2017

Swine blood transcriptomics: Application and advancement

Haibo Liu

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Bioinformatics Commons

Recommended Citation

https://lib.dr.iastate.edu/etd/15353

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations 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.
Swine blood transcriptomics: Application and advancement

by

Haibo Liu

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

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Christopher K. Tuggle, Co-major Professor
Peng Liu, Co-major Professor
Jack C. M. Dekkers
Dan Nettleton
Leslie Miller
Nicholas K. Gabler

Iowa State University
Ames, Iowa
2017

Copyright © Haibo Liu, 2017. All rights reserved.
DEDICATION

This work is dedicated to my wife, Yalin Rao, and my son, Austin Liu.
# TABLE OF CONTENTS

**NOMENCLATURE** .................................................................................................................. v

**ACKNOWLEDGMENTS** ......................................................................................................... viii

**ABSTRACT** .............................................................................................................................. xii

**CHAPTER 1. INTRODUCTION** .............................................................................................. 1

  Current Status of the Pig Reference Genome and Its Annotation ........................................ 1
  Porcine Transcriptomics: Tools and Their Applications ....................................................... 3
  Feed Efficiency in Pigs .......................................................................................................... 15
  Rationale for Developing Biomarkers for Breeding Feed Efficient Pigs ......................... 18
  Dissertation Organization .................................................................................................. 20

**CHAPTER 2. PIGS DIVERGENTLY SELECTED FOR RESIDUAL FEED INTAKE RESPOND TO SYSTEMIC INFLAMMATION TRIGGERED BY LIPOPOLYSACCHARIDE IN LARGELY SIMILAR WAYS AND SHARE A LARGE SET OF CORE INFLAMMATORY RESPONSE GENES WITH HUMANS** .......... 23

  Abstract ................................................................................................................................. 23
  Background ............................................................................................................................ 25
  Methods .................................................................................................................................. 29
  Results ...................................................................................................................................... 39
  Discussion ............................................................................................................................... 52
  Conclusions ............................................................................................................................. 57

**CHAPTER 3. POST-WEANING BLOOD TRANSCRIPTOMIC DIFFERENCES BETWEEN YORKSHIRE PIGS DIVERGENTLY SELECTED FOR RESIDUAL FEED INTAKE** ................................................................................................................................. 69

  Abstract ................................................................................................................................. 69
  Background ............................................................................................................................ 71
  Methods .................................................................................................................................. 74
  Results ...................................................................................................................................... 85
  Discussion ............................................................................................................................... 91
  Conclusions ............................................................................................................................. 98
### NOMENCLATURE

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D-PCA</td>
<td>3 Dimensional-Principal Component Analysis</td>
</tr>
<tr>
<td>ADFI</td>
<td>Average Daily Feed Intake</td>
</tr>
<tr>
<td>ADG</td>
<td>Average Daily Gain</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike Information Criterion</td>
</tr>
<tr>
<td>APC</td>
<td>Antigen Presenting Cell</td>
</tr>
<tr>
<td>BH</td>
<td>Benjamini–Hochberg method</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>BW</td>
<td>Body Weight</td>
</tr>
<tr>
<td>CAGE</td>
<td>Cap Analysis of Gene Expression</td>
</tr>
<tr>
<td>CBC</td>
<td>Complete Blood Count</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>CFA</td>
<td>Complete Freund’s Adjuvant</td>
</tr>
<tr>
<td>CGP</td>
<td>Chemical and Genetic Perturbation</td>
</tr>
<tr>
<td>CP</td>
<td>Canonical Pathway</td>
</tr>
<tr>
<td>CPB</td>
<td>Coverage Per Base</td>
</tr>
<tr>
<td>cpm</td>
<td>Count Per Million mapped reads</td>
</tr>
<tr>
<td>DC-megaBLAST</td>
<td>Discontiguous megaBLAST</td>
</tr>
<tr>
<td>DEG</td>
<td>Differentially Expressed Gene</td>
</tr>
<tr>
<td>DEP</td>
<td>Differentially Expressed Probeset</td>
</tr>
<tr>
<td>Drop-Seq</td>
<td>Drop-Sequencing</td>
</tr>
<tr>
<td>EC</td>
<td>Enzyme Commission</td>
</tr>
<tr>
<td>EDTA</td>
<td>EthyleneDiamineTetraacetic Acid</td>
</tr>
<tr>
<td>EST</td>
<td>Expressed Sequence Tag</td>
</tr>
<tr>
<td>FANTOM</td>
<td>Functional Annotation of the Mammalian Genome</td>
</tr>
<tr>
<td>FC</td>
<td>Fold Change</td>
</tr>
<tr>
<td>FDR</td>
<td>False Discovery Rate</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>FE</td>
<td>Feed Efficiency</td>
</tr>
<tr>
<td>FIRE</td>
<td>Feed Intake Recording Equipment</td>
</tr>
<tr>
<td>FPKM</td>
<td>Fragments Per Kilobase of exon per Million fragments mapped</td>
</tr>
<tr>
<td>GAGE</td>
<td>Generally Applicable Gene-set Enrichment for pathway analysis</td>
</tr>
<tr>
<td>GENCODE</td>
<td>Genome ENCylcopedia Of DNA Elements</td>
</tr>
<tr>
<td>GO-BP</td>
<td>Gene Ontology-Biological Process</td>
</tr>
<tr>
<td>GO-CC</td>
<td>Gene Ontology-Cellular Component</td>
</tr>
<tr>
<td>GO-MF</td>
<td>Gene Ontology-Molecular Function</td>
</tr>
<tr>
<td>GWAS</td>
<td>Differentially Expressed Gene</td>
</tr>
<tr>
<td>HFD</td>
<td>High-Fiber, low-energy Diet</td>
</tr>
<tr>
<td>HGNC</td>
<td>HUGO Gene Nomenclature Committee</td>
</tr>
<tr>
<td>hpi</td>
<td>Hour Post Injection</td>
</tr>
<tr>
<td>HRFI</td>
<td>High RFI</td>
</tr>
<tr>
<td>HSP</td>
<td>High-scoring Segment Pair</td>
</tr>
<tr>
<td>IACUC</td>
<td>Institutional Animal Care and Use Committee</td>
</tr>
<tr>
<td>ID</td>
<td>Identifier</td>
</tr>
<tr>
<td>IPA</td>
<td>Ingenuity Pathway Analysis</td>
</tr>
<tr>
<td>IPAD</td>
<td>Integrated Pathway Analysis Database</td>
</tr>
<tr>
<td>IsoSeq</td>
<td>Isoform Sequencing</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>LD</td>
<td>longissimus dorsi</td>
</tr>
<tr>
<td>LFD</td>
<td>Low-Fiber, high-energy Diet</td>
</tr>
<tr>
<td>IncRNA</td>
<td>long noncoding RNA</td>
</tr>
<tr>
<td>LPS</td>
<td>LipoPolySaccharide</td>
</tr>
<tr>
<td>LRFI</td>
<td>Low RFI</td>
</tr>
<tr>
<td>MBW</td>
<td>Metabolic mid-Body Weight</td>
</tr>
<tr>
<td>MDS</td>
<td>Multi-Dimensional Scaling</td>
</tr>
<tr>
<td>miRNA</td>
<td>microRNA</td>
</tr>
<tr>
<td>MHC</td>
<td>Major Histocompatibility Complex</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>misc_RNA</td>
<td>miscellaneous RNA</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>MsigDB</td>
<td>Molecular Signature DataBase</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>ncRNA</td>
<td>noncoding RNA</td>
</tr>
<tr>
<td>NR</td>
<td>NCBI Non-Redundant protein database</td>
</tr>
<tr>
<td>NT</td>
<td>NCBI NucleoTide database</td>
</tr>
<tr>
<td>ORF</td>
<td>Open Reading Frame</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PG</td>
<td>Putative Gene</td>
</tr>
<tr>
<td>PRRSv</td>
<td>Porcine reproductive and respiratory syndrome virus</td>
</tr>
<tr>
<td>PT</td>
<td>Putative Transcript</td>
</tr>
<tr>
<td>RBC</td>
<td>Red Blood Cell</td>
</tr>
<tr>
<td>RFI</td>
<td>Residual Feed Intake</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>RIN</td>
<td>RNA Integrity Number</td>
</tr>
<tr>
<td>RNA-seq</td>
<td>RNA-sequencing</td>
</tr>
<tr>
<td>ROS</td>
<td>Reactive Oxygen Species</td>
</tr>
<tr>
<td>scRNA-seq</td>
<td>single cell RNA-sequencing</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>STEM</td>
<td>Short Time Expression Miner</td>
</tr>
<tr>
<td>TCA</td>
<td>TriCarboxylic Acid</td>
</tr>
<tr>
<td>WBC</td>
<td>White Blood Cell</td>
</tr>
<tr>
<td>WGCNA</td>
<td>Weighted Gene Co-expression Network Analysis</td>
</tr>
</tbody>
</table>
First of all, I would like to thank my major professor, Christopher K. Tuggle. He is the best mentor I have had in my life. He always gives me trust, latitude, advice, opportunity, help and encouragement, not only in scientific research but also in my daily life. I appreciate his care about my future career. I enjoy my last five years of life as his student. I would like to give “Thanks a lot” to him for his time on mentoring me and on correcting and commenting on my manuscripts and other writing. I wish I had him as my mentor in my earlier academic life.

I would like to thank my co-major professor Peng Liu and my other program of study (POS) committee members, Jack C. M. Dekkers, Dan Nettleton, Leslie Miller, and Nicholas K. Gabler for their guidance and support throughout the years. I am lucky to have them as my POS committee. They are always there when I need advice and help from them. Also thanks should be given to them for their considerateness. I also have to say thanks to other faculty from multiple departments who taught me during the years. They made my learning experience enjoyable and valuable.

I am always grateful to Trish Stauble, our great program coordinator. I always remember the email from her, replying to my inquiry whether I was welcome to apply for the Bioinformatics and Computational Biology PhD program at Iowa State University (ISU). Her response was so encouraging and positive. She told me I even did not need to pay for the application fee, which is almost impossible for most programs around the country. She called me and invited me to have an on-site interview then offered me a wonderful fellowship. So I accepted the offer without hesitation. Without her, I would not be here in ISU today. Trish
was always the first person to get help and advice from when I was in trouble or have questions. She is also a great organizer of events. I enjoy going to the First Thursday Dinner she organized, where our BCBers meet each other and have fun. Without her, my life in ISU would not be so colorful.

I want to thank all former and present members of the Tuggle Lab during these years. I always learn from them, and get encouragement and feedback from them, especially Ellis Powell, before and after presentation especially my preliminary presentation. I still remember her hug after passing our preliminary exams. I am indebted to Kristina Feye who performed all wet lab work for Chapter 3, though it was originally her thesis project. Martine Schroyen was always helpful. She provided advice and helped with RNA preparation and globin depletion for Chapter 3 work. She also provided awesome support in supply order and lab management during the years. Ziyanda Mpetile’s help in collecting blood samples for Chapter 2 is also worth mentioning.

I want to thank the great collaborators at the Meat Animal Research Center (USDA-ARS). Susan Hauver and Kristen Kuhn there provided technical support to generate IsoSeq data. Dr. Timothy Smith provided us the valuable, newly-generated long read-based pig genome assembly and multi-tissue IsoSeq data, which were indispensible parts of Chapter 4. Dr. Timothy Smith and Dan Nonneman helped edit the manuscript, now as Chapter 4. Susan Hauver helped edit the Supplementary Methods presented in Appendix C.

I am grateful to Nick Boddicker and the Lauren Christian Memorial Swine Breeding farm staff for their assistance in collecting the blood samples and managing the pig lines for my research. I also thank PIC/Genus and Newsham Choice Genetics for donating FIRE
feeders, and Iowa Pork Producers Association for funding to support the RFI lines. All this made my research possible.

In addition, I would also like to thank my friends, other colleagues, faculty and staff over the campus for making my time at ISU a wonderful experience. Persons needed to be specifically mentioned are Ruolin Liu, Kokulapalan Wimalanathan, John Hsieh, and Melissa Herrmann. Kokulapalan is a good collaborator and a nice helper. I still remember that when I started my first BCBLab project, it was he and Ruolin that gave me a step-by-step guidance. Besides that, I also got help from him for several computational questions. He provided me his preliminary proposal as sample and helpful advice. John and Melissa provided me lots of help for editing and commenting on my preliminary proposal. Ann Shuey, Donna Nelson and other staff in Department of Animal Science are very helpful in providing support for research and study. In addition, many other BCBer and colleague in Department of Animal Science also help me with study and research. I would like to thank them all without listing their names one by one.

My research was funded by the Agriculture and Food Research Initiative competitive grant-NIFA-AFRI-2011-68004-30336 (awarded to Professor John F. Patience) from the USDA National Institute of Food and Agriculture. I was partly supported by the fellowship from the Plant Science Institute, ISU. I would also like to thank James Cornette Fellowship and Jorgenson Travel Grant for their support to my career development.

Last, but not least, I have to thank my family. Without their unwavering love and support, it would be impossible for me to go such a long journey towards my goals. Thanks to my wife, my strong follower. She spends so much time with my son and takes lots of responsibilities for my family. I appreciate their understanding of me and my work, though I
always feel sorry and guilty because I did not take all responsibilities I was supposed to do but did not have time or energy to do. I would like to thank my old parents for their understanding and support.
ABSTRACT

Improving swine feed efficiency (FE) by selection for low residual feed intake (RFI) is of practical interest. However, whether selection for low RFI compromises a pig’s immune response is not clear. In addition, current RFI-based selection for improving feed efficiency was expensive and time-consuming. Seeking alternative tools to facilitate selection, such as predictive biomarkers for RFI, is of great interest. The objectives of this thesis are as follows: (1) to investigate whether selection for low RFI compromise a pig’s immune response; (2) to develop candidate biomarkers applicable at early growth stage for predicting RFI at late growth stage; (3) to improve the annotation of the porcine blood transcriptome.

In Chapter 2, pigs of two lines divergently selected for RFI were injected with lipopolysaccharide (LPS). Transcriptomes of peripheral blood at baseline and multi-time points post injection were profiled by RNA-seq. LPS injection induced systemic inflammatory response in both RFI lines. However, no significant differences were detected in dynamics of body temperature, blood cell count and cytokine levels during the time course. Only a very small number of differentially expressed genes (DEGs) were detected between the lines over all time points, though ~ 50% of blood genes were differentially expressed post LPS injection compared to baseline for each line. The two lines were largely similar in most biological pathways and processes studied. Minor differences included a slightly lower level of inflammatory response in the low- versus high-RFI animals. Cross-species comparison showed that humans and pigs responded to LPS stimulation similarly at both the gene and pathway levels, though pigs are more tolerant to LPS than humans.
In Chapter 3, post-weaning blood transcriptomic differences between the two lines were studied by RNA-seq. DEGs between the lines significantly overlapped gene sets associated with human diseases, such as eating disorders, hyperphagia and mitochondrial disease. Genes functioning in the mitochondrion and proteasome, and signaling had lower and higher expression in the low-RFI group relative to the high-RFI group, respectively. Expression levels of five differentially expressed genes between the two groups were significantly associated with individual animal’s RFI values. These five genes were candidate biomarkers for predicting RFI.

Given limitations of current annotation of the porcine reference genome, a high-quality annotated transcriptome of porcine peripheral blood was built in the last study via a hybrid assembly strategy with a large amount of blood RNA-seq data from studies mentioned above and public databases.

Taken together, this work provides evidence that selection for low RFI did not significantly compromise pigs’ immune response to systemic inflammation, offers a few candidate biomarkers for predicting RFI to facilitate RFI-based selection, and significantly advances the structural and functional annotation of porcine blood transcriptome.
CHAPTER 1. INTRODUCTION

Current Status of the Pig Reference Genome and Its Annotation

The domestic pig is one of the most economically important livestock species. Globally, pork accounts for more than 40 percent of total meat consumed by humans and rank as the top one meat source (2015). The pig has also been recognized as an emerging biomedical model for human physiology and many diseases, because genomic, physiological and anatomic similarities of the pig to human are greater than those of classic animal models to human [1, 2]. For example, mini-pigs have played a valuable role in biomedical research. Besides naturally occurring mutants and pigs with artificially induced diseases, genetically modified pigs have also been used as biomedical models over the last two decades.

Deep understanding of the pig genome has been essential to improving pig genetics for better performance and developing useful models for human diseases. The pig genome is composed of 18 autosomes (SSC1-18) and two sex chromosomes (SSCX and SSCY), with a total size of 2.7 ~2.9 gigabases (Gb). The draft sequence of the pig genome was first released in 2009 by the Swine Genome Sequencing Consortium (SGSC), which was mainly based on BAC-by-BAC Sanger sequencing and low coverage of Illumina whole genome shotgun sequencing of DNA from 3 individuals of different genetic backgrounds [3]. The draft assembly was further refined and the current version of the draft assembly is Sscrofa10.2 [4]. This reference genome assembly includes 2.60 Gb placed onto 20 chromosomes consisting of 5,343 scaffolds and 212 megabases (Mb) unplaced sequences distributed among 4,562 scaffolds. Because of the sequence depth was not high enough, the draft assembly was still very fragmented, with contig N50 =
80,720 base pairs (bp) and scaffold N50 =637,332 bp. In the most recent release (Release 87) [5], 21,630 coding and 3,124 noncoding genes have been annotated. The non-coding genes include 2,804 small non-coding genes, 135 long non-coding genes and 185 miscellaneous non-coding genes. In addition, 568 psudogenes were also identified. Surprisingly, only 30,585 transcripts were annotated.

The draft reference genome has been shown to facilitate genetic improvement of pig for better performance [6-8]. The availability of the draft reference genome and its structural and functional annotation, and the development of genetic manipulation tools, such as somatic cell nuclear cloning, transgenesis and genome editing techniques especially the recent developed CRISPR/cas9 system, have been accelerating the use of genetically modified pigs as biomedical models for specific human diseases and a potential source of organ xenotransplantation [9-12].

However, further study showed that the current version of reference genome is neither well-assembled nor well-annotated [13]. For example, only 43% of the 25,510 transcripts of the 21,630 coding genes have both a 5’ and 3’ UTRs defined. On average, less than 1.2 isoforms per protein-coding gene and only 3,124 non-coding genes have been identified in SSC10.2. More recently, Warr et al. [14] reported that more than 33% of SSC10.2 is not correctly assembled or is otherwise unreliable. In addition, many known genes are missing from the reference genome [15, 16]. Therefore, improvement of the pig reference genome assembly is necessary. With the advent of PacBio long read sequencing technology, recently two groups, at the Roslin Institute, Edinburg University, and the Meat Animal Research Center, USDA-ARS have independently reassembled the pig reference genome by using this new sequencing technology [15, 17]. The new assemblies show great improvement over Sscrofa10.2, especially the assembly by the Roslin Institute, which has been released to NCBI GenBank as Sscrofa11.0, and is comprised of 19
chromosomes (SSC1-18, and SSCX), with DNA from the single female Duroc animal that was used for the Sscrofa10.2 assembly. There are only 122 contigs and less than 600 unplaced scaffolds in this version, with contig $N50 = 58.5$ Mbp and scaffold $N50 = 107.6$ Mbp. And the estimated base accuracy is 99.999%, meeting the gold standards of a finished genome. Further efforts are being taken to fill the remaining gaps by targeted sequencing. Compared to the Roslin assembly, the US-MARC assembly is less contiguous. The US-MARC assembly was based on a single castrated crossbred male ($5/8$ Yorkshire $\times 1/4$ Landrace $\times 1/8$ Duroc). Further work is needed to improve this assembly. These two new assemblies are to be annotated soon using ab initio algorithms, existing evidence, such as EST, cDNA, RNA-seq data and IsoSeq data (T. Smith and H. Liu et al., unpublished). RNA-seq short reads-based genome-guided and de novo transcriptome assemblies, such as the high-quality annotated blood transcriptome assembly presented in Chapter 4 will be useful in annotation of these reference genomes. These high-quality genome assemblies will further promote application of pigs in agriculture and biomedicine.

**Porcine Transcriptomics: Tools and Their Applications**

Transcripts derived from a genome are intermediate or end products of gene expression. Broadly speaking, all transcripts encoded by a genome as a whole are called the transcriptome (broad-sense transcriptome), including protein-coding mRNAs and non-protein-coding RNAs, such as microRNAs (miRNAs), small RNAs, and long non-coding RNAs (IncRNA). A narrow-sense transcriptome is the collection of RNA expressed in a specific cell or tissue type in a given condition at a given time. Hereafter, I will focus on the narrow-sense transcriptome because it is a building block of the broad-sense transcriptome. Transcriptomes are very dynamic, depending on various factors, including developmental stages and external conditions. For an organism, the
transcriptome plays important roles in controlling genetic information flowing from the genome to the phenome, by acting as messengers between genes and their protein products, components of translation machinery, regulators of gene expression, intercellular signal carriers and so on. The ultimate goal of transcriptomics is to determine the identity, abundance and function of each transcript in each cell or tissue type under a given condition, how they respond to internal and/or external changes and the biological effects of transcriptomic changes. Though the estimated correlation between the transcriptome and proteome is only moderate (reviewed by [18]), it is important to comprehensively study the transcriptome to better understand the genome. First, because the transcriptome can at least partially reflect the current and the immediately following states of a cell or tissue and can reveal molecular mechanisms underlying biological phenomena, especially by comparing transcriptomes under different conditions. Another reason that transcriptomes are extensively investigated is because many convenient, cost-effective, high-throughput tools are available or are being developed for transcriptomics. Over the last two decades, the technologies used for gene discovery, genome annotation and transcriptome profiling have evolved quickly from large scale EST sequencing to most recent RNA-seq and PacBio IsoSeq.

**Sanger method-based EST and full-length cDNA sequencing**

Large-scale EST sequencing came into being in the early 1990s [19]. It is a cost effective alternative to whole genome sequencing, with focus on the transcribed portion of a genome. Usually, by one-pass sequencing, reads of 200 to 800 bases are generated from randomly selected cDNA clones. These reads are useful for gene discovery, alternative splice isoform discovery, verifying predicted gene models, phylogenetics, transcriptome profiling, providing training data for \textit{ab initio} genome annotation, expression evidence for microarray probe design,
and facilitating protein discovery by proteomics approaches [20]. However, due to the high sequencing cost and limited throughput of the technology, expression analysis by EST sequencing was limited to shallow depth, and was thus considered as an only semi-quantitative approach. If the relative abundance of genes is not the main concern, using normalized or subtracted cDNA libraries can further boost the efficiency of gene discovery [21].

In pigs, the first EST sequencing effort was reported in 1994 [22], which determined the sequence of 14 pig skeletal muscle cDNAs. Since then, large scale EST sequencing for many various porcine tissues have been conducted, such as small intestine, anterior pituitary, backfat, brain, liver, reproductive tissues and so on [23-29]. By January, 2017, 1,676,489 EST sequences had been deposited to ESTdb (https://www.ncbi.nlm.nih.gov/nucest). The large number of ESTs has been further clustered to generate UniGenes. Based on the most recent statistics (https://www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=9823, Jun 19, 2012), 1,294,785 EST sequences are clustered into 50,106 UniGene clusters in the UniGene Build 42. ESTs from more than 40 categories of tissues have been sequenced. All the efforts of EST sequencing have greatly improved the discovery of pig genes, marker development, and genome annotation.

Full-length cDNA sequences are more valuable than fragmented ESTs to fully understand the fine structures, transcript diversity in pig transcriptome, and for functional studies. Full-length cDNA enriched libraries were constructed by using oligo-capping [30], vector-capping [31], SMART [32, 33] or more recently-developed TeloPrime [34] methods. To date, full-length cDNA sequences are available for 28 tissues [29].

A high throughput variant of EST sequencing is serial analysis of gene expression (SAGE) [35], which has been used to estimate relative gene expression abundance, determine differential expressed genes between different treatment conditions, and profile gene expression
patterns over development stages [36, 37]. Compared to the original EST sequencing, no cDNA clones or library construction is involved in SAGE. After synthesis of double-stranded cDNA, the cDNA is digested with enzymes, adaptors are ligated to both ends, and resulting fragments are concatenated and sequenced to get enough tags of 10 to 20 bp. The relative gene expression abundance is estimated based the count of tags after mapping the tags to the cDNA sequence of a gene.

**Gene expression microarray**

Large-scale EST sequencing and full-length cDNA sequencing for a broad range of tissues provided enormous amount of transcript sequences for design of cDNA-based microarrays and oligo probesets-based microarrays [38, 39]. This DNA microarray technology emerged in the middle of 1990s, which can simultaneously quantify the relative abundance of thousands of transcripts/genes by hybridizing labeled targets derived from transcripts to addressable probes that are fixed on a solid surface and imaging and quantifying the abundance of targets by fluorescence-based methods [38, 39]. This technology supplanted the EST sequencing and SAGE methods in many applications because it was more cost effective and higher throughput. The application of these platforms significantly sped up genome-wide gene expression profiling, identification of gene function and gene regulation, and understanding biological mechanisms that underlay a given biological phenomenon. Over the years, arrays of different profiling capacity from targeted to genome-wide were created based on the interest of researchers and the availability of knowledge about the porcine transcriptome, such as the Agilent-037880/INRA *Sus scrofa* 60K v1 Array, which was enriched with genes expressing in the immune system, muscle and adipose tissues, Affymetrix Porcine Genome Array based on UniGene Build 28 (2004), the Affymetrix PorGene-1_0-st-v1 and the Affymetrix Porcine
Snowball arrays (2010), both of which were designed based on Genome build 9, with the latter including nearly twice as many probes as the former, as well as probesets for noncoding RNAs, and the Agilent porcine gene expression microarray v1 and v2 (For more details, see review [40]). Using the Snowball platform, gene expression patterns of 65 tissues/cell types were profiled [41]. In addition, these platforms have been used to study differential gene expression of animals of different physiological states, health conditions, development stages, and genetic backgrounds. From July 21, 2006, to November 28, 2016, 355 transcription profiling data sets were deposited to ArrayExpress [42] (https://www.ebi.ac.uk/arrayexpress, EMBL EBI) using custom or commercial porcine microarrays with foci on different transcriptome targets, and profiling experiments using cross-species arrays excluded. However, microarrays are closed platforms. Probe design of microarrays depends on known knowledge about gene sequences and gene structural annotation, and, thus, it is only possible to study known genes using microarrays. In addition, this technology suffered from issues such as limited dynamics, cross-hybridization and noisy background [43].

**Semi-high throughput real time-PCR and Nanostring nCounter**

Real time PCR is usually considered as the gold standard for quantifying gene expression due to its sensitivity and accuracy and is often used to validate discoveries from high throughput assays, such as microarrays and RNA-seq (see below), due to its limited throughput and high cost. However, recent developments of the real time PCR platforms make it a useful and convenient technique for profiling a specific set of genes of interests (up to several hundred) in a large number of samples. Currently, several semi-high throughput real time PCR platforms are available, using 96-well, 384-well, and 1536-well plate formats. For example, the LightCycler® 1536 Real-Time PCR System (Roche Applied Science, Germany) can detect gene expression in
1536 reactions simultaneously [44]. The Microfluid Dynamic Arrays in different formats (Flex Six™, 48.48, 96.96 and 192.24 Dynamic Arrays12.765 and 48.770 Digital Arrays) (https://www.fluidigm.com/products/biomark-hd-system) provide very flexible choices for different high-throughput real time PCR [45-47]. The microchamber-based reaction significantly reduces the cost of consumables and the amount of biological materials needed. Single-cell high-throughput real time PCR assays using the Fluidigm arrays have been reported [48]. Fluidigm real time PCR has been applied to simultaneously detect host and pathogen gene expression, as well as miRNA and mRNA expression in pigs infected with bacteria and viruses [49, 50].

Another semi-high throughput platform for gene expression profiling is the NanoString nCounter system (http://www.nanostring.com). To detect the abundance of multiple transcripts, this system uses multiplexed probesets with two target-specific probes for each target of interest: a color-coded reporter probe for visualizing the target and a capture probe to immobilize the target-probe complex for counting [51]. Up to 800 targets, including miRNA, mRNA or lncRNA, can be detected in a single tube without enzymatic reaction or bias. The NanoString nCounter platform provides a digital readout for target abundance and its sensitivity is higher than that of microarrays and equivalent to that of real time PCR [51]. It can even handle samples of bad RNA quality, such as RNA extracted from formalin-fixed, paraffin-embedded (FFPE) tissues, and direct cell lysate, such whole blood lysate [52]. Assays (nCounter Plex®) for multiplexed single cell gene expression profiling, and for profiling hundreds of genes in multiple samples in a single tube are also available. To date, no application of Nanostring nCounter to gene expression analysis in pigs has been reported.
RNA sequencing

RNA sequencing (RNA-seq) is a more recently developed revolutionary, high-throughput, cost effective method for transcriptomics, based on next generation sequencing (NGS) technology. In one run, millions of short reads of length between 50 to 300 bases can be generated, so RNA-seq can be considered as a much higher throughput EST sequencing successor, providing nucleotide-level resolution but at much lower cost due to high throughput and multiplexing. RNA-seq suffers much less from background noise, provides a much greater dynamic range of detection of gene expression and higher reproducibility than microarray-based methods [53] and is not constrained by *a priori* knowledge of the genome or transcriptome of interest. So it can be applied to any organism with or without reference genome. The applications of this technology have been quickly expanded since its first use in profiling gene expression of yeast in 2008 [54]. Today, it has been used to determine differentially expressed genes between conditions, allele-specific expression, known and novel transcripts, novel RNA species, transcription starting sites (TSS), polyadenylation sites (PAS), alternative splicing, RNA editing, expressing variant, gene fusion, and pathways and gene networks involved in a specific biological processes [55-57]. In addition, RNA-seq has also been used to determine single-cell transcriptomes [58-60], RNA structure [61], and RNA interactions with other molecules [62-64].

Currently, RNA-seq can be implemented by using three mainstream NGS technologies: the SOLiD sequencing and Ion Torrent sequencing technologies (ThermoFisher), and the Illumina sequencing technology, with the first using a “sequencing-by-ligation” method and the other two using two different “sequencing-by-synthesis” methods (reviewed by [65]). Due to its high throughput, cost effectiveness, high accuracy and many open-source analysis tools, Illumina sequencing technology is currently the most popular for RNA-seq. A yearly updated comparison of NGS technologies and platforms is available at [http://www.molecularecologist.com/2016](http://www.molecularecologist.com/2016)
Hereafter, I will mainly focus on the Illumin RNA-seq technology and its applications. Depending on the species, the biological question of interest and the scale of projects, different Illumina platforms with different throughputs are available, from MiniSeq to Hiseq X (https://www.illumina.com/systems.html). Currently, the maximal read length is 300 bases on MiSeq series platforms. For more detailed information about all Illumina platforms, see the Illumina website (https://www.illumina.com/systems.html).

Illumina sequencing platforms can also be used to sequence miRNAs and other small RNAs, circular RNA and other long RNAs by adopting different strategies of library construction [67, 68]. Many variants of library construction methods have been developed [69-73]. Each strategy has its own advantages and disadvantages [74]. A typical procedure is as follows. First, mRNA is enriched by poly (A) capture or rRNA is selectively depleted from the total RNA before library construction for long RNA-seq. Then long RNA molecules are sheared into pieces of an average size between 100 and 600 bases by heat hydrolysis, with the actual optimal size range depending on the sequencing strategy. After that, the RNA fragments undergo reverse transcription by using random primers to obtain double stranded cDNA fragments. Subsequently, sequencing adaptors with or without barcodes are ligated to both ends and the resulting products are further linearly amplified by PCR. PCR products are size selected and purified to get sequencing-ready libraries. The libraries are loaded onto a flow cell to generate clusters for each fragment. Finally, each cluster is sequenced by using the cyclic reversible termination approach-based “sequencing-by-synthesis”.

However, RNA-seq technology is not unbiased in terms of read coverage along a given transcript or between transcripts. For long RNA-seq, the main biases include transcript length bias, GC content bias and dinucleotide frequency bias [75, 76]. For small RNA-seq, the main
bias is introduced by adaptor ligation due the sequence preference of RNA ligases [77].

Different library construction methods can bring different biases [78]. To reduce the biases introduced during library construction, several optimizations have been suggested [67, 75, 77-80].

Similar to the RNA-seq technology itself, the analysis tools and methods for RNA-seq data are still quickly evolving. Discovery of differentially expressed genes (DEG) between conditions is usually of interest to many researchers. In this paragraph, I will focus on this type of analysis. For RNA-seq-based DEG analysis, consensus on methods for count normalization and DEG discovery has not been reached yet. The actual analysis pipeline might have to be established case by case in a data-driven way [81]. However, several guideline papers on RNA-seq data analysis are worth mentioning [81-88]. A currently recommended practice for DEG analysis is as follows. First, quality of RNA-seq raw reads is checked. If necessary, sequencing adaptors and bases of low quality are trimmed. For species with a good-quality reference genome, RNA-seq reads are generally mapped to the reference genome using some splice-aware aligner, such as STAR [89], GSNAP [90], Novoalign (http://www.novocraft.com/products/novoalign/), CLC (http://www.qiagenbioinformatics.com/products/clc-genomics-workbench/), HPG aligner [88], and HISAT2 [91]. Gene- or exon-level count summary is generated by using featureCounts [92] or HTSeq [93], with the former more efficient and accurate than the latter. For species without a good-quality reference genome, it is necessary to perform de novo transcriptome assembly to get a rough reference transcriptome first. Then the RNA-seq reads are aligned to the de novo transcriptome using an aligner for DNA alignment, such as Bowtie2 [94] and HPG aligner [95]. Abundance per transcript is estimated by using expectation-maximization (EM) algorithm-based methods, such as RSEM [96], eXpress [97] and Mix² [98]. Gene-level
expression abundance can be obtained by summing the per-transcript abundance if the gene-transcript relationship is available, which is usually the case. After the read count summary at a feature (gene, transcript, or exon) level is available, the widely used R packages, such as edgeR [99], DESeq2 [100] and QuasiSeq [101], are used for a statistical analysis to identify differentially expressed genes. When the RNA-seq data includes some hidden variations that are not accounted for by known covariates, removing the hidden variations is recommended by using tools, such as svaseq [102] or RUVSeq [103].

To reconstruct full-length transcripts from RNA-seq short reads, de novo or genome-guided transcriptome assembly is essential [104]. All de novo assembly methods are based on the de Bruijn graph theory [105], although the detailed implementations are more or less different. While most methods, such as SOAPdenovo-trans [106], allow multiple k-mers for assembly, Trinity uses only 25-mers for assembly [107]. However, different theories have been adopted for genome-guided transcriptome assembly. The two most widely used assemblers are Cufflinks and StringTie [108, 109]. The former builds overlap graphs and determine the parsimony transcript isoforms that can explain the aligned reads, while the latter builds splice graphs and uses the network flow algorithm to determine the most likely transcript isoforms [108, 109]. Two genome-guided assembler which have been more recently developed but not yet widely used, which the authors claimed to be better performers are Bayesembler [110] and TransComb [111]. Florea and Salzberg [112] reviewed the details of different genome-guided transcriptome assembly approaches. Which transcriptome assembly strategy should be taken depends on the actual situation, because each has its own advantages and disadvantages [105]. Because de novo assembly is reference genome-independent, it is not misled by misassembled genome regions. However, it tends to assemble fragmented transcript pieces. On the other hand, genome-guided
assembly can take advantage of the well-assembled genome regions to assemble contiguous transcripts, but can also suffer from wrongly assembled genomic regions to report wrong transcripts. Both de novo and genome-guided transcriptome assembly strategies have difficulties in correctly assembling transcripts derived from very similar paralogs. If a high-quality reference genome is available, genome-guided assembly is usually proposed, otherwise de novo transcriptome assembly is the preferred choice or a hybrid assembly method that integrates de novo and genome-guided assemblies can be adopted if a decent reference genome is available [104, 113]. Some best practices for de novo transcriptome assembly have been suggested [114-117]. Because de novo assembly usually results in an unexpectedly large number of putative transcripts (PT), also known as “transfrags”, a significant portion of which are “transcriptional noise”, redundant PTs, and assembly artifacts. The redundancy, noise and artifacts can be filtered out based on following the following steps: (1) clustering and collapsing by using tools such as cd-hit-est [118] and EasyCluster2 [119]; (2) removing intronic unspliced PTs after aligning them to the reference genome, if possible; (3) removing sequences that are significantly similar to those of distantly related species in the NCBI NT database; (4) filtering sequences that are poorly supported by RNA-seq reads (Liu H. unpublished). Quality of the resulting transcriptomes can be evaluated in reference-free and reference-dependent ways [105, 120]. Subsequently, the large number of PTs should be functionally annotated by using tools such as BLAST2GO [121], TRAPID [122], Trinotate (https://trinotate.github.io/) and Annocript [123]. The assembled transcriptome can be used for genome annotation, gene expression analysis, expressed variant discovery, marker development, and phylogenetics and so on.

For pigs, data of 72 RNA-seq experiments on different tissue types of different breeds under different conditions have been collected in ArrayExpress [42] since 2009. These data are a
big asset for swine transcriptomics, which is still under-investigated, although the number of
tissue types used for RNA-seq is still very limited. The original interest of most of these
experiments was to identify differentially expressed genes between conditions. So far, very little
effort has been put to assemble these RNA-seq data [117, 124-128]. These studies all focused on
discovery of long noncoding RNA (lncRNA) based on genome-guided transcriptome assembly.
Given that the current pig reference genome Sscrofa10.2 is not well annotated yet and the newly
released Sscrofa11.1 is much better assembled than Sscrofa10.2, making use of these RNA-seq
data with and without assembly will be very helpful to annotate the emerging porcine reference
genome.

**IsoSeq: full-length cDNA isoform sequencing**

Due to the read length limitation of most next generation sequencing technologies except
Pacific Biosciences (PacBio) SMRT and Oxford NanoPore long read sequencing, determining
the full-length sequences of most transcripts (> 1kb) without assembly is impossible. With the
latest chemistry P6-C4, PacBio IsoSeq technology on RS II sequencing platform can sequence
full-length cDNAs more than 10kb in length with size fractionation [129]. The newly released
Sequel system from PacBio can generate nearly as 7 fold reads as the RS II platform per run at
half the cost of RS II, due to the use of the redesigned SMRT cell of 1 M zero-mode
waveguide (ZMW) wells. With modification to the solution for loading templates by adding
some additive, no size fraction is needed. Multiplexing samples for IsoSeq run on the Sequel
system resulted in no significant length bias of the output circular consensus sequences. The new
system will speed up transcriptome sequencing and reduce costs. The read length from the
Sequel system is currently shorter than that of the RS II, with an average polymerase-read-length
of 10 kb and a maximal subread length of 35 kb. Therefore, the Sequel system is good for IsoSeq
because most transcripts are shorter than 10 kb. However, due to the high error rate (~15%) of PacBio reads, error correction is necessary before the reads can be reliably mapped to the reference genome, and used for similarity search and functional annotation. Currently, two types of error correction approaches exist: (1) self-correction by using redundant PacBio long reads, and (2) hybrid correction by using Illumina RNA-seq reads.

The average length of reads generated on a MinION platform is ~ 10 kb, but reads more than 150 kb have been reported [130]. Direct RNA sequencing on a MinION platform using Oxford Nanopore technology has recently been reported [131]. By the technology, biases introduced by reverse transcription and PCR amplification are avoided; full-length, strand-specific reads are generated; and it is even possible to detect nucleotide analogues [131]. However, the error rate of reads generated by the Nanopore technology is still very high, ranging from 5 to 40% [130, 132, 133]. So far, only two studies using the Nanopore technology to determine transcript splicing isoforms of limited genes have been reported [134, 135]. With the maturation of the Nanopore technology, it is promising that in the near future the whole transcriptome of a complex organism could be directly sequenced without biases.

**Feed Efficiency in Pigs**

Feed efficiency (FE) is one of the important traits directly related to profitability, productivity, and sustainability in the pork industry [136, 137]. An increasingly popular measure of FE is residual feed intake (RFI), which is the difference between the actual and expected feed intake of an animal for given production parameters and maintenance [138]. While many non-genetic strategies have been developed to improve feed efficiency of pigs [139], improving feed efficiency by genetic selection could be a sustainable alternative, because RFI is moderately heritable and responds well to genetic selection in pigs [140, 141].
Over the last decade, significant efforts have been made to understand the molecular, genetic and physiological basis of RFI in pigs. Researchers have found many interesting differences between pigs with divergent RFI phenotypes [142, 143]. For example, compared to high RFI pigs (less efficient), low RFI pigs (more efficient) have altered feeding behaviors [144], slightly lower growth rate [145], less back fat [141, 146, 147], lower protein turnover rate in muscle [148], altered mitochondrial protein profiles [149, 150], less mitochondrial reactive oxygen species (ROS) production [151], and lower levels of leptin [152, 153] and juvenile IGF-1 in circulating blood [145]. To explore the genetic basis of RFI in pigs, several genome-wide association studies (GWAS) have been conducted [154-157]. Some chromosomal regions tagged by single nucleotide polymorphisms (SNPs) have been found to be associated with RFI, but these associations were not consistent across studies and explained only small portions of the genetic variance for RFI [154-157]. The wide range of differences between lines of pigs with divergent RFI and the lack of SNPs with major effects on RFI suggest that RFI is a highly polygenic, quantitative trait with multiple tissues contributing to its variation.

Global gene expression profiling technologies have also been used to explore the molecular basis of RFI in pigs. By profiling the transcriptomes of the adipose tissue of two lines of pigs divergently selected for RFI with gene expression microarrays, Lkhagvadorj et al. [158] found that genes involved in the lipid metabolic pathway were overrepresented among the differentially expressed genes (DEGs) that had lower expression in low versus high RFI pigs, and genes involved in carbohydrate metabolism and response to stress were overrepresented among the DEGs that had higher expression in low versus high RFI pigs [158]. They also reported that the leptin-related gene network to be different between the two lines. Recently, Vincent et al. [159] and Jing et al. [160] profiled the transcriptome of the longissimus dorsi (LD)
muscle from pigs with divergent RFI by using gene expression microarrays and RNA-seq, respectively. Using pigs from lines divergently selected for RFI, Vincent et al. [159] found that genes involved in protein synthesis and glycolysis, and genes associated with mitochondrial energy/oxidative metabolism had higher and lower expression, respectively, in the low versus high RFI lines. Using Yorkshire barrows with extreme phenotypes for RFI, Jing et al. [160] found that genes involved in glycolysis had lower expression in the low versus high RFI groups, while genes involved in muscle proliferation and differentiation had higher expression in the low versus high RFI groups. Surprisingly, these two studies shared no DEGs and proposed opposite differences in glycolytic activities in the low- versus high-RFI pigs. Therefore, in consideration of the complexity of RFI and the inconsistency from study to study, the molecular mechanisms underlying RFI in pigs are still largely unclear.

To incorporate RFI into animal breeding programs for improving feed efficiency, feed intake, body weight gain and back fat depth must be recorded on individual pigs. As it is very expensive and time-consuming to record feed intake on individual animals [140, 141], convenient and predictive biomarkers (for definition, see the next section) for RFI that can be measured at an early age are in demand. In cattle, Chen et al. [161] successfully used 14 DEGs identified in the liver of Angus bulls that were divergently selected for RFI to classify Angus steers from the same divergent RFI lines, and Al-Husseini et al. [162] developed a RFI predictor using 8 of these 14 DEGs and validated it in an independent Angus population. However, these liver biopsies for biomarker development were collected after feed efficiency test in their late life. Thus, whether these biomarkers could be used at an early growth stage to predict lifetime feed efficiency of cattle is not sure. Furthermore, these biomarkers are not very practical even if they are predictive because invasive liver biopsies are needed.
Rationale for Developing Biomarkers for Breeding Feed Efficient Pigs

Biomarkers are often used in both clinical research and in clinical practice for risk assessment, detection or screening, diagnosis, prognosis, predicting and monitoring treatment response and recurrence monitoring. Several definitions of biomarkers with different breadth have been given. A definition given by the National Institutes of Health is that a biomarker is “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” [163]. In clinic studies and practices, biomarkers may but are not necessarily correlated with the clinical endpoints, which generally refer to reduced morbidity and mortality and promote health and wellbeing. In clinical trials, “well-characterized biomarkers with well-evaluated clinical relevance” are considered to be surrogate endpoints, which “act as surrogates for clinically meaningful endpoints” [163]. Before biomarkers can be used as surrogate endpoints, their relevance and validity should be determined. That is, how relevant is a biomarker to the biological process of interest clinically? How effective and reproducible is a biomarker as a surrogate endpoint to predict the clinical endpoint? A caution that should be taken is that biomarkers as surrogate endpoints should be constantly reevaluated to reduce false positives and false negatives [164]. Mayeux gave a good review on the potential uses and limitations of biomarkers [165]. Currently, biomarkers have been suggested for breast cancer across the continuum of care from risk assessment to recurrence monitoring (https://www.provistadx.com/power-of-proteins), although very few cancer biomarkers are useful for cancer diagnosis and monitoring [166, 167].

Based on the usage of biomarkers, biomarkers are classified as prognostic biomarkers, predictive biomarkers, pharmacodynamic markers and efficacy-response biomarkers. Based on
the identities of biomarkers, one type of biomarkers is molecular biomarkers, which can further categorized into genomic biomarkers, epigenomic biomarkers, transcriptomic biomarkers, proteomic biomarkers, metabolomic biomarkers, and microbiomic biomarkers [168, 169]. Genomic biomarkers are DNA variants, which can be discovered by genome wide association studies (GWAS) and/or functional analysis [170]. Other types of molecular biomarkers are generally identified by differential analysis [171]. Based on how biomarkers are prioritized, biomarkers can be classified as gene biomarkers, gene set biomarkers, pathway biomarkers and network biomarkers. Because biological data are usually noisy, gene biomarkers are less stable and network biomarkers are most robust [172-175]. Based on the number of molecular types included, biomarkers can be classified as single-molecule-type biomarkers and multiple-molecule-type biomarker panels, with the former type in use and the latter still in a discovery stage [169]. Rafai et al. gave a review on discovery and verification of protein biomarkers, which should also be instructive for development of other types of molecular biomarkers [176].

Peripheral blood is an informative tissue, not only because it carries a variety of cells that are directly involved in immunity and inflammation, but also because it interacts with every organ and tissue in the body via bioactive circulating factors, such as nutrients, metabolites, cytokines, hormones and exosomal cargoes, which are released from the same or different organs or tissues [177]. Molecular signatures in circulating blood, including gene expression profiles, have been shown to reflect the physiopathological status, growth stage and lifestyle of subjects [177-181]. Due to its easy accessibility and informativeness, peripheral blood has become a popular sample for disease diagnosis, prediction, prognosis, and biomarker discovery [177, 178]. Interestingly, the concentration of IGF-1 in serum at a young age has been shown to differ between animals with divergent RFI in poultry and livestock, including pigs [145, 152, 182,
Blood cell profiles at early growth stages have also been found to be different between livestock with divergent RFI phenotypes, including pigs [184] and cattle [185]. In addition, the post-weaning abundance of three serum proteins, gesolin, vitronectin and serine protein inhibitor A3, are significantly higher in pigs with extremely low RFI than in pigs of extremely high RFI [186]. These results suggest that animals with divergent RFI phenotypes, measured later in life, have early physiological differences in circulating blood that may be reflected in blood gene expression profiles at these early stages.

**Dissertation Organization**

Genetic selection for low-RFI pigs can improve feed efficiency. However, it is not clear whether selection for improved feed efficiency compromises a pig’s immune response to pathogens or virulent components of pathogens. Chapter 2 of this thesis reports on results of investigating how similarly and/or differently the two ISU RFI lines responded to systemic inflammation induced by intramuscular injection of lipopolysaccharide (LPS), a typical inflammatory stimulant mimicking bacterial infection, by using RNA-seq profiling their peripheral blood transcriptome at baseline and multi-time points post LPS injection. Through this study, I found that the two lines that were divergently selected for RFI responded to the inflammation in very similar ways. No significant differences in dynamics of body temperature, complete blood count, and blood cytokine levels were detected between the two lines at baseline and post LPS injection. Only a very small numbers of DEGs were detected at baseline, and at 2, 6 and 24 hours post LPS injection (hpi) between the two lines. Powerful and sensitive gene set enrichment analysis suggested pigs from the low-RFI line generally had a lower level of inflammation response during the time course. Small differences in other pathways and biological processes were also identified. I further systematically compared response to LPS-
induced systemic inflammation in pigs to that of humans. I found pigs and humans shared a core set of more than 1,500 orthologous genes, many of which showed similar expression patterns in response to LPS stimulation and mainly function in immune response and inflammatory response, although pigs are more tolerant to LPS than humans.

In addition, it is very expensive and time-consuming to implement RFI-based genetic selection [140, 141]. As an alternative way, identifying convenient and predictive biomarkers for RFI that can be measured at an early age is of great interest. Thus, in this thesis, post-weaning blood transcriptomic differences between Yorkshire pigs from two lines that were divergently selected for RFI at ISU were profiled by RNA-seq. The ultimate goal was to develop practical biomarkers for feed efficiency based on the revealed transcriptomic differences in peripheral blood. Differentially expressed genes (DEGs) and gene clusters between post-weaning animal groups with extremely low RFI and with extremely high RFI were identified and annotated with gene ontology terms and pathways. These DEGs significantly overlapped expression signatures of human diseases, such as eating disorders, hyperphagia and mitochondrial diseases. I also identified five candidate DEGs whose expression level were significantly associated with RFI values. These five genes could be potential biomarkers for feed efficiency. This work is presented in Chapter 3.

When I worked on projects presented in Chapter 2 and 3, I realized the limitations and issues of the assembly and annotation of the current pig reference genome, *S. scrofa* 10.2. So I decided to improve the pig genome annotation by assembling a high-quality transcriptome for porcine peripheral blood with public RNA-seq data and RNA-seq data from studies described in Chapter 2 and 3. I built, validated and annotated a comprehensive porcine blood transcriptome, composed of more than 130,000 putative transcripts, with significant improvement over the
annotation of Ensembl *Sscrofa10.2* and the NCBI pig RefSeq mRNA collection. This work is presented as Chapter 4.

In Chapter 5, general conclusions were drawn based on the studies presented in Chapters 2 to 4; limitations of the current methodology for whole blood transcriptome profiling were discussed; perspectives of single cell-based blood transcriptomics were put forward with.

Supplementary methods, figures and tables supporting Chapters 2, 3, and 4 were provided as Appendices A, B and C, respectively, where appropriate. Large Excel spread sheets were provided as extra files hyperlinked to the texts for easy page layout.

In summary, this dissertation provides insight into the potential molecular mechanisms underlying feed efficiency in pigs and identified five candidate genes for developing biomarkers for swine feed efficiency in the future. It also provides evidence that selection for RFI has not compromised the pigs’ immune response to acute systemic inflammation. In addition, it offers significant advances to the structural and functional annotation of the porcine blood transcriptome.
CHAPTER 2. PIGS DIVERGENTLY SELECTED FOR RESIDUAL FEED INTAKE RESPOND TO SYSTEMIC INFLAMMATION TRIGGERED BY LIPOPOLYSACCHARIDE IN LARGE SIMILAR WAYS AND SHARE A LARGE SET OF CORE INFLAMMATORY RESPONSE GENES WITH HUMANS

A paper to be submitted to *BMC Genomics*

Haibo Liu, Kristina M. Feye, Yet T. Nguyen, Anoosh Rakhshandeh, Dan Nettleton, Crystal Loving, Jack C.M. Dekkers, Nicholas K. Gabler, Christopher K. Tuggle

Abstract

**Background:** Breeding pig for improved feed efficiency is of practical interest. However, it is not clear whether selection for improved feed efficiency by selection for low residual feed intake (RFI), a measure of feed efficiency, compromises a pig’s immune response to pathogens or inflammatory stimuli. In this study, we intramuscularly injected pigs from two lines divergently selected for RFI with lipopolysaccharide (LPS) to investigate whether these two lines responded to LPS stimulation differently. And we also aimed to determine to what extent transcriptomic response to endotoxemia in pigs and humans are similar.

**Results:** Intramuscular injection of LPS induced systemic inflammation in pigs of both lines \((n = 4\) per line), but no significant differences in dynamics of body temperature, complete blood count (CBC), and levels of several cytokines in plasma were detected between the lines during the first 24 hours post injection (hpi). Profiling peripheral blood transcriptomes at baseline and at 2, 6, and 24 hpi by RNA-seq revealed that LPS induced dramatic transcriptome-wide changes, with 5,871 genes differentially expressed for at least one time point post LPS injection compared to baseline for the average line \((|\log_2(FC)| \geq \log_2(1.2), q < 0.05)\). However,
we detected only a very small number of genes that were differentially expressed between the two lines at baseline, and at 2, 6, and 24 hpi, with negligible line-by-time interaction effects on gene expression. Gene set enrichment analyses suggested that at the pathway level the two lines were largely similar in transcriptomic response to LPS stimulation, although the low RFI line had a slightly lower level of inflammatory response than the high RFI line at each time point post LPS injection. Genes involved in protein translation, defense and immune response, and signaling formed individual co-expression clusters. Compared to the inflammatory response induced by intravenous LPS infusion in humans, swine inflammatory response observed in this pig study was much weaker, although a much higher dosage of LPS was used in pigs than in humans. While hundreds of key pathways and biological processes were shared between the two species during LPS response, any pathways and biological processes were specifically involved in human response, but only a few were pig-specific. In total, 1,543 porcine genes and their human orthologs were differentially expressed in response to LPS stimulation in both species. Nearly one third of these genes shared a very similar dynamic expression pattern over the time course in these species. These conserved genes mainly functioned in biological processes including translation, RNA catabolism, signaling, apoptosis, immune response and inflammatory response.

**Conclusions:** The pig lines divergently selected for RFI responded very similarly to systemic inflammation triggered by intramuscular LPS injection, though the low-RFI line had a lower level of inflammation response than the high-RFI line at every time point post LPS injection. Pigs and humans shared similarity in response to systematic inflammation triggered by LPS at both the gene and pathway levels, although the response to LPS in pigs was much weaker than previously reported in humans. But caution should be taken to interpret the differences seen
in humans and pigs due to differences in routes of LPS administration, dosage and sources of LPS, blood fractionation, expression profiling platforms, and completeness of genome annotation in both species. Our work indicates that selection for feed efficient pigs does not compromise a pig’s ability to handle acute systemic inflammation, and also offered another line of evidence that pigs can be used as models for human inflammatory disease.

**Keywords:** *Sus scrofa*, lipopolysaccharide, systemic inflammation, residual feed intake, RNA-seq

**Background**

Feed efficiency in pigs is a trait of economic, environmental and societal importance. One increasingly accepted metric of feed efficiency is residual feed intake (RFI), the difference of an individual animal’s observed and expected feed intake for growth and maintenance [138]. That is, pigs with low RFI are more feed efficient than those with high RFI. Pilot studies of divergent selection for RFI in pigs showed that RFI responds well to genetic selection [141, 187, 188].

Compared to high RFI pigs, low RFI pigs have reduced feed intake, but their growth rate is only slightly decreased [141, 187, 188]. So they are more efficient in allocating resources for production and maintenance. On the other hand, immune response is an energy-demanding biological process and directly related to pig health. Thus, one interesting question is whether improving feed efficiency by selection for low RFI affects the animal’s ability to respond to immune challenges. Based on the resource allocation theory [189], selection for low RFI is expected to compromise the animal’s capacity to handle immune stimulation, such as infectious
diseases. Importantly, selection of chicken and beef cattle for increased feed efficiency indeed negatively affected their immune system [189]. So far, several experiments have investigated the potential side effects of selection for divergent RFI phenotypes on immune response in pigs. First, based on results from an experimental infection of porcine reproductive and respiratory syndrome virus (PRRSv) on pigs from the two ISU RFI lines, Dunkelberger et al. [190] reported that low-RFI pigs tended to have lower levels of viral load in blood, a faster humoral immune response to PRRSv, and a less affected growth rate than high-RFI pigs. Second, to test the divergent RFI lines developed at INRA, piglets from both lines were challenged with Complete Freund’s Adjuvant (CFA) to induce a non-infectious pneumonia [191-194]. This work showed that both RFI lines handled the inflammatory challenge similarly but by adopting different metabolic strategies [191-194]. Finally, Vigros et al. [195] examined expression profiles of a set of target genes related to intestinal immunity of pigs with extremely divergent RFI phenotypes before and after an ex vivo LPS challenge of ileal and colonic tissue explants. No differentially expressed genes were found at the basal state. However, they found that the mRNA expression level of several cytokines (IL8, IL1, IL6, TNFα, IFNγ) and SOCS3 were lower in the low-RFI than the high-RFI explants following LPS challenge [195]. They proposed that low-RFI pigs may exert an energy saving mechanism during intestinal innate immune response to an immune challenge [195]. Taken together, although no significant negative side effects of a selection for lower RFI on immune response in pigs have been detected, little is known about the global transcriptomic response to acute systemic inflammation in pigs divergently selected for feed efficiency.

LPS is a major component of the outer membrane of gram-negative bacteria, which has been widely used in vertebrates to mimic bacterial infection. LPS can bind to TLR4 and its
associated factor MD2 with the help of LBP and CD14. MD2 facilitates dimerization of the LPS/TLR4/MD2 complex, which induces the assembly of myddosomes including MyD88 and several IRAK kinases with mediation of TIRAP [196-199]. The myddosomes activate transcription factors NFκB and AP1, leading to expression of pro-inflammatory genes, such as IL1, IL6 and IL10, in a MyD88-dependent manner. However, cellular response to LPS is predominantly mediated by MyD88-independent signaling [200], where the TLR4/MD2/LPS complex is internalized via endocytosis with the engagement of CD14 and other cell type-specific cofactors and then endosomal TLR4 activates IRF3-dependent production of Type I interferon, TNFα and co-stimulatory molecules through adaptor proteins TRAM and TRIF [201, 202]. Binding of secreted TNFα its receptor can also activate NFκB. TLR4-independent host response to LPS has also been identified [203], and cytosolic LPS can activate non-canonical inflammatory caspases to trigger pyroptosis [204].

Comparative studies have revealed that TLR4 in different mammalian species differs in affinity and specificity to its ligands, gene structure, cellular expression patterns and tissue distribution [205]. TLR4 genes are highly conserved across mammalian species, with the intracellular TIR domain most conserved and the ligand-recognition domain most divergent, which, along with divergent MD2 proteins, is believed to cause the TLR4 receptors in each species to have different affinity and specificity to their ligands [205]. Porcine TLR4 is more similar to human TLR4 than is the murine counterpart at the nucleotide level, including both the promoter and coding regions, and as well as at the protein sequence level [206-208].

Further, different mammalian species differ in their sensitivity to LPS. Among the studied mammals, humans and chimpanzees are most sensitive to LPS, while pigs and rabbits are more sensitive to LPS than mice and rats [205]. Previous comparative studies of murine and
human transcriptomic response to in vivo LPS treatment raised an unresolved debate whether mice are good models for human inflammatory diseases [209-212]. Given that pigs are more sensitive to LPS than mice, the porcine immunome is more similar to that of humans than that of mice [16], and that pigs share more similar physiological and genomic similarities with human than mice [213], it is of interest to study the porcine transcriptomic response to systemic inflammation induced by LPS, using a similar experimental design as reported in a human study [214] and compare the results between these species. So far, dozens of in vivo LPS challenge experiments have been performed in pigs (see review [40, 215]), but in most cases only one time point was investigated. An exception is one study [216], which reported a number of gene targets that were evaluated by RT-qPCR rather than a broad transcriptomics study [215]. In addition, several in vitro LPS challenge studies have been performed on porcine macrophage or PBMCs using gene expression microarray, mass spectrometry, or RT-qPCR [217-221]. Limited data support that porcine bone-marrow macrophages are very similar to the humans in their response to LPS [217]. However, how similarly pigs and human respond in vivo to systemic inflammation induced by LPS challenge has not been studied well.

The objectives of this study were twofold: (1) to determine whether divergent selection for RFI has significantly affected the pig’s immune response to mimic bacterial infection, (2) to determine whether pigs and humans share similar transcriptomic response to endotoxemia. Thus, we investigated the swine peripheral blood transcriptomic response to acute systemic inflammation induced by intramuscularly injection of LPS in the two ISU lines that were divergently selected for RFI by using RNA-seq of a time series of samples mimicking a previous human study [214]. Based on clinical observation and differential expression analysis, we detected a few minor differences between the two RFI lines in response to LPS stimulation.
Differential expression analysis, gene expression clustering analysis and gene set enrichment analyses suggested that humans and pigs share similar transcriptomic responses to LPS challenge at both the gene level and the pathway level, although more biological processes and pathways were involved in systemic inflammation in humans than in pigs. Our work indicates that feed efficient pigs are resilient to an acute inflammatory challenge, and also offered another line of evidence that pigs can be used as models for human inflammatory disease.

**Methods**

**Animals, experimental design and sample collection**

All animal-involved protocols were approved by the Institutional Animal Care and Use Committee (IACUC) at Iowa State University under permit number 11-1-4996-S. All pigs were from Generation 8 residual feed intake selection project at ISU [141, 154, 184] and the experiment was conducted in the summer of 2010.

Fourteen ($n = 7$ per line) gilts of initial body weight (BW) of $63 \pm 4$ kg from two lines of Yorkshire pigs divergently selected for low RFI and high RFI were randomly selected and utilized for either of two replicates. For each replicate, the following procedures were performed. All pigs were housed in individual metabolism pens, had free access to water and were fed a corn-soy-based diet twice daily, and feed restricted (1.5 kg/day) as previously described [222]. After a 9-day adaptation period pigs were randomly assigned within line to either a control ($n = 6$, three pigs per line) or LPS challenge ($n = 8$, four pigs per line) group. Pigs in the challenge group were then repeatedly challenged with an intramuscular injection of 30, 36, 39, and 42 $\mu$g/kg BW of LPS from *E. coli* O5:B55 (Sigma-Aldrich, St. Louis, MO) dissolved in saline solution at time points 0 (baseline), 48, 96, and 144 hours post initial injection (hpi). At each subsequent time point following the initial injection, LPS dosage was increased by 20, 30 and
40% from initial dose to compensate for LPS tolerance. Pigs in the control group were injected with an equivalent volume of saline solution at the equivalent time points. Rectal temperatures of individual pigs were measured immediately before the initial injection and at 2, 4, 6, and 24 hpi, and at 0, 4, 24 hours after each subsequent injection for all pigs. Blood samples were collected from the jugular vein into Tempus™ Blood RNA tubes (Life Technologies, Grand Island, NY) for long-term storage at -80 °C, into EDTA tubes (BD, Franklin Lakes, NJ) for CBC tests and into serum tubes for cytokine assays at 0, 2, 6, 24 and 168 hpi. At 168 hpi, all pigs were euthanized by barbiturate overdose, exsanguinated and tissue samples from the longissimus dorsi (LD) muscle, liver, spleen, and ileum were isolated, cleaned and frozen for later use. This study only focused on samples and data collected for the first 24 hours after initial injection to investigate the acute inflammatory response.

**RNA preparation**

Total RNA was extracted from the 32 blood samples of the treatment group in Replicate 2, collected at 0, 2, 6, and 24 hpi, by using preserved blood RNA purification kit I (Norgen Biotek Corp, Thorold, Ontario) per manufacturer’s instruction. On-column DNA digestion was performed as described using DNase I (Qiagen, Valencia, CA). Globin transcripts (HBB and HBA) were depleted by following an RNase H-mediated method [223]. The quantity and integrity of the RNA were monitored by using Nanodrop 2000 (Thermo Scientific, Waltham, MA) and Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA) before and after globin depletion [223]. The efficiency of globin depletion of each sample was checked by conventional RT-qPCR with β-actin and GAPDH as the internal controls. Globin depletion efficiencies for all RNA samples were above 85%. Metadata, including RINs and concentration of RNA post globin depletion, CBC, and sequencing batches are available in Table A.1.
Complete blood count (CBC) test and blood cytokine analysis

CBC tests were conducted as described [184, 224]. Eight cytokines (IFNα, IFNγ, IL1B, IL4, IL6, IL8, IL10 and TNFα) in the plasma sampled at 0, 2, 6, 24 and 168 hpi were assayed by using Aushon Ciraplex assays for cytokines (Billerica, MA) per manufacturer’s instruction.

Verification of transcriptional response to LPS stimulation by RT-qPCR

To verify the LPS stimulation did induce an acute inflammatory response in the pigs, expression profiles of 36 candidate genes (Table A.2), which were porcine orthologs of human and murine genes responsive to LPS stimulation or important for Gram-negative sepsis control and resolution, were assessed by using Fluidigm high throughput RT-qPCR. Briefly, primers were designed and synthesized by Fluidigm Corporation (San Francisco, CA) such that two primers of each pair were separated by exon-exon boundaries and could amplify all known isoforms of a target gene if possible. The specificities of the primer pairs were tested by conventional RT-qPCR on the DNA Engine Opticon 2 system (BioRad, Hercules, California) and only primer pairs that gave single peaks in melting curve analyses were kept. This resulted in 36 and 2 primer pairs for genes of interest and internal controls (RPL32 and GAPDH) meeting the requirement, respectively. The 32 RNA samples prepared from blood samples of the treatment group in Replicate 2 collected at 0, 2, 6, and 24 hpi were used for Fluidigm RT-qPCR without globin depletion. By following the Fluidigm user guide for Real-Time PCR analysis [225], real time-qPCR was done on a 48.48 dynamic array chip (Fisher Scientific, Pittsburgh, PA), using the Biomarker HD system (Fluidigm, San Francisco, CA). Data were analyzed with the Fluidigm Real-Time PCR analysis software with the default settings, to obtain raw C\(\text{t}\) (cycle of threshold) values. Since the expression levels of internal controls were not very stable from time point to time point for individual pigs, RT-qPCR data were analyzed by using the R
package “MCMC.qpcr” [226]. With this method, internal reference genes are not mandatory but can be incorporated as Bayesian priors or as trackers of global effects when template abundances correlate with conditions [226]. Briefly, \( C_t \) values were converted into the copy number of templates by incorporating the amplification efficiencies of the primers and then analyzed by using generalized linear mixed models, which assumed the copy numbers of transcripts of given genes follow lognormal-Poisson distributions. No line difference was detected for any time point post treatment compared to baseline for any selected gene, except for CXCL13. In the generalized linear mixed effects models, time or time and line (only for CXCL13) effects, and plate effects were modeled as fixed effects, and individual animal effects were treated as random effects, with GAPDH and RPL32 as internal controls by allowing 1.2 fold change across time points. The \( p \)-values associated with the line and time effects were adjusted by using the Benjamini-Hochberg (“BH”) method [227].

**RNA-sequencing**

The 32 RNA samples used for RT-qPCR assays as above were used for RNA-seq. Library construction and sequencing were performed by Beijing Genomics Institute (BGI, Hongkong). Briefly, the RNA-seq libraries were constructed using the Illumina TruSeq RNA Sample Preparation Kit v2 (Illumina, San Diego, CA) per manufacturer’s instructions. Individual libraries were diluted to 2 nM and pooled in approximately equimolar amounts with 8 libraries per pool. Paired-end sequencing (2 × 50 base) was run on an Illumina HiSeq2000 platform with one pool per lane. RNA-seq data will be available in ArrayExpress under accession number: [E-MTAB-5606, http://https://www.ebi.ac.uk/arrayexpress/].
Quality control, preprocessing and alignment of RNA-seq reads

Read quality was checked and filtered by BGI using their custom scripts. For a pair of reads, the whole pair was removed if either read met the following criteria: (1) either read had more than 50% of their bases aligned to the adapter sequences; (2) either read contained more than 10% of ‘N’ bases; (3) either read had more than 40% of bases with PHRED+64 quality scores lower than 20. The kept reads were aligned to the pig reference genome *Sscrofa* 10.2 (version 77, Ensembl) using 2-pass rna-STAR (version 2.3.0) with the default settings [228, 229]. Read counts per gene per library were summarized by using featureCounts (version 1.4.4) [92] with default settings and the pig genome GTF file (version 77, Ensembl) as the genomic annotation reference file. Prior to differential expression analysis, hemoglobin genes (HBA and HBB) and genes with few reads (mean read counts per gene across all 32 samples ≤ 8 or the number of samples with non-zero count per gene ≤ 3) were removed from the count table, resulting in a final count table of 11,911 genes. This count table was used for subsequent differential expression analysis and clustering analysis after further transformation and adjustment (see below).

Differential expression analysis of pig RNA-seq data

Because there is no generally accepted, applicable method to analyze RNAseq data for a short time series experiment in which within-individual measures are generally correlated, we used the R/Bioconductor package DESeq2 for differential expression analysis for two reasons: (1) DESeq2 adopts emperical bayes shrinkage estimation for dispersions and fold changes which improves stability and interpretability of estimates; (2) DESeq2 allows statistical tests of differential expression with a specified minimum effect size, which avoids the issue that *post hoc* filtering of differentially expressed genes (DEGs) based on a fold change threshold results in a
false discovery rate (FDR) that is not easy to interpret [100]. Briefly, the pig RNA-seq raw count table was first normalized by using the default normalization method of DESeq2 and then rounded to get a normalized count table. Although several nuisance variables, such as RNA preparation batch, RIN, body weight, and complete blood cell count were recorded, it was very difficult to account for the treatment-unrelated variations of the complicated blood transcriptome by using only those nuisance variables, several of which were not accurately measured. So we used six surrogate variables estimated by the svaseq function of the R/Bioconductor package sva (v3.20.0) [102] to account for the hidden, unwanted variations. The six surrogate variables were estimated by using a full model, including an intercept and effects for combinations of the RFI lines and sampling time points, and a reduced model, including only an intercept.

Differential expression analyses were conducted as follows. After performing normalization of the raw count data as above, dispersions were estimated by using a parametric method with maximum number of iterations of 500. log-transformed normalized count per gene was modelled by using a generalized linear model with the combinations of the RFI lines and sampling time points and the six estimated surrogate variables as covariates. The \textit{nbinomWaldTest} function was used to estimate and test the significance of coefficients of the covairates with the following explicit parameter settings: \textit{betaPrior}=FALSE, \textit{maxit} = 5000, \textit{useOptim} = TRUE, \textit{useT} = FALSE, \textit{useQR} = TRUE. Differentially expressed genes between conditions were identified by testing the significance of relavent contrasts by using the \textit{results} function with the following explicit parameters specified: \textit{alpha} = 0.01, \textit{contrast} = \textit{contrast.matrix}[i,], \textit{lfcThreshold} = \textit{log2}(1.2), \textit{altHypothesis} = "greaterAbs", that is, testing the absolute values of the \textit{log2} fold change between conditions were greater than \textit{log2}(1.2). Thus, the estimates of the fold changes were shrunken by performing empirical Bayes shrinkage and their
significance were tested by specifying a minimum effect size, which can improve stability and interpretability of the estimates [100]. Multiple testing correction was performed by using the “BH” method [227]. Genes with absolute values of the log_2 fold change greater than log_2(1.2) and q values less than 0.05 were considered as significant DEGs.

**Statistical analysis of data for body temperature, CBC and cytokine profiles**

Rectal temperature data were analyzed by using a general linear model, with temperature modelled by time, treatment and time-by-treatment interactions, and assuming an AR(1) correlation structure for repeated measures on the same individuals, since line and replicate effect were not significant (p > 0.1). Because the concentrations of some cytokines at some time points were zeros, a small constant was added to all concentrations of cytokines to make log-transforming possible. After log-transformation, the transformed cytokine and CBC data were analyzed by using SAS PROC MIX by assuming an unstructured (UN) covariance and heterogeneous variance compound symmetry (CSH) covariance matrices, separately. Starting with full models including line, treatment, time, two-way and three-way interaction between line, treatment and time, plate (not included for CBC response), cohortRoom, body weight before initial injection, age upon initial injection, with time specified as repeated measure, backward model selection was performed based on the Akaike information criterion (AIC). Models of the smallest AIC were preferred.

**Reanalysis of microarray data for human transcriptomic response to systemic inflammation**

Microarray data from profiling transcriptomic response in human peripheral blood to LPS stimulation and its metadata [214] were downloaded from [http://www.gluegrant.org/pubsupport/Nature_1/](http://www.gluegrant.org/pubsupport/Nature_1/). For details of the experimental design, see [214]. Briefly, eight humans
were assigned to two treatment groups, such that each group was assigned one female and 3 males. The treatment group was intravenously injected with 2 ng/kg BW of LPS dissolved in 0.9% NaCl, while the control group was injected with an equivalent volume of 0.9% NaCl saline. Blood samples were collected immediately before injection (called “0 hpi” for convenience) and at 2, 4, 6, 9 and 24 hpi. Transcriptomic responses were profiled by using 3' IVT expression analysis arrays (Human Genome U133 Set, Affymetrix, CA), which contains 44,924 probesets on the two sub-arrays (Hu133A and Hu133B) arrays. The latest annotation files (Release 36, July, 2016) for the array platform were downloaded from Affymetrix website (http://www.affymetrix.com/support/technical/annotationfilesmain.affx). Microarray .CEL files from the two sub-arrays were separately processed with DNA-Chip Analyzer (dChip, the latest build on July 19, 2010) [230, 231] to get the normalized signal intensities per probeset by model-based expression calculation with 5th percentile of PM signals as background subtraction and quantile normalization. Probesets identified as absent by dChip [230, 231] on more than 43 arrays were excluded from further analysis. The expression values for 100 probesets shared by the two sub-arrays were averaged for each sample and the original measures for these probesets were replaced with their averages for each sample when the normalized expression data from the sub-arrays were combined. Because no further metadata, such as RNA processing batch and array hybridization batch, were available [214], 8 surrogate variables accounting for hidden batch effects and other unwanted variation were identified by using the R/Bioconductor package sva [232] and considered in the linear models for differential expression analysis. Probesets showing significant differential expression with $|\log_2(\text{fold change})|$ greater than $\log_2(1.2)$ ($q < 0.05$) at each time point after LPS injection were identified by using the R/Bioconductor limma.
package [233] with time 0 as reference. Multiple testing correction was performed by using the “BH” method as above.

**Gene ontology (GO) term overrepresentation analysis**

GO annotation for pig genes was downloaded from Ensembl BioMart (Release 87)GO terms for human genes were extracted from the latest annotation files for Human Genome U133 Set provided by Affymetrix (Release 36, July, 2016). GO terms associated with less than 10 genes were excluded. For human microarray and pig RNA-seq data, 13,734 Entrez Genes IDs and 11,911 Ensembl Gene IDs with detectable expression in blood were used as reference lists, respectively. Hypergeometric tests of overrepresentation of GO biological process (GO-BP) and GO cellular component (GO-CC) terms by a given gene list were performed separately by using the Cytoscape v3.4.0 [234] package BiNGO (v3.0.3) [235] with a q value cutoff of 0.05. GO annotation was visualized using REViGO [236].

**STEM-based clustering of gene expression profiles**

Both pig RNA-seq data and human microarray data were clustered by using STEM (Short Time-series Expression Miner, v1.3.8) [237]. For RNA-seq data, the DESeq2 normalized count data were transformed to log (cpm) by using the voom function of the limma package [233, 238] with the model used for RNA-seq DEG analysis. The transformed gene expression levels were then adjusted for the nuisance parameters, i.e. the six surrogate variables. The adjusted log(cpm) for all individuals of the two lines were altogether used for clustering with STEM, which is a tool specifically developed for short time-series expression data mining [237]. For parameter settings, see Supplementary Methods. For human microarray data, preprocessed probeset-level expression data used for DE analysis were adjusted for the eight surrogate variables. The resulting expression values were used for clustering analysis by using STEM
Expression profiles with more genes assigned than expected ($q < 0.05$) were considered as significant profiles. Significant expression profiles were clustered if their profile correlation coefficients were no less than 0.6. GO term overrepresentation analyses and visualization were performed for genes in each cluster of profiles with a $q$ value cutoff of 0.05 by using BINGO (v3.0.3) [235] in Cytoscape (v3.4.0) [234], as above.

**Gene set enrichment analysis**

Gene set enrichment analysis for both the pig blood RNA-seq and human blood microarray data were performed using the R/Bioconductor (v3.3.1) GAGE package (v2.22.0), which implements a gene permutation-based algorithm for gene set and pathway enrichment analysis[239]. Expression data adjusted for surrogate variables were used as input. Hallmark gene sets (H), curated gene sets (C2, including CGP: chemical and genetic perturbations and CP: canonical pathways, with KEGG pathways excluded), transcription factor targets (C3:TFT) and immunologic signatures (C7) were downloaded from the Molecular Signature Database (MSigDB v5.2, [http://software.broadinstitute.org/gsea/msigdb](http://software.broadinstitute.org/gsea/msigdb)) [240, 241], the latest species-specific KEGG pathways and diseases, and GO terms-associated genes from the KEGG database (Release 81.0) and Ensembl Biomart (Release 87). Assuming porcine and human genes that have the same HGNC gene symbols share the same gene functions and belong to the same gene sets, gene sets from the MSigDB represented by HGNC gene symbols were used for analysis of the pig data, except the KEGG gene sets and GO gene sets. GAGE was run in a paired-comparison mode when compared sample were from the same individuals, otherwise it was run in a “group comparison” mode. For gene sets composed of only up- or down-regulated genes, tests for either direction of change was conducted for gene set composed of a mixture of up- and down-
regulated genes, tests unaware of change directions were performed. The $q$ value cutoff of significant enrichment was set at 0.05.

**Other statistical methods**

Surrogate variable-adjusted expression level of differential expressed genes or probesets were hierarchically clustered using the Ward.2 method with 1 minus Spearman correlation coefficient as distance to generate heatmaps using the gplots package (v3.0.1). Multidimensional scaling plots showing samples relationships were made by using the plotMDS function of the edgeR package (v3.14.0) [242]. 3D-PCA plots were generated using the R package pca3d (v0.8) (Weiner J. unpublished).

**Results**

**Intramuscular injection of LPS causes similar systemic inflammatory response in RFI lines**

We intramuscularly injected pigs from the two lines divergently selected for RFI with LPS or saline to investigate whether divergent selection for RFI for multiple generations affects response to systemic inflammation over a 24-hour time course. We did not detect significant line or replicate difference in rectal temperature profile ($p > 0.1$), but the effects of time, treatment, and time-by-treatment interactions were all significant ($p < 0.0001$). The average rectal temperatures of pigs challenged with LPS significantly increased at 2 hpi, peaked at 4 hpi, and then decreased and almost returned to baseline at 24 hpi. During the time course, the overall rectal temperature of the low-RFI animals tended to be lower than that of the high-RFI animals except at 24 hpi when the tendency reversed (Figure 2.1A). The maximal increase of mean rectal temperatures was respectively one and more than two degrees for the low- and high-RFI pigs treated with LPS compared to their baseline temperatures. On the other hand, the body
temperature of pigs in the control group only slightly fluctuated during the time course (Figure 2.1A). Increased rectal temperature indicates that intramuscular injection of LPS induced fever in pigs, which was one of the typical symptoms in systemic inflammation response.

The CBC profiles significantly changed in animals challenged with LPS during the time course, but no significant effects were detected for line, replicate, or line-by-time interaction (Fig 2.1B and Table A.3). Specific CBC parameters, including red blood cell count, hemoglobin and hematocrit decreased from 2 to 24 hpi in the control group, but their decrease were only seen by 24 hpi in the treatment group. The concentrations of platelets and white blood cells (WBCs) did not change in the control group except that of neutrophils, which slightly increased at 2hpi, then slightly decreased at 6 hpi and went back to baseline at 24 hpi. For the treatment group, the concentration of platelets and WBCs decreased at 2 hpi in response to the LPS challenge. The decreased concentration of neutrophils rebounded to baseline by 6 hpi and further increased at 24 hpi. The profile of eosinophils was very similar to that of neutrophils, but concentration of eosinophils rebounded more slowly by 24 hpi. The profiles of lymphocytes, basophils, monocytes and platelets were similar to each other. They all started rebounding slowly at 6 hpi, and were nearly on the halfway back to baseline by 24 hpi. Summary statistics of comparing the concentrations of the five major types of WBCs in peripheral blood at 2 and 24 hpi to that at baseline in the control and treatment groups are shown in Table A.3. The dynamic profiles of WBCs during this time course further indicated systemic inflammation was induced by LPS in pigs.

Another typical metric of inflammation is the levels of cytokines in the bloodstream. Response profiles of eight cytokines (IL1B, IL4, IL6, IL8, IL10, TNFα, IFNα and IFNγ) are shown in Figure 2.1C. Based on the measured cytokine level of high variations, no significant
effects were detected for line and line-by-time interaction. But animals treated with LPS tended to have higher cytokine levels in their blood at 2 hpi compared to their baseline level and compared to the control group at the equivalent time point. In the treatment group, the levels of IL6, IL1B and TNFα were generally still higher at 6 hpi than baseline and returned to baseline by at 24 hpi while the level of other cytokine almost went back to baseline by at 6 hpi. Summary statistics of comparing the concentrations of the eight cytokines in peripheral blood at 2 and 24 hpi to baseline in each experimental group are shown in Table A.4.

In addition, we assessed the changes in the transcript abundance of 36 inflammatory genes by RT-qPCR to verify the transcriptomic response triggered by LPS. The expression levels of 24 genes were significantly different from baseline levels for at least one time point after LPS challenge in pigs of the treatment group ($q < 0.05$) (Figure 2.2). Another two genes, TLR2 and TLR7, tended to be differentially expressed ($q < 0.09$) due to treatment. No line-by-time interaction effects were detected for any gene assayed. A line difference in gene expression was detected only for the CXCL13 gene (maximal $\log_2(FC) = 1.6$, $p = 0.004$). In summary, genes encoding nine cytokines/chemokines (IL1B, IL12A, IL15A, CCL20, CXCL2, CXCL8, CXCL10, CXCL13, and CCL20), seven receptors (TLR4, CD11B, CD14, CD97, CCR4, TNFRSF1A and IL1R2), and eight TLR4/NFκB signaling pathway components or effectors (IRF3, IRAKM, IKBNS, STAT4, S100A9, SOD2, CASP1 and IDO1) were significantly up-regulated in response to LPS stimulation for at least one time point (Figure 2.2). mRNA levels of proinflammatory cytokines and chemokines, including IL1B, IL12A, CXCL2, CXCL8, CXCL10 and CCL20, peaked at 2 hpi, while mRNA levels of anti-inflammatory cytokines IL1R2, TNFSF1A and IL10 peaked at 6 hpi. mRNA levels of several genes involved in initiating innate immune response, including TLR4, IRAK4, IRAKM, IRF3, NFκB, RELA and STAT1, and those involved in
limiting the innate immune response, such as IKBNS, reached or nearly reached their zeniths by 6 hpi. Thus, the RT-qPCR data indicated that the pigs did undergo typical inflammation response triggered by LPS, which progressed from an acute proinflammatory phase to an anti-inflammatory phase towards full recovery.

**RNA-seq reveals systematic transcriptomic response to LPS stimulation in pigs but only small differences between RFI lines**

We profiled the peripheral blood transcriptomes of pigs from the two RFI lines ($n = 4$ per line) in the LPS treatment group at 0, 2, 6 and 24 hpi using RNA-seq. On average, more than $19.4 \pm 2.5$ million (mean ± standard deviation) pairs of $2 \times 50$-base reads per library were obtained and more than $80.3 \pm 0.5\%$ of reads were uniquely mapped to the pig reference genome. After filtering genes of very low expression and hemoglobin genes (HBA and HBB), we had 11,911 genes whose expression level met the minimal abundance requirement for downstream analyses. Principal component analysis suggested that samples were well-separated by time, and samples of given time points were separated by line (Figure A.1). It also indicated that transcriptomes at 24 hpi were closer to those at baseline than those at 2 and 6 hpi, (Figure A.1).

We only detected 22, 32, 55 and 173 genes differentially expressed between the two RFI lines at 0 and 2, 6 and 24 hpi, respectively. Overall, we identified 216 unique genes which were differentially expressed between the two lines for at least one time point during the time course. Since differential expression analysis revealed there was significant line-by-time interaction effect for only four genes (INSIG1, PTX3, FBP1 and ENSSSCG00000015140), we tested for the main effect of line and identified 255 DEGs between the two lines across all time points. A heatmap showing expression profiles of these 255 DEGs are displayed in Figure 2.3. Ten of the
DEGs between lines at baseline were persistently detected as DEGs along the time course, as shown by the Venn diagram (Figure 2.4A).

By comparing within-line gene expression between time points post treatment and baseline, we identified many more DEGs (Figure 2.4B and 2.4C). DEGs of the two lines at a given time points compared to baseline largely overlapped, but DEGs at different time points only moderately overlapped (Figure 2.4B-G). Given the negligible line-by-time interaction effects, we performed average line differential expression analysis to increase the power of detecting DEGs at each time point post LPS treatment compared to baseline. The number of average line DEGs at 2, 6 and 24 hpi compared to baseline were 3,172, 4,137 and 1,239, respectively, and in total, 5,871 unique genes were differentially expressed at least for one time point post LPS stimulation relative to baseline, which included 34 of the 36 genes assayed by RT-qPCR as above (Figure 2.4D and Table A.5). Over the time course, the expression of many genes was transient, with only 508 genes persistently differentially expressed at all 3 time points post LPS stimulation compared to baseline (Figure 2.4D). Average line DEGs between each time point post LPS treatment and baseline included a large majority of DEGs for equivalent within-line comparison plus some extra DEGs specific to the average line analysis (Figure 2.4E-G). A full list of within-line and average line DEGs at 2, 6, and 24 hpi compared to baseline is available in Table A.5. Taken together, differential expression analyses suggested that the pigs’ blood transcriptomes were dramatically disturbed by LPS injection. Average line DEGs at least for one time point post LPS treatment compared to baseline were hierarchically clustered and are displayed as a heatmap (Figure 2.5).
DEGs in response to LPS stimulation functioned in inflammatory response and beyond

The functions of the DEGs between the two lines at each time point were annotated by using GO term overrepresentation analysis. No GO terms were overrepresented in the lists of DEGs between the two lines at 0, 2 and 6 hpi. However, one GO-BP term was overrepresented in the list of between-line DEGs at 24 hpi: antigen processing and presentation of peptide or polysaccharide antigen via MHC class II (GO:0002504). The expression levels of seven porcine MHC class II genes were lower in the low-RFI line compared to the high-RFI line ($q < 0.05$). Gene set enrichment analysis based on comparison of mean expression levels of genes of the low-RFI line to the counterparts of the high-RFI line indicated that cytokine-cytokine receptor interaction (KEGG pathway ssc04060) was slightly enriched by genes of higher expression levels in the high-RFI line than in the low-RFI line at 0 hpi ($q = 0.001$) and 24 hpi ($q = 0.03$) (Table A.6).

We also carried out gene set enrichment analysis by comparing gene expression at 2, 6 and 24 hpi to baseline for each line separately and then compared the enrichment level of each set between the two lines. These analyses are believed to be more powerful than the enrichment analysis mentioned above because the within-individual correlation in gene expression was considered [239]. Hallmark gene sets, including interferon gamma and alpha response, TNF$\alpha$ signaling via NFkB, inflammatory response, P53 pathway and complement, were of higher level of enrichment at 2, 6 and 24 hpi in relative to baseline in the high-RFI line than in the low-RFI line. Gene sets of IL2/STAT5 signaling and KRAS signaling down were more enriched in the low-RFI line than in the high-RFI line at 2 hpi compared to baseline (Figure 2.6). This analysis suggests the low-RFI line had a slightly lower level of inflammatory response at 2, 6 and 24 hpi than the high-RFI line. Besides inflammation response, this analysis also suggests genes involved in reactive oxygen species (ROS) pathway, apoptosis, hypoxia, glycolysis, xenobiotic
metabolism, apical junction and heme metabolism were disturbed post LPS stimulation in pigs (Figure 2.6).

Twenty-six KEGG signaling and metabolism pathways were significantly enriched by disturbed genes post LPS injection (Figure A.2), including cytokine-cytokine receptor interaction, neuroactive ligand-receptor interaction, hematopoietic cell lineage, TNF signaling pathway, NOD-like receptor signaling pathway, chemokine signaling pathway, complement and coagulation cascades, NFκB signaling pathway, Toll-like receptor signaling pathway, TGF-β signaling pathway, ribosome, and several amino acid metabolism pathways. Several KEGG signaling pathways was more highly enriched among disturbed genes in the high RFI at 6 and 24 hpi in relative to baseline, including cytokine-cytokine receptor interaction, NOD-like receptor signaling pathway and complement and coagulation cascades, although they were almost equally enriched in both lines at 2 hpi relative to their own baseline. The toll-like receptor signaling pathway was equally enriched by disturbed genes in both lines across all 3 time points. Ribosome component-encoding genes were more enriched among down-regulated gens in the high-RFI line than in the low-RFI line at 6 hpi. At 24 hpi, they were not enriched by disturbed genes in the high-RFI line but they were in the low-RFI line. KEGG gene sets of diseases were similarly enriched by disturbed genes at 2, 6, and 24 hpi in both lines (Figure A.3), including infectious diseases, autoimmune disease and cancer.

Transcription factor targets-based gene set enrichment analysis suggested targets of immune response-related transcription factors were significantly enriched by disturbed genes post LPS stimulation (Figure A.4), including AP1, E12/E47, PU1, several interferon-regulatory factors (IRFs), NFAT, CEBP, ISRE, NFκB and GATA6. Targets of transcription factors including IRF7, IRF8 (also called ICSBP), PU1 and GATA6 were more enriched, if
significantly, among disturbed genes in the high-RFI line than in the low-RFI line at 2, 6 and 24 hpi. But targets of transcription factors including AP1, NFκB, CDC5, GATA3, and CEBP were more enriched among disturbed genes in the low-RFI line than the high-RFI at 2 hpi, while the opposite was suggested for 6 and 24 hpi (Figure A.4).

Gene set enrichment analysis based on canonical pathways without KEGG pathways suggested that genes involved in the extracellular matrix were more affected by LPS treatment in the low-RFI pigs than in the high-RFI pigs at 2 hpi, but at 6 and 24 hpi genes involved in the extracellular matrix were more affected by LPS treatment in the high-RFI pigs than in the low-RFI pigs at 2 hpi (Figure A.5). In addition, enrichment analysis based on chemical and genetic perturbation-associated gene sets and immunological signatures suggested that genes that were disturbed post LPS stimulation in pigs were significantly similar to those detected in LPS- or cytokine-stimulation human and murine tissues or cell lines, while the high-RFI line had a slightly higher level of inflammatory response than the low-RFI line (data not shown).

GO-BP and -CC terms-associated gene set enrichment analysis gave more details about disturbed biological processes and affected cellular components (Table A.6 and A.7). GO-BP gene sets-based enrichment analysis suggested that at 2 hpi, LPS had induced a defense response, including both an anti-virus response and an inflammatory response, with GO:0051607 (defense response to virus) being the most enriched gene set among up-regulated genes, and more enriched in the high-RFI than the low-RFI lines. This analysis again suggested the inflammatory response was slightly higher in the high-RFI line than in the low-RFI line. At 6 hpi, GO-BP gene sets of I-κB kinase/NFκB signaling, positive regulation of JUN kinase activity and negative regulation of MAP kinase activity were more enriched in the low-RFI line than in the high-RFI line. Translation, the most enriched GO biological process was more repressed in the high-RFI
line than the low-RFI line at 6 hpi, but the opposite was suggested at 24 hpi. Antigen processing and presentation of peptide or polysaccharide antigen via MHC class II were more enriched among down-regulated genes in the low-RFI line than in the high-RFI line at 24 hpi. Other significantly enriched, up-regulated GO-BP gene sets included cell inflammatory response and chemotaxis, which was more enriched among up-regulated genes in the high-RFI line than the low-RFI line at both 6 and 24 hpi. Results of these GO-CC gene sets-based enrichment analyses were consistent with the enriched GO-BP (Table A.7).

Because we only detected line-by-time interaction effect for a few genes, we also performed average line gene set enrichment analysis. Average line differentially enriched gene sets between time points post LPS injection and baseline are shown in Table A.7, which were largely similar to those enriched gene sets in each individual line.

**Genes co-expressed in response to LPS stimulation showed enriched biological functions related to inflammatory response**

We then performed gene co-expression analysis using STEM to complement with the conventional differential expression analysis which considered each gene independently and the gene set enrichment analysis which considered groups of genes with known related functions or behaviors. Given that only few significant line-by-time interactions were detected, we performed an average line clustering analysis of gene expression to improve the stability of the clustered profiles. Thirty-one significant expression profiles were identified and were further merged into 8 clusters based on profile similarity (Figure 2.7). GO terms overrepresented among genes in each cluster are in Table A.8. Overrepresented GO-BP terms among genes in Cluster 1 included translation, ribosome biogenesis, fatty acid oxidation, cilium assembly and small molecular metabolism. Defense response, inflammatory response, response to chemical and biotic stimuli including bacterium and virus were overrepresented among genes in Clusters 2 and 3. Genes in
Cluster 4 were also involved in ribosome biogenesis. Among genes in Cluster 5, signaling and endocytosis were enriched. Phospholipid metabolic process and receptor complex were enriched among genes in Clusters 7 and 8, respectively (profile 13). No significant GO-BP or -CC terms were enriched among genes in Cluster 6. Expression of genes in Clusters 2 and 3 peaked at 6 and 2 hpi, respectively. Typical inflammatory genes were included in Cluster 2, including S100A8, S100A9, S100A12, MyD88, IRF3, JAK2, CXCL13 and CXCL16, and reactive oxygen species response genes including SOD2, while inflammatory and immune responsive genes were included in Cluster 3, including IL1B1, CXCL2, GBP4 and 5, IRF1 and 2, NFκBIB, STAT2, TNFSF4 and 10, and TGFβ2.

Pigs and humans shared conserved mechanisms responsive to systemic inflammation induced by LPS

We reanalyzed available microarray data for a time-series, repeated measures of peripheral blood transcriptomic response to LPS in human [214] to have an across-species comparison to the data of this study. A major modification we made was to include eight surrogate variables in the model for differential expression analysis to account for hidden variation of no interest, such as processing batch effects. MDS plots showing sample relationship based on gene expression abundance with and without adjusted for surrogate variables are displayed in Figure A.8A and A.8B. Adjusting for the hidden variation significantly improved the clustering of samples under the same conditions and separation of samples in different states. We identified 6,628 probesets (4,428 Entrez genes) that were significantly differentially expressed for at least one time point post LPS infusion compared to baseline, with 3,378 (2,378 DEGs), 4,810 (3,339 DEGs), 4,604 (3,118 DEGs), 3,582 (2,478 DEGs) and 0 differentially expressed probesets (DEPs) at 2, 4, 6, 9 and 24 hpi relative to baseline for the LPS treatment group, respectively. But we did not detect any significantly DEP for the control group over the
time course (Table A.9). A full list of DEPs/DEGs is provided in Table A.9. Notably, 3,955 (59.7%) of our DEPs were also detected as DEPs by Calvano et al. [214]. The number of DEGs as well the MDS plot suggested LPS induced systematic transcriptome changes by 2 hpi, and transcriptomic response at 4, 6 and 9 hpi were more similar to each other than to that at 2 hpi, while the blood transcriptome nearly returned to baseline at 24 hpi.

We performed gene set enrichment analysis for transcriptomic changes at 2, 4, 6, 9 and 24 hpi in relative to baseline. Enriched gene sets are listed in Table A.10. We then compared the gene set enrichment patterns in the humans to those in the pigs. Hallmark gene set enrichment in pigs and humans is shown in Figure 2.8. Generally, of the 39 Hallmark genes sets that were significantly enriched in at least one species for at least one time point, 22 were enriched in both species for at least one time point. This enrichment analysis suggests that pigs and human shared similar transcriptomic response to systemic inflammation induced by LPS, though gene sets of interferon α and γ response, TNFα signaling via NFκB, inflammatory response and IL2/STAT5 signaling were much more highly enriched in LPS-stimulated humans at 2, 4, 6 and 9 hpi compared to baseline, than in pigs at 2 and 6 hpi compared to baseline. Another strikingly and specifically enriched gene set in humans was Myc targets V1. Notably, 6 and 7 Hallmark gene sets were only significantly enriched in pigs and humans, respectively (Figure 2.8).

KEGG signaling and metabolism pathway enrichment analysis suggested gene sets of 19 KEGG pathways were enriched in both species, including ribosome, chemokine signaling pathway and cytokine-cytokine receptor interaction (Table A.10). It also showed that the gene set related to ribosome was more highly enriched among down-regulated genes in humans than in pigs post LPS stimulation. Gene sets associated with hematopoietic cell lineage and cytokine-cytokine receptor interaction were more highly enriched in humans and pigs, respectively, than
in pigs and humans in response to LPS challenge. Nineteen pathways were only significantly enriched in pigs post LPS stimulation, including several pathways for protein/amino acid metabolism and lipid/fatty acid metabolism, drug metabolism-cytochrome P450 and HIF-1 signaling, while 13 pathways were only significantly enriched in humans, including carbon metabolism, TCA cycle, oxidative phosphorylation and splicesome.

Comparison of the enriched targets of transcription factor suggested the two species shared many common transcription factors during LPS-induced systemic inflammation and resolution of inflammation, including NFκB, AP1, IRFs, STATs, CEBP, PU1, and CDC5. But it also suggested that species-specific transcription factors were involved in this process, such as Myc/MAX, USFs and E2F for humans, and ETS2 for pigs (Figure A.11).

Comparison of enrichment of gene sets of GO-BP and -CC terms suggested that there were many more biological processes and cellular components involved in the LPS-induced systemic inflammation in humans than in pigs, although tens of GO-BP and -CC gene sets were enriched in both species and a small number of GO gene sets were specifically enriched in pigs in response to LPS stimulation (Figure A.12 and A.13). “Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay”, SRP-dependent cotranslational protein targeting to membrane, translation initiation and viral transcription were specifically, persistently, and highly enriched in humans during the process of inflammation by 24 hpi (Figure A.12).

We also compared the clustered gene expression profiles in response to systemic inflammation in both species. For better comparison, only expression data for 0, 2, 6 and 24 hpi were considered. Based on the expression data of LPS-challenged humans at 0, 2, 6 and 24 hpi, we identified seven clusters, which consisted of 27 significant gene expression profiles (Figure A.14). Notably, no significant profiles were detected for the control group gene expression
profiles. Four human clusters were comparable to four porcine clusters in both dynamic patterns of gene expression, and enriched biological processes and cellular components (Figure A.14). The commonly enriched GO biological processes and cellular components in these four similar clusters included cytosolic and mitochondrial translation, ribosome, mitochondrial respiratory chain, inflammatory response, defense response, response to LPS, cell migration and signaling (Figure A.14). We also found that clusters that over-representing the biological process, “cytosolic and mitochondrial translation”, in humans and pigs shared 396 genes, including tens of ribosomal protein-encoding genes and several translation initiation factor-encoding genes.

Finally, we compared the dynamic expression profiles of porcine and human orthologs that showed differential expression in response to LPS in both species. Again, only expression data for 0, 2, 6 and 24 hpi were considered for better comparison. Figure 2.9 shows the time-course expression patterns of 1,543 porcine genes and 2,175 human probesets (1,548 human Ensembl genes), both of which were differentially expressed at least for one time point post LPS stimulation in pigs and human compared to baseline, respectively. One third of these common DEGs showed very similar expression dynamics over the time course (Figure 2.9, Green box 1). This group of genes was generally down-regulated during LPS-induced systemic inflammation. Only a very small number of genes showed opposite expression patterns in the two species (Figure 2.9, Blue box 2). The remaining shared more or less similar expression patterns over only some of the time points. GO term overrepresentation analysis indicated that these conserved genes mainly function in biological processes including translation, RNA catabolism, signaling, apoptosis, immune response and inflammatory response (Figure 2.10).
Divergent selection for RFI in pigs resulted in only minor differences in response to systemic inflammation triggered by LPS

In this study, we challenged pigs from two lines which had been divergently selected for RFI for multiple generations with LPS. Typical observations from pigs post LPS treatment induced a systemic inflammatory response, including fever response and dynamic changes of concentration of WBCs and cytokines in peripheral blood. However, no significant differences in these physiological traits were detected between the two lines. By transcriptome-wide profiling of peripheral blood, we only identified a very small number of DEGs between the two RFI lines at each time point. In addition, gene set enrichment analysis suggested that the toll-like receptor and NFκB signaling pathways, which are canonical pathways in response to LPS stimulation in mammals, were equally enriched by disturbed genes in both lines post LPS injection. All these lines of evidence support that the two lines responded to LPS-induced systematic inflammation in similar ways and indicated no distinguishable defect in inflammatory response in the low RFI line.

It is worth noting that we did find some minor differences between the two lines in response to LPS stimulation by using powerful and sensitive functional annotation tools. Both gene set enrichment analysis and GO term overrepresentation analysis of DEGs suggested that the low-RFI line may slightly differ from the high-RFI line at 24 hpi in antigen processing and presentation via MHC class II. Expression levels of seven SLA class II genes were significantly lower in the low-RFI line compared to the high-RFI line. Thus, it is possible that the low-RFI line has lower activity in processing and presenting exogenous antigen via MHC class II than the high-RFI line at the end of resolving phase of systemic inflammatory response.
The KEGG pathway ssc04060: cytokine-cytokine receptor interaction was found to be slightly more enriched among genes of higher expression levels in the high-RFI line than the low-RFI line at all time points studied. Consistent with this, the low-RFI line tended to have lower expression levels for many cytokines and cytokine receptors both at baseline and post LPS stimulation. We also found gene sets of inflammatory response, interferons gamma and alpha response, TNFα signaling via NFκB, and complement were of lower levels of enrichment at 2, 6 and 24 hpi relative to baseline in the low-RFI line than in the high-RFI line. This suggests that the low-RFI line had a slightly lower level of inflammatory response than the high-RFI line during the time course of inflammation.

Translational inhibition is widely observed from invertebrates to mammals in response to pathogen infection [243]. Without exception, we found genes that encode ribosome components and other translational machinery were down-regulated in response to LPS stimulation, including translation initiation factors. Notably, gene sets for ribosome and translation were more enriched by down-regulated genes in the high-RFI line than in the low-RFI line at 6 hpi. At 24 hpi, they were not enriched in the high-RFI line, but they still were enriched in the low-RFI line. It is unclear whether translation inhibition was a true direct sensor of bacterial pathogens or is an indicator of a more general host metabolic stress response to LPS [243]. A recent study reported that translational inhibition plays an important role in negative feedback of the inflammatory response in the macrophages [244].

The two lines were slightly different in the activity of transcription factors related to inflammatory and immune responses. Targets of transcription factors-based gene set enrichment suggested that transcription factor AP1 may be more active at 2 hpi and less active at 6 and 24 hpi in the low-RFI line than in the high-RFI line, while other transcription factors such as IRFs,
PU1 and CEBP may be more active over the time course in the high-RFI line than in the low-RFI. These potential differences in transcriptional network might explain the slightly different level of inflammatory response in the two lines.

Thus, the two lines divergent selected for RFI responded to the systemic inflammation triggered by LPS in similar ways, although the low RFI pigs has a slightly lower level of inflammatory response post LPS stimulation than the high-RFI line. A slightly lower level of inflammatory response might reduce the energy required for handling inflammation, which might be beneficial if the existing level of inflammatory response is enough for the host to remove the pathogen and recover from infection. Body temperature records for the 7-day time course indicated that the two lines responded equally to repeated LPS challenges (data not shown). Importantly, the gain of body weight of the two lines during the time course was not significant different (data not shown). Thus, it is reasonable to assume that the slightly lower level of inflammatory response in the low-RFI line than in the high-RFI line was not detrimental. It will be interesting to compare the level of inflammatory response post LPS stimulation in the low-RFI animals to that in a population not selected for feed efficiency in the future.

**Pigs and humans shared similar blood transcriptomic response to systemic inflammation induced by LPS**

Previous cross-species comparisons ended up with absolutely opposite conclusions in terms of similarity of inflammatory response in humans to that in mice based on different comparing metrics [211, 212]. The potential drawbacks in all previous comparative studies included, but were not limited to (1) Gene expression microarray data were generated on differently platforms in different labs with different experiment designs for different purposes; (2) Only data from single time points which were assumed to be the most comparable were compared, although inflammation is a very dynamic process; (3) Batch effects and other hidden
variations were not considered in differential gene expression analyses; (4) The same but arbitrary statistical stringency of cutoffs were used to determine DEGs for comparison; (5) Only common DEGs in both species or the union of DEGs in either species were compared. These and other limitations could compromise the reliability of the results of comparisons.

To avoid or minimize some of problems mentioned above, we compared blood transcriptomic response to systemic inflammation triggered by LPS in pigs in this study to that in humans from a similarly-designed experiment [214]. These two studies had 4 common time points: baseline, and 2, 6, and 24 hpi. After adjustment for hidden variations in the gene expression data from both studies, we not only compared the DEGs common in both species but also gene profiles without a hard threshold for identifying DEGs based on powerful and sensitive gene set enrichment analysis. We identified more than 1,500 orthologous genes that were differentially expressed in both species. These core sets of conserved genes mainly function in biological processes including immune response, inflammatory response, leukocyte migration, toll-like receptor signaling pathway, cytokine-mediated signaling, regulation of cytokine production, translation, protein targeting, and mRNA catabolism. Notably, nearly one third of this set of genes had very similar dynamic expression patterns in response to LPS stimulation over the common time points in the two species. Except for a handful of genes that showed opposite expression patterns, the rest of the genes showed more or less similar expression patterns over the time points in the two species. Clustering analysis of gene expression by STEM also identified gene clusters with similar temporal expression patterns, shared biological functions and gene contents.

More powerful gene set enrichment analysis further supported that the conclusion that pigs and humans shared many more biological pathways and processes than those mentioned
above, including those mentioned in the previous paragraph, in response to LPS-induced systemic inflammation. We found that most pathways and biological processes that were involved in porcine response to systemic inflammation were also involved in the human response. However, it also revealed many differences in biological processes and pathways that were involved in response to LPS stimulation in the two species, with many biological processes and pathway specifically involved in human response and only a few specifically involved in porcine response. This might be partly due to different level of inflammatory response triggered by LPS stimulation in pigs versus humans. Although 15,000-fold higher dosage of LPS was administered in pigs than in humans, pigs appeared to have a much lower level of inflammatory response than humans. This might be mainly because pigs are more tolerant to LPS than humans [205]. The differences in routes of LPS administration, and sources and serotypes of LPS could also contribute to the different levels of inflammatory response [210, 245, 246]. Differences in experimental procedures and platforms might also explain some of the detected differences in transcriptomic response to inflammatory stimulation. Total RNA extracted from whole blood and leukocytes were profiled in this study by RNA-seq and in the human study by gene expression microarray, respectively. In addition, all knowledge-based analyses for pigs were mainly based on functional annotation of pig orthologs of human genes. However, only nearly 7,000 of 11,911 genes that were detected in peripheral blood in this study have been projected with HGNC gene symbols. Thus, the poor functional annotation of porcine genes could exaggerate, but less likely underestimate, the interspecies differences in response to systemic inflammation induced by LPS.
Conclusions

Pigs from the two lines divergently selected for RFI responded to systemic inflammation triggered by LPS injected intramuscularly in similar ways, although the low-RFI animals had a slightly lower level of inflammation response than the high-RFI animals along the time points post LPS stimulation. Pigs and humans shared similar response to systemic inflammation induced by LPS at both pathway and gene levels, although many more biological processes and pathways were specifically involved in inflammatory response in humans and only a few were preferentially involved in pigs. But caution should be taken with the differences seen between humans and pigs, due to differences in routes of LPS administration, LPS dosage, sources and serotype of LPS, blood fractionation, and completeness of genome annotation in both species. Our work indicates that selection for feed efficient pigs did not compromise the pig’s capability to handle acute systemic inflammation triggered by LPS, and also offered another line of evidence that pigs can be used as models for human inflammatory disease.
Figure 2.1. Intramuscular injection of LPS induced systemic inflammation in pigs. (A) LPS, but not saline induced similar febrile response in pigs. Shown are profiles of mean body temperature for the two lines injected with LPS or saline at each time point. Error bars show standard errors of the mean body temperature of each line at each time point. (B) Dynamic CBC profiles in response to LPS or saline treatment in pigs showed LPS injection triggered inflammation in pigs, while saline not. Quantity of each CBC parameter was standardized by normal transformation and displayed as heatmap. (C) Cytokine dynamic profiles in response to LPS or saline injection in pigs. Cytokine concentrations were standardized by normal transformation and displayed as heatmap.
Figure 2.2. (Legend on next page)
Figure 2.2. Expression profiles of inflammation-related genes during inflammatory response determined by Fluidigm RT-qPCR. Shown are means of $\log_2$ abundance of gene expression ± 95% confidence intervals. Symbols of genes differentially expressed at least at one time point post LPS injection compared to baseline were colored in yellow. TLR2 and TLR7 tended to be differentially expressed at 6 hpi in relative to 0 hpi ($q < 0.09$). For reference, profiles of the two control genes, GAPDH and RPL32, are also displayed. The expression levels of the two control genes were not stable over the time course.
Figure 2.3. Heatmap showing expression profiles of 255 DEGs with line main effects. H, high RFI; L, low RFI; Tx, time point x hpi.
Figure 2.4. Venn diagrams showing between-line, within-line and average line DEGs at different time points. (A) Venn diagram showing 216 DEGs between lines at each time point. (B and C) Venn diagrams showing DEGs between time points post LPS treatment and baseline in the high (B) and low (C) RFI animals. (D) Venn diagram showing average line DEGs between time points post LPS treatment and baseline. (E-G) Venn diagrams showing relationship of within-line and average line DEGs at 2, 6, and 24 hpi, respectively. HRFI, high RFI; LRFI, low RFI. Time points without being qualified were for the averaged line.
Figure 2.5. Heatmap showing the expression profiles of 5,871 average line DEGs for at least one time point post LPS treatment relative to baseline. H, high RFI; L, low RFI; Tx, time point x hpi.
Figure 2.6. Hallmark gene sets enriched among disturbed genes in each line at time points post treatment compared to baseline. Shown are gene sets significantly enriched under at least one condition (q value < 0.01). HRFI, high RFI; LRFI, low RFI; Tx, time point x hpi.
Figure 2.7. Clusters of significant profiles of gene expression in response to LPS stimulation. Representative significantly overrepresented GO terms by each cluster are also displayed.
Figure 2.8. Hallmark gene sets enriched in disturbed genes in pigs and humans post LPS stimulation. Shown are gene sets significantly enriched under at least one condition (q value < 0.01). HRFI, high RFI; LRFI, low RFI; Tx, time point x hpi.
Figure 2.9. Expression profiles of genes differentially expressed in both species. Groups of genes showed very similar expression dynamics and opposite expression dynamics were labeled with green and blue boxes on the right side, respectively. H, high RFI; L, low RFI; Tx, time point x hpi.
Figure 2.10. GO-BP terms overrepresented by common DEGs in response to systemic inflammation in humans and pigs. GO-BP terms significantly overrepresented ($q < 0.05$) were summarized and visualized by using REViGO. GO-BP terms with semantic similarity in the measure of SemRel no less than 0.7 were collapsed. $\log_{10}$-based $p$-values associated with overrepresented GO terms were coded in a color gradient from red to blue.
CHAPTER 3. POST-WEANING BLOOD TRANSCRIPTOMIC DIFFERENCES BETWEEN YORKSHIRE PIGS DIVERGENTLY SELECTED FOR RESIDUAL FEED INTAKE

Modified from a paper published in *BMC Genomics* 2016;17:73.

Haibo Liu, Yet T. Nguyen, Dan Nettleton, Jack C. M. Dekkers, Christopher K. Tuggle

**Abstract**

**Background:** Improving feed efficiency (FE) of pigs by genetic selection is of economic and environmental significance. An increasingly accepted measure of feed efficiency is residual feed intake (RFI). Currently, the molecular mechanisms underlying RFI are largely unknown. Additionally, to incorporate RFI into animal breeding programs, feed intake must be recorded on individual pigs, which is costly and time-consuming. Thus, convenient and predictive biomarkers for RFI that can be measured at an early age are greatly desired. In this study, we aimed to explore whether differences exist in the global gene expression profiles of peripheral blood of 35 to 42 day-old pigs with extremely low (more efficient) and high RFI (less efficient) values from two lines that were divergently selected for RFI during the grow-finish phase, to use such information to explore the potential molecular basis of RFI differences, and to initiate development of predictive biomarkers for RFI.

**Results:** We identified 1,972 differentially expressed genes (DEGs) \( q \leq 0.15 \) between the low \( n = 15 \) and high \( n = 16 \) RFI groups of animals by using RNA sequencing technology. We validated 24 of 37 selected DEGs by reverse transcription-quantitative PCR (RT-qPCR) in a joint analysis of 24 (12 per line) of the 31 samples already used for RNA-seq plus 24 (12 per line) novel samples from the same contemporary group of pigs. Using an analysis of the 24 novel
samples alone, only 9 of the 37 selected DEGs were validated. Genes involved in small molecule biosynthetic process, antigen processing and presentation of peptide antigen via major histocompatibility complex (MHC) class I, and steroid biosynthetic process were overrepresented among DEGs that had higher expression in the low versus high RFI animals. Genes known to function in the proteasome complex or mitochondrion were also significantly enriched among genes with higher expression in the low versus high RFI animals. Alternatively, genes involved in signal transduction, bone mineralization and regulation of phosphorylation were overrepresented among DEGs with lower expression in the low versus high RFI animals. The DEGs significantly overlapped with genes associated with diseases, including hyperphagia, eating disorders and mitochondrial diseases \( (q < 1 \text{E-05}) \). A weighted gene co-expression network analysis (WGCNA) identified four co-expression modules that were differentially expressed between the low and high RFI groups. Genes involved in lipid metabolism, regulation of bone mineralization, cellular immunity and response to stimulus were overrepresented within the two modules that were most significantly differentially expressed between the low and high RFI groups. We also found five of the DEGs and one of the co-expression modules were significantly associated with the RFI phenotype of individual animals \( (q < 0.05) \).

**Conclusions:** The post-weaning blood transcriptome was clearly different between the low and high RFI groups. The identified DEGs suggested potential differences in mitochondrial and proteasomal activities, small molecule biosynthetic process, and signal transduction between the two RFI groups and provided potential new insights into the molecular basis of RFI in pigs, although the observed relationship between the post-weaning blood gene expression and RFI phenotype measured during the grow-finish phase was not strong. DEGs and representative
genes in co-expression modules that were associated with RFI phenotype provide a preliminary list for developing predictive biomarkers for RFI in pigs.

**Keywords:** *Sus scrofa*, feed efficiency, residual feed intake, blood transcriptome, RNA-seq

**Background**

Feed efficiency (FE) is one of the important traits directly related to profitability, productivity, and sustainability in the pork industry [136, 137]. While many non-genetic strategies have been developed to improve feed efficiency of pigs [139], improving feed efficiency by genetic selection can be a sustainable alternative. An increasingly popular measure of FE is residual feed intake (RFI), which is the difference between the actual and expected feed intake of an animal for production and maintenance [138]. Importantly, RFI is moderately heritable and responds well to genetic selection in pigs [140, 141].

Over the last decade, significant efforts have been made to understand the molecular, genetic and physiological basis of RFI in pigs. Researchers have found many interesting differences between pigs with divergent RFI phenotypes [142, 143]. For example, as compared to high RFI pigs (less efficient), low RFI pigs (more efficient) have altered feeding behaviors [144], slightly lower growth rate [145], less back fat [141, 146, 147], lower protein turnover rate in the muscle [148], altered mitochondrial protein profiles [149, 150], less mitochondrial reactive oxygen species (ROS) production [151], and lower levels of leptin [152, 153] and juvenile IGF-1 in circulating blood [145]. To explore the genetic basis of RFI in pigs, several genome-wide association studies (GWAS) have been conducted [154-157]. Some chromosomal regions tagged by single nucleotide polymorphisms (SNPs) have been found to be associated with RFI, but these associations were not consistent across studies and explained only small portions of the
genetic variance for RFI [154-157]. The wide range of differences between lines of pigs with divergent RFI and the lack of SNPs with major effects on RFI suggest that RFI is a highly polygenic, quantitative trait with multiple tissues contributing to its variation.

Global gene expression profiling technologies have also been used to explore the molecular basis of RFI in pigs. By profiling the transcriptomes of the adipose tissue of two lines of pigs divergently selected for RFI with gene expression microarrays, Lkhagvadorj et al. [158] found that genes involved in the lipid metabolic pathway were overrepresented among the differentially expressed genes (DEGs) that had lower expression in low versus high RFI pigs, and genes involved in carbohydrate metabolism and response to stress were overrepresented among the DEGs that had higher expression in low versus high RFI pigs [158]. They also found the leptin-related gene network to be different between the two lines [158]. Recently, Vincent et al. [159] and Jing et al. [160] profiled the transcriptome of the longissimus dorsi (LD) muscle from pigs with divergent RFI by gene expression microarray and RNA-seq, respectively. Using pigs from lines divergently selected for RFI, Vincent et al. [24] found genes involved in protein synthesis and glycolysis, and genes associated with mitochondrial energy/oxidative metabolism had higher and lower expression, respectively, in the low versus high RFI line. Using Yorkshire barrows with extreme phenotypes for RFI, Jing et al. [160] found that genes involved in glycolysis had lower expression in the low versus high RFI group, while genes involved in muscle proliferation and differentiation had higher expression in the low versus high RFI group. Surprisingly, these two studies shared no DEGs and proposed opposite differences in glycolytic activities in the low versus high RFI pigs. Therefore, in consideration of the complexity of RFI and the inconsistency from study to study, the molecular mechanisms underlying RFI in pigs are still largely unclear.
To incorporate RFI into animal breeding programs for improving feed efficiency, feed intake, body weight gain and back fat depth must be recorded on individual pigs. As it is very expensive and time-consuming to record feed intake on individual animals [140, 141], convenient and predictive biomarkers for RFI that can be measured at an early age are in demand. In cattle, Chen et al. [161] successfully used 14 DEGs identified in the liver of Angus bulls that were divergently selected for RFI to classify Angus steers from the same divergent RFI lines, and Al-Husseini et al. [162] developed a RFI predictor using 8 of these 14 DEGs and validated it in an independent Angus population. These biomarkers are, however, not very practical because invasive liver biopsies are needed.

The peripheral blood is an informative tissue not only because it carries a variety of cells directly involved in immunity and inflammation, but also because it interacts with every organ and tissue in the body via bioactive circulating factors, such as nutrients, metabolites, cytokines, hormones and exosomal cargoes, which are released from the same or different organs or tissues [177]. These bioactive factors interact with blood cells and thus might modify the gene expression profiles of the blood cells dynamically. Molecular signatures in circulating blood, including gene expression profiles, have been shown to reflect the physiopathological status, growth stage and lifestyle of subjects [177-181]. Due to its easy accessibility and informativeness, blood has become a popular sample (direct or as a surrogate) for disease diagnosis, prediction, prognosis, and biomarker discovery [177, 178]. Interestingly, the concentration of IGF-1 in serum at a young age has been shown to be different between animals with divergent RFI in poultry and livestock, including pigs [145, 152, 182, 183]. Blood cell profiles at early growth stages have also been found to be different between livestock with divergent RFI phenotypes, including pigs [184] and cattle [185]. These results suggest that
animals with divergent RFI phenotypes, measured later in life, have early physiological differences in circulating blood that may be reflected in blood gene expression profiles at these early stages.

In this study, our objective was to determine the blood transcriptomic differences between post-weaning pigs from two lines divergently selected for RFI, to explore potential molecular mechanisms underlying RFI in peripheral blood and to develop a list of candidate biomarkers for RFI prediction. We hypothesized that post-weaning expression levels of some genes in whole blood were correlated with RFI phenotype measured during the grow-finish phase. We identified 1,972 DEGs with \( q \leq 0.15 \) and 4 co-expression modules that were differentially expressed between the low and high RFI groups. A set of selected DEGs were validated by reverse transcription-quantitative PCR (RT-qPCR). Several interesting biological processes underlying DEGs and differential co-expression modules were suggested. We also identified several candidate biomarkers for RFI.

**Methods**

**Animals, blood sample collection and complete blood count (CBC) test**

The experimental protocols for this study were approved by the Institutional Animal Care and Use Committee (IACUC) at Iowa State University under permit number 11-1-4996-S. All pigs were from parity 2 of generation 9 of the two lines divergently selected for residual feed intake: the low and high RFI lines [141, 154, 184]. Bunter et al. found that the IGF-1 concentration in blood measured between 35 to 42 days of age differed between the two lines and were genetically correlated with RFI measured later during the grow-finish phase [145]. So blood samples were collected from the jugular vein into Tempus™ Blood RNA tubes (Life Technologies, Grand Island, NY) for long-term storage at -80 °C from 233 post-weaning piglets.
of the two lines in this age range. Meanwhile, blood from those animals was also collected into EDTA tubes (BD, Franklin Lakes, NJ) and kept at 4 °C before CBC tests, which were performed on the bleeding day, if possible, or the next morning as described [184, 224]. Differences in CBC profiles between the two lines have been published elsewhere [35]. At 107.0 ± 8.3 (mean ± standard deviation) days of age and 42.2 ± 7.2 kg of body weight (BW), 88 gilts and 78 barrows were randomly assigned to 12 mixed-line, mixed-sex finishing pens with electronic single-space feeders (FIRE, Osborne Industries Inc., Osborne, KS) for feed efficiency test, with 6 pens being randomly assigned to either of two diets: a high-fiber, low-energy diet (HFD) and a low-fiber, high-energy diet (LFD) [247]. The animals on feed efficiency test consisted of 21 barrows and 22 gilts from the low RFI line plus 14 barrows and 23 gilts from the high RFI line fed the HFD, and 23 barrows and 23 gilts from the low RFI line plus 20 barrows and 20 gilts from the high RFI line fed the LFD. All pigs had ad libitum access to feed and water. Individual feed intake was real-time recorded, body weight was recorded biweekly, and backfat (BF) depth above the 10th rib and loin muscle area were recorded at the end of the test, when the pigs were 227.0 ± 1.4 days of age and 127.7 ± 8.8 kg of body weight. RFI of individual pigs were calculated as described [154] with modifications and shown in Table B.1. Briefly, average daily feed intake (ADFI) was estimated by fitting a quadratic polynomial regression model of the daily feed intake from the on-test day to the off-test day on the number of days on test for each pig and individual average daily gain (ADG) was estimated as the slope from simple linear regression of bi-weekly BW on the number of days on test [141]. A single trait animal model was used to analyze ADFI with adjustments for fixed covariates of metabolic mid-body weight (average body weight during the test period raised to the power of 0.75, MBW), ADG, BF, deviation of the on-test weight from 50 kg, deviation of the off-test weight from 118 kg, and deviation of the on-test age from
90 days, and the random effect of pen. The RFI value for each pig was estimated as the residual of the fitted model (Young J. and Dekkers J.C.M., unpublished.). The growth performance and feed intake data of these animals have been published [248].

**RNA extraction and globin depletion**

Based on the RFI phenotype of individual pigs, the juvenile blood samples of 32 out of the 88 gilts were selected for RNA-seq. In detail, for the low RFI line, we selected eight samples from animals with extremely low RFI values on the LFD and eight samples from animals with extremely low RFI values on the HFD (designated as low RFI group) while for the high RFI line, we selected eight samples from animals with extremely high RFI values on the LFD and eight samples from animals with extremely high RFI values on the HFD (designated as high RFI group) (Figure 3.1). Total RNA extraction of the 32 samples was processed in four batches such that for each batch, two samples were randomly selected from each line by diet combination. Within each batch, the processing order of samples was randomized beforehand and followed in all subsequent procedures where blocking was necessary. The total RNA was extracted from blood samples using preserved blood RNA purification kit I (Norgen Biotek Corp, Thorold, Ontario) by following the kit’s manual. On-column DNA digestion was performed as described using DNase I (Qiagen, Valencia, CA). Globin transcripts (HBB [ENSSSCG00000007978] and HBA [ENSSSCG00000014725]) were depleted by following an RNase H-mediated method [223]. The quantity and integrity of the RNA were determined by using Nanodrop 2000 (Thermo Scientific, Waltham, MA) and Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA) before and after globin depletion. The efficiency of globin depletion of each sample was checked by conventional RT-qPCR with β-actin (ACTB) and glyceraldehyde 3-phosphate dehydrogenase (GAPDH) as the internal reference genes. The total RNA of one selected blood sample from the
low RFI line on the LFD was dropped from this study due to its low RNA quality (original RNA integrity number (RIN) was 7.5, lower than our criterion of RIN ≥ 8.0 before globin depletion). The average RIN for the 31 remaining samples before and after globin depletion were 9.05 ± 0.31 and 8.26 ± 0.30, respectively. Detailed information about the selected samples, including pedigree, RNA quality, batch, and CBC is available in Table B.1.

RNA-sequencing

Library construction and sequencing were performed by the DNA facility at Iowa State University. Briefly, the RNA-seq libraries were constructed using the Illumina TruSeq RNA Sample Preparation Kit v2 (Illumina, San Diego, CA) according to manufacturer’s instructions. For each sample, poly (A)-containing transcripts was enriched with oligo-dT-coated magnetic beads from 0.7 to 2 µg of total RNA with globin transcript depleted. The enriched RNA was fragmented by heat and reverse transcribed with hexamer random primers. For each sample, adapters with unique barcodes were ligated to the end-polished cDNA fragments. The libraries were linearly amplified by PCR, size selected and quantitated. The individual libraries were diluted to 2 nM and pooled in approximately equimolar amounts according to the processing batches mentioned above with 8 libraries per pool, except for one pool of 7 libraries. 100 base paired-end sequencing was run on an Illumina Hiseq2000 platform with one pool per lane on a flow cell. The RNA-seq data are available in ArrayExpress under accession number: [E-MTAB-4179, http://https://www.ebi.ac.uk/arrayexpress/].

Quality control, preprocessing and alignment of RNA-seq reads

Read quality was checked by using FastQC (version 0.10.1) [249]. Adapters and low quality bases were trimmed by running Trimmomatic (version 0.32) [250] in the paired-end mode with the following options, $\text{ILLUMINA\textunderscore CLIP:adapters\textunderscore fa:2:30:10:1: true LEADING:3}$
such that the average base quality was not lower than 15 for every sliding window of 4 bases and the minimum length of kept reads was 36 bases. For each set of raw paired-end reads, Trimmomatic outputted a pair of files for the kept paired-end reads, a file for unpaired forward reads (R1), and a file for unpaired backward reads (R2). Trimmed paired-end and unpaired reads were separately aligned to the pig reference genome Sscrofa 10.2 (version 77, Ensembl) using 2-pass rna-STAR (version 2.3.0), using the default settings [228, 229]. Read counts per gene per library were summarized by using featureCounts (version 1.4.4) [92], with the resulting SAM files for uniquely mapped, paired-end and unpaired reads as inputs, separately, and using the pig genome GTF file (version 77, Ensembl) as the genomic annotation reference file. The default settings for other featureCounts options were used [92]. A final table of read counts per gene of the 31 samples (designated as the count table) was generated by summing up the individual count tables for paired-end and unpaired reads. Prior to differential expression analysis, hemoglobin genes (HBA and HBB) and genes with few reads (average read count no larger than 8, or 28 or more zero read counts across the 31 samples) were removed from the count table to obtain a final count table with 12,280 genes. This count table was used for the subsequent differential expression analysis and weighted gene co-expression network analysis (WGCNA) after further transformation and adjustment.

**Differential expression analysis**

The statistical programming language R (version 3.1.0) was used for all statistical analyses, unless indicated otherwise. Differential expression analysis was carried out by using the R package “QuasiSeq” (version 1.0-4) [101]. For each of the 12,280 genes in the final count table, we used QuasiSeq to fit a full generalized linear model with a negative binomial response
and a \( \log \) link function that included an upper-quartile normalization offset \cite{43} and the fixed effects of RFI group (low and high RFI groups), diet (low and high fiber diets), batch (sample processing batches 1 to 4), and the linear covariates of RFI phenotype (estimated RFI value), pre_conc (RNA concentration before globin depletion), pre_RIN (RIN before globin depletion), post_conc (RNA concentration after globin depletion), post_RIN (RIN after globin depletion), and concentrations of neutrophils, lymphocytes, monocytes, eosinophils and basophils. Note that, although blood was collected before pigs were fed the different diets, a diet effect was included in our initial full model because diet affected the component traits that were used to estimate RFI and thus the estimation of RFI. Because not all the variables included in the initial full model may be associated with transcript levels, we used a backward variable selection algorithm to identify the most relevant variables \cite{251}. The final model included RFI group, batch, pre_conc, post_RIN, and the concentrations of neutrophils, lymphocytes, monocytes and basophils as independent variables. The default settings for arguments in all function calls were used unless specified otherwise. In the \( QL.fit \) function, the “method” argument was set to “optim”. The reported \( p \)-values, \( q \)-values and \( \log_2 \) (fold change) associated with all tests of significance were calculated by using the QLSpline method.

**Weighted gene co-expression network analysis (WGCNA)**

Before co-expression analysis using the R package “WGCNA” (version 1.46) \cite{252}, the expression levels for the 12,280 genes in the count table were adjusted for all independent variables in the final model used for differential expression analysis except RFI group. Briefly, \( \log \)-counts per million (designated as \( \log \text{-cpm} \)) were calculated using the \textit{voom} function of the Bioconductor package “limma” (version 3.20.9) with the upper-quartile normalized counts as input \cite{233, 253}. The \textit{lmFit} function was used to fit a linear model with \( \log \text{-cpm} \) per gene feature
as the responsible variable, and RFI group, batch, and the linear covariates of pre_conc, post_RIN, and concentrations of neutrophils, lymphocytes, monocytes and basophils as independent variables. Effects associated with relevant variables (batch, pre_conc, post_RIN, and the concentration of neutrophils, lymphocytes, monocytes and basophils) were subtracted from the original log-cpm, to create adjusted transformed gene expression values. The data matrix consisting of adjusted log-cpm per gene for the 31 samples (hereafter called the adjusted transformed gene expression matrix) was used as the input for WGCNA. WGCNA was performed by following tutorial I [254], with slight modifications as needed. In WGCNA, all correlation coefficients between gene pairs were calculated by using Pearson’s method. A soft-thresholding power of 7 was used by assuming the topology of the unsigned weighted gene co-expression network was scale-free. The average linkage method was used for all clustering procedures. The function cutTreeDynamic was used for identification of modules. Only modules with a minimum of 30 genes were considered. Modules with eigengene correlations no less than 0.75 were merged using the mergeCloseModules function with cutHeight = 0.25. The eigengene of a module is the first principal component of the gene expression values of that module and can be considered as a representation of the expression profiles of genes in the module [252]. We then fitted linear regression models with expression levels of the module eigengenes as the response variable and RFI group as the independent variable. The estimated effect of RFI group and the associated p-value for the null hypothesis that the RFI group term was not useful in explaining expression of the eigengene of the module were used to quantify the strength and significance of the association between the eigengene of a module and RFI group.
Hierarchical clustering, generation of heat map, and multi-dimension scaling analysis

The adjusted gene expression matrix was used for hierarchical clustering, heatmap generation, and multi-dimensional scaling (MDS) analyses. Spearman correlation coefficients for gene expression between samples were calculated and 1 minus this correlation coefficient was used as the distance between a pair of samples for both hierarchical clustering and heatmap construction. The Ward method was used in the function `hclust` for hierarchical clustering of the samples. A heatmap was generated with the `heat.map2` function to visualize the DEGs \((q \leq 0.05)\). MDS with the first two dimensions was used to visualize the relationships of samples with each other by using the function `plotMDS` from the Bioconductor package “limma” (version 3.20.9) [233]. For MDS, the distance between each pair of samples was the Euclidean distance between them based on the expression of all 12,280 genes.

Gene ontology term and pathway enrichment analyses

For all GO term and Ingenuity Pathway Analysis (IPA)-based pathway enrichment analyses, unadjusted \(p\)-values were reported [255]. We used Bioconductor package “topGO” (version 2.16.0) [255] to perform gene ontology (GO) term analysis, including GO biological process (GO-BP), GO molecular function (GO-MF) and GO cellular component (GO-CC). GO terms associated with each gene were downloaded from Ensembl Biomart (version 79). The “classical” algorithm, which treats all GO terms to be independent of each other, and Fisher’s exact test were used to estimate significance of such enrichment using the function `runTest`. For GO term enrichment analysis of DEGs \((q \leq 0.15)\), we analyzed DEGs with higher and lower expression in the low versus high RFI group separately, while for GO term enrichment analysis of WGCNA modules, genes in a whole module were analyzed together. The maximum subset of the 12,280 genes that was associated with at least one GO-BP, GO-MF or GO-CC term,
respectively, was used as the reference set (also known as background) in the corresponding GO term enrichment analyses. Significantly enriched GO terms associated with more than 10 annotated genes in the pig genome annotation (version 79) were reported.

We performed other enrichment analyses of the DEGs \((q \leq 0.15)\) using Ingenuity Pathway Analysis (IPA, 2015 spring release) and the Integrated Pathway Analysis Database (IPAD) for Systematic enrichment analysis [256]. For IPA-based analysis, 8,965 of the 12,280 genes could be mapped to IPA identifiers via the gene symbols of pig genes and these genes were used as the reference set. For networks and upstream regulator analysis, both direct and indirect relationships were considered. The options, “all data sources”, “confidence” and “mutation”, were checked. For species, “all mammals” was checked. For tissues and cell lines, only data on tissues and primary cells were considered. The cutoff for the \(\log_2\) (fold change) was set to 0 and the \(q\)-value cutoff was set to 0.15. All DEGs were analyzed together, with 1,488 of 1,972 DEGs mapped to IPA identifiers. For IPAD-based analysis of DEGs, the IPAD web server [257] was used. Of the 1,972 DEGs \((q \leq 0.15)\), 1,536 genes with human gene symbols were analyzed together. The default reference set was used, as the IPAD server does not allow the user to provide a reference set. The raw \(p\)-values were corrected for multiple testing using the “BH” method [227] for IPAD-based analysis.

**Validation by RT-qPCR**

We attempted to validate DEGs between the RFI groups with \(q \leq 0.15\), \(\mid \log_2 \) (fold change)\(\mid \geq 1\) and averaged FPKM (fragments per kilobase of exon per million fragments mapped) [258] within either RFI group \(\geq 1\), which resulted in 46 genes. All pairs of primers corresponding to the 46 DEGs and 6 internal reference genes were designed and synthesized by the Fluidigm Corporation (Fluidigm, San Francisco, CA), such that the two primers of each pair
were separated by exon-exon boundaries and could amplify all isoforms of the target gene if possible (see Table B.2). The efficiency of each primer pair was tested by conventional RT-qPCR on the DNA Engine Opticon 2 system (BioRad, Hercules, California) by using the standard curve method [259] and only primer pairs with amplification efficiency no less than 0.95 were further considered. As for specificity, only primer pairs that gave products with single peaks in melting curve analyses were used for the downstream RT-qPCR assays. This resulted in 37 pairs of primers for DEGs and two for internal reference genes (YWHAZ and RPL32) passing the test (Table B.2). The samples we used for RT-qPCR validation of the DEGs included 24 of the 31 RNA samples that we had sequenced by RNA-seq, and another 24 novel samples from gilts from the same contemporary group of pigs, with 6 samples from each line by diet combination. The 24 novel blood samples were selected such that the RFI phenotypes of the corresponding animals were evenly distributed across the distribution of RFI phenotypes that was not covered by the RFI phenotypes of the 24 animals originally selected for RNA-seq (Figure 3.1). Detailed information about the 48 samples used for validation of DEGs is in Table B.1. RNA was extracted from the novel samples as above in three batches, with two samples from each line by diet combination per batch. The concentration and quality of 24 novel RNA samples were determined as above. Total RNA without undergoing globin depletion was used for cDNA synthesis. By following the Fluidigm User Guide for Real-Time PCR Analysis [225], Real Time-qPCR was done on a 48.48 dynamic array chip (Fisher Scientific, Pittsburgh, PA) using the Biomarker HD system (Fluidigm, San Francisco, CA). Data were analyzed with the Fluidigm Real-Time PCR analysis software with the default settings, to obtain raw $C_t$ values. The raw $C_t$ values were corrected for differences in the amount of input RNA by using the geometric mean of the $C_t$ values of the two internal reference genes for the same RNA samples.
to get $-\Delta C_t = -(C_{t\text{ gene}} - C_{t\text{ reference}})$ [260, 261]. Differential expression analysis was performed by fitting linear models with $-\Delta C_t$ values as the response variables and RFI group, RNA extraction batch, and the linear covariates of the concentrations of neutrophils, lymphocytes, monocytes and basophils as independent variables. With the high RFI group as the reference, the estimated effects on $-\Delta C_t$ for RFI group were defined as the $-\Delta\Delta C_t$ values. The $p$-values associated with the effect of RFI group were adjusted to get $q$-values by using the “BH” method [227]. If the $q$-value of the significance test of the RFI group was less than 0.15, the corresponding gene was considered differentially expressed between the low and high RFI groups by RT-qPCR. Because the amplification efficiencies (see Table B.2) of the primers were close to 1, the fold changes of gene expression between the low and high RFI group were calculated as $2^{-\Delta\Delta C_t}$ [260].

**Association analysis of (eigen)gene expression with RFI phenotype**

For association analysis of the expression levels of eigengenes of modules identified by WGCNA with RFI phenotype, the eigengene expression levels were first analyzed with a linear model that included RFI group and diet as fixed effects and RFI phenotype as a covariate, along with the two-way and three-way interactions among these three factors. We did not include RNA processing batch, RIN, and concentrations of the blood cell types in the linear model because the eigengene expression levels were calculated based on gene expression values that had already been adjusted for these effects. Gene expression determined by RT-qPCR was also used to identify genes associated with RFI phenotype. The $-\Delta C_t$ values for the 37 target genes in the 48 samples were calculated as above and analyzed with a linear model that included RFI group, RNA extraction batch and diet as fixed effects, and RFI phenotype and the concentrations of neutrophils, lymphocytes, monocyte and basophils as covariates, along with the two-way and 3-
way interactions among RFI group, diet and RFI phenotype. However, none of the interaction terms were significant after correcting for multiple testing with the “BH” method ($q > 0.15$), thus interaction terms were removed from the models. The $p$-values associated with the regression coefficient(s) on the RFI phenotype covariate were adjusted to get $q$-values by the “BH” method [227]. If the $q$-value was less than 0.05, the association between the (eigen)gene and the RFI phenotype was considered significant.

**Results**

**Differentially expressed genes between low and high RFI groups**

Mauch *et al.* [248] showed that the low RFI line had significantly lower RFI than the high RFI line on the LFD ($p < 0.007$), but the low RFI line only tended to have lower RFI than the high RFI line on the HFD ($p > 0.05$). Thus, to maximize the contrast, we selected blood samples for RNA-seq from gilts of the low RFI line with extremely low RFI values when fed the LFD or the HFD (designated as low RFI group), and from gilts of the high RFI line with extremely high RFI values when fed the LFD or the HFD (designated as low RFI group) (Figure 3.1). The metadata, including RNA concentration, RINs before and after globin depletion, RFI phenotype and CBC test results are in Table B.1. The distribution of the RFI values for all gilts of each line by diet combination is in Figure 3.1. To increase the power to detect lowly expressed genes in the whole blood transcriptome, we depleted the most highly expressed transcripts, hemoglobin A and B (HBA and HBB), using the RNase H-mediated method [223]. Alignment of RNA-seq reads to the reference genome showed that globin transcripts had been effectively reduced (Table B.3). After globin depletion, on average, only $0.22 \pm 0.29\%$ and $1.86 \pm 2.70\%$ of the trimmed reads mapped to the HBA and HBB genes, respectively. The numbers of raw reads, trimmed reads, and mapped reads for each sample are in Table B.3. In summary, $20.4 \pm$
7.8 million pairs of 100-bp raw reads were sequenced per sample. After trimming, 81.4 ± 8.0% of raw paired-end reads were kept as paired-end reads, while 17.6 ± 8.0% of raw forward reads (R1) and 0.27 ± 0.05% of raw reverse reads (R2) were unpaired reads. 91 ± 1% of trimmed fragments were mapped to the pig reference genome, with 83.0 ± 1.2% (mean ± sd) of trimmed fragments being uniquely mapped. After removing genes with extremely low expression (See Methods), and HBA and HBB from the count table, we had expression data for 12280 genes for downstream analyses.

Mpetile et al. [184] showed that pigs at 35 to 42 days of age from the low RFI line had lower concentrations of lymphocytes, monocytes and basophils than those from the high RFI line in their peripheral blood. Therefore, we decided to account for the concentration of different blood cell types in our analyses so that we could adjust gene expression for these concentration differences. Although the animals were randomly assigned to one of the two diets after the blood samples were collected, we started with a full model that included diet and RFI phenotype (estimated RFI value) and performed backward selection to establish a final model for the following reasons: 1) the samples for RNA-seq were selected based on the RFI phenotype of the animals; 2) the RFI phenotype depended on both line and diet effects3) we were interested in identifying genes that might be associated with RFI phenotype. However, after accounting for RFI group (i.e. the low RFI and high RFI groups) and other variables identified via backward selection (batch, pre_conc, post_RIN, and the concentrations of neutrophils, lymphocytes, monocytes and basophils), we found the RFI phenotype, diet, pre_RIN, post_conc and the concentration of eosinophils were not significantly associated with the expression levels of most genes (p > 0.05). With our final selected model, we found expression levels of 836, 0, 3, and 42 genes to be significantly associated with the concentrations of neutrophils, lymphocytes,
monocytes and basophils, respectively \((q \leq 0.05)\) (Table B.4). However, we were most interested in DEGs whose expression differences were primarily dependent on line differences. Therefore, for all downstream analysis, we only considered DEGs between the low and high RFI groups, while accounting for the effects of relevant variables: batch, post_RIN, and the concentrations of neutrophils, lymphocytes, basophils and monocytes.

The numbers of DEGs between the low and high RFI groups based on different stringencies in terms of \(q\)-value and fold change cutoffs are in Table 3.1. A full list of the 1972 DEGs for \(q \leq 0.15\) is in Table B.4. Differential expression of the 454 DEGs with \(q \leq 0.05\) is shown in Figure 3.2A. As the volcano plot (Figure 3.2B) shows, the fold change between the low and high RFI groups was small for most genes. For a \(q\)-value cutoff of 0.05, only 50 DEGs had a fold change \(\geq 2\) or a fold change \(\leq 0.5\). However, the transcriptomes of the two groups of animals were collectively different, as the 31 samples were grouped into two clusters by RFI group based on hierarchical clustering and MDS analyses (Figure 3.2C and Figure B.1).

**Annotation of differentially expressed genes by GO term and pathway enrichment analyses**

To understand the biological differences between RFI groups based on the differential expression analyses, we carried out GO term analysis of DEGs using a less stringent cutoff of \(q\)-value \((q \leq 0.15)\) such that we could detect a broader group of significant GO terms. These analyses were conducted separately for DEGs with higher and lower expression in the low versus high RFI group. GO-BP term analysis showed that genes involved in small molecule biosynthesis, carboxylic acid biosynthesis, organic acid biosynthesis, steroid biosynthesis, antigen processing and presentation of peptide antigen via MHC (major histocompatibility complex) class I, and organic hydroxy compound biosynthesis were overrepresented among DEGs with higher expression in the low versus high RFI group \((p < 0.001, \text{Table 3.2})\). Among
the DEGs with lower expression in the low RFI group, genes involved in signal transduction, bone mineralization, regulation of phosphorylation, and phosphorylation were overrepresented \((p < 9\text{E}-05)\). Interestingly, GO-CC term analysis showed that genes functioning in the proteasome complex or the mitochondrion were enriched among the DEGs with higher expression in the low versus high RFI group \((p < 5\text{E}-06)\).

In addition, we conducted pathway enrichment analyses of the DEGs \((q \leq 0.15)\) by assuming pig genes have similar biological functions as their human orthologs. Detailed results from these IPA-based analyses are in Table B.5. In summary, IPA-based analysis suggested genes functioning in several canonical pathways tended to be enriched among the DEGs, including p53 signaling \((p = 2.18\text{E}-03)\), T cell receptor signaling \((p = 1.25\text{E}-02)\), antigen presentation \((p = 1.63\text{E}-02)\), IL-15 signaling \((p = 2.16\text{E}-02)\) and IL-9 signaling \((p = 2.21\text{E}-02)\). But after correcting for multiple testing, none of these pathways were significantly enriched among the DEGs \((q > 0.1)\). In addition, IPA-based analyses suggested that activities of genes involved in regulating the quantity of T lymphocytes \((p = 7.46\text{E}-05)\), consumption of oxygen \((p = 1.33\text{E}-04)\), insulin resistance \((p = 1.98\text{E}-04)\), cell survival \((p = 4.67\text{E}-04)\), glucose metabolism and transport \((p < 5.64\text{E}-03)\), respiration of mitochondrion \((p = 5.74\text{E}-03)\) and body mass index \((p = 1.03\text{E}-02)\) were different between the two RFI groups. IPA-based analysis also suggested several potential upstream regulators of the DEGs, including IL15, which had a higher expression in the low versus high RFI group \((p < 3.6\text{E}-03)\). The inferred regulatory network for IL15 is shown in Figure B.2.

Using the Integrated Pathway Analysis Database (IPAD) for Systematic Enrichment Analysis [256], we found genes involved in the immune system (pathway ID: 168256) \((q = 0.013)\) and metabolism (pathway ID: 1430728) \((q = 0.066)\) were enriched among the DEGs.
IPAD-based disease-associated gene enrichment analysis suggested genes involved in several diseases were overrepresented among DEGs ($q < 1E-05$), including taste disorder, eating disorder, anorexia, hyperphagia, obesity, insulin resistance, mitochondrial diseases and lymphocytosis. Details are in Table B.6. This is consistent with the phenotypic differences in feeding behavior [144], body composition [147], and growth rate [145] that have been observed between the two RFI lines.

**Validation of DEGs**

We used RT-qPCR to validate the differential expression of 37 DEGs that were selected based on average expression levels within RFI group (FPKM > 1), differential expression levels between RFI groups ($|\log_2\text{ (fold change)}| \geq 1$ and $q \leq 0.15$) and primer performance. To test whether the expression differences we detected were due to differences between the lines, rather than due to the comparison of pigs of the low RFI group to pigs of the high RFI group, we selected another 24 samples from the same parity of the 9th generation of the two RFI lines. These 24 samples were selected such that the RFI values of the corresponding animals were roughly evenly distributed across the ranges of RFI values for each line by diet combination, which were not covered by the RFI values of the 24 animals originally selected for RNA-seq (Figure 3.1). The RNA from these new samples as well as the 24 samples used for RNAseq analyses were used in RT-qPCR analyses. Data from the original 24 samples and the new 24 samples were analyzed independently as well as jointly. A total of 24 of the 37 DEGs were confirmed in the joint analysis of all 48 samples ($q < 0.15$, Figure 3.3), while 22 of the 37 DEGs were confirmed when analyzing the original 24 samples ($q < 0.15$). However, only 9 of the 37 DEGs were validated in the analysis of the novel 24 samples ($q < 0.15$, Table B.7).
Based on the RT-qPCR results of the 48 samples, we found the expression of 5 of the 37 DEGs (determined by RNA-seq) to be significantly associated with RFI phenotype ($q < 0.05$). They were LRP6 (low density lipoprotein receptor-related protein 6), ENSSSCG00000024900 (T-cell receptor beta chain), ENSSSCG00000008771, PDL1 (CD274 molecule) and ENSSSCG00000020945 (Table B.8).

**Co-expressed gene modules associated with RFI**

Since conventional differential expression analysis considers each gene independently and suffers from loss of power due to correction for multiple testing, we performed weighted gene co-expression gene network analysis (WGCNA) [252] based on the adjusted gene expression matrix (see Methods) to identify modules (groups of co-expressed genes) differentially expressed between the RFI groups and associated with RFI phenotype. We identified four modules, designated as C1-lightcyan, C2-darkturqoise, C3-skyblue3, and C4-black, whose eigengene expression levels were significantly differentially expressed between the low and high RFI groups (Table 3.3 and Table B.9). The eigengene of a module is a weighted average of the expression profiles of genes in the module, calculated as the first principal component score. The expression levels of the eigengenes for the 31 samples are shown in Figure 3.4A. Among the genes in module C1-lightcyan, those involved in lipid metabolism-related biological processes, such as lipid metabolic process, lipid biosynthesis and steroid biosynthesis were overrepresented ($p < 0.006$). In module C2-darkturqoise, genes involved in biological processes related to bone mineralization, immunity and stress response and lipid metabolism were overrepresented (Table B.10). We also found that DEGs identified by RNA-seq ($q \leq 0.15$) significantly overlapped with each of the four differentially expressed modules (Figure 3.4B), which suggests that the modules identified by WGCNA are not computational artifacts as
they were enriched for the DEGs between the RFI lines. In addition, we found that the expression levels of the eigengenes of modules C3-skyblue3 were significantly associated with RFI phenotype \((q < 0.05)\) (Table B.11).

**Discussion**

**Small but significant blood transcriptomic differences between the low and high RFI groups**

By transcriptomic analyses, we identified 454, 1185 and 1972 DEGs between the low and high RFI groups at \(q\)-value cutoffs of 0.05, 0.1 and 0.15, respectively. We validated 24 of the 37 selected DEGs by RT-qPCR in a joint analysis of 48 samples: 24 of the 31 samples used for RNA-seq plus 24 novel samples from the same population (Figure 3.1, Figure 3.3 and Table B.7). We validated 22 of the 37 selected DEGs by RT-qPCR when using 24 of the 31 samples that were used for RNA-seq, but only 9 of the 37 selected DEGs were validated when using only the 24 novel samples (Figure 3.1 and Table B.7). Our validation rate by RT-qPCR of DEGs detected by RNA-seq was low compared to the high validation rate reported by Wang *et al.*[262], especially when based on the 24 novel samples. There are several possible reasons for failure to validate some of the 37 DEGs that were detected by RNA-seq by RT-qPCR in this study. First, some of the selected DEGs could be false positives, especially as we selected DEGs with \(q \leq 0.15\) for validation, although we did not detect an obvious association between the validation rate and the \(q\) values for the DEGs. Second, the primers used for RT-qPCR maybe did not quantify all isoforms that were measured by RNA-seq. Although the primers were designed to amplify all isoforms, these designs were based on poorly annotated gene models: most of the 37 genes only have one isoform in the Ensembl pig genome annotation (Table B.2). Third, the 24 novel samples used for RT-qPCR were not the exact biological replicates of the 24 samples used
for RNA-seq given their different RFI phenotype. In addition, as observed in other differential expression analyses between low and high RFI groups conducted in poultry, pigs and cattle [158, 263-266], the magnitude of the differential expression in terms of fold change was generally small for most genes. This could be because the experimental units for the low and high RFI groups were genetically heterogeneous within each group, which cancelled out some differences between extreme individuals of the two groups. In addition, the experiments were conducted without strong external stimuli, thus the expression levels of genes were in their normal physiological range. It was also possible that the differences in gene expression in blood at the post-weaning stage were not as dramatic as during the grow-finish phase when feed efficiency was tested. However, it is possible that small changes in the transcript abundance of some genes involved in metabolism, transcription regulation, and signal transduction, could have significant effects on RFI phenotype. As our differential analysis indicated, a significant portion of genes involved in several biosynthetic processes, signal transduction, and regulation of phosphorylation were differentially expressed between the low and high RFI groups. Lastly, we cannot exclude the possibility that there is a weak or indirect relationship between early blood gene expression and later RFI phenotype. Nevertheless, based on the adjusted global gene expression profiles of all samples, hierarchical clustering and MDS analyses revealed that the post-weaning blood transcriptome was different between the two RFI groups.

Potential relationship of the DEGs and differentially co-expressed gene modules in blood with RFI

Although we did not attempt to validate most of the DEGs or the differentially co-expressed gene modules between the two RFI groups due to the generally small fold changes, we found some interesting biological processes and cellular components underlying them, which suggested their potential relationships with RFI phenotype. GO term enrichment analysis of
DEGs suggested that genes involved in small molecule (including organic acid, carboxylic acid, and alcohol) biosynthesis, antigen processing and presentation of peptide antigen via MHC class I, and steroid biosynthesis were enriched among DEGs with higher expression in the low versus high RFI group. DEGs involved in small molecular biosynthesis included FADS1, FADS2, and ELOVL3. FADS1 and FADS2 are important genes in regulating the synthesis of polyunsaturated fatty acids, which have pleiotropic influences on health and diseases by functioning in several pathways, including metabolism and immunity [267]. Of note, FADS2 also had a higher expression level in the liver in the low RFI versus high RFI group of Nelore steers [263]. ELOVL3 encodes one of the rate-limiting enzymes in elongation of very long chain fatty acids (with more than 17 carbon atoms). In mice, a proposed physiological function of ELOVL3 is to maintain lipid homeostasis by replenishing the intracellular pool of triacylglycerol [268].

MAT2B, is one of the genes associated with steroid biosynthesis but had lower expression in the low versus high RFI group. It encodes a regulatory subunit of methionine adenosyltransferase, MAT II, which catalyzes the synthesis of S-adenosyl methionine (SAM). SAM is a key methyl donor in transmethylation reaction and polyamine biosynthesis and also functions as a cofactor in key metabolic pathways. Down-regulation of MATIIB expression causes a 6 to10-fold increase in intracellular SAM levels [269]. Taken together, the expression difference of genes involved in small molecule biosynthesis and steroid biosynthesis might have significant effects on overall metabolism, immunity and beyond.

Seven of the 20 genes associated with GO term GO:0002474 (antigen processing and presentation of peptide antigen via MHC class I) had higher expression in the low RFI group. These included SLA-3 and TAP1. It has been shown that in post-weaning pigs resistant to *Escherichia coli* F18, SLA-3 has higher expression level than in susceptible pigs [270]. TAP1
transports antigens from the cytoplasm to the endoplasmic reticulum for loading antigen peptide onto MHC class I molecules [271]. In addition, we found 14 of the 15 DEGs associated with GO term GO:0000502 (proteasome complex) to have higher expression in the low versus high RFI group. The only exception was PSMD5, which had lower expression in the low versus high RFI group. It has been shown that overexpression of PSMD5 inhibits assembly and activity of 26S proteasome [272], which is an important component in processing and presenting intracellular peptide antigen via MHC class I to CD8+ T cells. Antigen presenting cells (APC) have both constitutive proteasomes and immunoproteasomes, both of which are important for antigen peptide processing [273]. Notably, an immunoproteasome subunit PSMB8 (also known as β5i) had the largest difference in expression level among the 15 DEGs that function in the proteasome complex. Given the increased expression levels of non-inhibitory proteasomal components in blood of the low RFI group, it is likely that proteasomal activity is higher in blood from the low versus high RFI group. Taken together, the expression differences of genes involved in antigen processing and peptide antigen presentation suggest that low RFI animals might have a more active system of antigen peptide processing and presentation, which might improve the robustness of their immune system of the low RFI animals. Consistent with this speculation, Dunkelberger et al. [274] recently showed that the low RFI line was less affected by experimental infection with porcine reproductive and respiratory syndrome virus (PRRSv), which mainly infects mature macrophages, derivatives of monocytes. In view of the lower concentration of monocytes in the low RFI line, the putative higher activity of antigen processing and peptide antigen presentation might lead to an increased efficiency of innate immunity in the low RFI pigs. It would be interesting to test the activity of the blood proteasome in more detail in these lines for a role in feed efficiency.
Another interesting finding was that genes functioning in the mitochondrion were significantly overrepresented among DEGs with higher expression in the low versus high RFI group. Of the 955 blood genes associated with GO-CC term mitochondrion (GO:0005739), 116 were DEGs with higher expression in the low versus high RFI group. Mitochondria play important roles in many biological processes, including fatty acid metabolism, amino acid metabolism, and steroidogenesis [275]. Importantly, mitochondrial function has been associated with feed efficiency in poultry, cattle and lambs [276, 277]. As well, it has been shown that the mitochondrial proteome profiles are different between these low and high RFI lines of pigs and the mitochondria from the liver and LD muscle of the low RFI pigs produced less ROS [149-151]. Interestingly, we found that four genes involved in detoxification of ROS, GPX3, SOD1 and CAT, had higher expression in the low versus high RFI group. Another interesting gene that functions in mitochondria is G0S2, which is a noncompetitive inhibitor of adipose lipase, a rate-limiting lipase of triglyceride hydrolysis [278]. G0S2 had higher expression in the low versus high RFI group. It has been shown that G0S2 can inhibit ROS production in endothelial cells [279]. In addition, we found higher expression in blood from the low RFI group for genes that encode 14 mitochondrial ribosomal proteins (MRPL3, 9, 10, 11, 18, 20, 24, 38, 40, 46MRPS10, 11,15, 27), 9 components of mitochondrial complex I (NDUFA4, 8NDUFB3, 7, 9NDUFS2, 7NDUFAB1, NDUFAF3), one component of complex II (SDHA), cytochrome c (CYCS) and four components of complex IV (COA3, COA6, COX5B, COX7C), two components of the complex V (ATP5D, ATP5G3), four key enzymes of the TCA cycle (ACO1, IDH2, IDH3A, SDHA, SUCLG1), four enzymes involved in mitochondrial DNA repair (APEX2, FEN1, OGG1, UNG), and two enzymes involved in nucleotide metabolism (DUT, MTHFD1), and three translocases of the inner/outer membranes (TIMM50, TOMM7, TOMM40). Taken together, it is
reasonable to speculate that the low RFI group has more efficient mitochondria and can more efficiently handle oxidative stress.

Among the 1972 DEGs ($q \leq 0.15$), 288 genes were associated with GO:0007165 (signal transduction), and genes involved in signal transduction were significantly enriched among DEGs with lower expression in the low versus high RFI group. DEGs associated with signal transduction were mainly cytokines, receptors and kinases. Signal transduction-related genes with much lower expression in the low versus high RFI group included LRP6, WNT10B, and FZD6. LRP6 belongs to the low-density lipoprotein receptor family and plays a key role in lipoprotein endocytosis and as a co-receptor in Wnt/β–catenin signaling. LRP6 is also involved in regulating lipid homeostasis and body fat mass [280], while LRP6, WNT10B and FZD6 can function together in the Wnt/β-catenin signaling pathway. Besides playing important roles in bone metabolism, the Wnt/β-catenin signaling pathway also takes part in glycolysis and regulates mitochondrial physiology and insulin sensitivity and is thus linked to metabolic diseases [281, 282]. On the other hand, we found a few genes involved in signal transduction but with much higher expression in the low versus high RFI group, such as IL15 and IRS1. IL15 plays important roles as a pleiotropic cytokine in innate and adaptive immunity [283]. IRS1 is an important player in both the insulin signaling and the IGF-1 signaling pathways and IGF-1 is an important anabolic hormone, playing a key role in growth. Note that juveniles from the low RFI line have been shown to have lower serum concentration of IGF-1 and the low RFI line grows only slightly slower than the high RFI line. The higher expression of IRS1 might partially compensate for the lower level of IGF-1 in the low RFI line. It is interesting to investigate the expression and activity of IRS1 in other tissues of the two lines, such as the muscle. The differential expression of many genes associated with signal transduction might be related to
differences in metabolism and immune response between the RFI lines. Lastly, we found four co-expression modules to be differentially expressed between the RFI groups, and these modules shared a significant portion of their genes with the list of DEGs. The eigengene expression of module C1-lightcyan was lower in the low RFI group than in the high RFI group. The top highly connected genes in this module were MAT2B, CPT1A, and LRP6. These genes were also differentially expressed between the low and high RFI groups. Since we have already discussed MAT2B and LRP6, we focus on CPT1A herein. CPT1A had lower expression in the low versus high RFI group and plays an important role in importing long chain fatty acids into mitochondria, by catalyzing the primary step of mitochondrial fatty acid oxidation [284]. Inhibition of hypothalamic CPT1A has been shown to decrease feed intake and glucose production in rats [285]. Thus it will be interesting to investigate its expression in other tissues including the hypothalamus. In addition, GO-BP term analysis showed that genes involved in lipid metabolism-related processes were enriched in module C1-lightcyan. Because low RFI pigs have less body fat and are leaner [147] and that deposition of energy as fat costs more energy than as protein[286], genes in this module might be relevant to differences in RFI.

Potential predictive biomarkers for RFI

DEGs whose expression levels are associated with the phenotype of interest, as well as genes whose expression profiles are similar to the eigengenes of co-expression modules associated with the phenotype are often considered as good candidate biomarkers for the phenotype. Based on our association tests, the best candidate biomarkers included LRP6, ENSSSCG00000024900, ENSSSCG00000008771, PDL1 and ENSSSCG00000020945. LRP6 and ENSSSCG00000024900 were ones of the DEGs between the two RFI groups and ones of the highly connected, eigengene-like genes in modules C1-lightcyan and C2-darkturquoise,
respectively (Table B.9). Notably, we only found a very few genes whose expression levels were significantly associated with RFI phenotype, which might be because the estimated RFI values were residuals which represented both random errors and true feed efficiency and thus were not accurate, or because there were not many genes whose post-weaning expression levels in blood were well associated with the RFI values measured during the grow-finish phase in blood, or because there may be little or no relationship of blood gene expression in blood and RFI phenotype regardless of age.

Conclusions

As far as we know, this is the first study to explore transcriptomic differences in blood between pig lines with divergent RFI by RNA-seq. We found that the blood transcriptome was clearly different between the low and high RFI groups, although only a small number of genes showed large fold changes of expression between the two groups. The two RFI groups may be different in mitochondrial and proteasomal activities, small molecule biosynthetic process, and signal transduction. These blood transcriptomic differences may be related to the difference in feed efficiency between these two groups, although the observed relationships of post-weaning blood gene expression with RFI phenotype measured during the grow-finish phase were not strong. The top candidate biomarkers for predicting RFI included LRP6, ENSSSCG00000024900, ENSSSCG00000008771, PDL1 and ENSSSCG00000020945.
Table 3.1. The number of differentially expressed genes identified at different $q$-value and fold change cutoffs.

| Cutoff of $q$-value | $|\log_2(FC)| > 0$ $^a$ | $|\log_2(FC)| \geq \log_2 1.5$ | $|\log_2(FC)| \geq 1$ |
|---------------------|------------------------|-------------------------------|---------------------|
| $q \leq 0.05$       | 454                    | 140                           | 50                  |
| $q \leq 0.1$        | 1185                   | 266                           | 80                  |
| $q \leq 0.15$       | 1972                   | 344                           | 93                  |

$^a$FC, fold change calculated as the ratio of mean gene expression in the low RFI group to mean gene expression in the high RFI group after accounting for the other relevant variables (see Methods).
### Table 3.2. GO terms overrepresented among differentially expressed genes.

<table>
<thead>
<tr>
<th>ID</th>
<th>Domain</th>
<th>Term</th>
<th>Annotated</th>
<th>Significant</th>
<th>Expected</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0044283</td>
<td>BP</td>
<td>small molecule biosynthetic process</td>
<td>170</td>
<td>33</td>
<td>13.77</td>
<td>1.70E-06</td>
</tr>
<tr>
<td>GO:0046394</td>
<td>BP</td>
<td>carboxylic acid biosynthetic process</td>
<td>113</td>
<td>22</td>
<td>9.15</td>
<td>8.70E-05</td>
</tr>
<tr>
<td>GO:0016053</td>
<td>BP</td>
<td>organic acid biosynthetic process</td>
<td>113</td>
<td>22</td>
<td>9.15</td>
<td>8.70E-05</td>
</tr>
<tr>
<td>GO:0006694</td>
<td>BP</td>
<td>steroid biosynthetic process</td>
<td>44</td>
<td>12</td>
<td>3.56</td>
<td>0.00014</td>
</tr>
<tr>
<td>GO:0046165</td>
<td>BP</td>
<td>alcohol biosynthetic process</td>
<td>48</td>
<td>12</td>
<td>3.89</td>
<td>0.00033</td>
</tr>
<tr>
<td>GO:0002474</td>
<td>BP</td>
<td>antigen processing and presentation of peptide antigen via MHC class I</td>
<td>20</td>
<td>7</td>
<td>1.62</td>
<td>0.00067</td>
</tr>
<tr>
<td>GO:1901617</td>
<td>BP</td>
<td>organic hydroxy compound biosynthetic process</td>
<td>60</td>
<td>13</td>
<td>4.86</td>
<td>0.00084</td>
</tr>
<tr>
<td>GO:0004298</td>
<td>MF</td>
<td>threonine-type endopeptidase activity</td>
<td>15</td>
<td>9</td>
<td>1.18</td>
<td>3.50E-07</td>
</tr>
<tr>
<td>GO:0005839</td>
<td>CC</td>
<td>proteasome core complex</td>
<td>15</td>
<td>9</td>
<td>1.23</td>
<td>5.10E-07</td>
</tr>
<tr>
<td>GO:000502</td>
<td>CC</td>
<td>proteasome complex</td>
<td>39</td>
<td>14</td>
<td>3.2</td>
<td>1.20E-06</td>
</tr>
<tr>
<td>GO:0005739</td>
<td>CC</td>
<td>Mitochondrion</td>
<td>955</td>
<td>116</td>
<td>78.35</td>
<td>4.60E-06</td>
</tr>
<tr>
<td>GO:0005689</td>
<td>CC</td>
<td>U12-type spliceosomal complex</td>
<td>17</td>
<td>8</td>
<td>1.39</td>
<td>2.50E-05</td>
</tr>
<tr>
<td>GO:0043227</td>
<td>CC</td>
<td>membrane-bounded organelle</td>
<td>5518</td>
<td>491</td>
<td>452.73</td>
<td>0.00039</td>
</tr>
<tr>
<td>GO:0032991</td>
<td>CC</td>
<td>macromolecular complex</td>
<td>2416</td>
<td>237</td>
<td>198.23</td>
<td>0.00041</td>
</tr>
</tbody>
</table>
(Table 3.2 Continued)

GO terms overrepresented among DEGs with lower expression in the low RFI group than in the high RFI group

<table>
<thead>
<tr>
<th>ID</th>
<th>Domain</th>
<th>Term</th>
<th>Annotated&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Significant&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Expected&lt;sup&gt;c&lt;/sup&gt;</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0065007</td>
<td>BP</td>
<td>biological regulation</td>
<td>4577</td>
<td>285</td>
<td>237.14</td>
<td>1.10E-06</td>
</tr>
<tr>
<td>GO:0050789</td>
<td>BP</td>
<td>regulation of biological process</td>
<td>4395</td>
<td>272</td>
<td>227.71</td>
<td>7.00E-06</td>
</tr>
<tr>
<td>GO:0007165</td>
<td>BP</td>
<td>signal transduction</td>
<td>2117</td>
<td>149</td>
<td>109.68</td>
<td>9.70E-06</td>
</tr>
<tr>
<td>GO:0030282</td>
<td>BP</td>
<td>bone mineralization</td>
<td>45</td>
<td>11</td>
<td>2.33</td>
<td>1.30E-05</td>
</tr>
<tr>
<td>GO:0023052</td>
<td>BP</td>
<td>Signaling</td>
<td>2276</td>
<td>156</td>
<td>117.92</td>
<td>2.50E-05</td>
</tr>
<tr>
<td>GO:0044700</td>
<td>BP</td>
<td>single organism signaling</td>
<td>2276</td>
<td>156</td>
<td>117.92</td>
<td>2.50E-05</td>
</tr>
<tr>
<td>GO:0007166</td>
<td>BP</td>
<td>cell surface receptor signaling pathway</td>
<td>979</td>
<td>79</td>
<td>50.72</td>
<td>2.90E-05</td>
</tr>
<tr>
<td>GO:0031214</td>
<td>BP</td>
<td>biomineral tissue development</td>
<td>49</td>
<td>11</td>
<td>2.54</td>
<td>3.10E-05</td>
</tr>
<tr>
<td>GO:0050794</td>
<td>BP</td>
<td>regulation of cellular process</td>
<td>4126</td>
<td>254</td>
<td>213.77</td>
<td>4.40E-05</td>
</tr>
<tr>
<td>GO:0007154</td>
<td>BP</td>
<td>cell communication</td>
<td>2328</td>
<td>157</td>
<td>120.61</td>
<td>5.70E-05</td>
</tr>
<tr>
<td>GO:0042325</td>
<td>BP</td>
<td>regulation of phosphorylation</td>
<td>525</td>
<td>48</td>
<td>27.2</td>
<td>7.10E-05</td>
</tr>
<tr>
<td>GO:0016310</td>
<td>BP</td>
<td>Phosphorylation</td>
<td>1024</td>
<td>80</td>
<td>53.05</td>
<td>8.10E-05</td>
</tr>
<tr>
<td>GO:0044763</td>
<td>BP</td>
<td>single-organism cellular process</td>
<td>5419</td>
<td>315</td>
<td>280.76</td>
<td>0.00021</td>
</tr>
<tr>
<td>GO:0018212</td>
<td>BP</td>
<td>peptidyl-tyrosine modification</td>
<td>125</td>
<td>17</td>
<td>6.48</td>
<td>0.00023</td>
</tr>
<tr>
<td>GO:001932</td>
<td>BP</td>
<td>regulation of protein phosphorylation</td>
<td>407</td>
<td>38</td>
<td>21.09</td>
<td>0.00027</td>
</tr>
<tr>
<td>GO:006468</td>
<td>BP</td>
<td>protein phosphorylation</td>
<td>774</td>
<td>62</td>
<td>40.1</td>
<td>0.0003</td>
</tr>
<tr>
<td>GO:0051716</td>
<td>BP</td>
<td>cellular response to stimulus</td>
<td>2670</td>
<td>171</td>
<td>138.33</td>
<td>0.00041</td>
</tr>
</tbody>
</table>
(Table 3.2 Continued)

<table>
<thead>
<tr>
<th>ID</th>
<th>Domain</th>
<th>Term</th>
<th>Annotated&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Significant&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Expected&lt;sup&gt;c&lt;/sup&gt;</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0060070</td>
<td>BP</td>
<td>canonical Wnt signaling pathway</td>
<td>86</td>
<td>13</td>
<td>4.46</td>
<td>0.00044</td>
</tr>
<tr>
<td>GO:0016477</td>
<td>BP</td>
<td>cell migration</td>
<td>405</td>
<td>37</td>
<td>20.98</td>
<td>0.0005</td>
</tr>
<tr>
<td>GO:0051239</td>
<td>BP</td>
<td>regulation of multicellular organismal process</td>
<td>841</td>
<td>65</td>
<td>43.57</td>
<td>0.00056</td>
</tr>
<tr>
<td>GO:0018108</td>
<td>BP</td>
<td>peptidyl-tyrosine phosphorylation</td>
<td>123</td>
<td>16</td>
<td>6.37</td>
<td>0.00058</td>
</tr>
<tr>
<td>GO:2000026</td>
<td>BP</td>
<td>regulation of multicellular organismal developement</td>
<td>526</td>
<td>44</td>
<td>27.25</td>
<td>0.00099</td>
</tr>
<tr>
<td>GO:0040011</td>
<td>BP</td>
<td>Locomotion</td>
<td>511</td>
<td>43</td>
<td>26.48</td>
<td>0.00101</td>
</tr>
<tr>
<td>GO:0060089</td>
<td>MF</td>
<td>molecular transducer activity</td>
<td>409</td>
<td>41</td>
<td>21.42</td>
<td>4.30E-05</td>
</tr>
<tr>
<td>GO:0044459</td>
<td>CC</td>
<td>plasma membrane part</td>
<td>537</td>
<td>45</td>
<td>27.48</td>
<td>0.00063</td>
</tr>
</tbody>
</table>

<sup>a</sup>Number of genes detected in the blood and associated with a given GO term.
<sup>b</sup>Number of DEGs associated with a given GO term.
<sup>c</sup>Expected number of DEGs associated with a given GO term.
Table 3.3. Summary of WGCNA modules differentially expressed between the low and high RFI groups.

<table>
<thead>
<tr>
<th>ID</th>
<th>Size</th>
<th>Reg. coef.(^a)</th>
<th>(p)-value</th>
<th>Adjusted (R^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1_lightcyan</td>
<td>198</td>
<td>-0.33</td>
<td>4.7E-13</td>
<td>0.83</td>
</tr>
<tr>
<td>C2_darkturquoise</td>
<td>142</td>
<td>0.30</td>
<td>2.9E-09</td>
<td>0.70</td>
</tr>
<tr>
<td>C3_skyblue3</td>
<td>89</td>
<td>0.29</td>
<td>2.7E-08</td>
<td>0.65</td>
</tr>
<tr>
<td>C4_black</td>
<td>786</td>
<td>-0.28</td>
<td>3.9E-07</td>
<td>0.58</td>
</tr>
</tbody>
</table>

\(^a\)Reg. coef., regression coefficient estimated by regressing the expression level of the eigengene of a module on RFI groups, with the high RFI group as the reference.
Figure 3.1. Distribution of RFI values and sample selection for RNA-seq and RT-qPCR assays. Post-weaning blood samples of 16 pigs (8 per diet) with extremely low RFI from gilts of the low RFI line (LRFI) and 16 pigs (8 per diet) with extremely high RFI from gilts of the high RFI line (HRFI) fed the high-fiber, low-energy diet (HFD) or the low-fiber, high-energy diet (LFD) were selected for RNA-seq. One sample in the LRFI-LFD group was excluded from RNA sequencing because of low quality RNA, leaving a total of 31 samples for RNA-seq. The green and red dots represent individuals selected for RNA-seq. Twelve samples from each line by diet combination were selected for RT-qPCR validation of DEGs such that the corresponding RFI phenotypes were representative. The red and blue dots represent the samples selected for RT-qPCR assays. The 24 novel blood samples were selected such that the RFI values of the corresponding animals were roughly evenly distributed across the ranges of RFI not covered by the RFI phenotypes of the 24 animals originally selected for RNA-seq. The distribution of RFI values of barrows from each line by diet combination is also shown for reference.
Figure 3.2. Differentially expressed genes and transcriptomic differences between the low and high RFI groups. (A) Heatmap showing 454 DEGs ($q \leq 0.05$) between low and high RFI groups identified by RNA-seq. Sample names are designated as RFI line followed by the pig identifier. LRFI, low RFI line; HRFI, high RFI line. Animals with sample names in blue were fed the high-fiber, low-energy diet (LFD), while animals with sample names in black were fed the low-fiber, high-energy diet (HFD). The relative orders of genes and samples were determined by two-way hierarchical clustering based on the adjusted transformed gene expression of the 454 DEGs. The adjusted gene expression was gene-wise
standardized to get the z-score as displayed. (B) Volcano plot showing the magnitude and significance of differential expression of genes between low and high RFI groups. Black vertical dash lines correspond to $|\log_2 (\text{fold change})| = 1$, while red horizontal dash line correspond to $q$-value of 0.15. FC, fold change calculated as the ratio of mean gene expression in the low RFI group to mean gene expression in the high RFI group after accounting for the other relevant variables. (C) Hierarchical clustering showing relationship of the 31 RNA-seq samples. The samples were hierarchically clustered by using Ward method with 1 minus Spearman correlation as distance. The Spearman correlations between pairs of samples were calculated based on the adjusted transformed expression of the 12,280 genes.
Figure 3.3. Validation of DEGs by RT-qPCR. 24 of the 37 selected DEGs between low and high RFI groups were confirmed by RT-qPCR when using the 48 samples, 24 of which were used for RNA-seq and another 24 of which were novel, as shown in this figure (q ≤ 0.15). 22 of the 37 selected DEGs were confirmed by RT-qPCR when using the 24 samples that were used for RNA-seq (q ≤ 0.15) but only 9 of the 37 selected DEGs were validated by RT-qPCR when using the 24 novel samples (q ≤ 0.15) (Table B.7). For comparison, the log₂ (fold change) of these genes determined by RNA-seq were also displayed. Genes were ordered based on their log₂ (fold change) as determined by RT-qPCR for display. Error bars for RT-qPCR assays show the standard errors of mean log₂ (fold change). DEGs not confirmed by RT-qPCR are labeled in red. Genes without corresponding human orthologs are labeled with the last 5 digits of their Ensembl gene IDs, with common prefix “ENSSSCG000000” omitted for simplicity. For example, the ID for “29500” should be ENSSSCG0000029500.
Figure 3.4. Co-expression modules differentially expressed between the low and high RFI group. (A) Distribution of eigengenes of modules highly associated with RFI groups. (B) Venn diagram showing overlapping between module genes and DEGs.
CHAPTER 4. A HIGH-QUALITY ANNOTATED TRANSCRIPTOME OF SWINE PERIPHERAL BLOOD

Modified from a paper submitted to *BMC Genomics*

Haibo Liu, Timothy P.L. Smith, Dan J. Nonneman, Jack C.M. Dekkers,
Christopher K. Tuggle

**Abstract**

**Background:** High throughput gene expression profiling assays of peripheral blood are widely used in biomedicine, as well as in animal genetics and physiology research. Accurate, comprehensive, and precise interpretation of such high throughput assays relies on well-characterized reference genomes and/or transcriptomes. However, neither the reference genome nor the peripheral blood transcriptome of the pig have been sufficiently assembled and annotated to support such profiling assays in this emerging biomedical model organism. We aimed to assemble published and novel RNA-seq data to provide a comprehensive, well-annotated blood transcriptome for pigs by integrating a *de novo* assembly with a genome-guided assembly.

**Results:** A *de novo* and a genome-guided transcriptome of porcine whole peripheral blood was assembled with ~162 million pairs of paired-end and ~183 million single-end, trimmed and normalized Illumina RNA-seq reads (~6 billion initial reads from 146 RNA-seq libraries) from five independent studies by using the Trinity and Cufflinks software, respectively. We then removed putative transcripts (PTs) of low confidence from both assemblies and merged the remaining PTs into an integrated transcriptome consisting of
132,928 PTs, with 126,225 (~95%) PTs from the *de novo* assembly and more than 91% of PTs spliced. In the integrated transcriptome, ~90% and 63% of PTs had significant sequence similarity to sequences in the NCBI NT and NR databases, respectively; 68,754 (~52%) PTs were annotated with 15,965 unique GO terms; and 7,618 PTs annotated with Enzyme Commission codes were assigned to 134 KEGG pathways. Full exon-intron junctions of 17,528 PTs were validated by PacBio IsoSeq full-length cDNA reads from 3 other porcine tissues, NCBI pig RefSeq mRNAs and transcripts from Ensembl *Sscrofa10.2* annotation. Completeness of the 5’ termini of 37,569 PTs was validated by public cap analysis of gene expression (CAGE) data. By comparison to the Ensembl transcripts, we found that (1) the deduced precursors of 54,402 PTs shared at least one intron or exon with those of 18,437 Ensembl transcripts; (2) 12,262 PTs had both longer 5’ and 3’ termini than their maximally overlapping Ensembl transcripts; and (3) 41,838 spliced PTs were totally missing from the *Sscrofa10.2* annotation. Similar results were obtained when the PTs were compared to the pig NCBI RefSeq mRNA collection.

**Conclusions:** We built, validated and annotated a comprehensive porcine blood transcriptome with significant improvement over the annotation of Ensembl *Sscrofa10.2* and the pig NCBI RefSeq mRNAs, and laid a foundation for blood-based high throughput transcriptomic assays in pigs and for advancing annotation of the pig genome.

**Keywords:** *Sus scrofa*, peripheral blood, *de novo* transcriptome assembly, genome-guided transcriptome assembly
Background

In higher animals, peripheral blood is a complex and informative tissue type, consisting of acellular plasma and multiple types of cells at various differentiation states [177, 287]. Peripheral blood cells can be roughly classified into anucleated red blood cells (RBCs) and platelets, and nucleated white blood cells (WBCs), with WBCs most transcriptionally active [287-293]. The WBCs can be further divided into neutrophils, eosinophils, basophils, B lymphocytes, T lymphocytes and monocytes. Each cell type is composed of continuously differentiating subtypes, leading to variability in RNA content of the overall peripheral blood transcriptome. Peripheral blood is also a highly dynamic tissue type with a high rate of cell turnover, resulting in high diversity of RNA content over time [294]. In addition, peripheral blood interacts with every organ and tissue in the body, and consequently presents gene expression profiles that can reflect the physiopathological status, behaviors, growth stage and lifestyle of subjects [177, 178, 180, 181, 295-298]. Combined with easy and minimally invasive accessibility, the reflective relationship to body status makes peripheral blood a highly desirable tissue type for disease prediction, diagnosis, monitoring, prognosis, and biomarker development [177, 299].

Due to the high similarity in anatomy, genetics and physiology of pigs and humans, pigs have been recognized as a more appropriate animal model for many human diseases than rodents [213, 300, 301]. However, the widely used swine reference genome *Sscrofa10.2* (SSC10.2 for short) [301] is neither well-assembled nor well-annotated [13]. For example, only 43% of the 25,510 transcripts of the 21,630 coding genes have both a 5’ and 3’ UTRs defined. On average, less than 1.2 isoforms per protein-coding gene and only 3,124 non-coding genes have been identified in SSC10.2. More recently, Warr reported that more than
33% of SSC10.2 is not correctly assembled or is otherwise unreliable [15]. In addition, there are many gaps and thousands of fragmented, unplaced contigs in the assembly, with many known genes missing from the reference genome [15, 16]. Furthermore, the approximate number of genes with detectable expression in porcine peripheral blood is still unknown. The lack of a detailed catalog and annotation of the pig genome and transcriptome hinders annotation-based high throughput studies [302, 303]. Fortunately, two independent assemblies of the pig genome have been recently initiated using a PacBio long read-based approach [15, 17] and are currently being annotated.

Public porcine RNA-seq data have rapidly accumulated over the past few years and are continuously being produced in labs around the world. These data are valuable resources for evidence-based genome annotation, due to their broad coverage, great depth and high resolution [304-306]. The unassembled RNA-seq reads can be used for genome annotation by mapping them directly to the reference genome, but an assembled transcriptome is more useful and less susceptible to mapping artifacts. In addition, RNA-seq read-based transcriptome assembly can reduce the amount of information to be tracked, help identify transcript boundaries and splice isoforms, and facilitate the discovery of novel transcripts whose genes are not included in the reference genome [306]. Currently, there are two popular strategies for transcriptome assembly: de novo assembly and genome-guided assembly, each with its own advantages and disadvantages [105, 307]. Integration of de novo assembly and genome-guided assembly, also known as a hybrid transcriptome assembly method, was proposed as an effective way to reconstruct comprehensive transcriptomes [105, 113, 308].

We here report the build and annotation of a comprehensive blood transcriptome by using blood RNA-seq data from several independent studies using pigs of different genetic
backgrounds and physiopathological conditions via the hybrid transcriptome assembly strategy. The resulting integrated transcriptome was assessed, validated and extensively annotated with different sources of knowledge and evidence. This integrated transcriptome lays a reliable foundation for future blood-based high throughput gene expression studies of pig diseases and physiology, and will also contribute towards annotation of the emerging improved porcine reference genome assemblies.

**Methods**

**RNA-seq data collection**

Publicly available porcine peripheral blood Illumina RNA-seq data were downloaded from the European Nucleotide Archive (ENA) and Ensembl FTP sites, while RNA-seq data from LPS- or saline-treated pigs were from an unpublished study in our lab (H. Liu, K. Feye *et al.*, unpublished). A detailed description of these RNA-seq data is available in Table 4.1. In summary, over 6 billion Illumina RNA-seq reads from 146 independent libraries made of blood RNA samples from 76 pigs of different breeds, age, treatments and health status were used for the blood transcriptome assembly [309-312].

**Pig reference genomes**

The sequence and annotation file of the pig reference genome, SSC10.2, were downloaded from the Ensembl FTP site [313]. Another pig reference genome, USMARCv1.0, was recently assembled using 65× coverage of PacBio long reads from the DNA of lung tissue of a cross-bred pig and is on the order of 100× more contiguous than SSC10.2 (T. Smith *et al.* unpublished).
Preprocessing and normalizing Illumina RNA-seq reads

Quality of the raw RNA-seq data was first assessed with FASTQC (v0.10.1) [314] for each library. Then, the Illumina sequencing adaptor sequences and low quality bases were trimmed from the raw reads by using the Trimomatic software (v0.32) [250] in a sliding window mode with options: `ILLUMINACLIP:adapters.fa:2:30:10:1:true LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 LEADING:3 TRAILING:3 MINLEN:25`, such that the average base quality score (Phred +33) was at least 20 for every 4-base sliding window, the quality scores of leading and trailing bases were at least 3 and the minimum length of kept reads was 25 bases. Because large amounts of reads were available for the transcriptome assembly, to reduce cost and runtime of de novo assembly, the trimmed reads were in silico normalized by using a k-mer abundance-based utility, `insilico_read_normalization.pl`, in the Trinity package (v2.1.1) [107, 315]. The paired-end and single-end reads were normalized separately with the following explicit option settings: `--max_cov 100 --pairs_together --PARALLEL_STATS` for paired-end reads and `--max_cov 100` for single-end reads.

Blood transcriptome assembly

The normalized paired-end and single-end reads from peripheral blood were combined and assembled by using the de novo Trinity software (v2.1.1), with an explicit option setting `--min_kmer_cov 2`, as described in [107]. By default, only assembled products longer than 200 bases were output. To supplement the de novo transcriptome, a genome-guided transcriptome assembly was also performed as follows. First, the normalized paired-end and single-end reads were separately aligned to the USMARCv1.0 reference genome by using the STAR (v2.4.1a) software [228] with Cufflinks-compatible options. The alignment output was subsequently sorted and merged after format conversion. The resulting BAM file
was used for genome-guided transcriptome assembly independent of genome annotation by using the Cufflinks software (v2.2.1) with the following option settings: 

```
-F 0.10 -u -j 0.25 -I 1000000 --min-intron-length 20 --overlap-radius 10 [258].
```

Assembled products of 200 bases or shorter were discarded before further analysis. Hereafter, the assembled products from both the de novo and genome-guided assemblies were called putative transcripts (PTs).

**Quality assessment, filtering, and annotation of the transcriptome assemblies**

The normalized RNA-seq reads were mapped back to the de novo transcriptome assembly and the mapping results were assessed using Trinity utilities, as described in [107, 315]. The number of full-length PTs in the de novo transcriptome was estimated based on the percentage of coverage of manually curated protein sequences and porcine RefSeq mRNA sequences in the Swiss-Prot and the NCBI nucleotide databases by PTs as determined by using BLASTX and BLASTN software (v2.5.0) with E-value cutoffs of $10^{-10}$ and $10^{-20}$, respectively [107].

The coding potential of each PT was predicted by using the PLEK software (v1.2) based on the default classification model for human transcripts [316]. The PLEK software is an efficient long noncoding RNA (lncRNAs) predicting tool, which can distinguish lncRNAs from mRNAs based on an improved k-mer scheme and a support vector machine (SVM) algorithm, independent of genomic sequences and annotation [316]. PTs with coding potentials less than zero were considered lncRNAs, with the remaining PTs considered as protein-coding PTs. The coverage per base (CPB) was calculated by using the Bedtools *genomecov* utility (v2.26.0) [317] based on the normalized reads mapped to the USMARCv1.0 genome assembly as above.
The genomic origins of the PTs in the de novo assembly were determined by mapping them to the two reference genomes, SSC10.2 and USMARCv1.0, using the GMAP software (version 2016-09-23) [90] with the following option settings: \(-n 0 -f samse\). Spliced versus unspliced transcripts were determined based on the alignment results. A uniquely mapping PT was considered spliced if there was at least one “N” in its CIGAR column of the SAM file; a translocation mapping transcript was considered spliced if it was mapped to two locations on two different scaffolds or more than 2 Mb apart on the same scaffold, and there was at least one “N” in its CIGAR column of the SAM file for either mapped part. The splice status of transcripts that mapped to multiple locations or that were not mappable was not determined. The splice status of PTs in the genome-guided assembly was determined based on the number of exons of each PT in the GTF file output by Cufflinks.

PTs not derived from the pig genome were identified for removal from the de novo assembly by comparing their sequences to the NCBI NT database through running DC-megaBLAST. Only alignments with E-value \(\leq 10^{-20}\) were considered significant to insure that true pig transcripts were not inadvertently removed. A PT was considered “contaminant” if its top match was a non-vertebrate sequence, unless it had a better scored alignment with the pig reference genomes than to the non-vertebrate sequence. The PTs in the de novo assembly were also aligned to sequences in the NCBI NR database with BLASTX, using the following explicit option settings: \(-outfmt 5 -evalue 1e-6 -word_size 5 -show_gis -num_alignments 10 -task blastx-fast -max_hsps 20\), which provided protein-level annotation of transcripts in the de novo assembly. Spliced PTs and unspliced non-intronic PTs in the genome-guided assembly were aligned with sequences in the NT and NR databases by using DC-megaBLAST and BLASTX with the same option settings as above.
One possible type of “transcriptional noise” is unspliced PTs that overlap intron intervals, either with or without exonic sequence (collectively called unspliced intronic PTs). 339,342 intron intervals were determined from (i) the splicing junctions of the spliced, uniquely or translocation mapping PTs of the de novo transcriptome, (ii) the splicing junctions of spliced PTs of the genome-guided assembly, and (iii) splice junctions supported by at least 3 uniquely mapped, spliced RNA-seq reads. The two assemblies were filtered to remove potential “contaminants” and “transcriptional noise” in the following order: (i) PTs with top megaBLAST hits on sequences derived from mitochondrial genomes or transcriptomes were excluded because the pig mitochondrial genome has already been well-annotated [318]; (ii) unspliced intronic PTs were discarded; (iii) PTs from the de novo assembly with top megaBLAST hits on sequences from non-vertebrates in the NCBI NT database and without better scored alignments against the two pig reference genomes were filtered out; (iv) unspliced non-intronic PTs within genomic regions for which the maximal CPB was below 50× were removed; (v) multi-mapping and non-mapping PTs from the de novo assembly without significant hits or with non-RNA sequences as top megaBLAST hits in the NT database were excluded.

Comparing and integrating the de novo and genome-guided transcriptome assemblies

Unspliced non-intronic PTs and spliced PTs in the genome-guided transcriptome that remained after removing “transcriptional noise” were compared to their counterparts in the de novo assembly. For unspliced non-intronic PTs, the comparison was simply done by using the Bedtools intersect utility (v2.26.0) [317] to look for overlapping PTs from both assemblies without considering strandedness, which was difficult to determine. On the other hand, because the strandedness of PTs assembled by de novo Trinity could be wrong, the
sense strand of the spliced transcripts was first identified based on the splice site consensus sequences, as determined by aligning them to the USMARCv1.0 assembly using GMAP (version 2016-09-23) [90]. Then, the Bedtools intersect utility (v2.26.0) [317] and custom Perl scripts were used to find intersections between the spliced PTs from the de novo assembly and from the genome-guided assembly from the same sense strands. Spliced PTs from the de novo assembly that did not overlap any spliced PT from the genome-guided assembly were considered unique to the de novo assembly, and vice versa. Transcripts from the same sense strand that shared at least one intron or exon were considered to arise from the same gene.

The integrated transcriptome included (i) all spliced PTs from the de novo assembly; (ii) spliced PTs that were unique to the genome-guided assembly; (iii) unspliced non-intronic PTs from the filtered de novo assembly; (iv) unspliced non-intronic PTs from the filtered genome-guided assembly that did not overlap counterparts in the de novo assembly; and (v) multi-mapping or non-mapping PTs from the de novo assembly that survived the filtering processes. The BLASTX results were then used to annotate these transcripts with GO terms, Enzyme Commission (EC) codes, and KEGG pathways, by using BLAST2GO with default settings [121, 319].

The integrated porcine peripheral blood transcriptome assembly has been deposited at DDBJ/EMBL/GenBank under the accession GFLN00000000. The version described in this paper is the first version, GFLN01000000.

**Similarity analysis of porcine blood transcripts to their homologous transcripts in humans**

Human transcripts were downloaded from the GENCODE database (v25) [320]. Reciprocal DC-megaBLAST was conducted to find the putative homologous transcripts of
pigs in humans [321]. Only the best reciprocal BLAST hits with E-value $\leq 10^{-20}$ were considered significantly similar.

**Validation of the PTs with PacBio IsoSeq full-length cDNA reads**

IsoSeq full-length cDNA sequences are assembly-free, so they can be considered as gold standards to validate the PTs from a transcriptome assembly, especially when the genome annotation is poor. Transcriptomes of three tissues (liver, thymus and spleen) of a single cross-bred pig, from which the USMARCv1.0 pig reference genome was assembled, were sequenced by using the IsoSeq technology (H. Liu and T. Smith *et al.*, unpublished). The resulting full-length, non-chimeric reads were error-corrected with preprocessed Illumina RNA-seq reads from the same RNA samples as those that were used for the IsoSeq. A detailed description of IsoSeq and error-correction processes is available in Additional file 1: Supplementary Methods. Error-corrected IsoSeq full-length cDNA reads and stranded RNA-seq data for porcine spleen, thymus and liver are available from the NCBI SRA under accession numbers: SRX2380861 and SRX2353179; SRX2342762 and SRX2353173; and SRX2342761 and SRX2353180, respectively. The 448,771 error-corrected IsoSeq full-length cDNA reads were first aligned to the USMARCv1.0 reference genome using GMAP (version 2016-09-23) [90] and the splice status and the sense strand of the reads were determined as above. To exclude potential genomic DNA contamination from the data, only the spliced, uniquely mapping transcripts from both the integrated transcriptome and IsoSeq reads were compared by using Bedtools (v2.26.0) [317] and custom Perl scripts, after determining the correct sense strands. Transcripts that shared the same exon-intron structures were considered identical. A PT was considered to be fully validated if we found that it had the same exon-intron junctions as at least one IsoSeq full-length cDNA read.
Assessing completeness of the 5' termini of the PTs

Published CAGE data from pig macrophages [13] were reanalyzed to estimate the proportion of PTs that have complete 5' ends. Briefly, adapter sequences and low quality bases were removed from the raw reads using Trimmomatic (v0.32) [250] with following settings: `ILLUMINACLIP:adapters.fa:2:30:10:1:true LEADING:3 TRAILING:3 SLIDINGWINDOW:4:10 LEADING:3 TRAILING:3 MINLEN:20`. Then, the trimmed reads were mapped to the USMARCv1.0 reference genome using STAR (v2.4.1a) [228]. The R/Bioconductor package CageR (v1.14.0) [322] was used to obtain transcription starting site (TSS) clusters, as described in its documentation. Sequences of 50 bases that flanked the dominant TSS of each TSS cluster were assumed to be proximal promoters. In addition, the pig genomic regions that matched the mouse and human promoters were retrieved from SSC10.2 based on the published information [13] and realigned to the USMARCv1.0 assembly by using GMAP (version 2016-09-23) [90]. All mapped human and mouse promoters plus the pig promoters identified in macrophages by CAGE were used to evaluate completeness of the 5' termini of the spliced, uniquely mapping blood PTs from the integrated transcriptome assembly by using the Bedtools `closest` utility (v2.26.0) [317] in consideration of strandedness of PTs and orientations of promoters. A PT was assumed to have a complete 5' terminus, if a promoter overlapped the first exon or the 500 bp of the 5' extremity of the first exon when the first exon was longer than 500 bp.

Comparison of PTs in the integrated transcriptome assembly with the annotated pig transcripts in SSC10.2 and pig RefSeq mRNA sequences

30,585 transcripts annotated in the Ensembl SSC10.2 (Ensembl transcripts) and 47,439 RefSeq mRNA sequences curated by the NCBI were separately aligned to the USMARCv1.0 assembly using GMAP (version 2016-09-23) [90]. Only uniquely mapping
spliced PTs from the integrated transcriptome assembly were compared to their counterparts from the Ensembl transcripts on the same sense strands by using Bedtools (v2.26.0) [317]. The spliced PTs that did not overlap any spliced Ensembl transcripts were considered as novel transcripts. PTs that shared at least one intron or exon with the Ensembl transcripts from the same sense strand were considered to originate from the same gene. Similarly, a comparison was conducted between uniquely mapping spliced PTs and pig RefSeq mRNAs. Gene loci for PTs were named after the HGNC gene symbols or Ensembl gene IDs of the transcripts of SSC10.2 and pig RefSeq mRNAs if they were considered to originate from the same gene loci.

**Tissue or cell type origin of PTs in the integrated transcriptome**

Lists of human genes specifically or preferentially expressed in given tissue or cell types were downloaded from the CTen database [323]. By assuming porcine and human orthologous genes have preferential expression in same tissues or cell types, the cell or tissue origins of the PTs were inferred.

**Results**

**A de novo RNA-seq read-based porcine blood transcriptome**

Our goal was to provide a transcriptome resource to facilitate blood-based, high-throughput gene expression studies and to simultaneously improve annotation of the pig genome. To this end, we first assembled a *de novo* porcine blood transcriptome that incorporates a broad range of RNA-seq data from 146 independent libraries from 5 studies of blood gene expression in 76 pigs of varied genetic backgrounds and physiological status (Table 4.1). Raw reads composed of 2.14 billion pairs of paired-end and 1.77 billion single-
end reads were trimmed to remove adaptor sequences and low quality bases. Table C.1 shows the effect of trimming on each library. The trimmed reads, consisting of 1.97 billion pairs of paired-end and 1.71 billion single-end reads, were further digitally normalized to reduce the cost and runtime of transcriptome assembly, which resulted in 162.29 million pairs of paired-end and 183.17 million single-end reads. The length distribution of the normalized and trimmed reads is shown in Figure C.1, with most reads of 49, 51 or 100 bases in length. Subsequently, the normalized reads were assembled into putative transcripts (PTs), by using \textit{de novo} Trinity [315] (Figure C.2). The raw \textit{de novo} transcriptome assembly included 490,209 PTs that were longer than 200 bases, which potentially originated from 397,560 putative genes (PGs). Figure C.3A shows the length distribution of these PTs; the longest PT was of 22,969 bases. PLEK [316] prediction indicated that the assembly potentially included 56,973 protein-coding PTs and 433,236 PTs without open reading frames (ORFs) that exceeded the minimum length requirements.

\textbf{Assessment, annotation, and filtering the \textit{de novo} blood transcriptome assembly}

The representation of the entire transcriptome predicted by RNA-seq reads in the \textit{de novo} assembly was assessed by mapping all normalized RNA-seq reads to the \textit{de novo} assembly. The results showed that 66.2\% of normalized reads could be mapped back to the \textit{de novo} assembly, demonstrating that the \textit{de novo} assembly is relatively comprehensive.

A \textit{de novo} transcriptome assembly usually results in many fragmented transcripts [107, 315]. Thus, the \textit{de novo} transcriptome were assessed from different aspects as summarized in Table 4.2. To assess the extent to which the assembled PTs were full-length, we first aligned the PTs to the protein sequences in the Swiss-Prot database with BLASTX. Based on the best high-scoring segment pairs (HSPs) of the top hits, the \textit{de novo} assembly
included 22,831 PTs that covered more than 80% of the full length of 10,097 protein sequences in the Swiss-Prot database. If the overall percentage of coverage of a protein sequence by its best matching PT was calculated based on all HSPs between them, then we obtained 35,541 PTs that covered 11,645 protein sequences in the database by more than 80% of their length (Figure C.3B). We also aligned all the PTs to the pig RefSeq mRNA using megaBLAST, resulting in 16,010 PTs that covered more than 80% of the full length of 9,228 pig RefSeq mRNAs (7,760 genes), which is 19.5% of the 47,439 RefSeq mRNA entries (24,122 genes) (Figure C.3C).

Whether the PTs were derived from the pig genome and whether they were “transcriptional noise” was initially determined by mapping them to the two available pig reference genomes, SSC10.2 and USMARCv1.0 (Figure C.2). This analysis indicated that 94.2% and 99.4% of the PTs could be mapped to SSC10.2 and USMARCv1.0, respectively, suggesting that nearly all of the PTs were of porcine origin and the USMARCv1.0 assembly had better representation of the pig genome. This result led us to rely on the USMARCv1.0 reference genome whenever a pig reference genome was used. As summarized in Figure C.2, mapping the PTs on the USMARCv1.0 reference genome identified 115,862 spliced and 359,012 unspliced PTs among the uniquely or translocation mapping PTs. Among the unspliced PTs, 331,774 (92.4%) PTs overlapped intron intervals, most of which were likely intronic sequences that were detected by RNA-seq. We also identified 19,184 unspliced non-intronic PTs which were mapped to genomic regions for which the maximal coverage per base (CPB) by the normalized RNA-seq reads was below 50×. We designated these unspliced intronic PTs and unspliced non-intronic PTs mapped to genomic regions of relative
low CPB collectively as “transcriptional noise”, which were to be removed from the de novo assembly.

An additional screen to identify non-porcine contaminants among the PTs was performed by searching for sequences with significant similarity to the PTs in the NCBI NT database using DC-megaBLAST. We found 340,286 (69.42%) PTs with significant top DC-megaBLAST hits on sequences from the NT database Figure C.2). There were also 2,567 (0.5%) PTs that had top BLAST hits on sequences from non-vertebrate species, including bacteria, plants, fungi, and viruses, and no better alignment against the pig reference genomes. These 2,567 PTs very likely did not belong to the blood transcriptome. Thus they were to be removed from the de novo assembly. The species distribution of the top BLAST hits on sequences in the NT database is shown in Figure C.3D. The distributions of percentage of identity, percentage of query coverage, bit score and E-value of top BLAST hits are shown in Figure C.3E. The biotypes of the top hits were mRNA/cDNA, ncRNA, misc_RNA, genomic DNA and mitochondrial DNA for 114937, 5736, 10970, 206024 and 52 PTs, respectively (Figure C.2).

The PTs were also aligned to protein sequences curated in the NCBI NR database using BLASTX to get protein-level annotation. This analysis identified 107,237 (21.9%) PTs whose putative proteins had significant similarity to sequences in the NR database, with E-values $\leq 10^{-6}$. The species distribution of the top BLASTX hits in the NR database is shown in Figure C.3F; over half were similar to pig protein sequences. The distributions of percentage of identity, percentage of query coverage, bit score and E-value of top BLASTX hits are shown in Figure C.3G, and show that, on average, 38% of the length of PTs was covered by the cognate proteins in the database and the percentage of identity was over 95%.
The \textit{de novo} transcriptome assembly was filtered to remove mitochondrial genome-derived sequences, “contaminants” and “transcriptional noise” as described in Materials and Methods. The filtered \textit{de novo} assembly consisted of 126,225 PTs (57,272 PGs) that included 115,859 spliced, uniquely or translocation mapping PTs; 8,042 unspliced non-intronic PTs; and 2,324 multi-mapping or non-mapping PTs with RNA sequences as top BLASTN hits (Figure C.2).

\textbf{Supplementing the \textit{de novo} assembly with a genome-guided transcriptome assembly}

To augment the \textit{de novo} Trinity assembly, a genome-guided assembly of the blood transcriptome was also performed using Cufflinks with the USMARCv1.0 reference genome as the guide (see schema in Figure C.4). The input was the same normalized RNA-seq reads as was used for \textit{de novo} Trinity assembly, and the output assembly contained 208,210 PTs, of which 162,294 (78%) were longer than 200 bases (see Figure C.5A for length distribution). The latter included 30,885 spliced and 131,409 unspliced PTs, which included 98,983 intronic PTs and 32,426 non-intronic PTs (Figure C.4). The genome-guided assembly after these steps includes only PTs that were not intronic and longer than 200 bases.

We validated the resulting genome-guided assembly by searching the NT and NR databases for sequences of significant similarity. Of the remaining 63,311 PTs, 53,436 PTs (84.4%) had significantly similar sequences in the NT database (Figure C.5B). The biotypes of the top hits were mRNA, ncRNA, misc_RNA, genomic DNA and mitochondrial DNA for 24864, 648, 1897, 256013, and 14 PTs, respectively (Figure C.4). On the other hand, 30,018 PTs (47.4%) potentially encoded protein sequences that were significantly similar to protein sequences in the NR database (Figure C.5C).
Finally, we removed 30,604 unspliced non-intronic PTs that mapped to genomic regions with relatively low CPB by normalized RNA-seq reads, and 14 mitochondrial genome-derived PTs from the genome-guided transcriptome assembly. The filtered genome-guided assembly that was used for comparison to the de novo transcriptome was thus composed of 32,702 PTs, including 30,885 spliced and 1,817 unspliced non-intronic PTs (Figure C.4).

Comparing and integrating the de novo and genome-guided porcine blood transcriptomes

The use of different assembly approaches for the de novo and genome-guided assemblies resulted in significant differences between the predicted transcripts. Integration of the two sets of transcripts was performed as shown in Figure C.6. Of the 115,859 spliced PTs in the filtered de novo assembly, 113,286 PTs had identifiable sense strands. Among the 30,885 spliced PTs in the filtered genome-guided assembly, 24,913 shared at least one intron or exon with 60,702 spliced PTs with identifiable sense strands from the filtered de novo assembly. Of the spliced PTs in the genome-guided assembly, 45 and 8,967 PTs had exactly the same exon-intron structures and same exon-intron junctions as 47 and 10,008 spliced PTs, respectively, in the filtered de novo assembly. On the other hand, among the 1,817 unspliced non-intronic PTs in the filtered genome-guided assembly, 1,086 PTs overlapped 1,885 of the 8,042 unspliced non-intronic PTs in the filtered de novo assembly.

The final integrated transcriptome was composed of 132,928 PTs, including (i) 115,859 spliced PTs from the filtered de novo assembly; (ii) 5,972 spliced PTs unique to the filtered genome-guided assembly; (iii) 8,042 unspliced non-intronic PTs from the filtered de novo assembly; (iv) 731 unspliced non-intronic PTs from the filtered genome-guided assembly, which did not overlap counterparts from the filtered de novo assembly; and (v)
2,324 multi-mapping or non-mapping PTs with known RNA sequences as BLAST top hits. Based on PLEK [316] prediction, 47,463 PTs (35.7% of total PTs; 40.0% of spliced PTs) potentially encode proteins.

The distributions of the length, number of exons, and number of isoforms of PTs in the integrated transcriptome are shown in Figure 4.1A-4.1C. In total, 119,463 (~90%) and 83,828 (63%) of the PTs shared significant similarities to sequences in the NCBI NT and NR databases, respectively. The biotypes of the top megaBLAST hits in the NT database were mRNA/cDNA, misc_RNA, ncRNA and genomic DNA for 80825, 6904, 1616 and 30118 PTs, respectively. The distributions of percentage of identity, percentage of query coverage, bit score and E-value of top megaBLAST and BLASTX hits against the NT and NR databases are shown in Figure 4.1D and 4.1E, respectively.

We then compared the PTs in the integrated transcriptome assembly to the human transcriptome to determine their similarity at the nucleotide level. The distributions of percentage of identity, percentage of query coverage, bit score and E-value of the best reciprocal DC-megaBLAST hits of 19,398 porcine blood PTs on 19,398 human transcripts are shown in Figure 4.1F. The percentage of identity between porcine blood PTs and their best human reciprocal DC-megaBLAST hits was above 65%, with a median of 85%.

The PTs in the integrated transcriptome were annotated with GO terms, EC codes, and KEGG pathways by using BLAST2GO. Schematic summaries of mapping and annotation of GO terms by BLAST2GO are shown in Figure C.7. The distribution of PTs that were annotated with GO terms and EC codes are shown in Figure 4.2A and 4.2B, respectively. In summary, 68,754 PTs were annotated with 15,965 unique GO terms (Table C.2) and 7,618 PTs associated with EC codes were further annotated with 134 KEGG
pathways (Table C.3). Consistent with the fact that blood was the source of the RNA samples, a GO-BP term, GO:0002376 (immune system process) ranked as a level-two BP term associated with 7,785 PTs (Figure 4.2A). More specific immune system process-related GO terms associated with PTs are displayed in Figure C.8. Among them, immune response, innate immune response, and inflammatory response were the top three GO-BP terms that were associated with many PTs. The purine metabolism KEGG pathway was associated with the largest number of PTs, as shown in Figure C.9.

In addition, we inferred the tissues or cell types where the PTs were expressed by assuming humans and pigs share the same lists of genes highly or specifically expressed in given tissues and cell types [323]. Transcripts of porcine genes whose human counterparts highly or specifically expressed in whole blood, T cells, B cells, monocytes, myeloid cells, NK cells, early erythroid, B lymphoblasts, and many other solid tissues were included in the integrated transcriptome assembly (Table C.4).

**Validation of PTs in the integrated blood transcriptome**

Many assembled PTs were validated by mapping the PTs to the reference genomes, and similarity searches in the NT, NR and Swiss-Prot databases, and in the NCBI pig RefSeq mRNA collection. However, due to the complexity of the pig genome and the short length of the RNA-seq reads, RNA-seq read-based transcriptome assembly is error-prone. Common errors include chimerism and artificial splice isoforms. Thus, the integrated transcriptome was validated from different perspectives (See summary in Table 4.2). To further validate the authenticity of the PTs, we compared 111,002 spliced PTs that uniquely mapped to the pig reference genome and that had identifiable sense strands in the integrated transcriptome to 448,771 error-corrected, assembly-free PacBio IsoSeq full-length cDNA reads with
determined sense strands from three porcine tissues: the liver, thymus and spleen (H. Liu, T. Smith et al, unpublished.). We found that (i) 15,303 PTs and 106,483 IsoSeq sequences had the same exon-intron junctions (Figure C.8A); and (ii) 63,845 uniquely mapping, spliced PTs shared at least one intron or exon with 390,943 IsoSeq reads (See an example in Figure 4.3A).

IsoSeq reads are generally considered as high-quality transcript standards for evaluating correctness and completeness of gene structural annotation. However, not all IsoSeq transcripts that have switching oligonucleotide and polyA tracts actually have complete 5’ and 3’ UTRs due to various artifacts of the library preparation process. In particular, the switching oligonucleotide that is used to identify the 5’ end may not accurately represent the 5’ terminal cap structure. This type of artifacts is more pronounced for long transcripts, for which the cDNA synthesis may not have reached the 5’ termini. Thus, we used available pig macrophage CAGE data [13] and human and mouse CAGE data from the FANTOM5 project [324] to verify the completeness of 5’ termini of the above-mentioned 111,002 spliced, uniquely mapping PTs. In summary, we found 8,529 proximal promoters in the pig macrophage data that were less than 50 bases from the 5’ termini of 19,172 PTs (See examples in Figures 4.3B and 4.4). Based on all promoters identified by CAGE in humans, mice and pigs, we found 43,845 proximal promoters that were less than 50 bases away from the 5’ termini of 37,569 PTs in total. The distribution of distance from the pig macrophage or human/mouse/pig proximal promoters to the 5’ termini of the nearest PTs is shown in Figure C.10A and C.10B, respectively.
Comparison of the putative blood transcripts with SSC10.2 transcripts and pig RefSeq mRNA sequences

The extent to which the integrated transcriptome PTs improved the existing annotation was evaluated by comparing the PTs with the transcript sets of the SSC10.2 and pig RefSeq mRNA collections, after mapping them to the USMARCv1.0 reference genome. To avoid ambiguity, we only compared the 111,002 spliced, uniquely mapping PTs with identifiable sense strands separately to the 23,565 spliced, uniquely mapping SSC10.2 pig transcripts and the 42,931 spliced, uniquely mapping RefSeq mRNA sequences. We found that (i) 4,155 and 6,641 PTs shared the same exon-intron junctions as 4,010 SSC10.2 annotated transcripts and 6,418 RefSeq mRNA sequences, respectively; and (ii) 54,402 and 60,180 PTs shared at least one intron or one exon with 18,437 SSC10.2 transcripts and 33,870 RefSeq mRNA sequences, respectively.

Notably, we also found that (1) 12,262 PTs that had both longer 5’ and 3’ termini than the maximally overlapping SSC10.2 transcripts; (2) 9,764 PTs had only longer 3’ termini; (3) 14,650 PTs had only longer 5’ termini; and (4) 17,726 PTs had both shorter 5’ and 3’ termini. Similar results were seen when the lengths of PTs was compared with those of RefSeq mRNA sequences (Figure 4.5). These length comparisons are shown in Figure C.11A and C.11B. A specific example in Figure 4.4 shows that PTs that arose from the Artemis gene (DCLRE1C, ENSSSCG00000011049), loss-of-function mutations of which cause SCID disorder in pigs [325], had an extended 3’ UTR and one more isoform relative to the SSC10.2 transcript, ENSSSCT00000012093. In conclusion, a large number of PTs were longer than their maximally overlapping transcripts in the two reference sets, although many PTs were shorter than their maximally overlapping transcripts in the two reference sets. These shorter PTs included both transcript fragments and real intact transcript isoforms.
which were shorter than their maximally overlapping transcripts in the two reference sets. Gene symbols or Ensembl gene IDs of Ensembl SSC10.2 transcripts or pig RefSeq mRNA sequences were assigned to PTs arising from the same gene loci (Table C.4). Interestingly, there frequently was a one-to-one relationship between overlapping PTs and pig RefSeq mRNAs; but in many cases, one PT corresponded to more than one overlapping SSC10.2 transcript, which might indicate some of these SSC10.2 transcripts were actually fragments of long transcripts.

In addition, we also found 41,838 and 35,738 spliced PTs that did not overlap any spliced, uniquely mapping SSC10.2 transcripts or with any spliced, uniquely mapping pig RefSeq mRNA sequences, respectively, by these comparisons. These PTs were potentially novel transcripts relative to the corresponding reference sets of transcripts used for comparison. The distribution of exon number of the two sets of potentially novel PTs is as shown in Figure C.11C.

**Discussion**

We assembled a comprehensive, verified transcriptome of porcine peripheral blood by using large amounts of RNA-seq short reads from multiple independent studies with pigs of different genetic backgrounds and physiopathological conditions. The final transcriptome, consisting of 132,928 PTs, was mainly composed of 126,225 PTs from the *de novo* transcriptome assembly, supplemented with 6,703 PTs from the genome-guided assembly. We discuss below our choices that led to this large number of PTs, which centered on maximizing the discovery of genes expressed in porcine whole blood, while minimizing the inclusion of contamination and technical artifacts. Primarily, we focus on two aspects of these results: (1) possible reasons why we obtained such a large number of assembled
transcripts and (2) our approaches to controlling the quality and validating the accuracy of these transcripts.

**Why so many assembled PTs?**

The most recent version of GENCODE (version 25) annotated 58,037 genes for the human genome, including 19,950 protein-coding genes, 15,767 long noncoding RNA genes, 7,258 small noncoding RNA genes, and 14,650 pseudogenes [326]. These genes are associated with 198,093 transcripts, which include a wide range of transcript types [320, 326]. Given the similarities of the porcine genome to that in human, we assume porcine and human genomes encode similar numbers of genes and transcripts. Thus we predict that this number of assembled PTs is much higher than the actual, yet unknown, number of expressed transcript isoforms in pig blood. We consider five sources that may have contributed to the high number of assembled PTs as follows.

a) **Intermediate splicing by-products and genomic DNA contamination**

More than 500 million normalized reads, representing more than 6 billion raw reads, were used to create both the *de novo* and genome-guided transcriptome assemblies. We took this assembly strategy because it has been shown that this strategy can significantly reduce the runtime and cost of transcriptome assembly and increase the chance of gene discovery when the post-normalization k-mer coverage was no less than 50× [107]. Cellular transcripts have a very broad range of abundance, such that some transcripts are so abundant that the other lowly expressed transcripts have a vanishingly small chance to be sequenced in a routine RNA-seq run, which makes it difficult to confidently reconstruct transcripts of such low expression levels from the resulting RNA-seq reads. In addition, gene expression depends on genetic background, developmental stage, and internal and external conditions.
By combining large amounts of RNA-seq reads from diverse studies that sampled animals with different genetic backgrounds, ages and treatment conditions, we alleviated the issue of poor read coverage of lowly expressed genes and increased the chances of detecting expressed genes.

However, a side effect of this strategy was that it increased the occurrence of “transcriptional noise” and genomic DNA contamination in the resulting assembly. For example, more than 67.7% and 61.0% of the PTs from the *de novo* and genome-guided assemblies, respectively, were unspliced and overlapped intron intervals. These might mainly arise from incompletely processed RNA or from spliced but not yet decayed intronic fragments, although most nascent transcription is co-transcriptionally spliced [327]. The portion of unspliced PTs resulting from DNA contamination as discussed in [328-330] is probably small, because we saw many more intragenic unspliced PTs than intergenic unspliced PTs, although the genic portion of the genome is much smaller than the intergenic portion. Although RNA samples usually undergo DNase I digestion and oligo-dT-coated bead selection, it is possible that some genomic DNA fragments carried over into library construction [328-330]. One interesting finding was that there was a considerable number of unspliced PTs that had at least one stretch (> 20) of A’s or T’s in their sequence, which suggested some genomic fragments and intronic pieces with a stretch of A’s might have been enriched by oligo-dT bead capture. In addition, given the dynamics of RNA biogenesis, for highly expressed genes, it is possible that some nascent transcripts containing intronic sequences due to incomplete splicing were captured during library construction and eventually sequenced, although they were poly(A)-free. Also, by visualizing the reads aligned to the reference genome in IGV [331], we found that often a long stretch downstream
of the 3’ termini of highly expressed transcripts was transcribed and sequenced. This might be due to alternative polyadenylation sites or because termination of transcription of RNA polymerase II occurred downstream, far away from the polyadenylation site. Reads from these regions often assembled into several fragmented PTs, which could account for a significant portion of the unspliced non-intronic PTs. Sequencing of the porcine blood transcriptome by using PacBio IsoSeq will help determine the real scenarios.

b) Some assembled sequences are from sources other than the porcine genome

Another side effect of using normalized reads from large amounts of raw RNA-seq data was that it increased the chance to detect contamination of non-pig genome origins. For example, we assembled 2,567 PTs for which a similarity search by using DC-megaBLAST suggested they were likely from feed, parasites, fungi, bacteria and viruses, and had poorer, if any, alignments to the pig reference genomes. These contaminants might have been from sample collection and/or manipulation processes, but could also derive from the microbiome in the peripheral blood of pigs and absorbed dietary RNA from feed the pig consumed. Traditional microscopy and current deep sequencing technology have clearly shown that there are parasites, fungi, bacteria and viruses in the bloodstream in diseased or non-diseased situations in humans [332-335]. As pigs are usually kept in dirty environments, they likely carry some parasites, fungi and other microbes in their bloodstream, which could come from blood exposure to the oral cavity, respiratory or intestinal tracts, although they may not show clinical symptoms. Notably, chronic inflammatory diseases can increase the chance for gut microbiome to translocate into the circulating blood, where they may be in a dormant state [334]. In two of the five independent studies that were used here, pigs suffered from chronic inflammation due to Salmonella infection [310] or chronic LPS stimulation (H. Liu and K. Feye et al. unpublished). In addition, diet-derived miRNAs have been detected in the
bloodstream and adipose tissue of animals [336-338]. So it is possible that some amount of dietary RNA could survive in the circulating blood in pigs, especially when gut integrity is compromised during inflammation.

c) PTs derived from porcine endogenous retroviruses, non-blood exosomal cargos, and beyond

Another source of PTs of non-blood origins is endogenous retroviruses in pigs. It is known that a variety of endogenous retroviruses can be active in pigs [339-341]. We identified 61 PTs derived from pig endogenous retroviruses in the de novo assembly. Furthermore, exosomal RNAs from non-blood cells and non-exosomal, cell-free RNAs in circulating blood, which have been identified in humans [342, 343], could further diversify the apparent peripheral porcine blood transcriptome.

d) Fragmented transcripts due to technical limitations

Another reason why so many PTs were assembled was that de novo transcriptome assembly based on RNA-seq reads tends to result in fragmented transcripts due to sequencing errors, genetic polymorphisms, uneven read coverage, and repetitive sequences. In this study, sequencing adapters and low quality bases were removed from reads and trimmed reads were digitally normalized before assembly, which can partly alleviate problems caused by sequencing errors and uneven read coverage for most transcribed genes in circulating blood. However, the use of RNA-seq reads from multiple studies with pigs of different genetic backgrounds could aggravate the issue of genetic polymorphisms, which are more common in non-coding regions, such as UTRs. In fact, we found that many 3’ UTRs, which were well-covered by reads, were assembled into fragments separate from the corresponding coding regions (data not shown). In addition, a small fraction of repetitive sequences in the
mature transcripts, as evidenced by multiple mapping RNA-seq reads, can contribute to fragmented assembly of some transcripts.

e) Complexity of the transcription landscape in the pig genome

Lastly, the large number of PTs in the transcriptome assembly can also be due to the complexity of the landscape of eukaryotic transcriptomes [326, 344-346]. 198,093 transcripts associated with 58,037 genes have been annotated for the human genome (GENCODE v25)[320]. Given the similarities between the pig and human genomes, it is very possible that the transcription landscape of the pig genome is also very complex and significantly underestimated in the current pig genome annotation [301]. Our de novo assembled blood transcriptome at least partly supports this inference, although it might overestimate the total number of transcripts due to technical limitations of de novo assembly as discussed above.

The initial transcriptome assemblies were filtered by mapping the PTs to reference genomes or by comparison to existing transcript databases, to remove PTs which were (i) most likely not derived from the pig genome, (ii) genomic DNA contamination or (iii) intermediate splicing transcription products. These filtering steps may remove some true porcine transcripts that happened to be mono-exonic and derived from intronic regions, such as intronic ncRNAs, or that by chance had high similarity to non-vertebrate nucleic acid sequences. However, on balance, such filtering is appropriate and conservative, and we believe that, as the majority of protein-coding and long ncRNA transcripts are spliced [347], the few true porcine transcripts that were removed by this procedure is outweighed by the potential for a large number of false transcripts that may be included in the assembly without filtering.
How accurate and comprehensive was the assembled final transcriptome?

Short RNA-seq read-based transcriptome assembly usually results in a large portion of transcript fragments, as discussed above, thus leading to overestimation of the number of transcripts or genes in the target tissue or cell. To assess the quality of the integrated transcriptome assembly, which mainly consisted of PTs from the de novo assembly, we evaluated the assemblies in various ways, which led to three major conclusions.

a) The de novo assembly represented a majority of the normalized blood RNA-seq data, and covered nearly the full length of many curated protein sequences, and porcine RefSeq mRNAs.

More than 66% of total input RNA-seq reads could be mapped to the de novo transcriptome, which was comparable to those reported mapping rates (65 -85%) for other similar assemblies [120]. In addition, we verified 35,541 and 16,010 PTs that covered more than 80% of the full length of the Swiss-Prot protein and pig RefSeq mRNA sequences, respectively. Note that this may underestimate the number of full-length PTs, as the transcriptome and proteome for porcine tissues are poorly annotated.

b) The short read-based transcriptome assembly was substantially verified by PacBio IsoSeq long-read data, CAGE data, as well as by existing porcine gene annotations.

A global sequence alignment-based method for assessing the quality of the assembled transcriptome may not reveal some structural details of the assembled transcripts, such as correctness of the exon-exon junctions of the isoforms and completeness of transcript ends. Therefore, we used the assembly-free, PacBio IsoSeq long cDNA reads from the three porcine tissues (liver, thymus and spleen), which were the IsoSeq data available, to validate the correctness of the fine structures of the PTs. These non-blood sources of cDNA reads are applicable because (i) different tissues usually share a significant portion of expressed transcripts, (ii) all three tissue biopsies contained some blood, (iii) the thymus and spleen are
primary immune tissues, with substantial proportions of cells found in blood such as monocytes and lymphocytes[348]. We found that 15,311 PTs (14% of the PTs in the integrated transcriptome assembly) and 106,484 IsoSeq reads shared the same intron arrangement. However, some peripheral blood-specific transcripts may not be validated by these PacBio IsoSeq reads. In addition, we made use of the extant human and mouse CAGE data from the FANTOM5 project [324] as well as pig macrophage CAGE data [13], to check the completeness of the 5’ extremities of the PTs. We found that 37,569 PTs had complete or nearly complete 5’ termini. These analyses indicated that many of the PTs were correctly reconstructed in terms of exonic architecture and completeness of 5’ termini. Completeness of the 3’ termini could be further verified by using the 3’ T-fill sequencing technology [349].

Comparison of PTs in the integrated transcriptome to SSC10.2 annotation identified 4,155 PTs with shared intron arrangements to 4,010 annotated transcripts, representing 17% of all SSC10.2 transcript annotations. Further, we found that 12,262 PTs had both longer 5’ UTRs and longer 3’ UTRs than the maximally overlapping counterparts in the SSC10.2 annotation. Notably, we also found 41,838 novel spliced PTs that were missing in the SSC10.2 annotation. Similar results were found when the PTs in the integrated transcriptome were compared to the pig RefSeq mRNA set. This suggests that our PTs were useful in expanding the annotation of the pig reference genomes.

c) The integrated transcriptome was well-annotated functionally.

Functional annotation of a transcriptome on a large scale currently depends mainly on sequence similarities between novel and annotated sequences, often across species. By searching the NCBI NT and NR databases, we identified significantly similar sequences for 119,463 (~90%) and 83,828 (63%) PTs, respectively, of the integrated transcriptome assembly; 80,825 (~61%) PTs were potentially mRNA, while the others were potentially
ncRNA, misc_RNA, or artifacts. PTs which encoded putative proteins significantly similar to protein sequences in the NR database were further annotated with GO terms, EC codes and KEGG pathways. In total, 68,754 (82%) and 7,618 (9%) of the PTs with significant BLASTX hits were annotated with 15,965 unique GO terms and 134 KEGG pathways, respectively. We found that a level-two GO-BP term, immune system process, and many more specific terms related to the immune system process were well represented by PTs in the integrated transcriptome, consistent with the whole blood origin of the RNA-seq data.

Conclusions

We constructed a comprehensive peripheral blood transcriptome for the pig by combining de novo and genome-guided assemblies of extensive RNA-seq data from multiple experiments. The resulting transcriptome has been carefully filtered to maximize useful information content, validated on a large scale, and well-annotated by using a series of methods. The transcriptome had significant improvement over the SSC10.2 annotation and the pig RefSeq mRNAs. Thus, the transcriptome we assembled can be used to analyze data from future blood-based high throughput gene expression studies and will help more fully annotate the newly assembled pig reference genomes.
Table 4.1. RNA-seq data used for the blood transcriptome assembly

<table>
<thead>
<tr>
<th>Study Accession</th>
<th>Sample Description</th>
<th>Read Length</th>
<th>Layout</th>
<th>Total Raw Read Count</th>
<th>Breed</th>
<th>Source</th>
<th>Run Accession</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRJEB5250</td>
<td>16 pigs (7 weeks old) were infected with <em>Salmonella enterica</em> serovar <em>Typhimurium</em>. Whole blood was sampled before infection (day 0) and day 2 post infection. Globin was depleted from the total RNA.</td>
<td>51</td>
<td>SE</td>
<td>1,611,707,563</td>
<td>Yorkshire</td>
<td>ENA</td>
<td>ERR413315-ERR413346</td>
<td>[310]</td>
</tr>
<tr>
<td>PRJNA189967</td>
<td>Whole blood from 3 individual healthy pigs</td>
<td>51</td>
<td>SE</td>
<td>106,157,368</td>
<td>unknown</td>
<td>ENA</td>
<td>SRR747924-SRR747926</td>
<td>[311]</td>
</tr>
<tr>
<td>PRJEB12300</td>
<td>Whole blood was sampled from 31 post-weaning (5 ~ 6 weeks old) pigs from lines divergently selected for residual feed intake. Globin was depleted from the total RNA.</td>
<td>100</td>
<td>PE</td>
<td>2 × 632,557,790</td>
<td>Yorkshire</td>
<td>ENA</td>
<td>ERR1199492-ERR1199522</td>
<td>[309]</td>
</tr>
<tr>
<td>PRJEB20136</td>
<td>28 pigs (~ 63 kg of body weight) from lines divergently selected for residual feed intake were muscularly injected with <em>E. coli</em> LPS or saline at time 0, and 48, 96 and 144 hours post the first injection. Whole blood was sampled immediately before first injection and 2, 6, 24 and 168 hours post first injection. Globin was depleted from the total RNA.</td>
<td>49</td>
<td>PE</td>
<td>2 × 1,509,173,972</td>
<td>Yorkshire</td>
<td>ENA</td>
<td>ERR1898446-ERR1898477</td>
<td>H. Liu, K. Feye et al., unpublished</td>
</tr>
<tr>
<td>Transcriptome assembly</td>
<td>Type of assessment</td>
<td>Purpose</td>
<td>Reference data</td>
<td>Software</td>
<td>Results</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>--------------------</td>
<td>---------</td>
<td>----------------</td>
<td>----------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>The de novo transcriptome assembly</strong></td>
<td>RNA-seq read representation of the assembly</td>
<td>To determine representation of RNA-seq reads</td>
<td>Normalized RNA-seq reads</td>
<td>Trinity [107, 315]</td>
<td>66.2% of normalized RNA-seq reads could be mapped back to the de novo assembly</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Representation of full-length assembled protein-coding transcripts</td>
<td>To assess the number of full-length PTs</td>
<td>All protein sequences in the Swiss-Prot database</td>
<td>BLASTX [350]</td>
<td>22,831 (nearly) full-length PTs covered more than 80% of the full length of 10,097 protein sequences in the Swiss-Prot database</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Representation of full-length assembled transcripts</td>
<td>To assess the number of full-length PTs</td>
<td>NCBI pig RefSeq mRNAs</td>
<td>DC-megaBLAST [350]</td>
<td>16,010 (nearly) full-length PTs covered more than 80% of the full length of 9,228 pig RefSeq mRNAs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Origin of assembled transcripts</td>
<td>To assess whether the assembled PTs were of porcine genomic origin</td>
<td>Pig reference genomes: SSC10.2 and USMARCv1.0</td>
<td>GMAP [90]</td>
<td>94.2% and 99.4% of the PTs could be mapped to SSC10.2 and USMARCv1.0, respectively</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Similarity-based assessment</td>
<td>To annotate the assembled PTs with known sequences of significant similarity</td>
<td>Sequences in the NCBI NT and NR databases</td>
<td>DC-megaBLAST and BLASTX [350]</td>
<td>~90% and 63% of the PTs shared significant similarities to sequences in the NCBI NT and NR databases, respectively</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Correctness of exon-intron splicing junctions of PTs</td>
<td>To validate the exon-intron splicing junctions of PTs</td>
<td>Porcine IsoSeq full-length cDNA read data from the liver, spleen and thymus, SSC10.2 transcripts and NCBI RefSeq mRNAs</td>
<td>Bedtools [317] and custom Perl scripts</td>
<td>15,303 PTs and 106,483 IsoSeq sequences had the same exon-intron junctions; and 63,845 uniquely mapping, spliced PTs shared at least one intron or exon with 590,943 IsoSeq reads; 4,155 and 6,641 PTs shared the same exon-intron junctions as 4,010 SSC10.2 annotated transcripts and 6,418 RefSeq mRNA sequences, respectively; 54,402 and 60,180 PTs shared at least one intron or one exon with 18,437 SSC10.2 transcripts and 33,870 RefSeq mRNA sequences, respectively</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Completeness of 5' termini of PTs</td>
<td>To validate the completeness of 5' termini of PTs</td>
<td>FANTOM5 CAGE data for humans and mouse, and porcine macrophage CAGE data</td>
<td>CAGEr [322], Bedtools [317] and custom Perl scripts</td>
<td>Completeness of the 5' termini of 37,569 PTs were verified by 43,845 proximal promoters determined by CAGE data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Length extension of existing transcripts</td>
<td>To determine to what extent the assembled PTs improved over the existing porcine annotation</td>
<td>SSC10.2 transcripts and NCBI pig RefSeq mRNAs</td>
<td>Bedtools [317] and custom Perl scripts</td>
<td>12,262 PTs had both longer 5' and 3' termini than the maximally overlapping SSC10.2 transcripts; 9,764 PTs had only longer 3' termini; and 14,650 PTs had only longer 5' termini</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Novelty of PTs</td>
<td>To determine novel PTs</td>
<td>SSC10.2 transcripts and NCBI pig RefSeq mRNAs</td>
<td>Bedtools [317] and custom Perl scripts</td>
<td>41,838 and 35,738 spliced PTs that did not overlap any spliced, uniquely mapping SSC10.2 transcripts or with any spliced, uniquely mapping pig RefSeq mRNA sequence were potential novel transcripts relative to the two reference sets, respectively</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.1. Characterization of the integrated transcriptome assembly. (A) Length distribution of the PTs of the integrated transcriptome; (B) Exon number distribution for the uniquely mapping spliced PTs of the integrated transcriptome; (C) Distribution of the number of isoforms per PT for PTs with at least two isoforms in the integrated transcriptome; (D, E) Boxplots showing the distributions of percentage of identity, percentage of query coverage, bit scores and E-values of the top BLAST hits of the PTs in the NCBI NT (D) and NR (E) databases by using DC-megaBLAST (D) and BLASTX (E), respectively. (F) Boxplots showing the distributions of percentage of identity, percentage of query coverage, bit scores and E-values of the best reciprocal DC-megaBLAST hits of the PTs in the human transcriptome (GENCODE v25). For clearer visualization, larger outliers of bit scores and E-values are not displayed in D-F.
Figure 4.2. GO terms and EC code annotation of the integrated transcriptome assembly. (A) Distribution of the top 20 GO terms at level two, where available; (B) Distribution of the six main EC classes.
Figure 4.3. IsoSeq full-length cDNA reads (A) and CAGE data (B) validate fine structures of PTs in the integrated transcriptome assembly. (A) An example showing one assembled PT shadowed gray in the “Assembly” panel was validated by one IsoSeq full-length cDNA read shadowed gray in the “IsoSeq” panel in terms of intron arrangement. For references, from top to bottom displayed are genomic coordinates, genome coverage by the normalized RNA-seq reads, aligned RNA-seq reads, pig RefSeq mRNAs, SSC10.2.
transcripts and IsoSeq alignments. In the panel labeled as “RNA-seq Cov”, heights of the gray or colored bars represent CPB by the RNA-seq reads. In the “RNA-seq” panel, purple and blue boxes represent reads mapped to the forward and reverse strands of the chromosome; while the thin segments represent introns spanned by spliced reads. In the panels labeled as “Assembly”, “RefSeq”, “SSC10.2” or “IsoSeq”, red boxes represent exons, and thin segments stand for introns, with arrows indicating the orientation of the sense strand. The “Assembly” panel shows PTs mapped to this genomic window. (B) An example showing a conserved proximal promoter among pigs, humans and mice, determined by CAGE, overlaps the 5’ termini of several assembled isoforms of a gene, indicating completeness of the 5’ termini of those PTs. Meaning of the symbols is the same as those in (A); in addition, the blue boxes stand for proximal promoters determined by CAGE in porcine macrophage, human and mouse cells in the three panels labeled with “Pig CAGE”, “Human CAGE” or “Mouse CAGE”.
Figure 4.4. The integrated blood transcriptome assembly improves the structural annotation of the porcine genome compared to the SSC10.2 annotation. The example shown is for the ARTEMIS gene locus, annotation of which in SSC10.2 was improved by extending the 3’ UTR and adding novel isoforms. Meaning of the symbols is the same as those in Figure 4.3. The two assembled ARTEMIS isoforms are verified by IsoSeq reads, which show many more isoforms for the ARTEMIS gene. For references, also shown are proximal promoters determined by CAGE data and a RefSeq mRNA of the Artemis gene.
Figure 4.5. Length comparison between PTs and transcripts in the reference sets. The lengths of uniquely mapping spliced PTs were compared with those of SSC10.2 transcripts and pig RefSeq mRNAs. The number of PTs with longer 5’ and 3’, only longer 3’, and only longer 5’ termini, or neither terminus than their maximally overlapping reference transcripts in SSC10.2 annotation (red) and RefSeq mRNA collection (blue), respectively is as displayed.
CHAPTER 5. GENERAL CONCLUSION

Summary and General Discussion

Effect of selection for RFI on the immune system in response to systemic inflammation

It was unclear whether selection for improved feed efficiency would compromise a pig’s immune response to pathogens or virulent components of pathogens. In a study presented in Chapter 2, we intramuscularly injected pigs from two lines divergently selected for RFI with an inflammatory stimulant, LPS. LPS injection induced systemic inflammation in pigs of both lines and excited very similar responses in terms of dynamics of body temperature, CBC, and levels of several cytokines in plasma. Both lines also showed a very similar transcriptomic response to LPS stimulation in terms of biological pathways and processes involved in this stimulation. Only a very small number of genes was differentially expressed between the two lines at all time points surveyed and negligible line-by-time interaction effects on gene expression were detected. Interestingly, the low RFI line had a slightly lower level of inflammatory response than the high RFI line at each time point post LPS injection. Genes involved in protein translation, defense and immune response, and signaling formed individual clusters in response to LPS stimulation.

Compared to human data from a similar experiment [214], pig’s inflammatory response observed in this study was much weaker, regardless of much higher dosage of LPS injected in pigs than in humans. Many pathways and biological processes were specifically involved in human response, but only a few were pig-specific. However, hundreds of key pathways and biological processes were shared between the two species. A core set of 1,543
Porcine genes and their human orthologs were differentially expressed in response to LPS stimulation in both species. Nearly one third of these genes shared very similar dynamic expression patterns over the time course. These conserved genes mainly functioned in biological processes including translation, RNA catabolism, signaling, apoptosis, immune response and inflammatory response. But caution should be taken with the differences seen in humans and pigs, due to differences in routes of LPS administration, dosage and sources of LPS, blood fractionation, and completeness of genome annotation in both species.

To conclude, pigs from the two lines divergently selected for RFI responded to systemic inflammation triggered by LPS injected intramuscularly in a very similar way, though the low-RFI animals seemed to have a lower level of inflammation across the time points post LPS stimulation. Pigs and humans shared similar response to systemic inflammation induced by LPS at both pathway and gene levels, though many more biological processes and pathways were specifically involved in inflammatory response in humans and only a few were preferentially involved in pigs.

Again, in this study bulk gene expression profiling was conducted, which can mask some differences existing in some special cell types or blood fractions. Surrogate variables instead of variables recorded as metadata were used to adjust for unwanted/hidden variations in detected gene expression because (1) the CBC parameters were not accurately measured; (2) no CBC tests were done for some samples; (3) RNA processing batches were not always followed from RNA preparation to sequencing. Removing unwanted/hidden variation by using surrogate variables has been shown to be successful and powerful [102, 351]. Sample clustering based on expression data with or without adjustment for surrogate variables, or with adjustment for only available nuisance parameters also indicated that surrogate
variables-based adjustment was best in terms of separation of clusters and compactness of clusters.

Another concern about the data analysis is that correlation in within-individual gene expression resulting from repeated measures was not taken into account in the differential expression analysis, although it was considered in gene set enrichment analysis run in a paired-comparison mode. These might decrease the power of detecting more DEGs between time points post LPS injection and baseline. Thus the number of conserved DEGs detected between humans and pigs might be different. However, the general conclusion will not be affected because that was based on paired-sample gene set enrichment analysis.

**Profiling blood transcriptome of RFI lines for developing predictive biomarkers for RFI**

Given the high cost in terms of money and time of current implementation of genetic selection for low RFI to improve feed efficiency in pigs, an exploratory study was conducted to search for predictive biomarkers for RFI in blood transcriptome. The results were presented in Chapter 2. DEGs between the two groups of animals with extreme RFI phenotypes were identified by using RNA-seq. Although the magnitudes of differential expression of these genes in the low-RFI group versus the high-RFI group were not large, collectively they did differentiate these two groups. Genes involved in small molecule biosynthetic process, antigen processing and presentation of peptide antigen via major histocompatibility complex (MHC) class I, and steroid biosynthetic process were overrepresented among DEGs that had higher expression in the low versus high RFI animals. Genes known to function in the proteasome complex or mitochondrion were also significantly enriched among genes with higher expression in the low versus high RFI animals. Alternatively, genes involved in signal transduction, bone mineralization and
regulation of phosphorylation were overrepresented among DEGs with lower expression in the low versus high RFI animals. The DEGs significantly overlapped with genes associated with disease, including hyperphagia, eating disorders and mitochondrial diseases. A weighted gene co-expression network analysis (WGCNA) identified four co-expression modules that were differentially expressed between the low and high RFI groups. Genes involved in lipid metabolism, regulation of bone mineralization, cellular immunity and response to stimulus were overrepresented within the two modules that were most significantly differentially expressed between the low and high RFI groups. We also found five of the DEGs and one co-expression module that were significantly associated with RFI phenotype of individual animals ($q < 0.05$).

To conclude, the post-weaning blood transcriptomes were globally different between the low and high RFI groups. The identified DEGs suggested potential differences in mitochondrial and proteasomal activities, small molecule biosynthetic process, and signal transduction between the two RFI groups and provided potential new insights into the molecular basis of RFI in pigs, although the observed relationship between the post-weaning blood gene expression and RFI phenotype measured during the grow-finish phase was not strong. DEGs and representative genes in co-expression modules that were associated with RFI phenotype provide a preliminary list for developing predictive biomarkers for RFI in pigs.

It is worth noting that due to possible reasons as discussed in Chapter 2, the validation rate of these DEGs by RT-qPCR was not high. Only 24 of 37 selected DEGs were validated in a joint analysis of 24 (12 per line) of the 31 samples that were used for RNA-seq plus 24 (12 per line) novel samples from the same contemporary group of pigs. Using an analysis of
the 24 novel samples alone, only 9 of the 37 selected DEGs were validated. So many of these genes may only be differentially expressed between the two groups of extreme animals, but not necessarily differentially expressed between the two RFI populations. Caution should be taken with validation of these candidate biomarkers in the future.

In this study, whole blood was used for bulk gene expression profiling. Although the gene expression data was roughly adjusted for the concentration of four types of WBCs in whole blood, the complexity of the peripheral blood in composition mentioned above might mask some differences that exist in special cell types/subtypes or blood fractions. Deconvolving the bulk gene expression profiling with sophisticated methods or performing cell-type specific gene expression profiling might be helpful in discovering some hidden differences between the two groups of animals for biomarker development. This will further be discussed later.

For the sake of biomarker development for predicting RFI, exosomal contents might be a focus of importance in the future. A large number of exosomes exist in bloodstream, which are dynamically released by different types of tissues, including tissues more related to feed efficiency, such as hypothalamus, liver, muscle, fat and intestine. mRNAs, miRNAs and proteins carried by those dynamic exosomes might better reflect the states of their source tissues ([352, 353], see reviews [354-356]).

A potential issue of the experimental design in this study was that the blood samples were collected at 5 ~ 6 weeks of age and before the pigs were fed diets, based on which the feed efficiency of individual animals was tested. The experimental design assumed the transcriptomic differences at a very early life stage could be significantly associated with future differences in feed efficiency. Given the dynamic gene expression during
development, this assumption might be risky. It might be more reasonable to collect a time-series blood samples before and during feed efficiency test. Then transcriptomic analysis of blood fractions or whole blood is conducted to find the earliest stage when the differences in blood can best predict RFI measured at the end of test.

**Advancing swine blood transcriptome**

In sight of the limitations of the annotation and assembly of current pig reference genome, a *de novo* and a genome-guided transcriptome of porcine whole peripheral blood was assembled a large amount of RNA-seq short reads from five independent studies by using the Trinity and Cufflinks software, respectively. After removing putative transcripts (PTs) of low confidence from both assemblies, an integrated transcriptome consisting of 132,928 PTs were complied, a majority of which was from the *de novo* assembly and were spliced transcription products. These PTs can be clustered into 27,782 transcription loci, of which 22,752 loci were composed of spliced PTs. Transcripts in the integrated transcriptome were further annotated via homolog search in the NCBI NT and NR databases. They were further annotated with GO terms, Enzyme Commission codes and assigned to 134 KEGG pathways. Full exon-intron junctions of 17,528 PTs were validated by PacBio IsoSeq full-length cDNA reads from 3 other porcine tissue types, NCBI pig RefSeq mRNAs and transcripts from Ensembl *Sus scrofa 10.2* annotation. Completeness of the 5’ termini of 37,569 PTs was validated by public CAGE data. By comparison to the Ensembl transcripts, we found the deduced precursors of 54,402 PTs shared at least one intron or exon with those of 18,437 Ensembl transcripts and 12,262 PTs had both longer 5’ and 3’ UTRs than their maximally overlapping Ensembl transcripts. 41,838 spliced PTs were totally missing from
the Sscrofa10.2 annotation. Similar results were obtained when the PTs were compared to the pig RefSeq mRNAs.

To conclude, we built, validated and annotated a comprehensive porcine blood transcriptome with significant improvement over the annotation of Ensembl Sscrofa10.2 and the pig RefSeq mRNAs, and laid a foundation for blood-based high throughput transcriptomic assays in pigs and for advancing annotation of the pig genome.

Because the blood transcriptome assembly was based on RNA-seq short reads there were definitely some errors, such as artifacts of fused transcripts formed by transcripts from adjacent genes on the genome, transcripts with retained introns due to sequencing of unprocessed transcript intermediates, transcripts with truncated 3’ and 5’ UTRs and fragmented transcripts. Manual editing of these assembled transcripts along with supporting RNA-seq reads, ESTs, cDNA and protein sequences by using tools like Web Apollo [357] or Otterlace [358, 359], will help correct some of the errors. Stranded RNA-seq data will be extremely helpful in distinguishing genomic DNA contamination from real mono-exonic transcripts and resolving artificially fused transcripts and antisense transcripts. Also due to the short reads used for assembly, some transcripts or transcript isoforms may not be assembled at all. Blood transcriptome sequencing by IsoSeq will complement this short reads-based sequencing.

Because whole blood was used for transcriptome profiling, transcripts only expressed in rare cell types might be not well represented in the assembly. And it is also impossible to ascertain which type of cells a transcript or transcript isoform was derived from. Recent progress in flow cytometric cell sorting, especially the index sorting technology, and microfluidics, makes it possible to isolate single cells for transcriptome profiling by single
cell RNA-seq sequencing. This will be further discussed later. A single cell IsoSeq technology has emerged [360]. Cell type-specific or single cell transcriptome will be expected in the near coming future for pigs.

**Current challenge in swine blood transcriptomics**

Although many blood-based transcriptomic analyses have been performed in the last decade, using a variety of gene expression microarrays or RNA-seq, to address diseases-related biological questions in pigs, all these analyses were actually bulk gene expression analyses, that is, averaged gene expression over tens of different blood cell types/subtypes at different states. DEGs detected by these experiments were a confounding result of change of cell proportions in blood and real expression level change in cells due to treatment. This makes it difficult or even impossible to discover biological mechanisms underlying some biological phenomena when the mechanisms only exist in special cell types/subtypes of low proportion or opposite effects exist on different cell types. It is even possible to get misleading results. In the past years, dozens of statistical methods have been developed, aiming at deconvolving gene expression in heterogeneous samples, such as blood and tumors. However, most of these methods depend on cell type-specific expression signatures, which are usually not available in most species and it is not trivial to develop such signatures; and most of them are not designed for differential expression analysis [361]. A few tools which could be used for differential expression analysis rely on known cell proportions in samples [362, 363]. A recent developed tool for deconvolving gene expression in complex samples, CellCODE, can be used for differential expression analysis with neither cell-type-specific expression signatures nor cell proportion information [361]. However, given the pronounced dynamic and heterogeneity of blood tissue, none of these deconvolution tools are
good enough to capture the real details hidden in blood fractions, or special rare cell types, such as circulating tumor cells and hematopoietic stem and progenitor cells.

Another limitation is the poor quality of the widely used pig reference genome Sscrofa10.2. Previously, gene expression microarrays were designed based on known gene sequences and poor genome annotation could compromise the design of microarray platforms and limit the functional annotation of probesets even if they were included in the platforms. Recently developed RNA-seq technology is not limited by *a priori* knowledge about the genome, but a well-annotated genome or transcriptome is still needed to accurately interpret the gene expression profiles under different conditions. New generations of the pig reference genome will facilitate a new round of discovery in swine biology, when their annotations are greatly improved by integrating all existing evidence, such as ESTs, RNA-seq data, cDNA, protein sequences from pigs and related mammals, and *ab initio* genome annotation, which is expected to be available in months.

**Future Direction: Blood Single Cell Transcriptomics**

As mentioned above, blood is strikingly heterogeneous and also very dynamic. Quickly evolving single cell RNA-seq (scRNA-seq) is a promising choice for blood transcriptomic studies. Low-throughput, high-cost scRNA-seq technologies have existed for several years. For these technologies, usually sequencing libraries are constructed for individual cells, which is labor-intensive and expensive. But the advantage is that 80 to 90% of the transcriptome of a single-cell can be sequenced. Single cell isolation and library preparation have been recently reviewed [364, 365]. They have been used in blood transcriptomic studies, such as discovering new subtypes of immune cells [366], tracing T cell clonality [367], and HIV infection [368].
With the development of the scRNA-seq technology in the past eight years [369], it is now of higher throughput and more automated and scalable. Tens of thousands of single cells can be isolated and sequenced in one run [370, 371]. More importantly, the cost of scRNA per cell has been greatly reduced due to the recently developed scRNA technology, Drop-Seq [370] and Seq-Well [371].

The Drop-seq technology can offer 16 million barcodes in its bead library. scRNA-seq based on this technology can analyze the transcriptome of a single cell at a cost of 6.5 cents [370]. Using the Seq-Well technology, the cost of scRNA-seq per cell is below $1 and the sample preparation system is highly portable; well loading is under easy control and cell lysis is more efficient than Drop-Seq [371]. These technologies have been used to study transcriptomic heterogeneity of macrophages and their response to tuberculosis infection [371] and HIV infection in humans, cell states during cell cycle and cell type identification in complex tissues [370].

A common limitation of these high-throughput, low-cost scRNA-seq technologies is their low capture capacity of transcripts. Currently, only 10 to 15 % of the total number of transcripts in a cell can be captured and analyzed with this process. And only polyA(+) mRNA can be captured. Thus only medium to highly abundant mRNA transcripts can be reliably profiled [370, 371]. However, the limited transcriptomic information captured by these methods is very valuable in providing answers to questions which were impossible to answer with bulk gene expression analysis.

In the future, it is expected that scRNA-seq will see broad application in discovering novel immune cell subpopulations, investigating differential gene expression, differential splicing and single cell gene regulatory networks, and understanding autoimmune, allergic
and infectious diseases at the single cell level in organisms including pigs. For example, the high-throughput scRNA-seq technology can be used to study the process of PRRSV pathogenesis in pigs by profiling single cell gene expression of porcine alveolar macrophages from pigs infected with PRRSV. All current gene expression studies of PRRSV infection have been based on bulk gene expression [372, 373], which have masked the cell-specific response to PRRSV infection and the dynamic process of viral infection. By single cell study, it is possible to identify new drug targets for drug therapy.
APPENDIX A. SUPPLEMENTARY MATERIALS FOR CHAPTER 2

Supplementary Methods

Parameter settings for clustering of gene expression data using STEM

Supplementary Figures

Figure A.1. 3D-PCA plots showing relationship of pig blood samples before and after LPS injection. H, high RFI; L, low RFI; Tx, time point x hpi.
Figure A.2. Enriched KEGG pathways among deregulated genes in RFI lines post LPS injection. Only gene sets significantly enriched under at least one condition \((q < 0.05)\) are displayed. HRFI, high RFI; LRFI, low RFI; Tx, time point x hpi.
Figure A.3. **Enriched KEGG disease gene sets among deregulated genes in RFI lines post LPS injection.** Only gene sets significantly enriched under at least one condition \((q < 0.05)\) are displayed. HRFI, high RFI; LRFI, low RFI; Tx, time point x hpi.
Figure A.4. Enriched transcription factor targets among deregulated genes in RFI lines post LPS injection. Only gene sets significantly enriched under at least one condition ($q < 0.05$) are displayed. HRFI, high RFI; LRFI, low RFI; Tx, time point x hpi.
Figure A.5. Enriched canonical pathways other than KEGG pathways among deregulated genes in RFI lines post LPS injection. Only gene sets significantly enriched under at least one condition (q < 0.05) are displayed. HRFI, high RFI; LRFI, low RFI; Tx, time point x hpi.
Figure A.6. Enriched GO-BP gene sets among deregulated genes in RFI lines post LPS injection. Only enriched gene sets with $q < 0.05$ are displayed. HRFI, high RFI; LRFI, low RFI; Tx, time point x hpi.
Figure A.7. Enriched GO-CC gene sets among deregulated genes in RFI lines post LPS injection. Only gene sets significantly enriched under at least one condition ($q < 0.05$) are displayed. HRFI, high RFI; LRFI, low RFI; Tx, time point x hpi.
Figure A.8. MDS plots based on gene expression abundance with (A) and without (B) adjustment for surrogate variables, showing relationship of human blood samples treated with LPS or saline (labeled as “Placebo”). Adjusting for surrogate variables dramatically improves separation of samples under different conditions and clustering of samples under the same conditions.
Figure A.9. Heatmap showing differentially expressed probesets in humans in response to LPS injection. Shown on in the left and right halves are for the LPS treatment group and the saline control group, respectively. Px_Ty, samples from human individual x at y hpi.
Figure A.10. Enriched KEGG pathways in deregulated genes in humans and pigs post LPS injection. Only gene sets significantly enriched under at least one condition ($q < 0.05$) are displayed. HRFI, high RFI; LRFI, low RFI; Tx, time point x hpi.
Figure A.11. **Enriched transcription factor targets in disturbed genes in humans and pigs post LPS injection.** Only enriched gene sets with $q < 0.05$ are displayed. HRFI, high RFI; LRFI, low RFI; Tx, time point x hpi.

Figure A.12. **Enriched GO-BP gene sets in disturbed genes in humans and pigs post LPS injection.** Only enriched gene sets with $q < 0.05$ are displayed. HRFI, high RFI; LRFI, low RFI; Tx, time point x hpi.

Figure A13. **Enriched GO-CC gene sets in disturbed genes in humans and pigs post LPS stimulation.** Only enriched gene sets with $q < 0.05$ are displayed. HRFI, high RFI; LRFI, low RFI; Tx, time point x hpi.
Figure A.14. Clusters of significant profiles in humans and pigs in response to LPS stimulation showed similar patterns and/or biological functions. Common GO terms overrepresented by similar clusters in humans and pigs are shown. The clusters of humans and pigs in the top row shared 396 common genes.
### Supplementary Tables

**Table A.1.** Metadata for pig RNA-seq samples.

**Table A.2.** Primers used to confirm transcriptional response to LPS stimulation in pigs by Fluidigm RT-qPCR.

**Table A.3.** Summary statistics of dynamic change of CBC parameters during inflammation in pigs treated with LPS or saline.

<table>
<thead>
<tr>
<th>Cell Type</th>
<th>Treatment group</th>
<th>Contrast</th>
<th>Effect</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutrophils*</td>
<td>Saline</td>
<td>0-2 hpi</td>
<td>-0.45</td>
<td>0.18</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.02</td>
<td>0.13</td>
<td>0.89</td>
</tr>
<tr>
<td>LPS</td>
<td>0-2 hpi</td>
<td>1.42</td>
<td>0.17</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>0-24 hpi</td>
<td>-0.97</td>
<td>0.12</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Lymphocytes</td>
<td>Saline</td>
<td>0-2 hpi</td>
<td>0.15</td>
<td>0.18</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.20</td>
<td>0.07</td>
<td>0.005</td>
</tr>
<tr>
<td>LPS</td>
<td>0-2 hpi</td>
<td>1.40</td>
<td>0.16</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>0-24 hpi</td>
<td>0.44</td>
<td>0.06</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Basophils</td>
<td>Saline</td>
<td>0-2 hpi</td>
<td>0.14</td>
<td>0.35</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.16</td>
<td>0.20</td>
<td>0.45</td>
</tr>
<tr>
<td>LPS</td>
<td>0-2 hpi</td>
<td>0.74</td>
<td>0.33</td>
<td></td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>0-24 hpi</td>
<td>0.42</td>
<td>0.18</td>
<td></td>
<td>0.03</td>
</tr>
<tr>
<td>Eosinophils*</td>
<td>Saline</td>
<td>0-2 hpi</td>
<td>0.07</td>
<td>0.14</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.00</td>
<td>0.12</td>
<td>0.98</td>
</tr>
<tr>
<td>LPS</td>
<td>0-2 hpi</td>
<td>1.89</td>
<td>0.12</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>0-24 hpi</td>
<td>-0.35</td>
<td>0.11</td>
<td></td>
<td>0.003</td>
</tr>
<tr>
<td>Monocytes</td>
<td>Saline</td>
<td>0-2 hpi</td>
<td>0.15</td>
<td>0.29</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.08</td>
<td>0.11</td>
<td>0.50</td>
</tr>
<tr>
<td>LPS</td>
<td>0-2 hpi</td>
<td>2.06</td>
<td>0.26</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>0-24 hpi</td>
<td>0.97</td>
<td>0.10</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

Notes: SE, standard error

* Unstructured covariance was considered for neutrophils and eosinophils analysis; For other cell types, heterogeneous compound symmetry covariance was considered.
### Table A.4. Summary statistics of dynamic change of cytokine levels during inflammation in pigs treated with LPS or saline.

<table>
<thead>
<tr>
<th>Cytokine</th>
<th>Treatment</th>
<th>Test</th>
<th>Effect</th>
<th>SE</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>IFNα</td>
<td>Saline</td>
<td>0 - 2 hpi</td>
<td>1.83</td>
<td>0.95</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.75</td>
<td>0.61</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>LPS</td>
<td>0 - 2 hpi</td>
<td>-0.06</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>-0.02</td>
<td>0.53</td>
<td>0.97</td>
</tr>
<tr>
<td>IFNγ*</td>
<td>Saline</td>
<td>0 - 2 hpi</td>
<td>0.68</td>
<td>0.71</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.12</td>
<td>0.80</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>LPS</td>
<td>0 - 2 hpi</td>
<td>-0.83</td>
<td>0.67</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>-0.52</td>
<td>0.70</td>
<td>0.46</td>
</tr>
<tr>
<td>IL10</td>
<td>Saline</td>
<td>0 - 2 hpi</td>
<td>-0.37</td>
<td>0.51</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.00</td>
<td>0.07</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>LPS</td>
<td>0 - 2 hpi</td>
<td>-2.48</td>
<td>0.50</td>
<td><strong>0.0003</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>-0.15</td>
<td>0.06</td>
<td><strong>0.02</strong></td>
</tr>
<tr>
<td>IL1b</td>
<td>Saline</td>
<td>0 - 2 hpi</td>
<td>0.73</td>
<td>0.89</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.27</td>
<td>0.55</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>LPS</td>
<td>0 - 2 hpi</td>
<td>-1.15</td>
<td>0.86</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.21</td>
<td>0.48</td>
<td>0.66</td>
</tr>
<tr>
<td>IL4</td>
<td>Saline</td>
<td>0 - 2 hpi</td>
<td>0.21</td>
<td>0.36</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>-0.03</td>
<td>0.06</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>LPS</td>
<td>0 - 2 hpi</td>
<td>-0.77</td>
<td>0.35</td>
<td><strong>0.04</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.12</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>IL6</td>
<td>Saline</td>
<td>0 - 2 hpi</td>
<td>-1.09</td>
<td>0.78</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.07</td>
<td>0.10</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>LPS</td>
<td>0 - 2 hpi</td>
<td>-3.87</td>
<td>0.74</td>
<td><strong>0.0003</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.07</td>
<td>0.09</td>
<td>0.44</td>
</tr>
<tr>
<td>IL8</td>
<td>Saline</td>
<td>0 - 2 hpi</td>
<td>-0.05</td>
<td>0.56</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.13</td>
<td>0.27</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>LPS</td>
<td>0 - 2 hpi</td>
<td>-1.56</td>
<td>0.57</td>
<td><strong>0.01</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.33</td>
<td>0.25</td>
<td>0.19</td>
</tr>
<tr>
<td>TNFα</td>
<td>Saline</td>
<td>0 - 2 hpi</td>
<td>-0.41</td>
<td>0.79</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.81</td>
<td>0.40</td>
<td><strong>0.05</strong></td>
</tr>
<tr>
<td></td>
<td>LPS</td>
<td>0 - 2 hpi</td>
<td>-2.55</td>
<td>0.74</td>
<td><strong>0.002</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-24 hpi</td>
<td>0.52</td>
<td>0.35</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Notes: SE, standard error

*Heterogeneous compound symmetry covariance was considered for IFNγ analysis; For other cytokines, unstructured covariance was considered.
Table A.5. A full list of between-line, within-line and averageline DEGs in pigs.

Table A.6. Gene sets enriched in differentially expressed genes between the two lines at each time point.

Table A.7. Gene sets enriched in average line deregulated genes at time points post treatment compared to baseline.

Table A.8. GO terms overrepresented by co-expression clusters.

Table A.9. Differentially expressed probesets and genes in humans in response to LPS infusion at each time point post treatment compared to baseline.

Table A.10. Gene sets enriched among deregulated genes in humans at each time point post LPS injection compared to baseline.
Figure B.1. MDS plot showing relationships of RNA-seq samples.
Figure B.2. Subnetwork consisting of differentially expressed genes regulated by the inferred upstream regulator IL-15. Nodes were colored based on the log ratio of averaged expression levels of genes between groups, with up-regulated genes in red and down-regulated genes in green.
Table B.1. **Meta data for RNA-seq and RT-qPCR assays.**

Table B.2. **Primers used to validate differentially expressed genes by RT-qPCR.**

Table B.3. **Statistical summary of raw, trimmed, mapped and residual globin reads.**

Table B.4. **Differentially expressed genes identified by RNA-seq.**

Table B.5. **Summary of Ingenuity IPA-based pathway analyses.**

Table B.6. **Summary of IPAD-based pathway and disease enrichment analyses.**

Table B.7. **Validation of differentially expressed genes by RT-qPCR assays.**

Table B.8. **Association of gene expression with RFI phenotype.**

<table>
<thead>
<tr>
<th>Gene.ID/Gene.symbol</th>
<th>Regression coefficient and significance</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RFI.phenotype</td>
<td>p-value</td>
<td>q-value</td>
</tr>
<tr>
<td>LRP6</td>
<td>1.082</td>
<td>0.000</td>
<td>0.005</td>
</tr>
<tr>
<td>ENSSSCG00000024900</td>
<td>-0.647</td>
<td>0.001</td>
<td>0.021</td>
</tr>
<tr>
<td>ENSSSCG0000008771</td>
<td>-1.333</td>
<td>0.002</td>
<td>0.021</td>
</tr>
<tr>
<td>PDL1</td>
<td>-1.320</td>
<td>0.003</td>
<td>0.023</td>
</tr>
<tr>
<td>ENSSSCG00000020945</td>
<td>-0.164</td>
<td>0.002</td>
<td>0.023</td>
</tr>
<tr>
<td>CLEC12B</td>
<td>-1.502</td>
<td>0.019</td>
<td>0.130</td>
</tr>
</tbody>
</table>

Table B.9. **Differentially expressed WGCNA modules between the RFI groups.**

Table B.10. **GO-BP terms enriched in the WGCNA modules.**

Table B.11. **Association of module eigengenes with RFI phenotype.**

<table>
<thead>
<tr>
<th>Module</th>
<th>Regression coefficient and significance</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RFI.phenotype</td>
<td>p-value</td>
<td>q-value</td>
<td>Adjusted R²</td>
</tr>
<tr>
<td>C1-lightcyan</td>
<td>0.11</td>
<td>0.14</td>
<td>0.19</td>
<td>0.84</td>
</tr>
<tr>
<td>C2-darkturquoise</td>
<td>0.11</td>
<td>0.26</td>
<td>0.26</td>
<td>0.69</td>
</tr>
<tr>
<td>C3-skyblue3</td>
<td>-0.27</td>
<td>0.0075</td>
<td>0.03</td>
<td>0.73</td>
</tr>
<tr>
<td>C4-black</td>
<td>-0.20</td>
<td>0.09</td>
<td>0.18</td>
<td>0.60</td>
</tr>
</tbody>
</table>
APPENDIX C. SUPPLEMENTARY MATERIALS FOR CHAPTER 4

Supplementary Methods

This appendix is to provide supplementary methods for generating of three-tissue IsoSeq and RNA-seq data, and error correction of the IsoSeq data with RNA-seq data for Chapter 4. The error-corrected IsoSeq data were used to validate the putative blood transcripts generated by de novo transcriptome assembly. The raw IsoSeq and RNA-seq data were generated by Dr. Timothy Smith in US-MARC, USDA-ARS. I established a pipeline to perform error correction of the IsoSeq data as described here. It is not tightly related to Chapter 4, so I presented it as an appendix. The other materials included in the appendix are supplementary figures and tables which are usually too big to include in Chapter 4.

Sequencing the transcriptomes of three porcine tissues by using the PacBio IsoSeq and Illumina RNA-Seq technologies

The transcriptomes of three tissues (liver, spleen and thymus) from a single cross-bred pig, from which the PacBio long read-based reference genome was assembled by using DNA of lung tissue, was sequenced by the Meat Animal Research Center (MARC, USDA, Clay Center, NE) using the Illumina NextSeq500 and PacBio RSII platforms for RNA-Seq and IsoSeq, respectively. Total RNA from each of the three tissues was extracted using Trizol reagent (ThermoFisher Scientific) and the provided protocol. Briefly, approximately 100 mg of tissue was ground in a mortar and pestle cooled with liquid nitrogen, and the powder was transferred to a tube with 1 ml of Trizol reagent added and mixed by vortexing. After 5 minutes at room temperature, 0.2 mL of chloroform was added and the mixture was shaken for 15 seconds and left to stand another 3 minutes at room temperature. The tube was
centrifuged at 12,000 x g for 15 minutes at 4°C. The RNA was precipitated from the aqueous phase with 0.5 mL of isopropanol. The RNA was further purified with extended DNase I digestion to remove potential DNA contamination. The RNA quality was assessed with a Fragment Analyzer (Advanced Analytical Technologies Inc., IA). Only RNA samples of RQN above 7.0 were used for library construction. PacBio IsoSeq libraries were constructed per the PacBio IsoSeq protocol. Briefly, starting with 3 μg of total RNA, cDNA was synthesized by using SMARTer PCR cDNA Synthesis Kit (Clontech, CA) according to the IsoSeq protocol (Pacific Biosciences, CA). Then the cDNA was amplified using KAPA HiFi DNA Polymerase (KAPA Biotechnologies) for 10 or 12 cycles followed by purification and size selection into 4 fractions: 0.8-2 kb, 2-3 kb, 3-5 kb and > 5kb. The fragment size distribution was validated on a Fragment Analyzer (Advanced Analytical Technologies Inc, IA) and quantitated on a DS-11 FX fluorometer (DeNovix, DE). After a second round of large scale PCR amplification and end repair, SMART bell adapters were separately ligated to the cDNA fragments. Each size fraction was sequenced on 4 or 5 SMART Cells v3 using P6-C4 chemistry and 6 hour movies on a PacBio RS II sequencer (Pacific Bioscience, CA).

Short read RNA-Seq libraries were prepared using TruSeq stranded mRNA LT kits and supplied protocol (Illumina, CA), and sequenced on a NextSeq500 platform using v2 sequencing chemistry to generate 2 x 75 paired-end reads. These data were then provided to the ISU Tuggle group for analysis.

**Error-correction and redundancy reduction of PacBio IsoSeq full-length cDNA reads**

The Read of Insert (ROI) were determined by using `consensustools.sh` in the SMRT-Analysis pipeline v2.0, with reads which were shorter than 300 bp and whose predicted accuracy was lower than 75% removed. Full-length, non-chimeric cDNA reads were
identified by running the classify.py command. Primer sequences as well as the poly(A) tails were trimmed prior to further analysis. Paired-end Illumina RNA-Seq reads from each tissue sample were trimmed to remove the adaptor sequences and low-quality bases using Trimmomatic (v0.32) [250] with explicit option settings: ILLUMINACLIP:adapters.fa:2:30:10:1:true LEADING:3 TRAILING:3 SLIDINGWINDOW: 4:20 LEADING:3 TRAILING:3 MINLEN:25, and overlapping paired-end reads were merged using the PEAR software (v0.9.6) [374]. Subsequently, the merged and unmerged RNA-Seq reads from the same tissue samples were in silico normalized in a mode for single-end reads by using a Trinity (v2.1.1) [315] utility, insilico_read_normalization.pl, with the following settings: --max_cov 50 --max_pct_stdev 100 --single. Errors in the full-length, non-chimeric cDNA reads were corrected with the preprocessed RNA-Seq reads from the same tissue samples by using proovread (v2.12) [375]. Untrimmed sequences with at least some regions of high accuracy in the .trimmed.fq files were extracted based on sequence IDs in .untrimmed.fa files to balance off the contiguity and accuracy of the final reads.

To reduce the redundancy of the PacBio IsoSeq reads of the three tissues, the error-corrected reads from all three tissues were combined after assigning unique identifiers to them, and mapping them to USMARCv1.0 using GMAP (version 2016-09-23) [90]. The sense strands of the spliced reads were determined on the basis of the splice site consensus sequences. Mapped reads were clustered based on chromosome coordinates and strand information using Bedtools [317] and custom Perl scripts. Reads in each cluster were collapsed if they met the following criteria: (1) they shared compatible intron-exon junctions; (2) they had compatible 5’ termini; (3) they differed in the 3’ termini by less than 100 bases

Supplementary Figures

Figure C.1. Length distribution of the normalized trimmed RNA-seq reads used for the porcine blood transcriptome assembly.
**Figure C.2. Flowchart for de novo blood transcriptome assembly, annotation and filtering.** The diagram shows the steps involved in construction and filtering of the de novo assembly, and includes the number of PTs that resulted from each step, where appropriate. Refer to the Materials and Methods section for details. Quality of the raw RNA-seq reads for
each library was first checked with FASTQC. Subsequently, sequencing adaptors and low quality bases were trimmed from the raw reads. These trimmed reads were then digitally normalized to reduce $k$-mer redundancy. Normalized reads were assembled into putative Trinity transcripts (PTs), which are collectively called “de novo transcriptome assembly”. This assembly was then analyzed in several ways. First, the coding potentials of the PTs were predicted by using PLEK, with PTs of coding potentials higher than zero considered as potentially protein-coding. Then all PTs were separately aligned to the two pig reference genomes, USMARCv1.0 and SSC10.2, by using GMAP. Finally, sequences for the top BLAST hits of the PTs were retrieved from the NCBI NT and NR databases by using DC-megaBLAST and BLASTX, with E-value cutoffs of $10^{-20}$ and $10^{-6}$, respectively. Because the alignment frequency of the assembly to the USMARCv1.0 reference genomes was much higher than to the SSC10.2 assembly, the de novo transcriptome was filtered based on the USMARCv1.0 mapping results. PTs with top megaBLAST hits on sequences from non-vertebrates and without better alignments with the two reference genomes were considered as “contaminants” and were filtered out. The potential biotypes of the PTs were determined based on the biotypes of their top megaBLAST hits if available. Other removed PTs were (i) PTs with top megaBLAST hits on sequences of mitochondrial genomes; (ii) unspliced intronic PTs; (iii) unspliced nonintronic PTs mapped to genomic regions of maximal coverage per base (CPB) lower than 50× (low-CPB regions); and (iv) multiple mapping or nonmapping PTs on the USMARCv1.0 assembly without top megaBLAST hits on RNA sequences in the NT database. The final filtered de novo transcriptome assembly was composed of 126,225 PTs.
Figure C.3. (Legend on next page)
Figure C.3. Characterization of the *de novo* transcriptome assembly. (A) Length distribution of PTs in the *de novo* transcriptome assembly. (B-C) Full-length assessment by using Swiss-Prot protein (B) and pig RefSeq mRNA (C) sequences as standards. In (B), percentage of coverage of the sequences as standards by PTs were calculated based only on the best high-scoring segment pairs (HSP) (“ungrouped percentage of coverage” calculation method) or based on all HSPs (“grouped percentage of coverage” calculation method) between the two aligned sequences. (D, F) Species distribution of top DC-megaBLAST (D) and BLASTX (F) hits of the PTs in the NT and NR databases, respectively. (E, G) Boxplots showing the distributions of percentage of identity, percentage of query coverage, bit scores and E-values of the top BLAST hits of the PTs in the NCBI NT and NR databases by using DC-megaBLAST (E) and BLASTX (G), respectively. For clearer visualization, larger outliers of bit scores and E-values are not displayed.
Figure C.4. Flowchart for genome-guided transcriptome assembly, annotation and filtering. The diagram shows the steps involved in construction and filtering of the genome-guided assembly, and includes the number of PTs that resulted from each step, where appropriate. Refer to the Methods section for details. The raw RNA-seq reads that were
preprocessed as above (see the legend of Figure C.2) were mapped to the USMARCv1.0 reference genome by using STAR, and then assembled into PTs by using Cufflinks. PTs of 200 bases or shorter in length were removed from the resulting genome-guided transcriptome assembly before further analysis. Splicing status of the 162,294 PTs was determined and unspliced intronic PTs were discarded. Sequences of top BLAST hits to the remaining 63,311 PTs were retrieved from the NCBI NT and NR databases by using DC-megaBLAST and BLASTX, respectively. The potential biotypes of the PTs were determined based on the biotypes of their significant DC-megaBLAST hits. To complete the filtering, we removed from the remaining PTs: (i) PTs with top DC-megaBLAST hits on sequences originating from mitochondrial genomes; and (ii) PTs mapped to genomic regions of maximal CPB lower than 50× (low-CPB regions). The final filtered genome-guided transcriptome consisted of 32,702 PTs.
Figure C.5. Characterization of the genome-guided transcriptome assembly. (A) Length distribution of PTs in the genome-guided transcriptome assembly. (B, C) Boxplots showing the distributions of percentage of identity, percentage of query coverage, bit scores and E-values of the top BLAST hits of the PTs in the NCBI NT and NR databases by using DC-megaBLAST (B) and BLASTX (C). For clear visualization, larger outliers of bit scores and E-values are not displayed.
Figure C.6. Integration of the de novo and genome-guided assemblies. This diagram shows the steps used to integrate the two assemblies that were described in Figures 1 and 2, and includes the number of PTs that resulted from each step, where appropriate; the overall goal was to identify those genome-guided assemblies that added information to the de novo
assembly. The sense strands of 113,286 spliced PTs from the *de novo* assembly (Set A) were first determined based on consensus splice site sequences after alignment to the USMARCv1.0 reference genome. The spliced PTs in the filtered *de novo* transcriptome were then compared to their counterparts in the filtered genome-guided transcriptome (Set B) by using the Bedtools intersect utility and custom Perl scripts. If a spliced PT from the *de novo* assembly shared at least one intron or exon, or all introns and exons with a spliced PT from the genome-guided assembly mapped on the sense strand, then they were considered overlapping or exactly the same, respectively. In addition, unspliced PTs from the *de novo* and genome-guided assemblies were directly compared, without considering their sense strands, by using Bedtools intersect utility. PTs from the *de novo* transcriptome assembly, which did not overlap PTs from the genome-guided transcriptome assembly were claimed to be *de novo* transcriptome assembly-specific, and *vice versa*. The final integrated transcriptome consisted of 132,928 PTs, which included all 126,225 PTs from the filtered *de novo* transcriptome assembly and 6,703 PTs specific to the filtered genome-guided transcriptome, including both 5,972 spliced and 731 unspliced PTs.
Figure C.7. Schematic summary of BLAST2GO mapping and annotation. (A) Distribution of source databases based on which GO terms were mapped to PTs in the integrated transcriptome assembly. (B) Distribution of the number of GO terms mapped to PTs of the integrated transcriptome assembly. (C) Distribution of evidence codes of the annotated GO terms mapped to PTs of the integrated transcriptome assembly. (D) Distribution of the number of GO terms mapped to PTs of the integrated transcriptome assembly. (E) Distribution of GO terms at different levels of the hierarchy of GO terms. (F) Distribution of annotation scores of GO terms.
Figure C.8. Distribution of GO terms related to immune system process.
Figure C.9. The purine metabolism KEGG pathway highlighted with enzymes encoded by PTs of the integrated transcriptome. Different enzyme codes are highlighted with different colors.
Figure C.10. The distributions of the distance between proximal promoters defined by pig macrophage CAGE data (A) or all available human/mouse/pig CAGE data (B) to the nearest 5' termini of spliced, uniquely mapping PTs in the integrated transcriptome assembly. The vast majority of these distances were less than 50 bp, which strongly suggested that the 5' termini of many PTs were completely assembled.
Figure C.11. Comparison of PTs of the integrated transcriptome assembly to the IsoSeq cDNA reads, transcripts annotated in Ensembl SSC10.2 and NCBI pig RefSeq mRNA.
sequences. (A) Venn diagram showing the number of PTs with their intron arrangements validated by the IsoSeq reads, pig RefSeq mRNA sequences and SSC10.2 transcripts. (B) Distribution of exon number of novel spliced transcripts not found in the SSC10.2 annotation or pig RefSeq mRNA set, which provides good evidence of significant extension of current genome annotation. (C) Length comparison between PTs and their maximally overlapping SSC10.2 transcripts. (D) Length comparison between PTs and their maximally overlapping RefSeq mRNA sequence.

Supplementary Tables

Table C.1 **Trimming effect on the raw RNA-seq data.**

Table C.2 **PTs annotated with GO terms.**

Table C.3 **PTs annotated with EC codes and KEGG pathways.**

Table C.4 **PTs assigned with HGNC gene symbols and Ensembl pig Gene IDs and tissue or cell types where their human counterparts specifically or preferentially expressed.**


24. Uenishi H, Eguchi-Ogawa T, Shinkai H, Okumura N, Suzuki K, Toki D, Hamasima N, Awata T: PEDE (Pig EST Data Explorer) has been expanded into Pig Expression Data Explorer, including 10 147 porcine full-length cDNA sequences. Nucleic Acids Research 2006, 35(suppl_1):D650-D653.


137. National program of swine feed efficiency [http://www.swinefeedefficiency.com/project.html]


190. Dunkelberger JR, Boddicker NJ, Seräo NVL, Young JM, Rowland RRR, Dekkers JCM: Response of pigs divergently selected for residual feed intake to experimental infection with the PRRS virus. Livestock Science 2015, 177:132-141.


224. Laboratory Procedure Manual for Complete Blood Test
   [http://www.cdc.gov/nchs/data/nhanes/nhanes_03_04/l25_c_met_complete_blood_count.pdf]
225. Fluidigm User Guide for Real-Time PCR Analysis
   [https://www.fluidigm.com/binaries/content/documents/fluidigm/resources/real-time-pcr-analysis-ug-68000088/real-time-pcr-analysis-ug-68000088/fluidigm%3Afile]


249. **FASTQC** [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/]


254. **WGCNA tutorials** [http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/]


257. **IPAD** [http://bioinfo.hsc.unt.edu/ipad/index.html]


multiple internal control genes. *Genome Biology* 2002, **3**(7):research0034.0031 - research0034.0011.


