was observed during infection, as 6 of 11 investigated Th1 associated genes, including TNF, IFNG, and some IFNG-signaling responsive genes (SOCS1, STAT1, WARS and IRF1), were strongly up-regulated at 24 hpi and/or 48 hpi (Fig. 4B). This is consistent with the results in porcine lung during *S. Choleraesuis* infection (Zhao et al., 2006) and in porcine MLN in both *S. Typhimurium* and *S. Choleraesuis* infection (Uthe et al., 2007; Wang et al., 2007). However, IL12A and IL12B, which are thought to also favor Th1 cell development, did not change their expression level during infection. Their low expression level in response to *S. Choleraesuis* has also been detected by QPCR analysis of infected porcine MLN by our group (Uthe et al., 2007). A low expression level of IL12 has been also observed in the pig in response to porcine reproductive and respiratory syndrome virus (Meier et al., 2004) and in pig MLN response to *S. Typhimurium* infection (Wang et al., 2007) and *Toxoplasma gondii* (Dawson et al., 2005). We speculate that lack of IL12 induction during *S. Choleraesuis* infection might stifle IFNG induction and negatively affect host defense against *Salmonella*.

Genes involved in innate immunity and inflammatory pathways, such as IL6, IL8, IL1B, SLC11A1 and TLR4, were up-regulated significantly at 48 hpi, which indicates that these pathways were strongly activated. The increase in expression for many of these genes was recently confirmed by Q-PCR analysis (Uthe et al., 2007). However, these innate immunity genes (IL6, IL8, IL1B, SLC11A1 and TLR4) showed no expression differences between non-infected and infected porcine lung during *S. Choleraesuis* infection (Zhao et al,
Because the innate immune response is concentrated at the site of bacterial entry and closely associated lymph nodes, this may indicate that the gut was the primary infection site, rather than the respiratory tract, in these nasally infected piglets.

Our microarray data shows that apoptosis pathway related genes displayed a strong induction at 24 hpi and reached a peak response at 48 hpi in response to *S. Choleraesuis* infection. Multiple genes within gene families involved in apoptosis pathways were often induced significantly, such as genes in the caspase family (CASP1, CASP3 and CASP4) and in the transglutaminase family (TGM1, TGM2 and TGM3). The caspases are a family of cysteine proteases important in not only initiating and executing apoptosis, but also processing and maturation of the inflammatory cytokines IL-1b and IL-18 (Nicholson, 1999). The CASP1 gene is known to be important in inflammation due to its role in maturation of IL-1b and IL-18. Even though it is unclear whether CASP1 plays a direct role in apoptosis, overexpression of CASP1 has been shown to cause apoptosis in a variety of cell lines (Monack et al., 2001). A recent study revealed that IPAF is an activator of CASP1 and IL-1b in *Salmonella* infected macrophages (Franchi et al., 2006); unfortunately, we did not find any oligonucleotide set representing the IPAF gene on the Affymetrix microarray. Like CASP1, CASP4 has been shown to process pro-IL-18 and IL-1F7b, albeit inefficiently, and to cleave CASP3 into its active form. One important characteristic of CASP4 is its robust induction by INFG and the NFκB complex (Chawla-Sarkar et al., 2003; Schauvliege et al., 2002).

Genes from the transglutaminase family can be significantly induced in porcine lung during *S. Choleraesuis* infection (Zhao et al., 2006) and in porcine MLN during *S. Typhimurium* infection (Wang et al., 2007). In this study, TGM1, TGM2 and TGM3 showed
exactly the same expression pattern: activation was initiated at 24 hpi and peaked at 48 h post S. Choleraesuis infection in porcine MLN. The TGM3 gene showed a very strong induction at 48 hpi, with the fold change about 600 compared to non-infected animals by Q-PCR. There is evidence that TGM2 is a NFκB dependent gene (Mirza et al., 1997), and researchers have shown that increased TGM2 activity can trigger NFκB activation through an unusual IκB polymerization reaction rather than IκB degradation through IKK signaling (Kim, 2006). Although transglutaminase genes have been demonstrated to elevate their expression level during inflammation and to play a physiological role in mediating defense against injury or infection in various cell types (Kim, 2006), their role in apoptosis is not yet clear.

Another important cellular immune response to infection that occurs in the MLN is antigen processing and presentation, as phagocytic cells in the MLN communicate with T cells for further immune activation and initiation of adaptive immune responses. Three of four antigen processing related genes (Fig. 4B) exhibited an increased expression level at 48 hpi, which is consistent with the gene expression patterns that were observed in porcine lung during S. Choleraesuis infection by Zhao et al (2006). Interestingly, two markers involved in the antigen presentation activation pathway, CD80 and CD86, did not show significant expression changes at 24 hpi and 48 hpi. These data tempt us to speculate that the host DC-mediated antigen presentation pathway was altered early after infection. Antigen presentation by murine DC cells can be inhibited by S. Typhimurium (Tobar et al., 2004; Cheminay et al., 2005), and we observed that both CD80 and CD86 were down-regulated at 8 hpi in S. Typhimurium infected MLN (Wang et al., 2007). Thus, we predict that lack of a strong DC-mediated antigen presentation might be a mechanism that permits S. Choleraesuis
to escape the porcine GALT and cause a systemic infection.

One characteristic of host MLN transcriptional response to *S. Choleraesuis* is that several groups of genes with annotations for calcium binding activities changed their RNA expression levels significantly during infection. Calcium binding proteins are important molecules in the transduction of calcium signaling, which evoke various cellular processes, such as cell migration, cell differentiation, cell death and cell growth (Foell et al., 2004). The S100 gene family is the largest group of calcium binding proteins; two members of this gene family, S100A9 and S100A12, were up-regulated significantly with a large fold change at 48 hpi relative to non-infected animals in our study. Induction of S100A9 has been also observed in porcine Peyer’s patch in response to *S. Choleraesuis* infection (Hyland et al., 2006). Overexpression of S100A9 and S100A12 at the site of inflammation has been well described in humans, and S100A8/S100A9 and S100A12 are used as clinical laboratory markers for inflammation (Foell et al., 2004). The exact biological function of these genes remains to be defined in greater detail, even though there is evidence that they have anti-microbial properties (Sohnle et al., 2000) and are involved in induction of apoptosis (Yui et al., 2003). Other genes with GO annotation for calcium binding activity, including annexins (ANXA1, ANXA5 and ANXA8) and transglutaminases (TGM1, TGM2 and TGM3), showed differential expression at 24 and/or 48 hpi compared to non-infected animals. Calcium is an important second messenger in cells and changes in calcium pathways can evoke various cellular processes (Foell et al., 2004). Our report is the first to describe the strong induction of multiple calcium binding protein genes in MLN of *S. Choleraesuis* infected pigs and adds new information to host transcriptional response to gram-negative bacteria infection in pig and other species.
**NFκB signaling pathway**

NFκB is a latent transcription factor held in the cytoplasm through binding by its inhibitors IκB. During stimulation or infection, IκB is phosphorylated and degraded, allowing NFκB translocating into the nucleus to induce expression of many genes (Janeway et al., 2005). The NFκB target genes control a variety of cellular processes. Both microarray and QPCR data analysis revealed that many known NFκB target genes were significantly up-regulated from 24 hpi to 48 hpi, indicating a strong NFκB response during acute *S. Choleraesuis* infection. Both the magnitude and timing of this response was different from the response to *S. Typhimurium*, where suppression of the NFκB pathway from 24 to 48 hpi was observed (Wang et al., 2007). Thus understanding the transcriptional profiles of NFκB target genes is an important step to further understand its regulatory function in bacterial infection. Different activation times for NFκB target genes during stimulation has been intensively studied (Tian et al., 2005), e.g., LPS-stimulated mouse macrophages demonstrated that NFκB binding occurs in two distinct waves due to different rates of NFκB recruitment (Saccani et al., 2001). Tian et al. (2005) identified “Early”, “Middle” and “Late” expression profiles of NFκB target genes during the TNF stimulation in epithelial cells, since these genes have peak response at 1 h, 3 h and 6 h, respectively. And also such NFκB responses were confirmed in MRC-5 fibroblasts stimulated with IL-1. Three mechanisms for these differential expression patterns of NFκB target genes were suggested; a) Early and Late gene promoters are bound by NFκB complexes containing different subunits; b) there are different environments in which the NFκB binding sites are located between the Early and Late gene promoters or c) genes in the Late group undergo an additional rate-limiting step necessary for promoter activation (Tian et al., 2005).
In our study, two groups of NFκB target genes were identified, an “Early group” and a “Late group”. To determine the biological functions of genes in the “Early” versus the “Late” groups, GO annotations were assigned. GO terms for cytokine activity or chemokine activity were enriched in the “Early group” genes (Fisher’s exact test p=0.053), while the “Late group” was predominantly composed of genes with signal transduction and cell metabolism annotations. This is consistent with the functional categorization of NFκB dependent genes in “early” and “late” groups of Tian et al (2005). The rapid induction of cytokines and chemokines during infection clearly be important for recruiting immune cells to infection sites.

Both microarray and Q-PCR data suggested that NFKBIA (which encodes IκBα) increased its RNA level slightly at 48 hpi. There is overwhelming evidence that demonstrates that expression of this inhibitory gene is activated by NFκB in a negative feedback loop, which provides an effective mechanism for controlling NFκB activity (Sun et al., 1993). It has also been shown that NFκB DNA binding activity is at its peak when NFKBIA is newly synthesized (Saccani et al., 2001). Thus, as we observe up-regulation of NFKBIA and strong activation of other NFκB target genes at 24 hpi, we speculate that NFκB activity was initiated at 24 hpi and peaked at 48 hpi. This contrasts to the response of NFκB target genes to S. Typhimurium infection, where NFKBIA was not up-regulated significantly at 48 hpi, and many genes were suppressed from 24 hpi to 48 hpi (Wang et al., 2007).

Identification of potential porcine NFκB targets

Using TFM-Explorer and Clover, it was possible to identify promoter regions containing
NFκB motifs in human genes orthologous to the porcine genes which had Early (within 24 hpi) and Late (by 48 hpi) response to *Salmonella* infection. An increase from 13 to 34 putative NFκB targets were found in the Early group genes by comparing computational results after removal of known NFκB target genes to those results with the known targets included. We have confidence that our analysis has identified strong candidate NFκB target genes, as the cutoff for identifying significant over-representation windows was stringent, given that the non-zero possibility that groups of randomly selected genes used as the background in the TFM Explorer analysis will contain some NFκB targets. Some of the known targets were missed by our TFM-Explorer analysis, but the input sequence window used was only 2,000 total bases. Thus for these known targets, there might be some binding sites outside of the input sequence range. However, we were able to identify 81% of the known targets in the E83 group, indicating that our sensitivity was adequate to find the large majority of known target genes.

Results for DNA regulatory motif analysis of human orthologs for the stimulated gene groups provided evidence for 58 putative NFκB targets in the Early group and 145 in the Late group. These putative targets were also analyzed by GO annotation to further develop our understanding of the regulatory response to *Salmonella* infection (data not shown), as performed on the known NFκB target genes (Figure 5A). These putative target genes, including SOCS1, SCARB2, CEBPD, CXCL16, IL1RAP, and CASP7, are involved in multiple cellular processes, including cell adhesion, regulation of transcription, immune response, and receptor activity. While there is no additional experimental evidence that these genes are direct NFκB targets, many of them have been previously reported to be involved in
the response to bacterial infection (Uthe et al., 2007; Fahy et al., 2006; Bruckmaier et al., 2005). Although further experimental work is needed to confirm that these genes are true NFκB targets, our analysis detecting NFκB regulatory motifs in these genes further supports the hypothesis that they constitute an important part of the early anti-bacterial infection response.

To check our PubMed literature search results that show no reports on these genes as NFκB direct targets, Pathway Studio (a PubMed text-mining software) analysis was run on the complete sets of Early and Late genes. Pathway Studio identified seven Early group genes having literature evidence of direct binding of NFκB to their promoter region (CEBPD, IL1B, IL6, IL8, FOS, EGR1, PTGS2), while for the late group only two were identified as having literature evidence of direct binding (TNF, IFNG). In addition, in the Early group, ten other genes were identified as having a connection with NFκB in the literature, usually sharing a common expression pattern (CSF3, CYCS, OAS1, CEBPB, CELP, IER3, SOD2, CCL2, HLA-A, CXCL2). For the Late group, an additional 11 were also found to have a connection with NFκB (LYZ, SP6, IL10, TLR2, BCL2, APP, HXB, CCR5, ANXA5, IL15, CDKN1A). TFM-Explorer and Clover found motifs at the promoters of all genes with direct binding evidence, all but one of the NFκB-connected genes (CYCS) identified in the Early group (94%), and all but two (CCR5, ANXA5) of the 13 (77%) in the Late group. Of the genes that were predicted to contain an NFκB binding site by both programs and had evidence from Pathway Studio as having a direct binding relationship with NFκB, only CEBPD and EGR1 were not previously on the “known” list of targets at available websites (see Methods). However, due to more recent literature evidence provided
by the Pathway Studio software, they should be considered “known” NFκB targets. Thus, this literature analysis indicates that the large majority of genes predicted to be NFκB targets based on co-expression and motif data have not yet been recognized as direct NFκB target genes.

Summary

This study investigated the host transcriptional response to S. Choleraesuis infection in porcine MLN using the Affymetrix porcine GeneChip® Transcriptome. A large number of differentially expressed genes \( (p<0.01, \text{FC}>2) \) were identified and functional analyses of these genes was performed by GO annotation enrichment calculations. Gene hierarchical cluster analysis and specific pathway analysis revealed several specific features of porcine host response to infection, and expanded gene targets for future genomic studies. Both microarray and QPCR data provided evidence of a strong NF-κB-dependent host transcriptional response during S. Choleraesuis infection. Promoter searching using TFM-Explorer identified 36 putative novel NF-κB direct targets in the Early group (8-24 hpi) of genes and 148 in the Late group (48 hpi). While experimental validation is needed to confirm these predicted NF-κB binding sites and regulatory relationships, our study adds new data towards an understanding of the NF-κB signaling pathway response to Gram-negative bacteria infection.

ACKNOWLEDGEMENTS

We thank Dr. Tom Stabel for collaboration in the production of these challenge populations. This project received support from the ISU Agriculture Experiment Station/Center for Integrated Animal Genomics, the USDA-ARS-NADC, the USDA-ARS-
BARC, and USDA-NRI 2004-35205-14202. O.C. acknowledges support of the USDA-Food and Agricultural Sciences-Multidisciplinary Graduate Education and Training Grant 2001-52100-11506.
REFERENCES


immune factors in the development of the host response to PRRSV vaccination. Vet Immunol Immunopathol. 102, 199-216.


porcine mesenteric lymph nodes to Salmonella enterica serovar Typhimurium. Genomics.


**TABLE 1.**

Genes showing differential expression of large magnitude ($p$-value<0.01 and estimated FC >10) at 48 hpi as compared to uninfected pigs.

<table>
<thead>
<tr>
<th>affyID</th>
<th>Gene Name</th>
<th>Human Refseq ID</th>
<th>$p$ value</th>
<th>8h/C</th>
<th>24h/C</th>
<th>48h/C</th>
<th>21d/C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ssc.2381.1.S1_at</td>
<td>S100A9</td>
<td>NM_002965</td>
<td>0.0004</td>
<td>2.25</td>
<td>9.81</td>
<td>226.80</td>
<td>5.90</td>
</tr>
<tr>
<td>Ssc.719.1.S1_a_at</td>
<td>CXCL5</td>
<td>NM_002993</td>
<td>0.0015</td>
<td>3.09</td>
<td>6.28</td>
<td>102.29</td>
<td>4.41</td>
</tr>
<tr>
<td>Ssc.16008.1.S1_at</td>
<td>FCN2</td>
<td>NM_002003</td>
<td>0.0041</td>
<td>1.36</td>
<td>7.90</td>
<td>52.98</td>
<td>1.40</td>
</tr>
<tr>
<td>Ssc.17573.1.S1_at</td>
<td>IL1B</td>
<td>NM_000576</td>
<td>0.0000</td>
<td>1.16</td>
<td>3.70</td>
<td>39.31</td>
<td>1.15</td>
</tr>
<tr>
<td>Ssc.8162.1.S1_at</td>
<td>PTX3</td>
<td>NM_002852</td>
<td>0.0009</td>
<td>-1.40</td>
<td>2.61</td>
<td>39.10</td>
<td>1.05</td>
</tr>
<tr>
<td>Ssc.9117.1.S1_at</td>
<td>S100A12</td>
<td>NM_005621</td>
<td>0.0040</td>
<td>1.46</td>
<td>4.29</td>
<td>38.85</td>
<td>4.89</td>
</tr>
<tr>
<td>Ssc.16250.1.S2_at</td>
<td>IL1RN</td>
<td>NM_173841</td>
<td>0.0002</td>
<td>1.20</td>
<td>6.20</td>
<td>33.79</td>
<td>1.81</td>
</tr>
<tr>
<td>Ssc.658.1.S1_at</td>
<td>IL8</td>
<td>NM_000584</td>
<td>0.0002</td>
<td>2.13</td>
<td>3.17</td>
<td>31.80</td>
<td>2.33</td>
</tr>
<tr>
<td>Ssc.4093.1.A1_at</td>
<td>IFNG</td>
<td>NM_000619</td>
<td>0.0000</td>
<td>-1.27</td>
<td>1.26</td>
<td>28.83</td>
<td>-1.19</td>
</tr>
<tr>
<td>Ssc.13769.1.S1_at</td>
<td>LTF</td>
<td>NM_002343</td>
<td>0.0024</td>
<td>2.55</td>
<td>3.47</td>
<td>26.31</td>
<td>1.89</td>
</tr>
<tr>
<td>Ssc.9114.1.S1_at</td>
<td>STEAP4</td>
<td>NM_024636</td>
<td>0.0011</td>
<td>-1.25</td>
<td>2.92</td>
<td>20.85</td>
<td>2.83</td>
</tr>
<tr>
<td>Ssc.11009.1.A1_at</td>
<td></td>
<td></td>
<td>0.0033</td>
<td>-1.12</td>
<td>5.59</td>
<td>20.81</td>
<td>2.95</td>
</tr>
<tr>
<td>Ssc.4871.1.S1_at</td>
<td>CXCL2</td>
<td>NM_002089</td>
<td>0.0000</td>
<td>1.17</td>
<td>9.02</td>
<td>17.40</td>
<td>1.48</td>
</tr>
<tr>
<td>Ssc.300.1.S1_at</td>
<td>SLC11A1</td>
<td>NM_000578</td>
<td>0.0004</td>
<td>1.32</td>
<td>2.49</td>
<td>16.79</td>
<td>1.82</td>
</tr>
<tr>
<td>Ssc.16228.1.S1_at</td>
<td>PPBP</td>
<td>NM_002704</td>
<td>0.0003</td>
<td>-1.52</td>
<td>-1.38</td>
<td>16.17</td>
<td>-1.06</td>
</tr>
<tr>
<td>Ssc.5053.1.S1_at</td>
<td>CD163</td>
<td>NM_203416</td>
<td>0.0002</td>
<td>2.01</td>
<td>1.75</td>
<td>16.02</td>
<td>4.69</td>
</tr>
<tr>
<td>Ssc.27433.1.S1_at</td>
<td>TGM1</td>
<td>NM_000359</td>
<td>0.0003</td>
<td>1.97</td>
<td>3.38</td>
<td>14.78</td>
<td>1.56</td>
</tr>
<tr>
<td>Ssc.30887.1.S1_at</td>
<td>TNFAIP6</td>
<td>NM_007115</td>
<td>0.0009</td>
<td>-9.58</td>
<td>2.42</td>
<td>14.44</td>
<td>1.01</td>
</tr>
<tr>
<td>Ssc.30027.1.A1_at</td>
<td></td>
<td></td>
<td>0.0020</td>
<td>1.64</td>
<td>2.26</td>
<td>14.30</td>
<td>1.18</td>
</tr>
<tr>
<td>Ssc.18261.1.S1_at</td>
<td>PSTPIP2</td>
<td>NM_024430</td>
<td>0.0000</td>
<td>-1.02</td>
<td>1.68</td>
<td>13.30</td>
<td>2.20</td>
</tr>
<tr>
<td>Ssc.16151.1.S1_at</td>
<td>CSF3</td>
<td>NM_172219</td>
<td>0.0032</td>
<td>1.22</td>
<td>2.78</td>
<td>13.15</td>
<td>-1.15</td>
</tr>
<tr>
<td>Ssc.62.2.S1_a_at</td>
<td>IL6</td>
<td>NM_000600</td>
<td>0.0005</td>
<td>1.18</td>
<td>3.03</td>
<td>13.12</td>
<td>1.41</td>
</tr>
<tr>
<td>Ssc.21663.1.A1_at</td>
<td>LIPG</td>
<td>NM_006033</td>
<td>0.0001</td>
<td>1.35</td>
<td>2.94</td>
<td>12.95</td>
<td>1.78</td>
</tr>
<tr>
<td>Ssc.5743.1.S1_a_at</td>
<td>HK3</td>
<td>NM_002115</td>
<td>0.0002</td>
<td>1.04</td>
<td>1.94</td>
<td>11.64</td>
<td>1.32</td>
</tr>
<tr>
<td>Ssc.24282.1.S1_at</td>
<td>CXCL1</td>
<td>NM_001511</td>
<td>0.0000</td>
<td>-1.34</td>
<td>2.26</td>
<td>11.40</td>
<td>1.26</td>
</tr>
<tr>
<td>Ssc.6797.1.S1_at</td>
<td>STXB1</td>
<td>NM_003165</td>
<td>0.0000</td>
<td>1.48</td>
<td>2.13</td>
<td>11.35</td>
<td>1.83</td>
</tr>
<tr>
<td>Ssc.11784.1.S1_at</td>
<td>TIMP1</td>
<td>NM_003254</td>
<td>0.0000</td>
<td>1.15</td>
<td>2.44</td>
<td>11.27</td>
<td>1.23</td>
</tr>
<tr>
<td>Ssc.7314.1.A1_at</td>
<td>PTGS2</td>
<td>NM_000963</td>
<td>0.0001</td>
<td>1.43</td>
<td>2.79</td>
<td>10.93</td>
<td>1.27</td>
</tr>
<tr>
<td>Ssc.30833.1.S1_at</td>
<td>CCL3</td>
<td>NM_021006</td>
<td>0.0033</td>
<td>2.36</td>
<td>3.01</td>
<td>10.67</td>
<td>1.35</td>
</tr>
<tr>
<td>Ssc.3706.1.S2_at</td>
<td>SOD2</td>
<td>NM_000636</td>
<td>0.0000</td>
<td>1.10</td>
<td>2.08</td>
<td>10.52</td>
<td>1.16</td>
</tr>
<tr>
<td>Ssc.7864.1.A1_at</td>
<td>IL1RAP</td>
<td>NM_002182</td>
<td>0.0000</td>
<td>-1.09</td>
<td>1.95</td>
<td>10.42</td>
<td>1.33</td>
</tr>
<tr>
<td>Ssc.16182.1.S1_at</td>
<td>ITPR1</td>
<td>NM_002222</td>
<td>0.0018</td>
<td>6.03</td>
<td>3.97</td>
<td>10.38</td>
<td>9.08</td>
</tr>
<tr>
<td>Ssc.18080.1.A1_at</td>
<td>C8orf46</td>
<td>NM_152765</td>
<td>0.0093</td>
<td>-1.45</td>
<td>-1.30</td>
<td>-10.34</td>
<td>-1.26</td>
</tr>
<tr>
<td>Ssc.10960.1.S1_at</td>
<td>CA3</td>
<td>NM_005181</td>
<td>0.0003</td>
<td>-1.17</td>
<td>-3.30</td>
<td>-10.50</td>
<td>-1.39</td>
</tr>
<tr>
<td>Ssc.8683.1.S1_at</td>
<td></td>
<td></td>
<td>0.0065</td>
<td>-1.11</td>
<td>-1.87</td>
<td>-10.61</td>
<td>1.26</td>
</tr>
<tr>
<td>Ssc.20230.1.S1_at</td>
<td></td>
<td></td>
<td>0.0036</td>
<td>-2.74</td>
<td>-1.53</td>
<td>-11.60</td>
<td>-1.37</td>
</tr>
<tr>
<td>Ssc.11065.1.A1_at</td>
<td>KRT17</td>
<td>NM_000422</td>
<td>0.0006</td>
<td>-2.06</td>
<td>-2.52</td>
<td>-11.93</td>
<td>1.15</td>
</tr>
<tr>
<td>Ssc.30225.1.A1_at</td>
<td></td>
<td></td>
<td>0.0007</td>
<td>-1.21</td>
<td>-2.54</td>
<td>-16.30</td>
<td>-1.13</td>
</tr>
<tr>
<td>Ssc.5227.1.S1_at</td>
<td>PLN</td>
<td>NM_002667</td>
<td>0.0034</td>
<td>1.06</td>
<td>-1.70</td>
<td>-18.57</td>
<td>1.80</td>
</tr>
<tr>
<td>Ssc.14164.1.A1_at</td>
<td></td>
<td></td>
<td>0.0000</td>
<td>-1.34</td>
<td>1.27</td>
<td>-22.76</td>
<td>1.32</td>
</tr>
<tr>
<td>Ssc.18150.1.A1_at</td>
<td></td>
<td></td>
<td>0.0027</td>
<td>-1.54</td>
<td>-4.65</td>
<td>-25.37</td>
<td>-1.25</td>
</tr>
</tbody>
</table>
TABLE 2.
Q-PCR results for gene expression (C_t values, SD) at each early response stage (8 hpi, 24 hpi and 48 hpi) in S. Choleraesuis infection.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>SC - Control</th>
<th>SC - 8h</th>
<th>SC - 24h</th>
<th>SC - 48h</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>SD</td>
<td>Stat</td>
<td>Average</td>
</tr>
<tr>
<td>IL15</td>
<td>26.5</td>
<td>0.67</td>
<td>A</td>
<td>25.7</td>
</tr>
<tr>
<td>CCL2</td>
<td>21.6</td>
<td>1.27</td>
<td>A</td>
<td>21.4</td>
</tr>
<tr>
<td>CCL3</td>
<td>29.0</td>
<td>1.03</td>
<td>A</td>
<td>28.2</td>
</tr>
<tr>
<td>CCR5</td>
<td>27.5</td>
<td>1.12</td>
<td>A</td>
<td>26.9</td>
</tr>
<tr>
<td>GBP1</td>
<td>22.8</td>
<td>0.97</td>
<td>A</td>
<td>22.2</td>
</tr>
<tr>
<td>GBP2</td>
<td>24.4</td>
<td>0.68</td>
<td>A</td>
<td>23.7</td>
</tr>
<tr>
<td>ICAM1</td>
<td>29.0</td>
<td>0.60</td>
<td>A</td>
<td>29.6</td>
</tr>
<tr>
<td>TAP1</td>
<td>22.2</td>
<td>0.57</td>
<td>A</td>
<td>22.4</td>
</tr>
<tr>
<td>TLR2</td>
<td>25.6</td>
<td>0.83</td>
<td>A</td>
<td>25.4</td>
</tr>
<tr>
<td>TGM3</td>
<td>31.8</td>
<td>0.67</td>
<td>A</td>
<td>31.3</td>
</tr>
<tr>
<td>GZMB</td>
<td>23.5</td>
<td>0.92</td>
<td>A</td>
<td>23.3</td>
</tr>
<tr>
<td>CD14</td>
<td>28.3</td>
<td>0.82</td>
<td>A</td>
<td>26.8</td>
</tr>
<tr>
<td>IL1A</td>
<td>28.9</td>
<td>0.85</td>
<td>A</td>
<td>27.9</td>
</tr>
<tr>
<td>PPBP</td>
<td>34.3</td>
<td>1.10</td>
<td>A</td>
<td>35.0</td>
</tr>
<tr>
<td>JUNB</td>
<td>30.8</td>
<td>0.68</td>
<td>A</td>
<td>30.1</td>
</tr>
<tr>
<td>IKBA</td>
<td>25.6</td>
<td>0.63</td>
<td>A</td>
<td>25.6</td>
</tr>
<tr>
<td>NFKBIZ</td>
<td>23.4</td>
<td>0.69</td>
<td>A</td>
<td>22.7</td>
</tr>
<tr>
<td>CD163</td>
<td>26.5</td>
<td>0.90</td>
<td>A</td>
<td>25.7</td>
</tr>
<tr>
<td>CXCL5</td>
<td>29.4</td>
<td>0.96</td>
<td>A</td>
<td>28.8</td>
</tr>
<tr>
<td>PTX3</td>
<td>28.7</td>
<td>0.71</td>
<td>A</td>
<td>28.6</td>
</tr>
<tr>
<td>TREM1</td>
<td>30.5</td>
<td>0.91</td>
<td>A</td>
<td>30.5</td>
</tr>
<tr>
<td>CD4</td>
<td>23.8</td>
<td>0.68</td>
<td>A</td>
<td>23.3</td>
</tr>
<tr>
<td>RPL32</td>
<td>16.9</td>
<td>0.34</td>
<td>A</td>
<td>16.9</td>
</tr>
</tbody>
</table>

Expression values not sharing the same letters are statistically different across time points.
FIGURE LEGENDS

**Figure 1.** Differentially expressed genes at each time point during infection with *S. Choleraesuis* (*p*<0.01, FC>2, *q*<0.26), compared to non-infected pigs.

**Figure 2.** Hierarchical clustering analysis of 1,853 genes from Affymetrix array analysis showing differential expression patterns during *S. Choleraesuis* infection. Differential expression criteria: *p*-value<0.01, estimated FC>2, *q*<0.26. The heat map was built by using Gene Cluster 3.0 software. Red color represents down-regulation and green shows up-regulation.

**Figure 3.** Functional annotation of genes in the porcine MLN transcriptome and of genes differentially expressed (DE) between infected and non-infected pigs.

A. Gene Ontology (GO) annotation of the porcine MLN transcriptome and differentially expressed (DE) genes (*p*<0.01, FC>2, *q*<0.26). All 16,046 probesets which were expressed in infected and non-infected porcine MLN (gray bars) and 1,853 DE genes (white bars) were annotated using our own specific GO-slim for biological processes and molecular functions relevant to immune response. The percentage of a given GO annotation across all annotations in the DE gene list, as compared to percentage annotation of the complete transcriptome, is shown. Statistical significance is denoted with an asterisk (*p*<0.05 and ** *p*<0.01).

B. GO annotations for up-regulated and down-regulated genes from induced and repressed gene clusters. Analysis details and statistics are denoted as in A.

**Figure 4.** Transcriptional profile of selected cell-type marker and immune response pathway genes by Affymetrix DNA microarray analysis. Expression patterns of specific marker genes for T cell, macrophage, granulocyte and dendritic cell types (A) or important immune
response pathways (B) are shown. The fold change from comparisons of infected pigs and non-infected controls at each time point were calculated from the Affymetrix array data using Genecluster. Statistical differences \( (p<0.05) \) between control and infected pigs are represented by an asterisk (*).

**Figure 5.** Gene Ontology analysis of Early and Late Groups of Genes Responding to Salmonella

The GO-slim described in Figure 3 was used for this analysis. The percentage of a given GO annotation across all annotations in the Early gene list, as compared to percentage annotation of the Late gene list, is shown.

**Figure 6.** Quantitative PCR analysis validates transcriptional profiling data for genes responding to *S*. Choleraesuis infection. Real-time Q-PCR data is presented as the fold change in gene expression in infected pigs compared to the negative controls. In the last column, gene expression from 48 hpi is compared to that at 24 hpi. Statistical significance \( (p<0.05) \) is denoted with an asterisk (*). † TGM3 showed dramatic increases in porcine lung during *S*. Choleraesuis infection, thus TGM3 was also selected for Q-PCR analysis. ‡ Even though TREM1 is not a known NFκB target gene, it was selected for Q-PCR analysis due to its similar expression pattern with NFκB known target genes.

**Figure 7.** Computational identification of NFκB binding site motifs at human orthlogs of genes from Early, Late, and a combined All groups predict putative NFκB genes responding to *S*. Choleraesuis using the intersection of Clover and TFM-Explorer. Each bar graph shows for each program, Clover or TFM-Explorer, and their intersection for each group: a) the number of known target genes whose promoter contains a NFκB motif (light gray), b) the number of known target genes whose promoter was not identified as having a NFκB motif.
(white), c) the number of unknown NFκB target genes whose promoter was identified as having a motif (dark gray), and d) the number of unknown target genes whose promoter was not identified as having a NFκB motif (black).
Figure 1. Wang, Couture et al., 2008.
Figure 2. Wang, Couture et al., 2008.
Figure 3A. Wang, Couture et al., 2008.
Figure 3B. Wang, Couture et al., 2008.
Figure 4. Wang, Couture et al., 2008.
Figure 5. Wang, Couture et al., 2008.
Figure 6. Wang, Couture et al., 2008.
Figure 7. Wang, Couture et al., 2008.
CHAPTER 5: CONFIRMATION OF BIOINFORMATICALLY PREDICTED NOVEL TARGETS OF NF-κB: H2-Eb1 AND TRIM26 IN MOUSE AND C2 AND UBD IN PIG

Modified from a manuscript to be submitted to BMC Genomics

Oliver Couture, Elizabeth Kenkel, Amanda Raimer-Tait, Jolita Utke, Curt Christian, Nowlan Freese, Michael Wannemuehler, Christopher Tuggle

ABSTRACT

Background

Salmonella enterica serovar Typhimurium (ST) is an important pathogenic Salmonella serovar. Its ubiquity and broad host range, which includes all vertebrate food animals, can lead to human illness from food borne contamination. ST also causes different diseases in different hosts, and thus to combat this zoonotic pathogen, it is important to understand the mechanisms underlying the disparate host immune responses to ST. NF-κB is one of the main transcriptional regulators of the innate immune recognition of Salmonella.

Results

In this paper we test the ability of NF-κB to regulate the mRNA expression level of 12 mouse genes and their predicted orthologs in pigs. These genes were previously predicted to be potential NF-κB targets, as their expression patterns following an experimental Salmonella infection in pigs clustered with known NF-κB target genes. We used an electrophoretic mobility shift assay (EMSA) to verify that predicted motifs can be bound by NF-κB. We then stimulated both murine and porcine macrophages with ST endotoxin with or without the NF-κB inhibitor SC-514 and used q-RT-PCR to determine those genes whose
RNA induction due to endotoxin stimulation is dependent on NF-κB signaling. We provide evidence that C2 and UBD in pigs and Trim26 and H2-K1 in mice are novel signal pathway targets of NF-κB. The known targets of NF-κB in this study, CFB, IER3, PSMB9, TAP1, and TNF, were also confirmed in the pig.

**Conclusions**

Our results additionally show that substituting human promoter sequences when genomic sequence in related species is unavailable can be useful for predicting regulatory interactions.

**BACKGROUND**

*Salmonella enterica* serovar Typhimurium (ST) can infect a broad range of hosts, including, but not limited to, humans, swine, and mice, yet causes different diseases in its different hosts. In humans and swine, it generally causes gastroenteritis, while in mice it causes a systemic infection that can lead to death (Mittrücker and Kaufmann, 2000). Although ST causes a mild disease in pigs, it is estimated that significant economic losses can be attributed to subclinical *Salmonella* infection in pigs due to growth inefficiency (Fedorka-Cray et al., 1995). The presence of *Salmonella* in the United States was sampled in 2005 and found to be present on approximately 58% of farms across five states (CAHFSE, 2005). Such broad zoonotic ability also leads to ST infections in humans due to the consumption of infected eggs and meat, including pork (Muir et al., 1998 and Callaway, 2010). Of the different *Salmonella* spp., ST is one of the most commonly reported and isolated serotypes in human infections since 1996, with a substantial portion of these isolates showing resistance to multiple drugs. While there are immunizations for swine for ST (Roesler et al., 2006 and 2004, and Fagan et al., 2001), the rate of human *Salmonella*
infection in the United States has remained stable for the past 20 years (Hall-Baker et al., 2010).

In addition to the role of pigs as a source of food borne human disease, porcine immunity is important to understand due to the use of swine as a medical model species, such as in tissue repair using autoallographs (Shinmomura et al., 2010), immune recognition in the embryo (Clark, 2010), and effects of H1N1 virus infection (Khatri et al., 2010). Potentially important molecular differences in the immune systems of human and rodents, such as complement response to bacterial components, have been identified (Mestas and Hughes, 2004), while the pig has been identified as a preferred species for modeling the innate immune system (Butler et al., 2009). Swine are also a good candidate species for xenotransplantation due to their similar size and physiology to humans, and several genetic modifications have been made already in swine to help facilitate this (Lai and Prather, 2002; van der Windt et al., 2009).

One of the central regulatory components to innate immunity is the transcription factor NF-κB (Li and Verma, 2008). NF-κB responds to a variety of signals, including the presence of bacterial lipopolysaccharide (LPS), and activates genes involved in multiple intracellular signaling pathways, such as inflammation, cell growth, and apoptosis (Gilmore, 2006 and Perkins, 2007). Lipopolysaccharide, a component of endotoxin, activates an intracellular cascade through the binding of the TLR4/CD14/MD2 receptor complex (Gangloff et al., 2005 and Rallabhandi et al., 2006) that, working through MAPK cascades, results in activation of the IKK complex. One of the subunits of IKK, IKKβ, phosphorylates the main inhibitor of NF-κB, IκBα, which in the unphosphorylated state binds to NF-κB and sequesters it in the cytoplasm. Phosphorylation by IKKβ of IκBα results in IκBα eventually
becoming ubiquinated and ultimately degraded. Degradation of IκBα allows the nuclear signal on NF-κB to be exposed, which results in its translocation to the nucleus, allowing NF-κB to activate target genes, such as its own intracellular inhibitor, IκBα (Braiser 2006). The small molecule, SC-514, can selectively inhibit IKKβ. This inhibition results in a decreased phosphorylation and degradation of IκBα, increasing the binding of newly synthesized IκBα to NF-κB and decreasing translocation of NF-κB into the nucleus. SC-514 has also been shown to increase the export of p65 from the nucleus, further reducing its ability to direct cellular processes (Kishore et al., 2003).

In earlier work, we predicted specific genes were novel targets of NF-κB by finding porcine genes that had a similar temporal pattern of increased mRNA levels due to Salmonella enterica serovar Choleraesuis (SC) infection of the mesenteric lymph node (MLN, Wang et al., 2008). These patterns were identified by hierarchical clustering of the microarray expression data taken at 8 hours post infection (hpi), 24 hpi, 48 hpi, and 21 days post infection (dpi). Genes were identified as potential NF-κB target genes by clustering with known targets of NF-κB. We provided further evidence for NF-κB-dependent regulation of these genes by calculating an over-representation of NF-κB binding motifs in the promoter regions of human homologs to these genes, as compared to random sets of human promoters (Wang et al., 2008).

While there have been many predictions of regulatory interactions made based on clustering of expression data (i.e. Segal et al., 2003, Hudson et al., 2009, Reverter et al., 2010, Seok et al., 2010), few have confirmed these predictions through experimental methods, such as in Gilchrist et al. (2006), Quattrini et al. (2007), Ramsey et al. (2008), and
Litvak et al. (2009). In this paper, to experimentally verify previous bioinformatic predictions, we first test the ability of NF-κB to bind to predicted NF-κB binding motifs within the promoters of specific target genes from the Wang et al. (2008) gene lists through electrophoretic mobility shift assays (EMSAs) using nuclear extracts from a mouse macrophage cell line. Genes were selected based on their presence in the swine major histocompatibility complex (MHC), as the MHC has been sequenced and motif sequences were available. We also obtained the equivalent motifs for the orthologous mouse promoters, some of which are verified direct targets of NF-κB. We then induced an inflammatory response in murine and porcine macrophages by the administration of ST endotoxin, and measured the RNA levels of these known and predicted targets of NF-κB at several times post-stimulation. In parallel experiments using pretreatment with the NF-κB pathway inhibitory drug SC-514, in both mouse and pig macrophages we were able to provide evidence for NF-κB dependency for endotoxin response in novel targets and verify previously known targets of NF-κB.

MATERIALS AND METHODS

Bioinformatic prediction and selection of NF-κB targets

The predictions of the NF-κB targets tested in this paper were made in Wang et al. (2008). Briefly, pigs were treated with SC and the mesenteric lymph node was assayed for changes in gene expression using the Affymetrix porcine genome GeneChip®. The differentially expressed genes were then clustered using hierarchical clustering. Sub-clusters were chosen based on two criteria: having known NF-κB target genes as members and an increase in gene expression due to the infection. Human homologues were identified using
BLAST, and the homologues were then used to extract promoter regions consisting of the proximal region around the transcription start site (TSS) and consisted of the 1,500 bases upstream and 500 bases downstream of the TSS. The human homologues were used due to the poor availability of known porcine TSS, which is required for some algorithms to function properly. These sequences were then run through two algorithms, TFM-Explorer (Defrance and Touzet, 2006) and Clover (Frith et al., 2004), to predict if the NF-κB binding motif was over-represented within the sub-clusters, and which members of the clusters had an NF-κB motif. Both of these algorithms used the NF-κB matrices found in TRANSFAC (Matys et al., 2003) and Jaspar (Sandelin et al., 2004) as their queries.

In this paper, we focused on the candidate NF-κB target genes that mapped to the major histocompatibility complex (MHC), as it is one of the few regions of the porcine genome that has sufficient annotation (Renard et al., 2006) to know the location of the TSS for the genes found within the region, as well as, being an important genomic location for innate immunity. These sequences were extracted for both the pig and mouse genes as described above. The MATCH software (Matys et al., 2003), using TRANSFAC matrices, was used to locate the possible NF-κB binding sites within each sequence; using a MATCH score greater than 0.75 as the criterion for selecting binding sites.

**3D4/31 culture, endotoxin isolation, and 3D4/31 endotoxin treatment**

The swine macrophage-like cell line 3D4/31 was obtained from the American Type Culture Collection (ATCC, Manassas, VA) and grown as previously described by Weingartl et al., 2002. All cell culture reagents were obtained from Mediatech Inc, Manassas, VA, unless otherwise noted. Briefly, 3D4/31 cells were grown in RPMI supplemented with 1X
non-essential amino acids, 5 µg/ml gentamycin, 10 mM HEPES and 10% heat-inactivated fetal bovine serum (FBS; Atlanta Biologicals, Lawrenceville, GA) in T75 tissue culture flasks at 37°C in a 5% CO₂ humidified atmosphere. Thirty-six hours before activation, 3D4/31 cells were split and cultured in medium that contained RPMI supplemented with 1X non-essential amino acids, 5 µg/ml gentamycin, 10 mM HEPES and 10% heat-inactivated porcine serum (Sigma, St. Louis, MO). For activation, semi-confluent 3D4/31 cells were exposed to either 1 or 10 µg/ml of *Salmonella enterica* serovar Typhimurium (*S. Typhimurium*) endotoxin (STE) for 1 or 4 h. *S. typhimurium* endotoxin was prepared utilizing the aqueous butanol-1 extraction procedure as described in Morrison and Leive (1975). All cells were stimulated using the same isolation of STE.

**RAW 264.7 culture and endotoxin treatment**

The murine macrophage-like cell line RAW 264.7 was originally obtained from ATCC. RAW cells were maintained at sub-confluence at 37°C in a 5% CO₂ humidified atmosphere in Dulbecco’s Modified Eagle’s Medium (DMEM) supplemented with 10% heat-inactivated FBS, 100 IU/ml penicillin and 10 µg/ml streptomycin. For inhibition assays, cells were incubated with either 25 µm SC-514 (Calbiochem, La Jolla, CA) dissolved in DMSO, DMSO only, or PBS for 1 hour prior to activation. Cells were treated with 0 or 100 ng/ml STE for 1, 2, and 4 hours. The solution by time treatments created six groups for each time point, PN (PBS and no STE), PY (PBS and STE), DN (DMSO only and no STE), DY (DMSO only and STE), SN (SC-514 in DMSO and no STE), and SY (SC-514 in DMSO and STE). Following treatments, all cells were washed with PBS, harvested by scraping, pelleted at 1,500 x g for 3 minutes in microcentrifuge tubes, snap-frozen in liquid nitrogen, and stored at -80°C.
Porcine PBMC derived macrophage differentiation, culture, isolation, and STE treatments

To obtain porcine monocyte-derived macrophages (MΦ), peripheral blood mononuclear cells (PBMC) were isolated from whole blood by density centrifugation. Briefly, peripheral blood was diluted 1:2 (v:v) in sterile PBS, overlayed onto Lymphocyte Separation Media (LSM; Mediatech, Manassas, VA) and centrifuged at 500 x g for 30 minutes. PBMC were obtained from the LSM/plasma interface and washed twice with sterile PBS. Monocytes were further enriched to > 95% purity by positive magnetic bead selection as previously described by Bimczok et al. (2007) with some modifications. Briefly, PBMC were labeled with a primary antibody against CD172a (SWC3a; clone 74-22-15A; BD Biosciences, San Jose, CA) at a final concentration of 0.5 µg/ul for 30 minutes. After two washes, PBMC were labeled with anti-mouse IgG MicroBeads (Miltenyi Biotec, Bergisch Gladbach, Germany) according to manufacturer’s recommendations and sorted with an AutoMACS automated magnetic cell sorter (Miltenyi Biotec). Monocytes were cultured in 6-well tissue culture plates at 37°C and 5% CO₂ for 6 days in (DMEM) that contained 4.5 mg/ml glucose, 2 mM L-glutamine, 100 U/ml penicillin, 100 µg/ml streptomycin, 25 mM HEPES, 0.05 µM 2-mercaptoethanol, 10% heat-inactivated porcine serum (Sigma-Aldrich, St. Louis, MO) and 30% conditioned media from confluent cultures of L929 fibroblasts to serve as a source of colony stimulating factor. Approximately 70% of the culture media was renewed at day 3. For activation at day 6, macrophages were exposed to 10 µg/ml STE for 1, 2, 4 or 8 hours. For inhibition assays, macrophages were incubated with either 25 µM SC-514 dissolved in DMSO or with DMSO only for 1 hour prior to activation. This created only four treatment groups per time point: DN (DMSO only and no STE), DY (DMSO and STE),
SN (SC-514 in DMSO and no STE), and SY (SC-514 in DMSO and STE). Following treatments, all macrophages were washed with PBS, harvested by scraping, pelleted at 1,500 x g for 3 minutes in microcentrifuge tubes, snap-frozen in liquid nitrogen and stored at -80°C.

**Probe and Competitor Design for Electrophoretic Mobility Shift Assays**

Gene-specific probes and competitors for the electrophoretic mobility shift assays (EMSAs) were designed using the predicted motifs from section 2.1, and placed in the middle of a 24 bp oligonucleotide. If the gene had at least one known binding site, that site was used in preference to predicted sites, and used to help narrow the region to locate the most appropriate NF-κB binding site in the other species. Gene-specific probes were made by synthesizing one of the two complementary oligonucleotides with a 5’ Cy5 fluorophore (Ruscher et al., 2000 and Forwood and Jans, 2006). Three gene-specific probes were made for porcine targets: AIF1, TAP1, and UBD. In addition to these, two generic probes were designed based on a NF-κB consensus sequence shown to be useful in EMSA assays by Ajuwon et al. (2004). The generic strong (GS) probe is the published sequence; while the generic weak (GW) probe has two nucleotides changed to two bases that are seen less frequently in NF-κB-bound motifs, and results in a reduced MATCH matrix score from 1.000 to 0.886, without altering the core score. Specific competitors (SCs) were identical to that of the relevant probe. The non-specific competitors (NSCs) had one or more base pair changes, which altered highly conserved bases in the NF-κB binding motif to decrease the matrix score, or remove the binding site altogether (see Supplemental Table 1 for all probe and competitor sequences). All probes and competitors were synthesized and labeled at Integrated DNA Technologies (IDT; Coralville, IA).
**Nuclear extractions for NF-κB protein**

Nuclear proteins were extracted from RAW 264.7 cells treated with STE, according to the method of Ajuwon and colleagues (Ajuwon et al., 2004) with slight modification. The RAW 264.7 cells were thawed and then centrifuged at 3,300 x g for 15 minutes at 4°C. The supernatant was decanted and the pellets were re-suspended in 400 µl of ice-cold lysis buffer (10 mM HEPES, pH 7.9, 10 mM KCl, 0.1 mM EDTA, 0.1 mM EGTA, 0.5 mM dithiothreitol (DTT), 0.5 mM phenylmethanesulfonyl fluoride (PMSF), 2.0 µg/mL leupeptin and aprotinin, 2.5 µg/mL pepstatin; PMSF was added just prior to use) and placed on ice for 15 minutes with intermittent agitation. Samples were then centrifuged at 3,300 x g for 15 minutes at 4°C. The supernatant was decanted and the pellets re-suspended in 50 µL of extraction buffer (20mM HEPES, pH 7.9, 0.4 M NaCl, 1mM EDTA, 1 mM EGTA, 0.5 mM DTT, 1mM PMSF, 2 µg/mL leupeptin and aprotinin, 2.5 µg/mL pepstatin; PMSF was added just prior to use) and placed on ice for 30 minutes with intermittent agitation. Samples were then centrifuged at 20,000 x g for 30 minutes at 4°C and the supernatant was stored at -80°C as the nuclear extract. The protein concentration of the nuclear extracts was determined using a Bradford Protein Assay (Bio-Rad, Carlsbad, CA).

**EMSA**

The protein-DNA complexes were resolved on 4% polyacrylamide gels made according to the Promega (Madison, WI) Technical Bulletin TB110 with a volume of 50 ml volume instead of 20 ml. Binding of nuclear proteins to the labeled probes was performed by incubating 12.5 µg of nuclear proteins with 1 µl of 0.5 pmol/µl of labeled probe for 30 minutes on ice in 2.5 µl 5X binding buffer (Ajuwon et al., 2004) with 0.5 µl of 1 µg/µl poly
(dI-dC), and adding water to create a final volume of 12.5 μL. For competition assays, 5 μl of 10 pmol/μl of the competitor was added to the labeled probe before extract addition to create a 100:1 ratio of competitor:probe. For the supershift assays, 1 μl of 0.2 μg/μl of anti-p65 SC-372, Santa Cruz Biotechnology®, Inc., Santa Cruz, CA) polyclonal antibody were incubated with 12.5 μg of nuclear proteins for 15 minutes on ice, after which binding buffer and poly (dI-dC) were added to the reaction as above and incubated on ice for 15 additional minutes. All images were taken with a Typhoon 8600 scanner (Amersham Pharmacia Biotech, GE Healthcare, Piscataway, NJ).

A method similar to Day et al. (1992) was used to test for binding competition in the EMSAs, using the generic sequences as the probes and the gene competitors to test for binding. To test for competition, first the generic strong probe was used with the gene-specific competitors, and if no competition was observed, then the generic strong competitor was tested using the same gene-specific competitors. This was done to eliminate use of radioactivity as well as save on reagents. In addition, for genes for which a specific probe was designed (AIF1, TAP1, and UBD), the corresponding competitors were used to test for binding.

**Identification of common motifs among EMSA probes**

After the EMSAs were run, MEME 4.5.0 (Bailey and Elkan, 1994) was used to create *de novo*, common motifs among different groups of the designed probes and competitors. Three groups of probes and competitors were created and tested: all (29 sequences), those which showed competition (22 sequences), and those that did not (7 sequences). All groups were run using the same parameters in MEME: allowing the reverse complement to be tested
independently and allowing MEME to create up to 10 motifs per group. Once MEME identified the common motifs, TOMTOM 4.5.0 (Gupta et al., 2007), a subset of the MEME software suite, was used to identify if known transcription factor binding sites were within them using Pearson’s correlation to both matrices found in TRANSFAC and JASPAR.

**RNA extractions, quality control, and reverse transcriptase reactions**

RNA was extracted using the RNeasy Mini Kit (QIAGEN, Valencia, CA) following the large buffer volume protocol and the optional 2-mercaptoethanol as written in the manual. After extraction, the samples were checked for DNA contamination by regular polymerase chain reaction (PCR) using GoTaq Flexi (Promega, Madison, WI; 2.5 µl 5X Green Buffer, 2.5 µl 25 mM MgCl₂, 0.75 µl 10 pmol/ul (600 nM final concentration) of each primer, 1 µl extract, 0.25 µl 25 mM (0.5 mM final concentration) dNTP, 0.06 µl GoTaq enzyme, and 4.69 µl water per reaction for a 12.5 µl total volume, see below for the corresponding programs). If DNA was present, the samples were treated with either the DNase I kit (QIAGEN) followed by cleaning up with the MinElute Kit (QIAGEN) or the DNA-free kit (Agilent Technologies, Santa Clara, CA) using the rigorous clean up protocol. The PCR test and administration of DNase I was repeated until there was no amplification of DNA. After extraction and DNA clean up, RNA quality was determine with a BioAnalyzer 2100 using the RNA Nano 6000 chip (Agilent Technologies), and quantified using a Nanodrop (Thermo Scientific, Wilmington, DE).

For the RAW264.7 cells and the 3D4/31 cells, a total of 1600 ng of RNA was used per sample, while for the porcine PBMCs, a total of 750 ng of RNA was used, in the reverse
transcriptase reactions using the SuperScript II (Invitrogen, Carlsbad, CA) enzyme and following the manufacturer’s protocol.

**Quantitative real-time PCR**

Primers were designed by using primer3, available at [http://frodo.wi.mit.edu/primer3](http://frodo.wi.mit.edu/primer3). All PCR amplicons were designed to span an intron, and to be between 100-150 bp in length (see Supplemental Table 2 for all primer sequences). The salt correction formula used was the SantaLucia 1998 method (SantaLucia, 1998), as per the recommendation of the provider, and the monovalent cation concentration was set to 5 mM as per the final concentration of MgCl₂ in the PCR reactions. All primers were synthesized by IDT (Coralville, Iowa).

Primers were designed for mouse Aif1, C2, Cfb, H2-Eb1, H2-K1, Hspa1a/b, Ier3, Tap1, Tnf, Trim26, and Ubd. Separate mouse Hspa1a and Hspa1b primers could not be optimized due to the similarity of the gene sequences, so a primer pair was designed in a common region to quantify expression from both genes. In addition to the genes that were tested on the mouse, PSMB9, IL-8, CD14, and TLR4 were added for the pig gene set and separate primers were optimized for both HSPA1A and HSPA1B based on sequences downloaded from the Vega site at Sanger, as NCBI lacked separate RefSeq sequences for HSPA1A and HSPA1B (both listed under the HSP70.2 sequence at the time of oligonucleotide design). The corresponding homologues between the two species, when the gene symbol is different, H2-Eb1 and SLA-DRB, and H2-K1 and SLA-1 and all gene symbols are current with NCBI nomenclature (see Supplementary Table 2 for primer sequences).

All samples were placed in opaque, 96 well ABgene® PCR plates (Thermo Scientific) and fluorescence levels were read by an iCycler (Bio-Rad, Hercules, CA). For both the murine RAW 264.7 cells and the porcine 3D4/31 cells, a cDNA amount equivalent
of 50 ng RNA was used per PCR reaction, with each reaction carried out in duplicate.

SYBR® Green PCR Master Mix (Applied Biosystems, Carlsbad, CA) was used in 25 µl total volumes (12.5 µl 2X SYBR Green master mix, 1.5 µl 10 pmol of each primer (final concentration of 600 nM), 50 ng RNA, and water to 25 µl). The two cells lines used the same program: 95°C for 10 minutes, followed by 35 cycles of 95°C for 15 seconds, 55°C for 30 seconds, and 72°C for 45 seconds, with a final extension step of 72°C for 5 minutes. A melt curve was run on every plate at the end of the extension step to verify specificity of amplification. For the RAW 264.7 cells, Aif1, C2, Cfb, H2-Eb1, H2-K1, Hspa1a/b, Ier3, Tap1, Tnf, Trim26 and Ubd were tested. For the 3D4/31 cells, only TNF, IL-8, TLR4, and CD14 were tested.

For the porcine PBMCs, a cDNA amount equivalent to 25 ng RNA was used per reaction, with each reaction being done in duplicate, using 2X SYBR Green master mix (QIAGEN), in the same total volume and concentration of primers as above. The program for these samples was: 95°C for 10 minutes, followed by 35 cycles of 95°C for 30 seconds, 55°C for 30 seconds, and 72°C for 30 seconds, with a final extension step of 72°C for 5 minutes. An analysis of the melt curve was done at the end for every reaction. The primer sets included AIF1, CFB, C2, HSPA1A, HSPA1B, IER3, PSMB9, SLA-1, SLA-DRB, TAP1, TNF, TRIM26, and UBD. All expression values were normalized to RPL32 (porcine, Rpl32 mouse), which was run for each treatment on the same plates.

Statistics

All statistics were carried out in R 2.10 using the ΔΔCt method (Pfaffl, 2001), with RPL32 (or Rpl32) used to correct the Ct. The lm() function, using the formula y ~ replication
+ endotoxin + solution + endotoxin*solution + error, was used to fit a linear model for each combination of time point and gene. F-tests were conducted as part of each linear model analysis to test for differences between treatments. Least square means were calculated from triplicate RNA samples for each treatment at each time point.

Heatmaps were generated by first normalizing the least square mean of the control corrected ΔCt within the gene, then running hierarchical clustering, with centroid linkage on the Euclidean distance, in Cluster3 (de Hoon et al., 2003). The result was then visualized in Treeview (Eisen et al., 1998) to create the heatmap.

RESULTS

Development of a method to test function of bioinformatically predicted NF-κB motifs based on competition against a common known motif

Predicted NF-κB motifs at the proximal promoter for three porcine MHC genes were tested using EMSA and nuclear extracts from RAW267.4 cells stimulated with endotoxin. The motifs at AIF1, UBD, and TA_P1 all were bound by nuclear proteins, and unlabeled competitors with the same sequence as the probe, but not mutated competitors, diminished such complex formation (Figure 1a, lanes 2-16). We then determined that NF-κB p65 is present in the observed sequence-specific complexes through using a p65 antibody to bind to and cause a supershift of the NF-κB-DNA complex (Figure 1b, indicated by the arrow). A supershifted band is seen for every probe except for the AIF1 specific probe. In some cases, we also observe decreased binding of the probe in reactions with antibody, likely due to the interference by the antibody in complex formation. Supershifts were not observed using a non-NF-κB antibody (data not shown).
Similar results were also observed for the known NF-κB-binding motif (Ajuwon, et al., 2004), which we call “generic strong” as well as a second motif called “generic weak” which is based on the generic strong motif but whose sequence has been altered to decrease the theoretical binding affinity of NF-κB. These two confirmed generic strong and weak probes were then used for analysis of the rest of the predicted target gene motifs, so that a uniform assay could be used and to avoid the high experimental cost of individual fluorescent probes for each motif to be tested.

When testing competition of each predicted motif against the generic probes, most of the murine motifs caused a decrease in binding (Figure 2). Specifically, nine of the 12 murine motifs, Aif1, C2, Cfb, Hspa1a, Ier3, H2-K1, Tap1, Trim26, and Ubd, all competed against either the generic strong (GS) or the generic weak (GW) probe, but Hspa1b, H2-Eb1, and Tnf failed to show any competition against either of the generic probes. Similarly, eight of the 12 porcine motifs decreased binding to the labeled generic probes. The porcine motifs for CFB, C2, HSPA1B, IER3, SLA-DRB1, TAP1, TNF, and UBD showed competition (Figure 3), while AIF1, HSPA1A, SLA-1, and TRIM26 did not (see Table 3 for summary).

**Predicted NF-κB Motifs grouped by Binding Activity have common sequence elements**

Based on the results of the EMSAs, three groups of motifs were identified: all motifs, motifs that showed evidence of sequence-specific recognition by NF-κB, and finally those that did not show any NF-κB binding activity. The MEME algorithm was used on each of these groups to identify different common motifs within the three groups. For the “all motifs” group, MEME created a single, 21 base pair long motif. When this motif was analyzed by the TOMTOM software, to identify similarity to previously known matrices,
NF-κB was found to have the highest correlation out of all the different matrices found in TRANSFAC or JASPAR.

When only the group of motifs that showed competition in the EMSA assays were analyzed for common sequence elements, MEME identified three. The longest of these is a 15 base long subset of the 21 bp motif above. Similarly to the previous MEME analysis, the most highly correlated matrices contained within the 15 bp motif were NF-κB matrices. The other two motifs were both shorter (eight bases long), and contained only two contributing sequences, and neither correlated to NF-κB matrices.

When the final group of oligonucleotides, those that did not show any competition in the EMSA, were analyzed by MEME, two common motifs were identified. The longer motif was 15 bases long and is similar to the one identified in the first group and to the longer motif identified in the second group and was created from all the contributing sequences. However, unlike the longer motif found in the previous two groups, this motif was more similar to a STAT3:STAT3 dimer binding site than to NF-κB, although NF-κB matrices did correlate with the motif. The second motif found is eight base pairs in length, contributed by only two sequences, and does not contain any NF-κB regions.

Effect of *Salmonella enterica* serovar Typhimurium on 3D4/31 porcine macrophage-like cells

The results thus far indicated that many of the predicted NF-κB motifs in *Salmonella*-induced genes could be bound by this transcription factor. To develop an in vitro system to test the regulatory function of these promoter motifs, we tested the ability of the 3D4/31 porcine macrophage-like cell line (Weingartl et al., 2002) to induce gene expression upon ST
endotoxin (STE) stimulation. Surprisingly, treating the 3D4/31 cells with 1 or 10 µg/ml STE for 1 or 4 hours did not elicit a significant change in the level of RNA for either TNF or IL8. Both of these genes have been shown to respond to LPS in primary porcine macrophages (Chen et al., 2003). We also tested for and showed that the RNA for both CD14 and TLR4 was present in these cells (see Supplemental Table 3 for full results). One explanation for this lack of response is that the STE used is inactive and could not bind to TLR4, thus not activating the downstream intracellular pathways. However, as the same STE caused a large increase in expression for TNF in both the RAW 264.7 and the porcine PBMC MΦ (see below), this lack of a response is more likely due to the inability of 3D4/31 cells to recognize STE. Based on this finding, we discontinued using this porcine cell lines as a model of host response to Salmonella.

Minimal effect of DMSO alone or SC-514 alone on RNA expression in RAW 264.7 cells for selected genes

In these experiments using the mouse macrophage cell line RAW 264.7, we performed the following treatments (categories): PN (PBS and No STE); PY (PBS and STE); DN (DMSO and No STE); DY (DMSO and STE); SN (SC-514 and No STE) and SY (SC-514 and STE). The solvent for SC-514 is DMSO, and since DMSO is a free radical scavenger and has been shown to reduce inflammation on its own (Colucci et al., 2008), it is important to test whether or not it has any effects on the gene expression. When comparing RAW 264.7 cells that were untreated by either DMSO or SC-514 (groups PN and PY, respectively) to cells that were treated with just DMSO (DN), only Ubd showed a significant difference in expression at 1 hpi (p = 0.015, fold change (fc) = 5.06), meaning that Ubd was
down regulated due to DMSO. At every other time point for every other gene, no significant effect was seen.

Similarly, to look at the potential independent effect of SC-514 on basal gene expression by the inhibition of NF-κB, DN results were compared to the samples with both DMSO and SC-514 added (group SN). Of all the genes tested, only Ier3 at 2 hpi (p = 0.015, fc = 0.67) showed a significant effect. It should be noted that at 1 hpi for Ubd, when comparing DN to SN, there was not a significant difference (p = 0.081, fc = 0.34), meaning that the SC-514 pretreatment has a higher expression than with only DMSO, returning Ubd to a more basal level of expression. Thus as DMSO and SC-514 (in DMSO) causes minimal changes on the expression level of target genes, we focused on comparing DN, DY and SY treatments in all future analyses, and used DMSO in all porcine cell work (see below).

ST Endotoxin (STE) treatment of RAW 246.7 murine cells causes changes in RNA expression at most of the predicted and known NF-κB target genes

When comparing the effects of DN and DY treatments in RAW 264.7 cells, seven of the 12 genes tested had a significant change in the expression of their RNA at one or more time points. At 1 hpi, Ier3 (p = 1.22x10^{-6}, fc = 16.53, fc is DY/DN) and Tnf (p = 1.61x10^{-6}, fc = 15.46) had increases in their expression, while Ubd (p = 0.011, fc = 0.21) had decreased expression. At 2 hpi, in addition to Ier3 (p = 1.73x10^{-7}, fc = 8.06) and Tnf (p = 1.28x10^{-10}, fc = 25.19) still being up regulated, Hspa1a/b (p = 5.93x10^{-4}, fc = 1.61) also increased in expression. The decrease in expression for Ubd was no longer significant. Hspa1a/b (p = 0.006, fc = 1.64), Ier3 (2.19x10^{-5}, fc = 15.67) and Tnf (p = 2.10x10^{-10}, fc = 37.27) were still up at 4 hpi, and Cfb (p = 6.60x10^{-8}, fc = 10.88), H2-Eb1 (p = 0.016, fc = 1.97), Tap1 (p =
2.33x10^{-5}, fc = 2.93), and Trim26 (p = 6.35x10^{-4}, fc = 3.06) also had a significant increase in RNA expression (see Table 1 and Figure 4).

**Effect of SC-514 pretreatment on the effect of STE on predicted target genes in mouse RAW 264.7 macrophages**

To determine the dependency on NF-κB for the changes in gene expression caused by STE seen in the RAW 264.7 cells, the cells were pretreated with SC-514 to inhibit IKKβ function (group SY). At one or more time points, this pretreatment affected six of the seven genes that showed changes in their gene expression due to STE (see Table 1). At 1 hpi, Ier3 had a significantly reduced increase in expression upon pretreatment of SC-514 when compared to STE alone (p = 0.002, fc = 0.49, fc is SY/DY), and Tnf tended (0.1 ≤ p < 0.05) toward a significant decrease in expression (p = 0.085, fc = 0.60). At 2 hpi, the SY data for Cfb, which showed a nearly significant increase in expression due to STE (p = 0.059, fc = 1.45), also showed a significant decrease in STE induction of expression due to SC-514 pretreatment (p = 0.033, fc = 0.65). Neither Hspa1a/b (p = 0.384, fc = 0.87), Ier3 (p = 0.615, fc = 1.09), nor Tnf (p = 0.955, fc = 0.99) at 2 hpi had a significant difference between the two STE treatments. At 4 hpi Cfb (p = 1.49x10^{-5}, fc = 0.27), Tap1 (p = 0.002, fc = 0.55), and Tnf (p = 0.021, fc = 0.68) had a significant decrease in expression due to SC-514 (see Table 1, 3 and Figure 4).

In addition to several genes exhibiting statistically significant differences due to SC-514 treatment, several genes with non-significant SC-514 effects nevertheless showed decreases in their response level (fold change) with SC-514 pre-treatment, which provides evidence of their dependency on NF-κB for response to STE. At 1 hpi, the fold change for Tnf drops from 15.45 (DY/DN) to 9.16 (SY/DN); both of which are significantly different
from the DN level. At 4 hpi, there was no significant effect of SC-514 treatment for Ier3, Trim26 and H2-Eb1, but the fold change for Ier3 dropped from 15.67 to 12.18, Trim26 dropped from 3.06 to 2.25, and H2-Eb1 dropped from 1.97 to 1.28 (see Tables 1, 3 and Figure 4).

The effect of STE on RNA expression of predicted NF-κB target genes in pig PBMC derived Macrophages

To compare the changes in RNA expression observed in porcine peripheral blood mononuclear cell (PBMC) derived macrophages (MΦ) of the same genes in response to STE to that seen in the mouse, a similar but expanded experimental design was used. In addition to the 12 genes assayed in the RAW 264.7 cells, the RNA expression level for PSMB9 was evaluated to determine if porcine cells responded similarly to that reported for mouse cells at this gene (Marqués et al., 2004). No significant difference existed between genes in cells treated with DMSO alone (DN) to those in cells treated with both DMSO and SC-514 (SN, data not shown), thus we focused on response to STE treatment and the effect of pre-treatment with SC-514. Of the 13 porcine genes tested for STE response, ten showed a significant change in RNA expression (see Table 2). At 1 hpi for the porcine MΦ, IER3 (p = 0.002, fc = 46.85, fc is DY/DN) and TNF (p = 9.22x10^{-4}, fc = 73.26) mRNA expression level increased, while HSPA1A (p = 0.044, fc = 0.47), SLA-DRB1 (p = 0.049, fc = 0.67) and TAP1 (p = 0.040, fc = 0.48) mRNA expression level decreased. At 2 hpi TAP1 and SLA-DRB1 mRNA level returned to baseline, however HSPA1A (p = 0.044, fc = 0.39) was still significantly down regulated, and IER3 (p = 0.002, fc = 86.22) and TNF (p = 5.47x10^{-4}, fc = 35.84) were still up regulated. CFB (p = 0.031, fc = 2.43), IER3 (p = 0.002, fc = 126.68), PSMB9 (p = 0.023, fc = 2.88), TNF (p = 0.001, fc = 54.70), and UBD (p = 0.009, fc = 9.59)
were all up-regulated, and HSPA1A was no longer significantly down regulated at 4 hpi. At the final time point of 8 hpi, CFB (p = 0.039, fc = 2.96), C2 (p = 0.035, fc = 2.02), IER3 (p = 5.21x10^{-4}, fc = 40.04), PSMB9 (p = 0.005, fc = 3.33), TAP1 (p = 0.049, fc = 3.14), TNF (p = 2.93x10^{-4}, fc = 20.23), TRIM26 (p = 0.037, fc = 1.64), and UBD (p = 0.001, fc = 35.63) were all up regulated (see Tables 2, 3 and Figure 5).

The inhibitory effect of SC-514 on STE induced RNA expression in primary porcine macrophages

As performed for the mouse cells, the role of NF-κB was evaluated on the porcine PBMC derived MΦs by SC-514 pretreatment prior to STE administration. SC-514 had an inhibitory effect at almost every significant response due to STE treatment, although most of these effects were not statistically significant. At 1 hpi, IER3 (p = 0.403, fc = 0.52) and TNF (p = 0.070, fc = 0.21) were not significantly different between SY/DY. However, when comparing the fold changes of DY/DN and SY/DN, the fold change for IER3 dropped from 46.85 to 24.31 and TNF dropped from 73.26 to 15.37, providing evidence that NF-κB is required for the highest response. The fold changes in response to STE for the other genes were not altered by SC-514 at 1 hpi. At 2 hpi, only TNF had a significant decrease in its expression (p = 0.018, fc = 0.18). At 4 hpi, two genes CFB (p=0.051) and TNF (p=0.071) were close to significance for a SC-514 effect, and CFB had a fold change drop from 2.43 to 1.12 and the fold change due to STE treatment for TNF dropped from 54.69 to 11.26. Two other genes also had large numerical decreased in the response to STE: IER3 dropped from 126.68 to 62.54, and UBD dropped from 9.59 to 5.57. Finally at 8 hpi, CFB (p = 0.047, fc = 0.36) and TNF (p = 0.021, fc = 0.29) had a significant difference in RNA expression when using the SC-514 pretreatment. As well, several additional genes had a decrease in STE.
response with SC-514 pre-treatment, including TAP1 (p = 0.211, fc = 0.52), whose response dropped from 3.14 to 1.64, and UBD (p = 0.106, fc = 0.293) which dropped from 35.63 to 10.43 (see Tables 2, 3 and Figure 5).

**DISCUSSION**

*Effectiveness of using a generic motif sequence and competition to assay for NF-κB binding for a series of putative regulatory sequences*

Nine of the 12 murine motifs and eight of the 12 porcine motifs, which were bioinformatically predicted based on porcine expression clustering data, competed against either the labeled generic strong (GS) or the generic weak (GW) probes. It is not surprising that some of the predicted motifs did not compete with the GS probe for binding by NF-κB protein, due to the GS having a very high similarity to the conserved binding sequence of NF-κB, having a MATCH score of 1.000 out of 1.000. It is for this reason the GW probe was designed to include base changes from the GS sequence that resulted in a lower MATCH, 0.886, indicating it has a less well conserved binding sequence and a lower affinity for NF-κB (Zabel et al., 1991). For the motifs that were not able to compete for binding by NF-κB, the highest matrix score was 0.901 (for porcine AIF1), more conserved than the GW probes, and the lowest matrix score for the competitors that did compete was 0.842 (mouse Aif1). This overlap in scores might indicate that MATCH scores alone are not a sufficient discriminator to predict which potential NF-κB binding sites would compete against the generic probes. One of the murine motifs that failed to show competition, Tnf, is designed from the Tnf gene that is a known target of NF-κB in humans (Goldfeld et al., 1990). The binding site used for Tnf, about –70 bp relative to the TSS in C7/BL6 mice is near the
equivalent location in the human TNF gene for the binding site that was required for viral and LPS induction of TNF, which is approximately –90 bp, and has a Match core score of 1.000 with a matrix score of 0.808. The lack of competition by the Tnf motif may be due to its higher similarity to a STAT3:STAT3 binding site, as shown by MEME and TOMTOM. This could mean that in the presence of STAT3, whose physical association with NF-κB is not fully understood (Yoshida et al., 2004 and Han et al., 2010), the generic probes would be bound by NF-κB, while this motif could be bound by STAT3. Hence, this motif may not be able to compete for NF-κB effectively in vitro, though it may still be bound by NF-κB in vivo.

3D4/31 cells unresponsive to STE

After treating 3D4/31 cells with STE, we did not observe an increase in RNA abundance for IL8 or TNF (which showed a fold change range of 20-73 in the porcine PBMCs in response to the same preparation of STE). While both IL8 and TNF are known to increase in response to LPS (Chen et al., 2003), this lack of response in 3D4/31 cells could be consistent with the results in Volf et al. (2007), where they showed little to no response to a noninvasive mutant of ST, but increased RNA expression in response to an invasive ST strain. This could be interpreted as follows: while external signals of ST do not elicit a response by 3D4/31 cells, these cells can respond once they receive signals of an internal pathogen. In contrast, 3D4/31 cells have been shown to increase RNA expression of IL8 and TNF in response to external bacterial signals from various strains of E. coli (Pavlova et al., 2008), and 3D4/31 cells responded to LPS inoculation by translocation of NF-κB to the nucleus (de Greeff et al., 2010). Hence, the lack of a response may only be limited to an
external signal consisting of only ST endotoxin. In any case, it appears that 3D4/31 cells
would not make a good cell line for modeling of ST infection using ST endotoxin
stimulation.

**Inhibition of NF-κB signaling through SC-514 pretreatment shows variable repression
of STE induced genes**

Based on the results of the effect of SC-514 on gene expression, three different
groups distinguished by their degree of response to STE can be constructed. The first group
(G1) is defined for genes that show a significant effect of STE (DY vs. DN) as well as a
significant effect of SC-514 on the STE effect (SY vs. DY). The second group (G2) is when
there is a significant effect of STE treatment, and SC-514 creates an intermediate response
where the SY treated cells are not significantly different from DY or from DN. This
category was created because this type of result shows SC-514 has an effect of preventing a
significant induction of expression by STE. Finally, we created a third group (G3), which is
defined for genes that show a significant effect of STE treatment and show a large (but not
statistically significant) difference of expression between DY and SY; our threshold for this
difference was set as a fold change difference for SY/DY of \( \leq 0.75 \). This latter group has the
weakest evidence for NF-κB regulatory dependency for a gene of interest.

Several genes for both mouse and pig fell into G1. CFB and TNF were both
differentially expressed due to both treatments in both species but at different time points:
CFB at 8 hpi for pig; 2 and 4 hpi for mouse, while TNF showed such effects at 2 and 8 hpi
for pig; 4 hpi for mouse. In addition to these genes, the mouse genes Ier3 (at 1 hpi) and Tap1
(at 4 hpi) also fall into this group. CFB at 8 hpi in pig and 2 hpi in mouse seems to be almost
entirely dependent on NF-κB signaling, as the fold change between the SC-514 plus STE
treated cells and the STE untreated cells was close to 1 for both of these time points. Other
genes tested showed only a partial decrease in fold change due to the loss of NF-κB signaling
prior to STE administration.

Most of the genes that show a reduction in expression due to SC-514 inhibition of
NF-κB belong to G2, including CFB for pig at 4 hpi, the same time point that the mouse Cfb
showed a significant effect for both STE and SC-514 treatment. In addition to CFB, C2 (8
hpi), PSMB9 (4 hpi), TAP1 (8 hpi), and TRIM26 (8 hpi) also fall into this group for the pig.
Only two of the mouse genes, H2-Eb1 (SLA-DRB1 homologue) at 4 hpi and the Hspa1a/b at
2 hpi belong to G2.

Included in G3 are several known targets of NF-κB. Included are IER3 (at 1 and 4
hpi), PSMB9 (at 8 hpi), TNF (at 1 and 4 hpi) for pig and Tnf (at 1 hpi) for mouse. In
addition, UBD for both pig and mouse (at 4 and 8 hpi for pig and 1 hpi for mouse) and
Trim26 for only mouse (at 4 hpi) fall into this group.

**Integrating the results of EMSA and q-RT-PCR analyses provide evidence of direct
regulation of known and predicted NF-κB target genes**

The results from the SC-514 inhibition of STE response experiments show whether
NF-κB activity is required to control the normal response of genes to STE, while the EMSAs
show if it is possible for NF-κB to bind to their promoters at predicted NF-κB motifs. If SC-
514 inhibits the effect of STE on the expression of a gene, and the EMSA shows it is possible
for NF-κB to bind to their promoter, then NF-κB is inferred to be a direct regulator of that
gene. For all porcine and murine motifs tested, except for the murine motifs for Tnf and H2-
Eb1, whenever there was a SC-514 effect on a gene’s response to STE treatment, the EMSA
results showed a competition for one of the two generic probes. In the following sections, we describe specific examples of these results.

**HSPA1A/B affects the NF-κB pathway**

Reductions in the activity of NF-κB have been associated with increases in expression of Hspa1a/b (Guzhova et al., 1997, Sun et al., 2005, and Dokladny et al., 2010), though it is usually Hspa1a/b regulating NF-κB. Guzhova et al. (1997) used a mild heat shock to activate Hspa1a/b, Sun et al. (2005) used sodium arsenite in Kupffer cells, and Dokladny et al. (2010) used either an adenovirus followed 5 days later by LPS in rats. All these studies showed increased expression of Hspa1a/b affecting NF-κB by either preventing the degradation of IκBα or by binding the p65 subunit directly and preventing its translocation to the nucleus, thereby decreasing NF-κB signaling. We showed that the expression of Hspa1a/b increased due to STE at 2 and 4 hpi in mouse, with SC-514 at 2 hpi causing a non-significant decrease in the fold change compared to just STE. If NF-κB is regulating the Hspa1a/b family of genes in mouse, then it could be another feedback loop similar to that of IκBα, in which NF-κB up regulates its own inhibitor (Sun et al., 1993 and Brasier, 2006). Hagawara et al. (2007) showed in RAW264.7 cells that, at normal temperatures, Hspa1a/b is not up regulated by LPS after 6 hours, results which we also observed, so this could be an early, transient effect of Hspa1a/b.

Alternatively the amount of STE used may have been sufficient to initiate apoptotic pathways in some of the RAW264.7 cells, in which the Hspa1a/b genes have a SC-514 sensitive increase at 2 hpi. In the pig, we did not see an increase in expression for either HSPA1A or HSPA1B; rather we saw a significant decrease in HSPA1A expression at 1 and
2 hpi. This could be a real difference between the two species, because ST causes a more severe disease in mouse (Mittrücker and Kaufmann, 2000) than in pig (Bearson and Bearson, 2011). For the porcine MΦ, the toxicity of the STE may not be enough to signal the onset of the apoptotic pathways from which HSPA1A/B protects cells. A similar response to that of mouse cells was seen in mesenteric lymph node tissues from pigs infected with SC (Wang et al. 2008), and the response we saw in the porcine MΦ for HSPA1A/B genes is different from that observed by Wang et al. (2008). This difference is most likely due to the different serovar used and the different times post inoculation/treatment studied, as well as differences in the in vivo versus in vitro design, which introduces many differences such as the removal of other cell signals (like IFNγ), and the loss of a fever, which can cause the activation of HSPA1A/B through the activation of their main transcription factor during inflammation: HSF1 (Sistonen et al., 1994).

**NF-κB Signaling and STE effects on antigen processing and presentation**

Several genes involved in antigen processing and presentation were tested, including SLA-1/H2-K1 (MHC Class I peptide binding), SLA-DRB1/H2-Eb1 (MHC Class II peptide binding), as well as TAP1 and PSMB9 (both MHC Class I antigen processing). Neither SLA-1 nor H2-K1 genes were affected in either cell source due to STE, but a potential NF-κB binding site was identified in their promoters. The lack of response, which is different from the results of Wang et al. (2008) for SC infection, is not surprising as STE lacks antigenic proteins. Without antigenic protein, the cells may not activate the genes for antigen presentation. This lack of response is similar to Gao et al. (2010), in which SLA-1 was not affected due to LPS treatment in PBMCs, and to Wang et al. (2007) in which SLA-1 did not
show any up regulation. While Schmid et al. (2006) showed that in renal biopsies with diabetic nephropathy HLA-A, the human homolog of SLA-1, did increase in expression with an increase in activity of NF-κB. They described HLA-A as containing a predicted NF-κB binding motif, but they did not show that NF-κB activity nor the binding site was directly responsible for the expression of HLA-A.

The MHC Class II genes, H2-Eb1 and SLA-DRB1, showed differences between the two species in their response to endotoxin. In the mouse, H2-Eb1 was not only up regulated by STE, but it was also affected by pretreatment of SC-514, though the motif tested did not compete in the EMSA. SLA-DRB1 showed down regulation due to STE in porcine macrophages at 1 hpi, but SC-514 had no effect on its expression level, though the motif showed competition for NF-κB binding in the EMSA. This could indicate, as for Tnf, that the lack of competition could be a due to selection of the wrong motif for H2-Eb1, or that the regulation of H2-Eb1 acts through a different, NF-κB dependent, transcription factor in the mouse. The difference between the species could indicate a difference in the recognition and presenting the antigens for STE, due to the different effects S. Typhimurium has on the two organisms. In the mesenteric lymph node of pigs infected with SC for 24 hours, there was an increase in expression for SLA-DRB1. Because the entire pathogen was present for a more extended period of time, it is not surprising that MHC Class II antigen presentation was activated in vivo.

On the other hand, the two antigen processing genes, PSMB9 and TAP1, are previously known NF-κB targets (Wright et al., 1995) that share a bi-directional promoter and regulatory region. Both genes in the pig and Tap1 in mouse (Psmb9 was not tested in the mouse) showed increases due to STE and inhibition of that increase due to SC-514
pretreatment, which agrees with Wright and colleagues in that they are both regulated, at least in part, by NF-κB. Taken together with the antigen presenting results, it could be that antigen processing genes are unaffected by the lack of a whole organism and will be activated with an initial TLR4 signal, but the actual antigen presenting genes require additional signals to be activated.

**Complement gene C2 regulatory differences between mouse and pig**

Two complement genes, C2 and CFB, were also among the list of NF-κB targets, with mouse CFB as a known target of NF-κB (Nonaka and Hwang, 1990 and Rothgiesser et al., 2010). While CFB showed similar results in both mouse and pig macrophages, C2 differed between the two species. C2 in the porcine MΦ is up regulated at 8 hpi, and this effect was inhibited by SC-514 pretreatment, making it a confirmed novel NF-κB target in the pig. In mouse, STE failed to cause an increase in the expression of C2 RNA in the time points that we measured (1, 2 and 4 hpi). These results are similar to Bolger et al. (2006), which showed that in bronchoalveolar lavage the level of C2 protein increased due to inhaled LPS in humans, though in mice and rats it did not increase. The results for CFB were different from those of C2 in that an increase in expression, followed by inhibition of SC-514, was seen in both cell lines, which was also reported by Bolger et al. (2006). This might indicate that the porcine complement system is more similar to human than rodents, at least for the complement components tested in this study.

**Other effects of NF-κB on immune related genes**

Several other genes not belonging in one of the above systems were also predicted to be targets of NF-κB: AIF1, IER3, TNF, TRIM26, and UBD. Of these five, IER3 and TNF
are known targets of NF-κB (see Wu et al., 1998 for IER3 and Goldfeld et al., 1990 for TNF). We observed dependence on NF-κB for these two genes, the extent of which depended upon the time point. Both were up-regulated by STE at every time point in both cell lines. Similarly, AIF1 had the same response in both species, but differed from IER3 and TNF in that STE failed to induce an increase in AIF1 expression. This is consistent with reports that show, while AIF1 expression does increase though induction by INFγ, (which was shown to be induced in a Salmonella infected animal [Uthe et al., 2009]), but AIF1 has been shown not to respond to treatment by IL-2, IL-4, IL-8, TNF, LPS, or E. coli infection of Kupffer cells in rats (Nagakawa et al., 2004).

In the mouse, a novel target gene of NF-κB was detected: Trim26. The molecular function of TRIM26 is currently not entirely known, but it does contain both metal ion and DNA binding regions (Chu et al., 1995) and appears, in humans (but not mouse), to interfere with the entry and release of HIV and MLV from HEK293 cells in culture (Uchil et al., 2008). Currently there is no published evidence in for any member of the TRIM family to be regulated by NF-κB. There are four other TRIM genes directly upstream of TRIM26 in the MHC (TRIM31, TRIM40, TRIM10, TRIM15). Little functional data is available on any of these four genes, except that TRIM10 seems to be required for the finial differentiation of erythrocytes (Harada et al., 1998), and Uchil et al. (2008) showed that TRIM15 and TRIM31 have strong anti-viral properties similar to TRIM26. Another member of the TRIM family, TRIM21, has been shown to aid in immunity by binding to intracellular virus particles coated with either IgM or IgG. This binding seems to mediate the ubiquitinization of the viral particles, so certain members of the TRIM family appear to have important intracellular immune related functions (Mallery et al., 2010).
The expression of UBD has not previously been shown to be affected by changes to NF-κB signaling, though Ubd in mouse has been shown to have a predicted NF-κB binding site within its promoter (Zhang et al., 2006). At 4 and 8 hpi in porcine PBMC MΦ cells, UBD was up regulated by STE, and this response was inhibited by SC-514. However, in the mouse, there was no induction of Ubd due to STE, so this may be another important difference between the pig and mouse. UBD has been shown to selectively bind IKKγ which then causes the IKK complex to become active, resulting in the release of NF-κB from IκB for translocation to the nucleus, creating a positive feedback loop to release NF-κB for translocation and preventing it from being bound in the cytoplasm (Rahighi et al., 2009, Lo et al., 2009, and Gong et al., 2010).

**Similarity of porcine results to previous studies on genome-wide RNA responses to Salmonella**

In addition to the Wang et al. (2008) paper from which the NF-κB targets were predicted, we compared our results to two microarray experiments which reported on porcine immune tissues or cells responding to bacteria, or bacterial components: Wang et al. (2007), where pigs were infected with ST and response measured in mesenteric lymph node using the Affymetrix GeneChip®, and Gao et al. (2010), where porcine PBMCs were treated for 24 hours with *E. coli* LPS and the response was measured using the SLA-RI/NASRP8-13K microarray. In general, our results are more similar to the SC response seen in the MLN than to the other two studies. Specifically, porcine C2, CFB, IER3, PSMB9, TAP1, TNF, TRIM26, and UBD all showed a statistically significant increase in expression due to the STE treatment (Table 2). Porcine AIF1, HSPA1A/B, SLA-1, and SLA-DRB1 failed to show the same increase seen as SC; however, all except HSPA1A did show competition in the
EMSAs. It is possible that other conditions, besides recognition and response to only the endotoxin component of bacteria, are required for induction of these genes. Of the genes that did not show an increase in expression due to ST endotoxin alone, only AIF1 and TRIM26 had an increase in mRNA expression due to live ST in the MLN, further documenting that the live pathogen compared to just the endotoxin component can elicit different responses. Of the 13 genes in this study, Gao et al. (2010) saw only a decrease in SLA-DRB1 after 24 hours of *E. coli* LPS stimulation, which, while similar to the result at 1 hpi, is a time point outside of what was measured here.

CONCLUSION

Selecting the proper cell line to model whole organism responses is a critical step in designing an experiment. The overall response seen in the porcine PBMCs, where eight out of the 13 tested genes displayed an increase in mRNA expression due to STE, is similar to that seen in SC infected MLN. When comparing these tissue culture results on STE stimulation with the two live bacterial infections in pigs reported in Wang et al. (2007) and Wang et al. (2008), typically for a gene in which there is an effect of both STE and SC-514 plus STE on gene expression *in vitro*, there is an effect of SC infection only on gene expression, such as that seen for porcine C2, CFB, IER3, PSMB9, TAP1, TNF, and UBD (Table 3). We interpret this as providing further evidence that ST can inhibit NF-κB signaling to increase its survival (Negrate et al., 2008). It also indicates that such *in vitro* studies can be a good method to identify new targets of transcription factors that have been identified in more costly animal studies. Our study did not match the Gao et al. (2010) study except for the decrease seen in SLA-DRB1 at 1 hpi. However, they looked at a much later time point that would miss the early events in STE recognition and processing within the MΦ
that we have identified through emphasizing such early responses. We also differentiated the PBMCs to MΦ, which might also explain some of the differences seen. While the STE is a more similar to the SC infection for these genes and STE with SC-514 is more similar to ST for this subset, a better model would be one that would induce the HSPs, as they are important aspect to the immune response. Since it does not appear that STE is sufficient to induce them through NF-κB, treating the cells at fever temperatures, similar to Hagiwara et al. (2007), or treating cells with IFNγ in addition or prior to STE, might be a better model for the MLN responses reported. However, using these two cell lines as models for infection, a total of four novel targets of NF-κB have been identified: H2-Eb1 and Trim26 in mouse, and C2 and UBD in pig.

AUTHORS’ CONTRIBUTIONS

This article was written primarily by OC under the supervision of CT. JU and NF contributed to the lab work, MW provided suggestions about the cell culture experiments, and EK and ART provided help with lab work, as well as suggestions on the analysis and manuscript preparation, and sections of text for the manuscript.

ACKNOWLEDGMENTS

This project was supported by National Research Initiative Competitive Grant no. 2009-35205-05192 from the USDA National Institute of Food and Agriculture. Support is also gratefully acknowledged from the USDA-CSREES Special Grant to the Food Safety Consortium and the ISU Center for Integrated Animal Genomics. A USDA MGET 2001-52100-11506 Fellowship to OC is gratefully acknowledged. CC gratefully acknowledges support from the USDA National Needs Graduate Fellowship Competitive Grant No. 2005-38420-15810 from the National Institute of Food and Agriculture. The NSF Research
Experience for Undergraduates for both EK and NF, and the ISU Undergraduate Research Assistant Program for EK are also acknowledged for their support in this project.
REFERENCES


TABLE 1

SC-514 inhibits induction by STE treatment for several murine genes.

Several murine genes activated by S.Typhimurium endotoxin show significant decreases in induced expression due to STE (DY/DN column) after pretreatment of SC-514 (SY/DY or SY/DN column). DN is STE and SC-514 untreated cells, DY is STE treated cells, SY is SC-514 pretreated STE treated cells. All samples received DMSO.

<table>
<thead>
<tr>
<th></th>
<th>1 hpi</th>
<th>2 hpi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DY/DN</td>
<td>SY/DY</td>
</tr>
<tr>
<td>Aif1</td>
<td>p-value</td>
<td>0.227</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>1.621</td>
</tr>
<tr>
<td>Cfb</td>
<td>p-value</td>
<td>0.135</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>2.099</td>
</tr>
<tr>
<td>C2</td>
<td>p-value</td>
<td>0.436</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.541</td>
</tr>
<tr>
<td>Hspa1a/b</td>
<td>p-value</td>
<td>0.892</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.964</td>
</tr>
<tr>
<td>Ier3</td>
<td>p-value</td>
<td>1.22E-06**</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>16.526</td>
</tr>
<tr>
<td>H2-K1</td>
<td>p-value</td>
<td>0.156</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.749</td>
</tr>
<tr>
<td>H2-Eb1</td>
<td>p-value</td>
<td>0.416</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.681</td>
</tr>
<tr>
<td>Tap1</td>
<td>p-value</td>
<td>0.942</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>1.045</td>
</tr>
<tr>
<td>Tnf</td>
<td>p-value</td>
<td>1.61E-06**</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>15.456</td>
</tr>
<tr>
<td>Trim26</td>
<td>p-value</td>
<td>0.106</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>1.488</td>
</tr>
<tr>
<td>Ubd</td>
<td>p-value</td>
<td>0.019**</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.211</td>
</tr>
</tbody>
</table>

* 0.05 < p ≤ 0.1, ** p ≤ 0.05, † fc ≤ 0.75 in the SY/DY comparison
### TABLE 1 (CONTINUED)

<table>
<thead>
<tr>
<th>Gene</th>
<th>p-value</th>
<th>fc</th>
<th>p-value</th>
<th>fc</th>
<th>p-value</th>
<th>fc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4 hpi DY/DN</td>
<td>SY/DY</td>
<td>4 hpi SY/DN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aif1</td>
<td>0.906</td>
<td>0.330</td>
<td>0.279</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.979</td>
<td>0.839</td>
<td>0.822</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cfb</td>
<td>6.60E-08**</td>
<td>1.49E-05**</td>
<td>9.49E-05**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10.878</td>
<td>0.266†</td>
<td>2.894</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>0.056*</td>
<td>0.223</td>
<td>0.409</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.329</td>
<td>0.601†</td>
<td>1.401</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hspa1a/b</td>
<td>0.006**</td>
<td>0.875</td>
<td>0.008**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.640</td>
<td>0.977</td>
<td>1.602</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ier3</td>
<td>2.19E-05**</td>
<td>0.511</td>
<td>4.94E-05**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15.671</td>
<td>0.777</td>
<td>12.182</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H2-K1</td>
<td>0.795</td>
<td>0.464</td>
<td>0.631</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.064</td>
<td>0.837</td>
<td>0.891</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H2-Eb1</td>
<td>0.016**</td>
<td>0.093*</td>
<td>0.320</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.972</td>
<td>0.648†</td>
<td>1.277</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tap1</td>
<td>2.33E-05**</td>
<td>0.002**</td>
<td>0.008**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.928</td>
<td>0.551†</td>
<td>1.613</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tnf</td>
<td>2.10E-10**</td>
<td>0.021**</td>
<td>6.46E-10**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>37.271</td>
<td>0.677†</td>
<td>25.223</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trim26</td>
<td>6.35E-4**</td>
<td>0.206</td>
<td>0.005**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.060</td>
<td>0.734†</td>
<td>2.244</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ubd</td>
<td>0.405</td>
<td>0.711</td>
<td>0.240</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.519</td>
<td>1.202</td>
<td>1.826</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TABLE 2

SC-514 inhibits induced by STE treatment for several porcine genes.

Nearly every up-regulated porcine gene requires NF-κB for full expression of their response to STE (DY/DN column) as shown by decreases in expression due to the pretreatment of SC-514 (SY/DY or SY/DN columns) on genes up regulated by treatment of S. Typhimurium endotoxin. DN is STE and SC-514 untreated cells, DY is STE treated cells, SY is SC-514 pretreated STE treated cells. All samples received DMSO.

<table>
<thead>
<tr>
<th>Gene</th>
<th>1 hpi</th>
<th>2 hpi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DY/DN</td>
<td>SY/DY</td>
</tr>
<tr>
<td>AIF1</td>
<td>p-value</td>
<td>0.901</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.972</td>
</tr>
<tr>
<td>CFB</td>
<td>p-value</td>
<td>0.485</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.705</td>
</tr>
<tr>
<td>C2</td>
<td>p-value</td>
<td>0.531</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.862</td>
</tr>
<tr>
<td>HSPA1A</td>
<td>p-value</td>
<td>0.058*</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.471</td>
</tr>
<tr>
<td>HSPA1B</td>
<td>p-value</td>
<td>0.109</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.522</td>
</tr>
<tr>
<td>IER3</td>
<td>p-value</td>
<td>0.002**</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>46.851</td>
</tr>
<tr>
<td>PSMB9</td>
<td>p-value</td>
<td>0.149</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.766</td>
</tr>
<tr>
<td>SLA-1</td>
<td>p-value</td>
<td>0.299</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.788</td>
</tr>
<tr>
<td>SLA-DRB1</td>
<td>p-value</td>
<td>0.049**</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.671</td>
</tr>
<tr>
<td>TAP1</td>
<td>p-value</td>
<td>0.040**</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.481</td>
</tr>
<tr>
<td>TNF</td>
<td>p-value</td>
<td>9.22E-04**</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>73.262</td>
</tr>
<tr>
<td>TRIM26</td>
<td>p-value</td>
<td>0.374</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>0.799</td>
</tr>
<tr>
<td>UBD</td>
<td>p-value</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>fc</td>
<td>1.653</td>
</tr>
</tbody>
</table>

* 0.05 < p ≤ 0.1, ** p ≤ 0.05, † fc ≤ 0.75 in SY/DY comparison
<table>
<thead>
<tr>
<th></th>
<th>4 hpi</th>
<th>8 hpi</th>
<th>8 hpi</th>
<th>8 hpi</th>
<th>8 hpi</th>
<th>8 hpi</th>
<th>8 hpi</th>
<th>8 hpi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DY/DN</td>
<td>SY/DY</td>
<td>DY/DN</td>
<td>SY/DN</td>
<td>SY/DN</td>
<td>SY/DN</td>
<td>SY/DN</td>
<td>SY/DN</td>
</tr>
<tr>
<td>AIF1</td>
<td>p-value 0.760</td>
<td>0.940</td>
<td>0.818</td>
<td>0.940</td>
<td>0.818</td>
<td>0.940</td>
<td>0.818</td>
<td>0.940</td>
</tr>
<tr>
<td></td>
<td>fc 1.083</td>
<td>1.062</td>
<td>1.062</td>
<td>1.062</td>
<td>1.062</td>
<td>1.062</td>
<td>1.062</td>
<td>1.062</td>
</tr>
<tr>
<td>CFB</td>
<td>p-value 0.031**</td>
<td>0.051*</td>
<td>0.039**</td>
<td>0.051*</td>
<td>0.039**</td>
<td>0.051*</td>
<td>0.039**</td>
<td>0.051*</td>
</tr>
<tr>
<td></td>
<td>fc 2.434</td>
<td>1.124</td>
<td>2.945</td>
<td>1.124</td>
<td>2.945</td>
<td>1.124</td>
<td>2.945</td>
<td>1.124</td>
</tr>
<tr>
<td>C2</td>
<td>p-value 0.351</td>
<td>0.726</td>
<td>0.202</td>
<td>0.351</td>
<td>0.726</td>
<td>0.202</td>
<td>0.351</td>
<td>0.726</td>
</tr>
<tr>
<td>HSPA1A</td>
<td>p-value 0.220</td>
<td>0.988</td>
<td>0.161</td>
<td>0.220</td>
<td>0.988</td>
<td>0.161</td>
<td>0.220</td>
<td>0.988</td>
</tr>
<tr>
<td>HSPA1B</td>
<td>p-value 0.452</td>
<td>0.535</td>
<td>0.281</td>
<td>0.452</td>
<td>0.535</td>
<td>0.281</td>
<td>0.452</td>
<td>0.535</td>
</tr>
<tr>
<td>IER3</td>
<td>p-value 0.002**</td>
<td>0.003**</td>
<td>7.20E-04**</td>
<td>0.002**</td>
<td>0.003**</td>
<td>7.20E-04**</td>
<td>0.002**</td>
<td>0.003**</td>
</tr>
<tr>
<td></td>
<td>fc 126.676</td>
<td>62.538</td>
<td>0.806</td>
<td>126.676</td>
<td>62.538</td>
<td>0.806</td>
<td>126.676</td>
<td>62.538</td>
</tr>
<tr>
<td>PSMB9</td>
<td>p-value 0.023**</td>
<td>0.082*</td>
<td>0.292</td>
<td>0.023**</td>
<td>0.082*</td>
<td>0.292</td>
<td>0.023**</td>
<td>0.082*</td>
</tr>
<tr>
<td></td>
<td>fc 2.878</td>
<td>1.079</td>
<td>1.079</td>
<td>2.878</td>
<td>1.079</td>
<td>1.079</td>
<td>2.878</td>
<td>1.079</td>
</tr>
<tr>
<td>SLA-1</td>
<td>p-value 0.762</td>
<td>0.659</td>
<td>0.166</td>
<td>0.762</td>
<td>0.659</td>
<td>0.166</td>
<td>0.762</td>
<td>0.659</td>
</tr>
<tr>
<td></td>
<td>fc 1.049</td>
<td>1.072</td>
<td>1.042</td>
<td>1.049</td>
<td>1.072</td>
<td>1.042</td>
<td>1.049</td>
<td>1.072</td>
</tr>
<tr>
<td>SLA-DRB</td>
<td>p-value 0.557</td>
<td>0.737</td>
<td>0.176</td>
<td>0.557</td>
<td>0.737</td>
<td>0.176</td>
<td>0.557</td>
<td>0.737</td>
</tr>
<tr>
<td></td>
<td>fc 0.948</td>
<td>1.030</td>
<td>0.418</td>
<td>0.948</td>
<td>1.030</td>
<td>0.418</td>
<td>0.948</td>
<td>1.030</td>
</tr>
<tr>
<td>TAP1</td>
<td>p-value 0.418</td>
<td>0.438</td>
<td>0.327</td>
<td>0.418</td>
<td>0.438</td>
<td>0.327</td>
<td>0.418</td>
<td>0.438</td>
</tr>
<tr>
<td></td>
<td>fc 1.553</td>
<td>1.523</td>
<td>1.641</td>
<td>1.553</td>
<td>1.523</td>
<td>1.641</td>
<td>1.553</td>
<td>1.523</td>
</tr>
<tr>
<td>TNF</td>
<td>p-value 0.001**</td>
<td>0.015**</td>
<td>2.93E-04**</td>
<td>0.001**</td>
<td>0.015**</td>
<td>2.93E-04**</td>
<td>0.001**</td>
<td>0.015**</td>
</tr>
<tr>
<td></td>
<td>fc 54.695</td>
<td>11.262</td>
<td>5.816</td>
<td>54.695</td>
<td>11.262</td>
<td>5.816</td>
<td>54.695</td>
<td>11.262</td>
</tr>
<tr>
<td>TRIM26</td>
<td>p-value 0.427</td>
<td>0.239</td>
<td>0.106</td>
<td>0.427</td>
<td>0.239</td>
<td>0.106</td>
<td>0.427</td>
<td>0.239</td>
</tr>
<tr>
<td></td>
<td>fc 1.330</td>
<td>1.549</td>
<td>1.421</td>
<td>1.330</td>
<td>1.549</td>
<td>1.421</td>
<td>1.330</td>
<td>1.549</td>
</tr>
<tr>
<td>UBD</td>
<td>p-value 0.009**</td>
<td>0.029**</td>
<td>0.011**</td>
<td>0.009**</td>
<td>0.029**</td>
<td>0.011**</td>
<td>0.009**</td>
<td>0.029**</td>
</tr>
</tbody>
</table>
### TABLE 3.

Summary of results for RNA responses of mouse or porcine macrophages after various treatments *in vitro*.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Mouse</th>
<th>Pig</th>
<th>Known</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC</td>
<td>STE</td>
<td>SC514</td>
<td>EMSA</td>
<td>STE</td>
</tr>
<tr>
<td>AIF1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>CFB</td>
<td>4, 6</td>
<td>4</td>
<td>2*, 4</td>
<td>+</td>
</tr>
<tr>
<td>C2</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>HSPA1A</td>
<td>–</td>
<td>2, 4</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>HSPA1B</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>–</td>
</tr>
<tr>
<td>IER3</td>
<td>ALL</td>
<td>ALL</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>PSMB9</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>+</td>
</tr>
<tr>
<td>SLA-1/H2-K1</td>
<td>0.5(d)</td>
<td>–</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>SLA-DRB1/H2-Eb1</td>
<td>–</td>
<td>4</td>
<td>4</td>
<td>–</td>
</tr>
<tr>
<td>TAP1</td>
<td>–</td>
<td>4</td>
<td>4</td>
<td>+</td>
</tr>
<tr>
<td>TRIM26</td>
<td>ALL</td>
<td>ALL</td>
<td>1, 4</td>
<td>–</td>
</tr>
<tr>
<td>UBD</td>
<td>N/A</td>
<td>1(d)</td>
<td>1</td>
<td>+</td>
</tr>
</tbody>
</table>

Numbers in each column indicate a statistically significant effect at that time point in hpi. For the Time Course (TC) and the STE columns this effect is an increase in expression and in the SC-514 (SC514) columns it is a return in the direction of the untreated cells.

Y indicates that the gene is a known target of NF-kB.

C indicates there was a significant increase in expression due to S. Choleraesuis (Wang et al., 2008)

T indicates there was a significant increase in expression due to S. Typhimurium (Wang et al., 2007)

G indicates a significant decrease in expression in Guo et al. (2010)

Z indicates a gene differentially expressed in Zhao et al. (2006).

(d) denotes a decrease in expression.

* C2 had a significant difference between STE and SC-514 pretreated STE treated cells at 2 hpi, however the STE was not significantly different when compared to untreated cells.
FIGURE LEGENDS

**Figure 1a.** Antibody for NF-κB p65 subunit shows presence of active subunit in NF-κB complexes

Cy5 labeled probes specifically designed for three porcine genes, AIF1, UBD, TAP1, and two generic probes, strong (GS) and weak (GW) designed from the NF-κB consensus sequence show formation of protein:probe complexes, indicated by the black arrow.

**Figure 1b.** Antibody for NF-κB p65 subunit shows presence of NF-κB in protein-DNA complexes

Cy5 labeled probes specifically designed for three porcine genes, AIF1, UBD, TAP1, and two generic probes, strong (GS) and weak (GW) designed from the NF-κB consensus sequence show NF-κB specific binding as indicated by supershift or deletion of NF-κB band with anti-p65 antibody (Ab-p65). Black arrow indicates location of the shifted NF-κB:probe complex. Dark gray arrows indicate the location of the supershifts.

**Figure 2.** Nine of 12 murine motifs compete with generic probes for NF-κB binding.

The majority of the murine competitors worked against either the GS or GW probe. Only Hspa1b, H2-Eb1, and Tnf failed to show competition and C2 had week competition to the generic probes.

**Figure 3.** Eight of 12 porcine motifs compete with generic probes for NF-κB binding.

All competitors except for HSPA1A showed competition against either the GS or GW probe, with AIF1, SLA-1 and TRIM26 showing weak competition.

**Figure 4.** SC-514 has variable effects on the response of murine genes to STE.
The effect of SC-514 has different degrees of repression on STE induced mRNA expression of the murine genes. Some genes, represented by Tap1 at 4 hpi in 4b, have both a significant effect due to STE (DY/DN) as well as a significant difference between STE with and without SC-514 (SY/DY); termed Group 1 (G1) genes. Other genes, represented by H2-Eb1 at 4 hpi in 4c, have a significant difference between DN and DY, but not between DN and SY nor DY and SY; termed Group 2 (G2) genes. Some other genes have similar results to Tnf at 1 hpi, having a significant effect between DN and DY, a significant difference between DN and SY, and having a fold change difference between DY and SY of SY/DY ≤ 0.75. DN in the figure indicates untreated cells, DY are the STE only treated cells, and SY are the SC-514 pretreated STE treated cells. The number following is the hour post infection. Bar graphs in 4b, 4c, and 4d are the calculated by subtracting the Rpl32 adjusted Ct values from 40.

**Figure 5.** SC-514 has variable effects on the STE response by porcine genes

The effect of SC-514 has different degrees of repression on STE induced mRNA expression of the porcine genes, similar to the murine. Some genes, represented by TNF at 2 hpi in 5b, have both a significant effect due to STE (DY/DN) as well as a significant difference between STE with and without SC-514 (SY/DY); the Group 1 (G1) genes. Other genes, represented by TRIM26 at 4 hpi in 5c, have a significant difference between DN and DY, but not between DN and SY nor DY and SY; Group 2 (G2) genes. Some other genes have similar results to UBD at 8 hpi in 5d, having a significant effect between DN and DY, a significant difference between DN and SY, and having a fold change difference between SY/DY of ≤ 0.75. DN in the figure indicates untreated cells, DY are the STE only treated cells, and SY are the SC-514 pretreated STE treated cells. The number following is the hour post infection.
post infection. Bar graphs in 5b, 5c, and 5d are calculated by subtracting the RPL32 adjusted Ct values subtracted from 40.
**Figure 1a.** Couture et al., in prep.

<table>
<thead>
<tr>
<th>Lane</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>NSC</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>NE</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

**NF-κB Complex**
**Figure 1b.** Couture et al., in prep.

<table>
<thead>
<tr>
<th>Lane</th>
<th>GS</th>
<th>AIF1</th>
<th>UBD</th>
<th>TAP1</th>
<th>GW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ab-p65</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
</tr>
</tbody>
</table>

Supershifts

NF-κB Complex
Figure 2. Couture et al., in prep.

<table>
<thead>
<tr>
<th>Lane</th>
<th>Aif1</th>
<th>Cfb</th>
<th>C2</th>
<th>Hspa1a</th>
<th>Hspa1b</th>
<th>Ier3</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>NS</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
</tr>
</tbody>
</table>

NF-κB Complex

The table above shows the expression levels of different genes across various lanes, indicating the presence (+) or absence (−) of specific proteins or transcripts.
**Figure 2 (CONTINUED).** Couture *et al.*, in prep.

<table>
<thead>
<tr>
<th>Lane</th>
<th>H2-K1</th>
<th>H2-Eb1</th>
<th>Tap1</th>
<th>Tnf</th>
<th>Trim2</th>
<th>Ubd</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>NS</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

Diagram showing NF-κB Complex.
**Figure 3.** Couture *et al.*, in prep.

<table>
<thead>
<tr>
<th>Lane</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>NS</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
</tr>
</tbody>
</table>

**NF-κB Complex**
Figure 3 (CONTINUED). Couture et al., in prep.

<table>
<thead>
<tr>
<th>Lane</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>NS</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–</td>
<td>+</td>
</tr>
</tbody>
</table>

Lane 13-24: Experimental conditions across different lanes.

- **NF-κB Complex**: Indicates the presence of NF-κB complex in the experiment.
Figure 4a. Couture et al., in prep.
Figure 4b. Couture et al., in prep.

![Tap1 4HPI](image)

Figure 4c. Couture et al., in prep.

![H2-Eb1 2HPI](image)

Figure 4d. Couture et al., in prep.

![Tnf 1HPI](image)
Figure 5a. Couture et al., in prep.
Figure 5b. Couture et al., in prep.

Figure 5c. Couture et al., in prep.

Figure 5d. Couture et al., in prep.
<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>Specific and Non-specific competitor sequences</th>
<th>MATCH Core Score</th>
<th>MATCH Matrix Score</th>
<th>COMPETITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic Strong</td>
<td>*GGTGGAGGGGACTGCGAACGAGGCC</td>
<td>1.000</td>
<td>1.000</td>
<td>+</td>
</tr>
<tr>
<td>Generic Weak</td>
<td>GGTGGAGGGGACTTCAAGGAGGCC</td>
<td>1.000</td>
<td>0.866</td>
<td>+</td>
</tr>
<tr>
<td>AIF1 Probe</td>
<td>CTAGCTGGGAACCTCTGTATGCC</td>
<td>1.000</td>
<td>0.886</td>
<td>+</td>
</tr>
<tr>
<td>UBD Probe</td>
<td>TGCACTGGAAATCCCGACCCTGCT</td>
<td>1.000</td>
<td>0.886</td>
<td>+</td>
</tr>
<tr>
<td>TAP1 Probe</td>
<td>GATCTGGGAAATCCCGGAGACC</td>
<td>1.000</td>
<td>0.886</td>
<td>+</td>
</tr>
<tr>
<td>P_AIF1</td>
<td>TTTTGTAGAGATCCTGGTGAGG</td>
<td>1.000</td>
<td>0.901</td>
<td>-</td>
</tr>
<tr>
<td>P_CFB</td>
<td>AATAACTGGGAAATCCCGGC</td>
<td>1.000</td>
<td>0.903</td>
<td>+</td>
</tr>
<tr>
<td>P_C2</td>
<td>ACAGATCGGAACCTCTAAAGA</td>
<td>1.000</td>
<td>0.886</td>
<td>+</td>
</tr>
<tr>
<td>P_HSPA1 A</td>
<td>CCCCAGGAAACAGCGAGCCCGG</td>
<td>1.000</td>
<td>0.843</td>
<td>-</td>
</tr>
<tr>
<td>P_HSPA1 B</td>
<td>AACCTCTGGAAATCCCGGACCCTG</td>
<td>0.977</td>
<td>0.843</td>
<td>+</td>
</tr>
<tr>
<td>P_IER3</td>
<td>TAAACCTGGGAAATCCCGGACCC</td>
<td>1.000</td>
<td>0.986</td>
<td>+</td>
</tr>
<tr>
<td>P_SLA-1</td>
<td>CCGGGGTGGGGCGTTCCTGGGCC</td>
<td>0.929</td>
<td>0.825</td>
<td>-</td>
</tr>
<tr>
<td>P_SLA-DRB1</td>
<td>CGGCTGGAGGGACTTCCGCTGT</td>
<td>1.000</td>
<td>0.959</td>
<td>+</td>
</tr>
<tr>
<td>P_TAP1</td>
<td>GATCTGGGAAATCCCGGCAACCG</td>
<td>1.000</td>
<td>0.894</td>
<td>+</td>
</tr>
<tr>
<td>P_TNF</td>
<td>CCGGTGTGGGAAATCCCGGACCC</td>
<td>1.000</td>
<td>0.986</td>
<td>+</td>
</tr>
<tr>
<td>P_TRIM2 6</td>
<td>AGGCCTGGGAATCTCGAGGAC</td>
<td>1.000</td>
<td>0.847</td>
<td>-</td>
</tr>
<tr>
<td>P_UBD</td>
<td>TGCACTGGGAAATCCCGGACCCTGCT</td>
<td>1.000</td>
<td>0.859</td>
<td>+</td>
</tr>
</tbody>
</table>
## SUPPLEMENTAL TABLE 1 (CONTINUED)

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>Specific and Non-specific competitor sequences</th>
<th>MATCH Core Score</th>
<th>MATCH Matrix Score</th>
<th>COMPETITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>M_Aif1</td>
<td>TTCTGAGGGCAGATTCAGAGTTT</td>
<td>0.973</td>
<td>0.842</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>--C------------------------------------------</td>
<td>0.973</td>
<td>0.671</td>
<td></td>
</tr>
<tr>
<td>M_Cfb</td>
<td>TTCACACGGATTTCCAGGTATG</td>
<td>1.000</td>
<td>0.986</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>--G------------------------------------------</td>
<td>0.973</td>
<td>0.843</td>
<td></td>
</tr>
<tr>
<td>M_C2</td>
<td>TTTTCAGGAAAGGACCCAGAAAG</td>
<td>1.000</td>
<td>0.951</td>
<td></td>
</tr>
<tr>
<td></td>
<td>--C------------------------------------------</td>
<td>0.899</td>
<td>0.810</td>
<td></td>
</tr>
<tr>
<td>M_Hspa1a</td>
<td>GGGAGGCGGAGGAAAGCTCCACCAGA</td>
<td>1.000</td>
<td>0.945</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>C------C--G------</td>
<td>0.899</td>
<td>0.810</td>
<td></td>
</tr>
<tr>
<td>M_Hspa1b</td>
<td>AACTAAGAAAATGGCTTCTCTCTG</td>
<td>0.952</td>
<td>0.739</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T-------------------------</td>
<td>0.952</td>
<td>0.696</td>
<td></td>
</tr>
<tr>
<td>M_Ier3</td>
<td>TGCGAAGGGGAGTCCCCATCTCC</td>
<td>1.000</td>
<td>0.911</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>C--------C--------</td>
<td>1.000</td>
<td>0.763</td>
<td></td>
</tr>
<tr>
<td>M_H2-K1</td>
<td>CAGGGCTGGGGATCCCCATCTCC</td>
<td>1.000</td>
<td>1.000</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>C--------C--------</td>
<td>0.973</td>
<td>0.853</td>
<td></td>
</tr>
<tr>
<td>M_H2-Eb1</td>
<td>TCAGAGGGGAGACCCTGAAACTGAA</td>
<td>1.000</td>
<td>0.798</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C--------C--------</td>
<td>0.973</td>
<td>0.853</td>
<td></td>
</tr>
<tr>
<td>M_Tap1</td>
<td>GGTTCCGGGACTTTAACGCGACG</td>
<td>1.000</td>
<td>0.889</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>C--------C--------</td>
<td>0.788</td>
<td>0.754</td>
<td></td>
</tr>
<tr>
<td>M_Tnf</td>
<td>TGAATGAGACTTTCCGCCGCTC</td>
<td>1.000</td>
<td>0.808</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C--------A--G--G--</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>M_Trim26</td>
<td>TGCTGGGATTTGCCAGTGATC</td>
<td>1.000</td>
<td>0.951</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>C--------C--------</td>
<td>0.899</td>
<td>0.839</td>
<td></td>
</tr>
<tr>
<td>M_Ubd</td>
<td>TGAACCTGGGATTTCCAAACCAG</td>
<td>1.000</td>
<td>1.000</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>C--------C--------</td>
<td>1.000</td>
<td>0.980</td>
<td></td>
</tr>
</tbody>
</table>
### SUPPLEMENTAL TABLE 2.

q-RT-PCR Primer Sequences

<table>
<thead>
<tr>
<th>Gene</th>
<th>Organism</th>
<th>Forward</th>
<th>Reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIF1</td>
<td>Pig</td>
<td>ATCCAGTGGCTCTGGGGAGA</td>
<td>AGGACCCGTTGGCTTTTCTCT</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>CAGCAATGATGAGGATCTGC</td>
<td>CCAGCATTCGCTCTCAAGG</td>
</tr>
<tr>
<td>CFB</td>
<td>Pig</td>
<td>CCATCAAGGTCAAGCTGGT</td>
<td>TCCCTACCCTACTGTTCC</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>GCTATGATGGTTACGTTCTCC</td>
<td>GAATACCCGGGATTGGGACAGTAC</td>
</tr>
<tr>
<td>C2</td>
<td>Pig</td>
<td>AGCGACAGTGGCCAGGACAAC</td>
<td>TTGTGGCTCCAAGGAGGTGG</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>GGTGTTCTGCAAGGACTTC</td>
<td>GAACGCTCAGAGTATTAGACAGG</td>
</tr>
<tr>
<td>HSPA1A</td>
<td>Pig</td>
<td>TGCAAAAGTTGTACTGCGCTCTC</td>
<td>AGGCAGTGCAATAACAGGGAA</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>GAGGAGTTCAGAGAGAAGCACAAG</td>
<td>CGTGATGGATGTTGTAAGATCG</td>
</tr>
<tr>
<td>HSPA1B</td>
<td>Pig</td>
<td>CCACCATCGAGGAGGTGGAT</td>
<td>ACAAGTCATTATGCGATMCAACCT</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>GAGGAGTTCAGAGAGAAGCACAAG</td>
<td>CGTGATGGATGTTGTAAGATCG</td>
</tr>
<tr>
<td>IER3</td>
<td>Pig</td>
<td>AGCCCTCCGGACTACGCTTTG</td>
<td>TTCTTTTTGCGGGGTTGTCG</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>GCCCTGAACCTGACCTCGG</td>
<td>CGGCCTGTCACATAACCGTA</td>
</tr>
<tr>
<td>PSMB9</td>
<td>Pig</td>
<td>CTGGGACCAACGTTAAGG</td>
<td>GCCGCGTCCACATAACCGTA</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>SLA-1</td>
<td>Pig</td>
<td>CTGCCCCCTGGGGACTGAGG</td>
<td>CCACCTCCTCTCAGGATGC</td>
</tr>
<tr>
<td>(H2-K1)</td>
<td>Mouse</td>
<td>CAATGACGAGATTTCTGGAGT</td>
<td>CAGCCATATATCACCTGAAATG</td>
</tr>
<tr>
<td>SLA-DRB1</td>
<td>Pig</td>
<td>GCTCTTTGTGTCGTCGGG</td>
<td>TGAAGACGCTGHCAGCAGG</td>
</tr>
<tr>
<td>(H2-Eb1)</td>
<td>Mouse</td>
<td>GACCTGCGCTCCTCTGATTC</td>
<td>ACAAAAGGCAAGGACAAAAAGG</td>
</tr>
</tbody>
</table>
### SUPPLEMENTAL TABLE 1 (CONTINUED)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Organism</th>
<th>Forward</th>
<th>Reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAP1</td>
<td>Pig</td>
<td>GGTCTGTCTCTGCGCAAGAA</td>
<td>GCAAGAGGCTCAGCTCCGAA</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>ACGATTTCATCTCTGGGTTC</td>
<td>GCACCTGGTGCCATCC</td>
</tr>
<tr>
<td>TNF</td>
<td>Pig</td>
<td>CTTCGAGGTTATCGGCCCC</td>
<td>AGCTGTCCCTCGGCTTTGAC</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>GTGGAACCTGCAGAGGAGGC</td>
<td>AGACAGAAGAGCGTGTTGGC</td>
</tr>
<tr>
<td>TRIM26</td>
<td>Pig</td>
<td>CGAGGCCTGCGAGAAATTCCA</td>
<td>CTGGTACAGGGCGCTGTAGG</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>AAATGACTCCCCCTCTTTCCTGAACAC</td>
<td>GGCACCTGGCTAAACAACCTGAACC</td>
</tr>
<tr>
<td>UBD</td>
<td>Pig</td>
<td>AGATGGCTGCTGGCGTCTG</td>
<td>CGGCACAGGGACCTTGGTCT</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>CTTTCTGAGAACACACTGCAC</td>
<td>CAAGAGATGTGATTGGTAATAGC</td>
</tr>
<tr>
<td>RPL32</td>
<td>Pig</td>
<td>TGGGAAGAGAGCGTTGTGAGCAA</td>
<td>CGGAAGTTTCTGGTACACAAATGTAA</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>CTTACTGTGCTGAGATTGC</td>
<td>GGATTTGGTGACTCTGATGG</td>
</tr>
<tr>
<td>Genomic</td>
<td>Pig</td>
<td>CCTGCACCTCTGCTCTTGTAGT</td>
<td>CAGGGAGGTTAGGAAGTATCC</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>TTGAGCAGCTGCAGAACAGG</td>
<td>GAGTGTGTCATAGACATCCAG</td>
</tr>
</tbody>
</table>
SUPPLEMENTAL TABLE 3.

3D4/31 cells show lack of response to STE.

After treating with 1 or 10 ng/ml STE, 3D4/31 cells did not show an increase in TNF or IL-8.

<table>
<thead>
<tr>
<th></th>
<th>1 hpi</th>
<th>2 hpi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 vs. 1</td>
<td>0 vs. 10</td>
</tr>
<tr>
<td>TNF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>0.708</td>
<td>0.346</td>
</tr>
<tr>
<td>fc</td>
<td>1.038</td>
<td>1.573</td>
</tr>
<tr>
<td>IL8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>0.366</td>
<td>0.146</td>
</tr>
<tr>
<td>fc</td>
<td>0.708</td>
<td>0.615</td>
</tr>
</tbody>
</table>
CHAPTER 5: GENERAL CONCLUSIONS

DISCUSSION

To help understand and better use porcine expression data a database and associated web pages, ANEXdb (Chapter 2), was created to have two sides: an expression data repository, expressDB, and a sequence annotation repository, annotDB. Through the mapping of various sources of data, such as probe sets from arrays, GO and KEGG functional information, and sequence variation, to an ITC a central location has been created to look at both changes in expression as well functional information about the gene in question. This centralized sequence will allow users to go from seeing changes in their data directly to putative single nucleotide polymorphisms (SNPs) that could be targets for association studies, which could lead to finding genes that could influence meat quality, growth, innate immunity, any system that has a genetic control of variability.

Currently, two main parts of the annotation side of the database is being used: the annotation of the Affymetrix porcine GeneChip® and the variations predicted in the sequences from the assembly. The annotation stored in ANEXdb has been used and published in eleven papers to date from groups in the United States and Europe (Engrav et al., 2011, Uthe et al., 2011, Barb et al., 2010, Lkhagvadorj et al., 2010, Tomás et al., 2010, Towfic et al., 2010, Lkhagvadorj et al., 2009, Ross et al., 2009, Wang et al., 2008, Zhu et al., 2008, and Wang et al., 2007). Since the annotation is now available freely online, it is unknown how many groups are currently using the annotation. In addition to using the annotation for information about the Affymetrix platform, Uthe et al. (2011) utilized SNPs predicted from the assembly in sequences mapped to the Affymetrix
GeneChip® that showed differential expression between two groups of animals, which showed different responses to infection, verifying several of the predictions through sequencing of several populations.

The annotation was used in another way in Wang et al. (2008) to predict novel targets of NF-κB. To do this, homologous humans genes were identified, since the porcine genome is not currently annotated enough to run algorithms like TFM-Explorer, which requires known transcription start sites. However, we did show that even using surrogate sequences from human to model the pig, we were still able to identify two novel targets of NF-κB in both pig and mouse, TRIM26 and UBD, while being able to verify several known targets in pigs in response to endotoxin recognition: CFB, IER3, PSMB9, TAP1, and TNF and showing the more human response of C2 to endotoxin in pigs. While we were not able to get the exact same results as the data that was used to predict the set tested in Chapter 4, the difference in results are explainable due to the difference in using a cell line at limited time points with a bacterial component versus a live animal at extended times with a living pathogen. It was also possible to show that most of the targets that showed changes in expression with the loss of NF-κB function by showing it is possible for NF-κB to bind to a region of these genes promoters through EMSA competition. The method for the EMSA used in Chapter 4 is a previously unpublished method that is an effect and quick way to run an EMSA compared to other methods. However, care must be made when designing the probes and competitors, since it appears that if the probe is much more similar to the binding sequence than the competitor, this method may not be sensitive enough to show competition, which can be
avoided through careful designing of the probes and competitors, reducing the necessity of multiple probes.

MATURATION OF ANEXDB SINCE PUBLICATION

Since the publication of the article in 2009, ANEXdb has been updated to include the ability to have two color array data submitted to it in addition to the Affymetrix based data for which it was originally designed. To accomplish this, novel tables in expressDB were added and new web pages were coded in PHP to handle the different kinds of inputs, the main differences being two samples instead of one per array and the greater variation in data handling. To handle the multiple samples per slide a new hybridization table was created to link to two separate sample IDs within the database, and a new table was created to link the experimental factors and levels to the individual samples rather than the hybridization itself. As for the data analysis portion of the two-color array, only a single option exists from the multiple available methods and array scanners (see Appendix A for details).

There are currently several types of spot readers for two-color arrays, each producing their own formatted output file, unlike Affymetrix based output files, which are all standard. To help simplify this, ANEXdb allows the user to upload one file per slide, similar to how a single .CEL file represents an Affymetrix chip. Contained in this file are five columns: the name for the probe (either an ID or a gene name), the first colors foreground intensity, the second colors foreground intensity, the first colors background intensity, and the second colors background intensity. These values are then fed to R via PHP, just as is done for the Affymetrix data, which uses function from the limma package to run the most common background correction and normalization as
used by Steibel (2010). The data is corrected for the background noise, then lowess normalized, and finally has the intensities and the ratio of the two colors calculated. The data, along with the experimental information, is then outputted to the new GEO format and zipped together, ready for submission to GEO, just like the Affymetrix data.

The annotation portion of the database is also undergoing an update. Both additional species and a new porcine assembly are being added (see Appendix B for details on an assembly). No additional tables are needed for this; however, by adding additional species to the current tables, the search time will be extended, as MySQL will have to search through additional rows. The current search time is already quite long to go through the BLAST result tables, so to work around this, fewer BLAST results for each sequence are planned to be uploaded, up to a maximum of 25, instead of the current maximum of 250. By catching the best 25 BLAST hits, it should still be possible to obtain results from multiple species, which provides better evidence that the porcine sequence is that gene than if it hits one species only.

FUTURE RESEARCH

Future work in this area will be possible with the release of a stable porcine genome. While it would be preferred to have the genome also be well-annotated, this is not necessary as the assembled sequences and left over singletons could be aligned to the genome and transcription start sites could be inferred. However, with a well-annotated genome, this process would be much easier, especially in cases where a sequence might align twice to genome, with both alignments being strong. With a well annotated genome, the more advanced clustering methods to look for regulatory modules could be
ran on current and future high throughput expression experiments, furthering out understanding of porcine genetics.

Other methods can also replace the use of EMSAs for seeing if NF-κB can bind to the promoter regions of predicted genes. Two of these, chromatin immuno precipitation (ChIP) and promoter cloning, are also more sensitive than EMSA, they were not done for Chapter 4 due to certain limitations due to the use of NF-κB. The first limitation imposed by NF-κB is due to the number of possible binding sites its various subunits and heterodimers can bind. This promiscuous behaviour made it had to locate specific sites and regions of most of the promoters to knock out for promoter cloning, as the removal of regions with NF-κB binding sites potentially could contain other, important binding sites for other transcription factors contributing to a genes expression. The other limitation is the antibodies for NF-κB. While they work in EMSAs and have been shown to work in ChIP in some instances, lot-to-lot variability in binding affinity makes it hard to replicate results. New antibodies are becoming available and tested, but none were available at the time of Chapter 5.

Advancements to the database can also be done through the addition of more functionality in several areas. One of these would be to adapt a sequence viewer to display the sequences comprising the assembly along with tracks of annotation, such as Pfam domains, open reading frames, and if it is a contig, the contributing sequences and sequence variants. Another part that could be improved is the data analysis for two color datasets by allowing users to submit the files directly produced by the various spot readers instead of forcing a standard format on the users through modifying both the R scripts and PHP. Additional R command batch scripts can also be created to run different
preprocessing options, however, due to the nature of microarray data, it would always be recommended to check the results after any preprocessing stage to make sure it was the correct way to process it. By making the two sides of the database independent from the other, addition, non-microarray expression measures can also be included in the future by either creating additional databases or addition, or modified, tables in the current database.

CONCLUSIONS

In conclusion, it has been shown that through the use of bioinformatics followed by experimental manipulation that novel targets of NF-κB still exist. Two were identified in Chapter 5 but in other treatments, or combination of treatments, it may be possible to discover more. This is will be easier, and more direct, when a well-annotated porcine genome exists. These methods can also be applied to other transcription factors or other signal transduction members to help understand porcine genetics, which is important for a variety of reasons, including food safety, the use as a disease model, and the potential for being used as an organ contributor for xenotransplantation.
REFERENCES


APPENDIX A. STRUCTURE OF THE ANEXDB WEB PAGES AND UNDERLYING MYSQL DATABASES

USER SUBMISSION TO ANEXDB

From the home page, at www.animalgenome.org/anexdb/ (abbreviated as w~/anexdb or on the server at /home/users/tuggle/pub/anexdb/index.php or ~/index.php), users first have to either log in by clicking the ‘Login’ link or create an account by clicking the ‘Register’ link. Clicking the ‘Login’ link takes the user to w~/anexdb/login.php (~login.php) where they will either enter their username and password or click the link ‘Create new user account’ to make one. Clicking the ‘Create new user account’ link with take the user the same page as clicking ‘Register’ on the home page: w~/anexdb/register.php (~register.php). The data entered gets put into the anexdb.Users table after the user clicks the ‘Register’ button.

Once registered or logged in, the user gets taken to w~/dashboard.php (~dashboard.php). From here the user can view which groups they belong to, their account information, the submissions and experiments previously entered, and they can start a new submission. By clicking on “Start New Submission” the user will be taken to w~/submit/experiment.php?new=1 (~submit/experiment.php) where they can fill out the information pertaining to the entire experiment. This includes the experiments title, summary, a written summary of the overall design, the GEO accession number if available, a more detailed experimental design, which includes the factors and levels of the experiment as keywords, the number of colors on the array, any publications or links they want associated with the experiment, and if they want to make the experiments visibility private or public, or group if it is available to the user. This page uses multiple
anexdb_tmp tables: the experiment title, summary, overall design, GEO accession number, number of colors, and visibility, as well as the users ANEXdb ID get stored in anexdb_tmp.Experiments (Figure 1, Figure 2). The experiment is also assigned an ANEXdb experiment ID and a temporary ANEXdb experiment accession when the user hits “Save” or “Save and Continue.” The detailed experiment design section, which sets up the factors and levels of the experiment, gets split into multiple tables. The actual text data is stored in anexdb_tmp.Factors and anexdb_tmp.Levels and each factor and each level is given a unique ANEXdb ID. To link the levels with the factors, the factor ID is stored in anexdb_tmp.Levels. The levels and factors are not directly associated with the experiment, but rather with the hybridizations, which are associated with the experiment (see below, Figure 2). This was done because different experiments could have the same factor. The publication information is stored in anexdb_tmp.Publications and linked to the experiment through anexdb_tmp.PublicationExperiment table using the publication ID and the experiment ID (Figure 1, Figure 2). Links only uses one table, the anexdb_tmp.Links table, in which the experiment ID is stored (Figure 1, Figure 2). The publication and link data are all optional, and no rows are created in the database if none are provided. Once the user "inputs all their experiment data, they hit either the “Save” or “Save and Continue” button to move on.

Pressing the “Save and Continue” button takes them to the file upload page, w~/submit/upload.php (~/submit/upload.php). This page uses a flash program to upload the files from the user’s computer. Since it uses flash, Linux users will sometimes have difficulty getting it to work, with one of the easier solutions being to install the Windows version of Firefox or another browser through Wine to get it to work. For Affymetrix
data, the user needs to upload both the .CEL and .CHP files, for two color data the user needs to upload a file with a single header line followed by the correct rows: Name, F635 Median, B635 Median, F532 Median, and B532 Median. It is possible to rewrite the R batch files in order to make this not needed, but due to the number of available option, this one was chosen to simplify the submission to the database. Once all the files have uploaded, the user will then click the “Continue” button. When this button is pressed, 
~/submit/uploads.php calls ~/submit/uploadfiles.php to have Apache create the appropriate directory and to put the files in it. The files are created as owned by Apache in the Apache group.

Once clicking on “Continue” the user is taken to one of two pages depending on the number of colors inserted in the experiment page, either w~/submit/hybrids1.php or w~/submit/hybrids2.php (both in ~/submit/). These two pages do the same function but for either one or two color arrays, respectively, and uses different tables. Hybrids1.php uses anexdb_tmp.Files, anexdb_tmp.Hybridizations, anexdb_tmp.LevelHybridization, and anexdb_tmp.Samples. The main information for each hybridization, including the sample ID for the hybridization, is stored in anexdb_tmp.Hybridizations, the link between the hybridization and the levels for the array are stored in anexdb_tmp.LevelHybridization by linking the level ID with the hybrid ID, anexdb_tmp.Files is used to link the hybridizations with the .CEL and .CHP file, and the anexdb_tmp.Samples table is used to store the information concerning the sample which was hybridized on the array (Figure 1, Figure 3). For two color arrays, anexdb_tmp.Files2 replaces anexdb_tmp.Files, anexdb_tmp.Hybridizations2 replaces anexdb_tmp.Hybridizations, and anexdb_tmp.LevelSample replaces
anexdb_tmp.LevelHybridization (Figure 1, Figure 4). This is done because two samples are hybridized to a single array and they may represent very different treatments. The user will then click on “Add Another Hybridization” until all their hybridizations are loaded into the database. This does not have to be accomplished in one session and everything will be saved by clicking the ‘Save’ button. In addition to the ‘Save’ button, each time the user clicks “Add Another Hybridization” or “Save & Continue” the entered data is loaded into the corresponding tables.

Once all the hybridizations are entered, the user will be taken to w~/submit/review.php to look over the information they have entered so far. From here they can go back and edit or add more hybridizations. Once they are finished, they hit “Complete Submission” which then takes the experiment out of their hands, and they can no longer edit the information, leaving the curation up to the administrator.

ADMINISTRATOR FINALIZATION OF SUBMISSIONS

The administrator can then look over the submitted experiments by clicking on their link in the admin page, taking the administrator to the w~/admin/exp_review.php?ex=<exp> page where <exp> is the experiment ID for the submitted experiment. At this page the administrator can look over and make changes to the hybridizations without having to log into MySQL to do it directly. Once everything is acceptable, the administrator will then click ‘Approve Submission’ to finalize the process.

When the “Approve Submission” button is pressed, several things happen. The first is the experiment gets a static ANEXdb accession number by looking up how many previous experiments are in the database, similarly for the hybridizations, they too get
static accession numbers. The data is also migrated from the temporary database to the static database. This is done by having PHP call the move.cgi script in ~/cgi-bin/. If this is successful, a data entry will be made for each hybridization and put into anexdb.Datataks and will be used by _cron.php (~admin/_cron.php) to run the appropriate R script by calling _dataloader.php (~includes/_dataloader.php). However, there seems to be problems with _dataloader.php finishing as it appears the R does not close the datafile it creates in time for PHP to use it, and cannot upload the data to the appropriate tables in the database. To work around this, the administrator has to have permission to use Apache owned/group files and mimic the dataloader.php file. To facilitate this, dataloader_exp.cgi was created and can be run with the following command:

~/cgi-bin: perl dataloader_exp.cgi <expID> <process_cel> <process_mas> <process_rma> <process_2c> <colors>

where <expID> is the experiments ID, <process_cel>, <process_mas>, <process_rma>, and <process_2c> are either 1 or 0 depending if they want to be run (1) or not (0), and <colors> is either 1 or 2, depending on the number of colors on the array. While this script has the same problem the PHP version does, if run twice, it will successfully load the data into the databases, with the jobs being mapped to the anexdb.Datasets table and the data stored in anexdb.CelProbes for the CEL data, anexdb.MASData for MAS5, anexdb.RMADATA for the RMA data, anexdb.TwoC for the raw two color data, and anexdb.TwoCRG for the adjusted two color array data generated by R (Figure 1). The batch R scripts have been tested and work at the command line. Once the data is in the database, soft.cgi (~cgi-bin/soft.cgi) can then be run by the following command:

~/cgi-bin: perl soft.cgi <expID> <colors>
with the same arguments as above to create the soft file. For GEO submission, the soft file, along with the .CEL and .CHP files for Affymetrix data or the datafiles created by R for two color data need to be zipped together and can then be batch submitted to GEO.

**OTHER ADMINISTRATOR DUTIES FOR MANAGING EXPRESSION DATA**

In addition to finalizing the submissions, the administrator can also set up groups (using the anexdb.Groups and anexdb.UserGroup tables, and will use the anexdb_tmp.GroupExperiment and anexdb.GroupExperiment tables to link the group to the experiment during submission), enter in common protocols (saved in the anexdb.Protocols table), and load CDF files for Affymetrix data (stored in the anexdb.ArrayDesigns table). See Figure 1 for general relationships. All the group information can be added directly in MySQL.

In order for a user to properly enter all their data, an administrator must ensure three tables within in the database are up to date. These tables are: anexdb.OrganismPartNames, anexdb.ArrayDesigns, and anexdb.ProbeSets. The anexdb.OrganismPartNames table is used by the administrator to control the vocabulary of the various parts of all the organisms being submitted to the database in an attempt to reduce the multiplicity of terms to make searching the database easier. The anexdb.ArrayDesigns and anexdb.ProbeSets work together to define the arrays available in the database for submission. Updating these three tables vary, with both the anexdb.OrganismPartNames and anexdb.ArrayDesigns for two color arrays being done directly in MySQL using an INSERT statement, while being done via the web for Affymetrix based arrays. Due to the large number of probes, however, a Perl script was created to update anexdb.ProbeSets for two color arrays and a PHP page was created for
Affymetrix based platforms, which is run only after the array is loaded into the anexdb.ArrayDesigns table.

To update anexdb.ProbeSets for two color arrays, the Perl script 2c_probe_insert.pl file in ~/arrays was created. This links the probe sets to the correct array ID stored in anexdb.ArrayDesigns. To run the script type at the command line:

```bash
~/admin: perl 2c_probe_insert.pl <file> <array_id>
```

where `<file>` is a comma separated values file with a header line and the desired probe name or ID in the first column and `<array_id>` is the corresponding ID for the array in anexdb.ArrayDesigns. To insert the probe information for an Affymetrix based array, the corresponding cdf file must be uploaded into ~/affy/cdf. Once there the admin can go to w~/admin/ and clink on the link for the “CDF Loader”, which directs the admin to w~/admin/cdf_loader.php. Once the appropriate information is entered on the web page, both anexdb.ArrayDesigns and anexdb.ProbeSets gets updated with the appropriate information.
**FIGURE LEGENDS**

**Figure 1.** EER diagram of ANEXdb submission process. From left to right general relationships are grouped. First user logs in, creates a new submission and enters data for the Experiments, Levels, Factors, and Publications. Next the files that will populate Files(2) are uploaded. Then the Experiment, Levels, and Files, along with hybrid and samples specific data is entered, and Hybridizations(2) and Samples are populated. Once all hybridization information is entered, Datatasks is set during the admin step of finalization. The data tasks are then turned into the appropriate data sets once R is called to run the various statistics.

**Figure 2.** User entered experiment related information in ANEXdb. Relationship of tables pertaining to the uploaded experimental information during the submission process to ANEXdb. No foreign keys were used in the actual database.

**Figure 3.** User entered hybridization related information in ANEXdb for single color arrays. Relationship of tables pertaining to the uploaded experimental information during the submission process to ANEXdb for single color arrays, i.e. Affymetrix GeneChip®. No foreign keys were used in the actual database.

**Figure 4.** User entered hybridization related information in ANEXdb for two color arrays. Relationship of tables pertaining to the uploaded experimental information during the submission process to ANEXdb for two color arrays, i.e. spotted oligonucleotide arrays. No foreign keys were used in the actual database.
Figure 2
Figure 3
Figure 4
APPENDIX B. CONSTRUCTION AND ANNOTATION OF THE ASSEMBLY BY USING TGICL

GENERAL ASSEMBLY PROCEDURE

Sequence Download From NCBI

To obtain the sequences for the assembly, the three main databases at NCBI are first queried then the sequences extracted. For dbCore use NCBI’s eutils in dbCORE-extractor.pl to pull out sequences. If <COUNT> does not equal the number of sequences in the output file, truncate the search_results file to the number in <COUNT> and run the edited file using dbCORE-extractor-2.pl. Clean up files by deleting up to first ‘>’ to make the first sequence the first line in the file. To get just the number of results that are in <COUNT> use limit_search.pl. For dbEST the same method as dbCORE is used, but uses dbEST-extractor.pl, which points to dbEST instead of dbCORE. Due to an update at NCBI, the web pages can be used to select only the new sequence, then be saved into FASTA file available for download, which is much easier than using the above eUtils method.

The final database, TRACE, and for this database first get the number of sequences by using NCBI’s query_tracedb Perl file by entering at the command line:

```
./query_tracedb "query count species_code='SUS SCROFA' and source_type='NON GENOMIC'
```

Then retrieve by entering at the command line:

```
(echo -n "retrieve fasta 0b"; ./query_tracedb "query page_size 40000 page_number x binary species_code='SUS SCROFA' and source_type='NON GENOMIC'") | ./query_tracedb > pagey.fa
```

where x = {0…total number of sequences/40000} and y = x + 1. Repeat to get quality values by switching ‘fasta’ with ‘quality’ and ‘.fa’ with ‘.qual’ for the file extensions, can also get both at the same time using ‘single,’ but this is not recommended for later steps.
Each separate file then needs to be compiled into one. I did this by database, so first combine within database (I used Perl to ensure FASTA format was maintained, cat will work as well), and then combine the three main files into one. Due to the overlap between the databases, it is recommended to remove all the duplicate sequences. This can be done through the TRACE annotation, which usually includes the GenBank sequence ID for its corresponding sequence in either dbCORE or dbEST. However, since individual labs annotated their sequences and NCBI alters the annotation line without changing previous submissions, there can be a variety of different formats for this annotation so before each assembly, it has to be checked on how to obtain the GenBank ID from the TRACE annotation.

**Pre-assembly checks**

Once all the unique sequences are put into a single file, there are a few steps prior to the actual assemble: first they must be cleaned of bad sequence, then aligned, and finally comparing the alignments to pick out the consensus nucleotide. TGICL provides different components to do each step. SeqClean is the part of the TGICL package that will clean the sequences. It is a standalone addition to the core package and uses separate commands to run. To use, first download UniVec from NCBI (ftp://ftp.ncbi.nih.gov/pub/UniVec/) and create a file of species specific LINEs and SINEs (do nucleotide search for ‘sus scrofa’ and LINE/SINE). Once those have been downloaded, run the following command:

```
./seqclean <seqfile> -v <vectors> -o <outfile> -c <num_CPUs> -s <screens>
```

where `<seqfile>` is the compiled file, `<vectors>` is UniVec, `<outfile>` is the name of the output file desired (`<seqfile>_clean.fa` is what I usually use), `<num_CPUs>` is the number of CPUs to use (I usually use maximum), `<screens>` are the files to screen against, like the
LINEs and SINEs, but can also be sequences like mitochondria DNA or ribosomal sequences if they want to be excluded from the assembly. If using quality files, to clean them use the seqclean cln2qual to clean them. This uses the output file from seqclean to remove the base quality scores at the positions that were removed in the original sequence. However, it would be easier to clean TRACE files separately from other two databases, and then combine after cleaning, as TRACE is the only database of the three to have quality scores. On this file, it is recommended to use pattern matching to find sequences that are full-length clones and add the ‘et|’ prefix to them for use in the clustering phase (see below).

**Clustering**

Once the sequences are cleaned, clustering the sequences comes next. To perform the clustering, run TGICL on cleaned file using the command:

```
perl tgicl <fasta_db> -c <num_CPUs> -s <maxsize> -X
```

where `<fasta_db>` is the cleaned file of sequences, `<num_CPUs>` is the number of CPUs to use, and `<maxsize>` is the maximum number of sequences allowed in a cluster (typically around 60,000). This option will first run tclust, followed by sclut on the large clusters. By adding in the ‘et|’ prefixes earlier, sclust will require a higher match percentage as well as requiring shorter overhangs than tclust, thereby helping to separate out gene family members or alternative splicing. Other options include: if previous clustering was attempted, but failed, on the same sequences, using the –I option as well to skip making the database indices. If building a different cluster set after alignments are already created, use the -I and -K options to only rebuild the clusters clusters. This generally takes about 2 weeks using four, 3 GHz processors and 4 GB RAM.
Assembly

Finally, for the actual assembly, use the output file from clustering step with the following command:

```
perl tgicl <fasta_db> -c <num_CPUs> -a <cluster_file> -l
```

where `<cluster_file>` is the output file from the clustering step. The `-l` saves a little time by not rebuilding the file indexes, which is only used for the clustering stages. This feeds the clusters, one per node at a time, through CAP3 with the `-p 93` argument, meaning that the sequences must be at least 93% identical over their overlap to be included in the assembled sequence. For sequences that do not finish, usually do to a limitation in RAM, run them individually on CAP3 allowing more RAP per processor (making sure to run with it writing out to an `.ace` file: `./cap3 inputfile -p93 > inputfile.ace`) CAP3 is a slower, but stricter assembly program and this step usually takes about 2 to 3 weeks to complete with four, 3 GHz processors and 4 GB RAM. For the clusters that do not assemble, transfer to a 64 bit machine as large clusters can take over 8 GB RAM apiece to cluster, so have at least 8 GB RAM/processor on the machine, preferably 16 GB RAM or more. There is a new version of TGICL (TGICL2) available for 64-bit machines, and does appear to speed up the process, with the same commands, by about a week. However, it still has the RAM limitation.

ANNOTATING, PARSING, AND LOADING DATA INTO ANNOTDB

Assembly

On completion of the assembly, TGICL should create three files, a contig file, a singleton file, and an `.ace` file. There will also be a contig, singleton, and `.ace` file for each cluster that had to be run separately through CAP3 if they did not get through the first pass. To start the annotation, run `seq_rename_cat.pl` to pull all the contigs, singletons and `.ace` files
together while renaming all the contigs to match the previous IPA annotation (ITC.contig.member) as well as make a mapping of the CAP3 sequence name and the ITC name. After running this, to enter the contigs and sequences making up the contigs into their respective files run cluAss-parser-DBI-annotDB.pl, which will populate annotdb.consensus and annotdb.seqcon in the same directory as the contig and ACE file, with the ACE file as the first argument and the contig file as the second when running the script. To enter in all the sequences, run fastaParser-DBI-annotDB.pl in the same directory as the mapping file made from seq_rename_cat.pl with the name of the sequence file as the first argument. This will update the annotdb.sequences table.

**BLAST Annotation**

Previous annotation involved mainly used nucleotide BLAST for a quick initial run. For nucleotide blasts, download desired databases in FASTA format (previously ran RegSeq only, can download in ready to run BLAST format using NCBI's update_blast.pl script, which allows the skipping of the following command) and run formatdb using the command:

```
formatdb -i <input file> -p F -o T
```

Then run blastall with the following command:

```
blastall -p blastn -i <input file> -d <target database> -e <evalue cut off> -a <number of processors> -o <output file>
```

For an E-value cutoff I used 1e-5 to decrease run time and capture everything more stringent. For newer versions of BLAST, a different command will have to be entered, as blastall is no longer included from NCBI. For protein blast follow the above, but leave out the –p option for formatdb, and change the –p portion of blastall for whichever blast program desired. I used an E-value cutoff of 1e-3 for all protein blasts, as the databases are shorter. AnnotDB is dependent on the NM numbers obtained from NCBI's RefSeq database, so it is recommended
to use it. It is also recommended to download the gene2go, gene2refseq and gene_info files from the FTP site for gene at NCBI as well as animal specific <animal>_ncbi-geneid.list file from the ‘genes’ section of the FTP site at KEGG at the same time the RefSeq database is downloaded for blasting to ensure the NM and Gene IDs are kept current with each other. It would also be recommended to download the animal specific KEGG pathway list file, <animal>_gene_map.tab, from the ‘pathway’ portion of the FTP site, which map the KEGG ID to the pathway name. To locate the right KEGG files, look in the organism specific folders where <animal> is the three letter abbreviation of the animal’s name.

Open Reading Frames and SNPs, and Physical Location Uploading

Independently of the blast annotation, other annotations can be made to the sequences, such as the open reading frame, nucleotide polymorphisms, and genomic location. ORFinder at NCBI was used previously to predict all possible reading frames in the sequence. This was done by submitting the sequences online and then parsing the returned HTML pages from NCBI using a set of Perl scripts: ncbiORF.pl, to submit the sequences one at a time and save the resulting HTML file, and orfParser-DBI-annotdb.pl, which iterated through each HTML file and pulls out all the ORFs for each sequence to store in the annotdb.conpredictedprotein MySQL table. The resulting scripts need multiple folders to put the downloaded web pages in, each one containing 2,500 sequences and their results. The polymorphisms are predicted using several Perl script that uses the .ace file created by CAP3 and the other script to upload it to the database. However, the .ace file has to be split apart into its separate contigs for it to work properly, using the ace_exploder.pl script, creating several thousand files in different folders (currently set up to print 10,000 ACE files per folder). To iterate through them, SNP_finder.pl was created which will search through
folders of .ace files and call ace2snp-v2g.pl to find the SNPs and write them to a file (one file per .ace file, for a total of 20,000 files per folder), which SNP_DB1_annotdb.pl goes through and uploads the sequence variation information to the annotdb.seqvariation database table.

To get chromosomal location, exonerate was used previously, using the e2g (est to genome) model. While this take much longer than BLAST, it handles long introns better due to not penalizing for extended gaps if the sequence 5’ to the gap matches the conserved sequence for splicing. To speed up the process, on a multi-node computer, have one node act as the exonerate server and submit jobs on the other nodes. This does require more RAM to run, since the server will store the target database, in this case the genome, in RAM and will still need more to run the alignments. Two tables have been set up to handle the exonerate output, the annotdb.exoneratealn table stores each section of the alignment, and the annotdb.exoneraresum stores the overall alignment as well as the description for the sequences.

**BLAST Uploading**

Similar to the exonerate tables; there are several tables for the blast results depending on the kind of blast run. This was done to reduce the search time for finding the blast hits for a sequence. Currently the cutoff used is and E-value ≤ 1e-5 for nucleotide blasts and ≤ 1e-3 for protein blasts. In the future, while this E-value cutoff can still be used to shorten the blast runs, cutting off the number of results inserted into the tables will also reduce the search time. There are four sets of two tables, for a total of eight tables for the blast results. The ones ending in ‘results’ (i.e. annotdb.blastnresults) store the overall results, like annotdb.exoneratesum, while the tables ending in ‘hsp’ hold the information for high scoring pairs, like annotdb.exoneratealn. The letter following the blast table indicates what that set
of blast tables is for: ‘a’ for array (currently for the Affymetrix chip, can also be extended to two color chips), ‘f’ for function (usually pfam alignments), ‘n’ for nucleotide alignments to RefSeq nucleotide, and ‘x’ for jobs to RefSeq protein. To upload the blast results, blastParser-DBI-annotDB-v4-k.pl is the current version of the blast parser used. To run, type at the command line:

```
perl blastParser-DBI-annotDB-v4-k.pl <input file> <TAX> <query database>
```

where `<TAX>` is the taxonomy of the target database, or if the target database has multiple taxonomies, use 0. This will open the names.dmp file from NCBI and will use it to find the appropriate taxonomy ID for the hit. It is recommended to update this when the RefSeq databases are update as new organisms might get placed in it. The `<input file>` has to have the standard blast output in a flat file.

**KEGG, NCBI, Pfam, GO**

Parsers for annotdb have already been written for KEGG, NCBI, and Pfam as well. To update KEGG, download the three files previously mentioned, or download the KEGG general files (which are the ones currently being used): genes_ncbi-geneid.list, genes_pathway.list, and map_title.tab. Once these are downloaded, run keggParser-DBI-annotDB-v2.pl, which has to be in the same directory as the files. This will update both the annotdb.kegg2ncbi and annotdb.kegggenepath files simultaneously. For NCBI, download the gene2refseq, gene_info, and gene2go files and run their corresponding parsers:
gene2refseqParser-DBI-annotdb.pl (updates annotdb.gene2refseq), geneInfoParser-DBI-annotdb.pl (updates annotdb.geneinfo), and gene2goParser-DBI-annotdb.pl (updates annotdb.gene2go). For Pfam updates, download Pfam-A.full from Pfam and pfam2go from Gene Ontology and run pfamParser-DBI-annotDB-v2.pl to update annotdb.pfam and
annotdb.pfam2go respectively. For all of these, the files have to be in the same directory as the program. The Pfam-A.full file contains all the protein sequences representing all the Pfam consensus sequences and is what should be the target database to obtain Pfam data for the assembly.

Once the Pfam and GO information is updated, running blast2function.pl will gather all necessary data from annotdb and populate annotdb.blast2kegg and annotdb.blast2go tables. However, this generally takes more than 2 GB of RAM and can not be run on a Windows machine due to the limit of no more than 2 GB of RAM allocated to a single process. To get around this a series of Perl scripts have been created if they are needed: b2r.pl (links blast results to RefSeq ID), g2r.pl (links Gene ID to RefSeq ID), gi.pl (links just Gene ID to Gene Symbol), g2g.pl (links Gene ID to GO ID and term), kpn.pl (links KEGG Gene ID to KEGG Pathway ID, and KEGG Pathway ID to KEGG Pathway Name), k2n.pl (links Gene ID to KEGG Gene ID) and prints the data to flat files, which can then be used by an edited version of blast2function.pl. However, just using it on a Linux based system will work.

WEB PAGES

The web pages for viewing this data are found in on the server at /home/users/tuggle/pub/anexdb/annotdb/ (~/). To get there online, the user has to go to www.animalgenome.org/anexdb/search/ and run a query on the gene side of the page. If searching by an ITC, the user will get taken to the ~/gene_results.php where they can select on a result to go to ~/gene.php where summary information about the sequence can be found, such as the actual sequences (annotdb.consensus), how many sequences contribute to it (annotdb.seqcon), if it is on an expression array (annotdb.blastaresults), if it has sequences
variation (annotdb.seqvar), if it has an ORF (annotdb.conpredictedprotein), the top five
RefSeq nucleotide blast hits (annotdb.blastnresults), the top five RefSeq protein hits
(annotdb.blastxresults), if there is a chromosomal alignment (annotdb.exoneratesum), if it
belongs to a KEGG (annotdb.blast2kegg) or GO (annotdb.blast2go) pathway or term
associated with it, and if there is a Pfam alignment (annotdb.blastfresults). All of these have
link outs to different web pages: ~/blast_results.php displays all the blast hits,
~/go_results.php display all the GO terms based on the blast results, ~/kegg_results.php
shows all the KEGG pathways based on the blast results, ~/orf_results.php shows the ORFs
predicted by ORFinder, snp_results.php displays all the sequence variation, ~/chr.php shows
all the chromosomal alignments from exonerate, and ~/probe.php shows all the information
about the probe that the sequence aligns to.

In addition to pointing at ~/probe.php, ~/probe.php also points back at ~/gene.php and
can be gotten to on the search page by entering in the Affymetrix Probe ID. Right now this is
set up only for Affymetrix probes, but can be edited to work with any probes. Also, the
gene_name_results.php page, which can also be a result from the search page, hyperlinks to
gene.php by searching the annotdb.blast[nx]results.br[nx]Description field for the term
entered in the search page.
FIGURE LEGENDS

Figure 1. EER diagram of AnnotDB with similar tables grouped together by function.
Figure 1
ACKNOWLEDGEMENTS

First I would like to thank the people who brought me into this world and spent the better part of their lives putting up with me, my parents. From getting my head stuck in a chair or jumping down wells and the other close calls I’ve had, you have always been there for me, trying to make me do my best at whatever I wanted to do, and believing in me. Your support has been instrumental in my life and has helped me be the person I am now. I would also like to thank Scott for also always being available when needed, and for all the Dave Barry stories, I wish he could be here today. Also, to the rest of my family, my brothers and sister, grandparents, aunts, uncles, and cousins, thank you for all the years and years of memories.

I would also like to thank all the teachers and mentors along the way who made a difference. Mrs. Hunaffa, Mrs. Sievers, Mr. Ayers, Dr. Stiles, Mr. Hewins, Mr. Mauler, Mr. Long, Mr. Treman, Mr. Pfander, Mrs. Dilts, your classes in elementary through high school helped move and keep me interested in science, and without you, who knows that path I would have taken. To Drs. Charisse Buising and Mike Myszewski, thank you for being amazing advisors while at Drake, as well as to Drs. Steven Faux, Bill Kilpec, Brian Sanders, Maria Clapham, Donald Stratton LaRhea Henderson, and the rest of the amazing faculty at Drake, thank you all for treating me as more than just a student and being there in the good and not so good times. And the members of my POS committee, Drs. Vasant Honavar, Jack Dekkers, James Reecy, and Dan Nettleton, than you all for taking time out of your busy schedules when I had questions and for your suggestions and guidance. And of course I am very greatful to Dr. Chris Tuggle for all the advice and help over the years, without which I wouldn’t be here today. I am so grateful to Drs. Matt Abbott, Milan Joksimovic, and
Shuhong Zhao for all the time you spent training me when I was just starting out, and to the undergrads, Beth Kenkel, Nowlan Freese, and Keith Callenberg, it was a privilege to work with you on the many various project. I would also like to thank Fadi Towfic, Jolita Uthe, and Curt Christenson for their suggestions and insights on various projects over the years.

I would also like to thank my friends who have stuck with me, mostly by putting up with me for all these years. Jeff, and the entire Greenberg family, I really ask for a better friend over the many, many years. To Christy, thank you for finally coming around and letting me escape to Chicago while going to Drake, and to Kristy and Robert for being there through Drake and now. To Josh, and the rest of the Moore family, a special thank you for all the great trips for some rest and relaxation down in Florida. And to Clayton, Ed, Wez, and Nate thank you for all the good times in Iowa City and now. Also to Allison and Eric for sticking through the years and being there whenever I needed someone to talk with. A big thanks for Jamie, James, Dawn, Colin, Ruth, Adam, and Jason for all the great times and memories at the Ames bars, lunches, and dinners over the years, a guy’s gotta eat. And very special thanks to Benny and Brandi for giving me a place to rest my head at nights and putting up living with me.