PhD Thesis

Victor Adriano Okstoft Carmelo

DTU Compute
Department of Applied Mathematics and Computer Science

Systems Genomic and Transcriptomic approaches for simultaneous improvement of feed efficiency and production in Danish Pigs.

Supervisor: Haja N. Kadarmideen

Co-supervisor: Claus Ekstrøm

Submitted on the 14th of February 2020

Summary

Feed efficiency (FE) is the most important phenotype in commercial pig production. It is of high economic value, but also important for sustainable production. The goal in this thesis, was to further our biological understanding of FE in pigs, using metabolomic, transcriptomic and genomic data. Beyond biological understanding, we aimed to develop potential biomarkers that could be applied in pig production for FE in all omics data types.

Metabolomics is one of the key links in the connections between genetics, environment and phenotypes. Metabolomics analysis can predict underlying phenotypes with non-invasive techniques. We performed metabolomics analysis of blood plasma on 109 performance tested young boars, of the DanBred Duroc (Duroc) and DanBred Landrace (Landrace) breeds, with 59 and 50 boars of each breed, respectively. This was the first study applying metabolomics analyses of FE phenotypes in pigs. As an addition, we also analyzed daily gain (DG) at different growth stages. The results showed significant overall relation between both FE phenotypes and the DG phenotypes, based on mixed and linear modelling. This identified 67 metabolites significantly associated with DG phenotypes and 1 with FE at a false discovery rate (FDR) < 0.05. Based on metabolites network analysis, we identified several modules, which were correlated both with DG and FE phenotypes, respectively. Pathway enrichment analysis and gene-metabolite networks identified several putative key hub metabolites.

If we view the metabolites as the most external link to our phenotypes, the next link in the chain are the proteins. An effective way of doing genome wide analysis of protein activity is by doing analysis of the expression of the genes associated with the proteins through transcriptomics. In pig production, and for FE, muscle is a key organ. Thus, we performed muscle transcriptomics on a sub-population of 41 pigs from the metabolomics study, mainly focusing in feed conversion ratio (FCR). Similarly to the metabolomics results, we were able to demonstrate an overall relation between gene expression and FCR. We identified 14 differentially expressed (DE) genes (FDR < 0.1). Pathway analysis revealed enrichment of mitochondrial genes in the top FCR genes. Gene-gene interaction analysis identified top interactive genes among potential FCR genes. Network analysis revealed two modules correlated to FCR, which contained enrichment of mitochondrial and nucleic acid metabolism genes, respectively. Finally, a novel possible link between the effect of exercise on human muscle, and the muscle of efficient pigs was established.

The deepest layer underlying all the causal mechanism in organisms, is genetics. We thus aimed to establish a link between genes whose expression might affect FCR, and the genetic control mechanisms behind them. This was done through expressed quantitative trait loci (eQTL) analysis. We identified 15 potential individual eQTLs (FDR < 0.1), and in agreement with previous studies, we observed that the overall distribution of p-values in our analysis were significantly left-skewed towards lower values. We applied targeted pathway enrichment to trans-eQTLs, demonstrating significant enrichment of genomic context-based gene ontologies.

Overall, based on the work in this thesis we identified many potential FE biomarkers, and found strategies for analyzing the complex and statistically challenging phenotype of FE. This has given us new insights in the biological background of FE, and acts as a stepping-stone for future work in the subject.

Dansk Resumé

Foderudnyttelse (FDU) er den vigtigste egenskab i kommerciel svineproduktion. FDU har høj økonomisk værdig, men er også vigtig for bæredygtig produktion. Målet i denne afhandling var at fremme vores biologiske forståelse for FDU i grise, ved brug af metabolomiske, transkriptomiske og genomiske data. Udover biologisk forståelse, havde vi også som mål at udvikle potentielle biomarkører for FDU til brug i svindeproduktion baseret på alle omics data typer.

Metabolomics er en af nøgle forbindelserne mellem genetik, miljø og fænotyper. Metabolomics analyser kan forudse underliggende fænotyper med ikke-invasive teknikker. Vi udførte metabolomiske analyser af blod plasma fra 109 ydelses teste orner fra racerne Danbred Duroc og Danbred Landrace, med henholdvis 59 og 50 fra hver race. Dette var det første studie som anvendte metaboliske analyser af FDU fænotyper i grise. Vi udførte også analyser af daglig tilvækst (DT) ved forskellige vækstfaser. Resultaterne viste signifikant samlede relation mellem både FDU fænotyper og DT fænotyper baseret på mixed og lineær modellering. Dette identificerede 67 metaboliter som var signifikante for DT og en for FDU, med en false discovery rate (FDR) < 0.05. Baseret på netværk analyse, identificerede vi adskillige moduler som var korreleret med FDU og DT fænotyper. Pathway og gen-metabolit analyse identificerede adskillige mulige nøgle metaboliter.

Hvis vi ser metaboliter som det mest eksterne link til vores fænotype, så er det næste link i kæden proteiner. In effectiv måde at udføre helgenom undersøgelse af protein aktivitet er gennem transkriptomics. I griseproduktion, og for FDU, der er muskel et vigtigt organ. Derfor, udførte vi muskel transkriptomiske analyser a en sub-population på 41 grise fra det metabolomiske studie, med hoved focus få feed conversion rate (FCR). I en gengivelse af de metabolomiske resultater, kunne vi vise en samplede relation mellem gen ekspression og FCR. Vi identificerede 14 gener som var differentielt udtrykte (FDR < 0.1). Pathway analyse viste berigelse af mitokondriske gener i top FCR gener. Gen-gen interaktion analyse identificerede top interaktive gener i blandt potentielle FCR gener. Netværks analyser fandt to moduler med correlation to FCR, som henholdsvis indeholdte berigelse af mitokondriske og nukleinsyre

metabolisme gener. Endelig, så fandt vi en mulig ny forbindelse mellem effekten af træning på humane muskler, of muskler fra effektive grise.

Det dybeste underliggende biologiske lag i organismer er genetik. Derfor havde vi som mål at forbinde gener med en mulig forbindelse til FCR og underliggende genetisk kontrol. Dette blev gjort gennem expressed quantitative trait loci (eQTL) analyse. Vi identificerede 15 individuelle eQTLS (FDR < 0.1), og ligesom i de tidlgere studier, der så vi at vi havde signifikant berigelse af lave p-værdier i vores modeller. Vi anvendte målrettet pathway analyse på trans-eQTLer, som viste signifikant berigelse af gen ontologier baseret på genomisk kontekst.

Sammenalgt, identificerede vi mange mulige FDU biomarkører på baggrund af arbejdet i denne these, og fandt strategier for at analysere det komplekse og statisk udfordrerne fænotype, FDU. Dette har givet os ny biologisk viden om FDU, og kan agere som et startskud for flere analyser.

Publications Included

The PhD thesis is based on the following three articles:

A. **Carmelo, V.A.O.**, Banerjee, P., da Silva Diniz, W.J. *et al.* Metabolomic networks and pathways associated with feed efficiency and related-traits in Duroc and Landrace pigs. *Sci Rep* **10,** 255 (2020). https://doi.org/10.1038/s41598-019-57182-4

B. **Victor A. O. Carmelo**, Haja N. Kadarmideen, Genome regulation and gene interaction networks inferred from muscle transcriptome underlying feed efficiency in Pigs, under *submission.*

C. **Victor A. O. Carmelo**, Haja N. Kadarmideen , Eqtl and pathway enrichment analysis on FCR and mitochondrial genes of Danish performance tested pigs, *in preparation.*

Additional Publications

I. Sabino, M., **Carmelo, V.A.O.**, Mazzoni, G. *et al.* Gene co-expression networks in liver and muscle transcriptome reveal sex-specific gene expression in lambs fed with a mix of essential oils. *BMC Genomics* **19,** 236 (2018). https://doi.org/10.1186/s12864-018-4632-y

II. **Carmelo, V.A.O.**, Kogelman, L.J.A., Madsen, M.B. *et al.* WISH-R– a fast and efficient tool for construction of epistatic networks for complex traits and diseases. *BMC Bioinformatics* **19,** 277 (2018). https://doi.org/10.1186/s12859-018-2291-2

III. Banerjee, P., **Carmelo, V.A.O**., Kadarmideen, H.N. Genome-wide epistatic interaction networks affecting feed efficiency in Duroc and Landrace pigs. *Front. Genet,* Provisionally Accepted, doi: 10.3389/fgene.2020.00121

Table of contents

Preface

This PhD project was performed under the FeedOMICS project, funded by the Independent Research Fund Denmark (DFF) – Technology and Production (FTP) grant (grant number: 4184-00268B). The PhD student received partial stipends from the University of Copenhagen and Technical University of Denmark. The PhD was performed at the Department of Veterinary and Animal Sciences of Copenhagen University, at the Danish Technical university at DTU Bioinformatics and the section for statistics and Data Analysis at DTU compute. The project was performed under the main supervision of Professor Haja N. Kadarmideen and with co-supervision from Professor Claus Ekstrøm, from the department of Public Health of Copenhagen University.

Aims

Improvement of FE is one of the most important goals in pig production. There has been consistent increase in price of pig feed, making feed costs by far the largest cost of raising pigs (around 60%). Furthermore, it is a necessity to increase the environmental sustainability of pig production. As Denmark is one of the largest producers and exporters of pigs in the world, producing around 30 million pigs a year, there is strong motivation to improve efficiency. This Project investigates if pigs that are efficient and inefficient for feed utilization also differ in their whole genome-wide genetic make-up, global gene expression levels and metabolites levels. Ultimately, we want to identify non-invasive biomarkers which could be used in genomic selection programs and develop deeper understanding of the systems of biological mechanisms of FE.

To reach these goals, we collected both blood and tissue samples from performance tested pigs and extracted metabolomic, transcriptomic and genomic profiles, and analysed these, on the basis of collected feed intake and growth data.

The aims of this thesis were:

1. Plan and perform collection of blood and muscle samples, and generate transcriptomic, metabolomics and genomic data.

2. Identify significant individual metabolites related to FE phenotypes, based both on individual metabolite analysis and network-based strategies. Identify pathways underlying significant metabolites.
3. Identify differentially expressed genes for FE phenotypes and create gene expression networks related to FE phenotypes. Find underlying functional mechanisms based on gene ontology and enrichment.
4. Identify the connection between expression and genetics for genes that are significantly involved in FE traits.
5. In all analysis, identify potential biomarkers for FE.
6. Based on experiences learned during data analysis, develop novel methodologies and analysis strategies.

Background

The Genome

All known life relies on DNA and RNA to function. DNA, or deoxyribonucleic acid, is the memory of life, coding all the necessary information for generating all the diverse life form we see on earth. DNA is made of two chains coiled into a helix structure, which is made of individual building blocks. These individual building blocks are called nucleotides, and DNA is solely made of four specific nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). The two strands are connected by covalent bonds that form between nucleotides, according to base pairing rules: A with T and C with G. Thus, the two strands are completely complementary. In eukaryotes, the DNA is located in the nucleus of the cell, and organized into chromosomes, which are large individual DNA molecules. As an example, the pig genome is divided into 19 chromosomes, comprising a total of roughly 2.6 billion base pairs (bp) [1].

If DNA is the storage of the cell, then RNA, or ribonucleic acid, is the messenger. RNA differs from DNA in two key ways: it has ribose sugar backbone instead of deoxyribose, and instead of using T, it uses uracil as a nucleotide. The change in the molecular structure makes RNA more chemically active and unstable. RNA is generally a single stranded molecule, although it can still generate complementary base pairs with itself or other RNA molecules. RNA is synthesized with DNA as a template by RNA polymerases, which are specialized proteins that are conserved in all life forms.

Transcription and Translation

For the genetic information of genes to be activated, it first needs to be transcribed into RNA, by the aforementioned RNA polymerases. This is a complex process involving many factors, including the un-wounding and opening of DNA, the recruitment of RNA polymerases by transcription factors to the promoter area of genes, transcription elongation and transcription termination [2]. As the information on each DNA strand is complementary to the other, the actual transcription happens on the antisense strand of a gene, thus generating a copy of the sense strand in RNA. A particular feature of eukaryotes, including pigs, is the exon-intron structure of genes. Thus, genes are largely split into two types of regions – exons, which are the protein-coding regions of genes, and introns, which are not [3]. While the whole gene is transcribed, introns will be spliced and removed co-transcriptionally, ensuring that mature messenger RNA (mRNA) will only contain the exons [4]. mRNA is the class of RNAs that is destined to be translated into protein. While not all RNA is spliced, even long non-coding RNA (lncRNA) can be spliced [5], even though it will not be translated into protein. This pattern, of exons and introns allows for alternative versions of a gene to be generated, ensuring protein diversity [6].

Once DNA has been transcribed into mRNA, the mRNA is translated into protein, which is the actual functional unit in organisms. RNA can however, be functional, such as in the case of the lncRNAs [7]. Translation into protein is mediated by the ribosome together with transfer RNAs (tRNAs). Essentially, mRNAs bind to the ribosome, and then tRNAs select and transfer the correct amino acids based on the RNA sequence, which are linked together into a peptide chain of amino acids, which will eventually form the complete protein [8]. A schematic of the process is found in *Figure 1*.

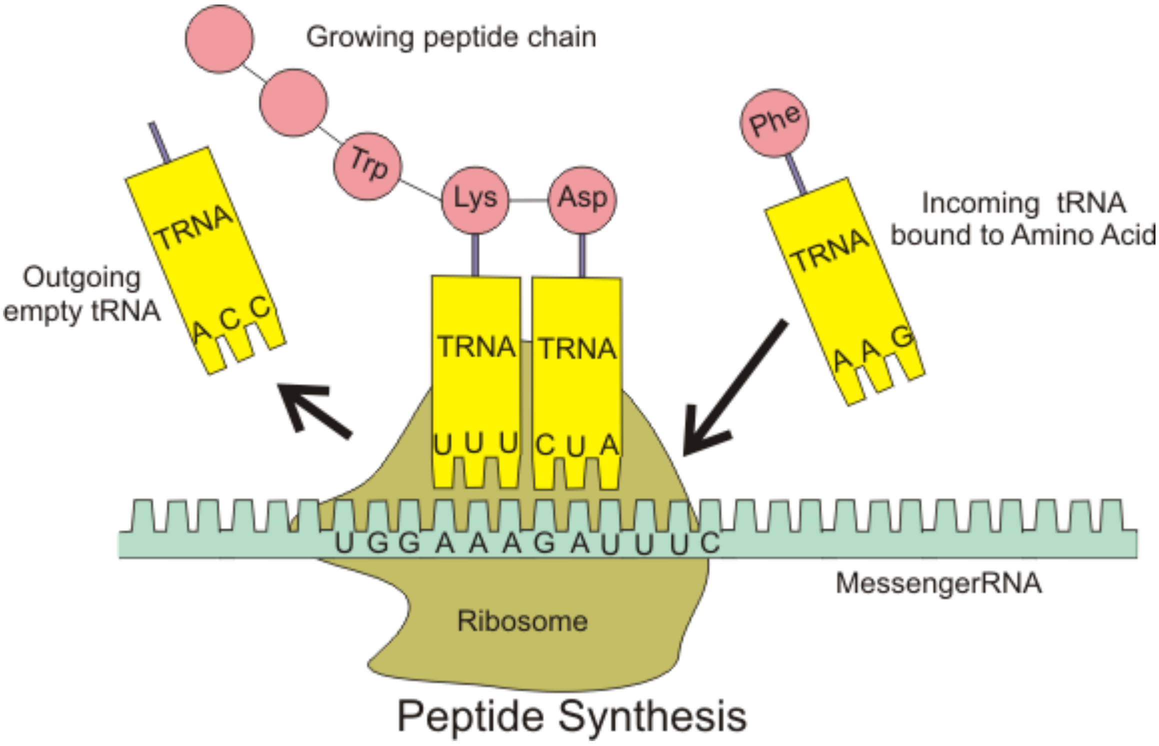


Figure 1. Schematic of RNA translation and protein synthesis. Figure by Boumphreyfr licensed under CC BY-SA 3.0 https://creativecommons.org/licenses/by-sa/3.0/

Genetic Variation

When observing a given species, it is obvious to a casual observer that most individuals of a given species are quite similar – most humans have two hands, two eyes, etc. It is also obvious, that similar does not mean identical – we see variation in the color of the hair or eyes, difference in size and proportion and more. Some of these differences will come down to environment, or epigenetics (a topic we will not discuss further in this thesis), but many of the differences are of genomic origin. While most of the genome within a species is identical, there is variation between all individuals, due to mutations that have happened over time in different individuals and populations. A large meta-analysis of 50 years of twin-studies including 17804 phenotypes reported an average heritability of 0.49 over all phenotypes studied [9]. Heritability is the proportion of the variation in a given phenotype explained by genetics effects. The initial definition is largely attributed to Jay L. Lush, who actually developed the idea in an animal breeding context [10]. More rigorously, a given phenotype P is expressed as follows:

$$P = G + E$$

Where G is the genetic component, and E the environmental effect. The heritabilityH^2, is then the variance in G over the variance in P:

$$H^2 = \frac{\mathrm{Var(G)}}{\mathrm{Var(P)}}$$

This is the broad sense heritability. When using heritability in a selection context, it is often more important to take into account the contribution of additive genetic effects to the variance, denoted as Var(A). The additive variance represents the contribution of parents in the differences of their offspring [11]. Using this, one can calculate the narrow sense heritability h^2:

$$h^2 = \frac{\mathrm{Var(A)}}{\mathrm{Var(P)}}$$

How is genetic variance expressed in practice, on a molecular level? When defining genetic variation, one must typically define it as the difference to a reference, as a sequence of DNA cannot be different from itself. These differences are called genetic variants, which are separated into three overall categories: Single Nucleotide Polymorphisms (SNPs), insertion and deletions (indels) and structural variants. SNPs are point mutation, that is, single positions in the genome where different members of a species differ in the nucleotide found at that position [12]. Indels are short deletions or insertions of nucleotides at a given position ranging from 1 to 10,000 bp [13]. Structural variants are essentially indels, but larger and sometimes harboring specific properties, such as repeated elements [14]. Based on human studies, SNPs are by far the most common variants [15], which undoubtable extends to pigs. A cursory search of “pig snp” of Pubmed (https://www.ncbi.nlm.nih.gov/pubmed/) revealed 1256 hits, while “pig indel” only has 56 matches. Given that mammalian genomes are diploid, a given genomic position has two versions, or two alleles. Thus, one can have different variants of a SNP on each allele, which then defines one overall genotype.

Given the role and function of DNA, one can easily realize how genetic variants lead to different phenotypes. Changes in nucleotides can lead to changes in the final protein product, complete inactivation of genes, or affect the level of expression of a given gene, all of which can affect measurable phenotypes.

Linkage Disequilibrium

When determining the causal impact of a given genetic variant, one cannot simply expect that a correlation between a phenotype and the variant signifies and underlying functional impact. One of the main reasons for this, is that genetic variants will invariably be linked with their surrounding genomic context, meaning that the true underlying effect may be caused by unseen genetic effects. The reason for these effects is that DNA is passed from parents to offspring in chromosomal units, with some degree of recombination between maternal and paternal chromosomes. Effectively, most regions in the genome will not be recombined, and thus there will be stretches of DNA, which are highly correlated among parent and offspring, leading to high correlations in neighboring variants, which is called linkage disequilibrium (LD) [16]. Over time, in a given population, more chromosomal segments will be recombined, resulting in shorter stretches of correlated genome. One useful measure of LD is r^2, which represents the squared correlation between two alleles [17]:

$$r^2(p_a, p_b, p_{ab)} = \frac{(p_{ab} - p_a p_b)^2}{p_a(1 - p_a)p_b(1 - p_b)}$$

Where p_a is the probability of allele a, p_b is the probability of allele b, and p_{ab} is the joint probability of allele a and b.

The Danish pig production

Denmark is one of the world's leading producers of pork and has the highest per-capita pig production of any nation. Consequently, most of this production is exported. In 2017 alone, 31.8 million pigs were produced, and a total of 1.908.017 tons of pork products were exported (Landbrug of Fødevare, Statistik 2017 grisekød). Therefore, it has been of key importance to make more effective the production as much as possible, without hampering quality. One key method of improvement is selective breeding of economically important phenotypes. The Danish production pig is a crossbred, with the boars being purebred Duroc and the sows being a cross of Landrace and DanBred Yorkshire (Yorkshire). Given this mixture, the purebred pigs have slightly different breeding goals. The Duroc does not include female fertility phenotypes, such as LG5 (live piglets at day 5) and mother effects, but has a higher focus on FE. In *Figure 2*, we can see the relative economic gain achieved by the different phenotypes in the breeding program of Danish pigs in the purebred lines. Overall, we see that FE is the single phenotype which contributes the most economically through improvement. Combined with daily gain from 30kg-slaughter, they cover over half of the increased value from improvement. In practice, it is impractical and expensive to measure FE in the general production population. Therefore, in Denmark, testing for FE is done at the core testing facility at Bøgilgård.

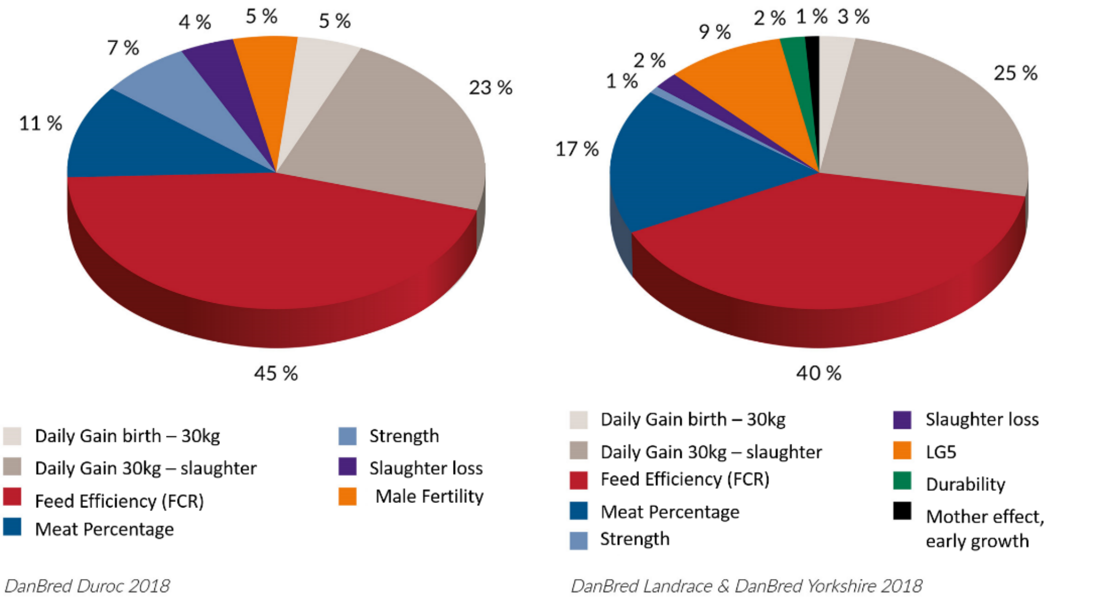


Figure 2 Overview over relative economic gain of improvement of the various trait in the Danish pig breeding program. Overall, FE represents the highest importance, and together with daily gain 30kg- slaughter represent over half of value of the improvement. Source: SEGES svineproduktion 2018.

Here, approximately 5000 potential breeding boars for all three pure breeds are tested yearly. The pigs arrive from 28 breeders around the country at around 4 weeks of age, and testing begins after an acclimatization phase of about 5 weeks. The testing phase happens from a weight of ~30kg to ~100kg, where feed consumption and weight gain is accurately measured.

Breeding values

The procedures, methods and techniques for calculating and estimating breeding values of pigs for the various phenotypes in the breeding goal are outside the scope of this thesis. However, a short explanation of the process serves as a background for how the pigs have been selected over time, and thus their genomic background. Since 2010 in Duroc, and 2011 in Yorkshire and Landrace, genomic selection has been used in the breeding program to estimate breeding values. Estimated breeding values (EBVs) are estimates of an animal's value for a given phenotype in the breeding program. Breeding values are used in the selection process to select the best animals for breeding. In practice, they are estimated based on known phenotypes, and the phenotypes of related animals, which is extended with genomic information in genomic selection. This means EBVs are a sum of both a polygenic and a genomic random effect. The genetic random effect is correlated with a SNP based genomic relationship matrix, and the polygenic random effect is correlated with a relationship matrix. The parameters in the model are estimated using mixed models, based on average information REML, and the EBVs are best linear unbiased predictions [18]. Overall, using genomics offers more accurate breeding values, even without the need of phenotyping an animal or close relatives [19]. Beyond the increased accuracy, the improved possibility of prediction without phenotypes allows for shorter turnover time in breeding in general, as accurate predictions can be made as soon as genetic data is available regardless of the availability of phenotypes and extended to animals with no testing.

Feed efficiency

Feed Efficiency is the most important phenotype in commercial pig breeding, as feed represents the single highest cost in pig production [20, 21]. Beyond commercial interests, as the need and demand for more environmentally friendly food production is increasing, higher efficiency leads to more sustainable production. There are two main metrics used for feed efficiency in pigs. The first one is feed conversion ratio (FCR).

FCR is simply the ratio between growth and feed intake. The second method is Residual Feed Intake (RFI) which was first suggested by Koch et al in 1963 [22]. RFI is based on the difference between the expected feed intake and the actual feed intake for a given animal. This can be calculating by linearly regressing feed intake as a response to growth, and including any relevant covariates for a given animal, such as fat percentage[23]. The general theory behind RFI is that it represents overall metabolic efficiency, independent from growth rate or animal size [21, 24, 25]. In contrast, it is reported in the literature that selection for FCR will result in co-selection for daily gain and overall growth [21, 24, 25]. This is also supported by a relatively old simulation study, showing that when basing selection of a ratio term, one would actually be selecting for the components of the ratio [26]. It should be noted that these co-selection effects are not negative in pig production, as demonstrated in figure 1, where we saw that increase in daily growth was the phenotype with the second highest positive economic impact. FCR is based on a simple calculation, where RFI relies on individual production and population factors [27, 28]. RFI and FCR have a reported correlation above 0.7, and both RFI and FCR have a low to medium heritability [28]. FCR has been by far the most common measure used in pig production [21]. Inconsistent feed measurements due to feeder loss, varying weight range of animals or differences in feed affect the accuracy of FCR calculations[29]. These issues are not present in our studies given the well-designed nature of the breeding setup in Danish pig production, as feeder loss has largely been removed, feed is standardized and the period of FCR testing is weight based, not time based.

The feed efficiency in pigs is affected by many factors, such as choice of feed, environment, and inherent biological factors[29]. Of these, our interest lies solely in the biological part. It has been estimated that around one third of the variance in FCR between pigs is independent from growth rate and animal composition[30]. This variation can generally be split into protein turnover rate, metabolic rate, activity, thermoregulation and immune function [31-34]. While this accounts for the soruce of variation, the actual molecular background is still somewhat elusive[35].

Metabolomics

Metabolites are small molecules that take part in metabolic processes, and can often be seen as the link between the internal organism and the environment[36]. The metabolome can then be defined as the ensemble of metabolites that partake in maintenance growth and general function of living organisms[37]. Metabolomics is then the study of a comprehensive set of metabolites from a given source, such as blood or urine.

Mass Spectrometry Liquid Chromatography

Mass spectrometry Liquid Chromatography (LC-MS) is an untargeted technique for the identification metabolites (and other molecules or atoms). The LC part refers to the solution of the sample into a fluid mobile phase which is passed through a stationary phase. Different molecules will then move at different speeds through the stationary phase, ensuring different retention times, which in turn will create separation of metabolites[38]. The MS part is an analytical technique that can measure the mass of the metabolites. A MS device includes an ion source, a mass analyzer and a detector. The ion-source is necessary for the ionization of molecules ensuring the generation charged molecules, as charge neutral molecules cannot be detected. The mass analyzer separates metabolites according to their m/z ratio, and the detector is then able to detect the relative abundance by measuring the strength of current or charge at a given m/z ratio, leading to a spectrum[39, 40]. In *Figure 3* we can see a schematic of LC-MS. Here it is important to emphasize that the abundances are relative, meaning that the intensities can only be compared of peaks with the same m/z and retention time. This means that we cannot say anything about the absolute abundances of different metabolites in relation to each other

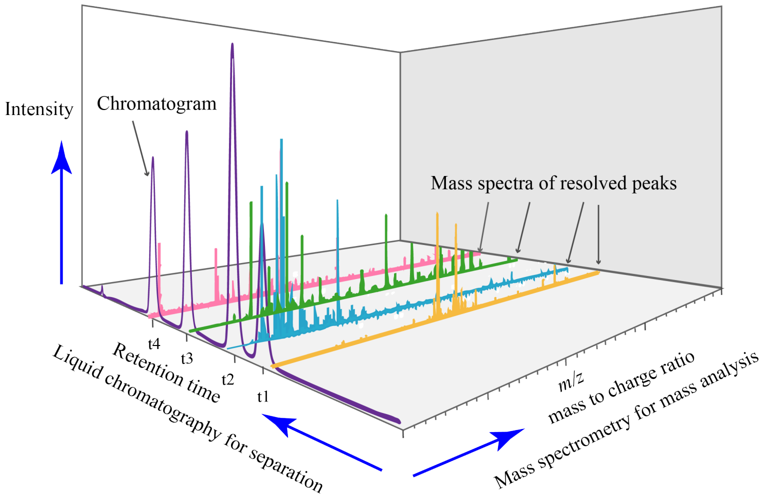

Intensity
Chromatogram
Mass spectra of resolved peaks
t4
t3
t2
t1
Retention time
Liquid chromatography for separation
m/z
mass to charge ratio
Mass spectrometry for mass analysis

Figure 3 Schematic visualization of LC-MS. Figure by Daniel Norena-Caro CC by 1.0 https://creativecommons.org/licenses/by/1.0/.

The Metabolome

The main advantages of the metabolome in relation to the genome or transcriptome is that no reference is needed in principle. We can simply analyze the spectrum of any animal or source without needing to have a species-specific reference. The big challenge in metabolomics is the actual identification of metabolites in the metabolome. In LC-MS analysis of metabolites, typically only under <2% of potential metabolites can actually be identified [41]. Furthermore, less than 10% of known human metabolites have properly validated spectra [42] While genomic and transcriptomic sequencing technologies generally require a priori knowledge for best use, we can easily and relatively cheaply accurately identify the specific sequence of RNA and DNA molecules. In contrast, there is no current solution to this problem in metabolite analysis, and it is estimated that it would cost billions of dollars and decades of research to develop technology that can directly identify specific molecules [43]. Therefore, *in silico* techniques are the main method of identification and annotation of metabolites currently, and there are various accurate open source tools available for this [44, 45]. Using *in silico* approaches and manual curation, the human metabolome database 4.0 contains over a 100.000 known, expected and predicted metabolites [46]. This database is a valuable resource for animal science, as many basic metabolic processes are shared between higher organisms. This means, that if we have a set of m/z values from an animal metabolomics experiments, we can annotate our metabolites using the human metabolome database.

In animal science, and thus pig research, metabolomics can be a valuable tool for non-invasive phenotypic prediction. In animal breeding and selection, one often wants to quantify various phenotypes as early and efficiently as possible. FE is generally expensive and time consuming to monitor, and other phenotypes, such as carcass phenotypes, require animal slaughter, thus invalidating potential breeding. Non-invasive metabolomics have been shown to have the power to detect small differences in phenotypes [47-49], making them potentially useful for subtle phenotypes such as FE. There have been a number of papers linking the metabolome to important commercial phenotypes, such as RFI [50, 51], fertility[52], milk quality.

Decoding the transcriptome

The transcriptome represents a snapshot of all the RNA transcripts and their quantity in a cell. This is seen as a proxy for the activity of the genes associated with the RNA transcripts, thus allowing one to understand which genes, and ultimately, proteins, are active in a tissue or cell type. Furthermore, RNAs themselves can have catalytic activity, as seen in lncRNA [7]. To quantify the transcriptome, one can use RNA sequencing (RNA-Seq) [53]. RNA-Seq is the process of quantifying the RNA molecules including a reading of specific nucleotides, which is generally called sequencing. The applications and studies using RNA-seq are more numerous than one can easily summarize, and a search for RNA-seq at Pubmed gives 25134 hits. This could range from cancer studies [54] to temperature stress response in freshwater fish [55]. Here we present the workflow and methods applied to the analyses of RNA sequencing data, and thus, the transcriptome. The generation of transcriptomic data always starts with a source, which could be anything from a simple cell to a complex tissue. After RNA has been extracted from the sample, the first step is to sequence our RNA. After this sequencing, we must do quality control (QC), mapping and quantification to get to a stage were the data is usable. Once we have properly processed our data, we can perform differential expression analysis to find genes that are associated with our phenotypes of interest, or other more complex analysis methods.

From Sanger to Next Generation Sequencing

The first generally used untargeted sequencing method was Sanger sequencing, but this was expensive, low-throughput and not quantitative [53, 56, 57]. There were also hybridization techniques, relying on the hybridization of specific targets with pre-prepared cDNA or oligo probes, using fluorescence to quantify RNA molecules[53]. These approaches could not analyze untargeted or unknown sequences, and suffered from high background level and lack of dynamic range [53, 58, 59]. However, as the technology matured, probe arrays were able to target thousands of sequences [60]. This means that probe array are still useful today for SNP based genotyping, as the dynamic range issues does not matter for SNPs. Furthermore, in genetics, given the LD structure, it is not a necessity to do full genome sequencing in all cases. As probe arrays are cheap, they can also be used as a low cost alternative for analysis of RNA, or in cases where exact abundance are less important.

While the options above presented some methods for analyzing RNA and DNA sequences, they still posed severe limitations to what was possible. As an example, the initial sequencing of the human genome took 13 years and almost 3 billion dollars to complete [61]. At these costs and time frames, the amount of genomes one could sequence and analyze is severely limited, and we would likely not have a pig genome today, much less other genomes with even less economic impact, such as the sequencing of rare local pig breeds.

NGS was the second generation of sequencing methods after Sanger sequencing, and started to appear in the mid-2000s[62]. There are several platforms for NGS, but they all have some important factors in common: they are cell free systems not relying on bacterial cloning, they perform thousands to millions of sequencing reactions in parallel and the base detection is performed in cycles in parallel [62]. One such example is the Illumina platform, a widely used platform for NGS. Illumina sequencing can be separated into the following general steps:

1. DNA/RNA extraction - First, the genomic DNA (gDNA) or RNA of choice must be extracted and isolated
2. Library preparation - DNA and RNA are first fragmented into random small sequences. RNA must then be transformed into cDNA using reverse transcriptase. After this, adapters must be ligated to each end of the c/gDNA.
3. Cluster Amplification - The library of fragments are then loaded into a flow cell and hybridized to the surface. Each fragment is then amplified through bridge amplification, generating cluster of identical fragments
4. Sequencing – Fluorescently labeled nucleotides are then added in the flow cell, and incorporated sequentially at each position in each fragment. After each round of incorporation, the wavelength and intensities of all sequence clusters are imaged, allowing for the identification of the nucleotides.

A typical Next generation sequencing experiments produce millions of reads, and a lengths of the read length available for a given platform range from 25 kb all the way to 15kb[63, 64]. The reads must be aligned for further analysis, a topic we will cover in the section on the transcriptome.

Data QC

In sequencing experiments there will always be a certain amount of base calling uncertainties and unwanted events, such as sequencing of artificial adapter sequences. While modern computational mapping techniques are quite robust to errors, it is still best practice to properly QC data. RNA-seq experiments can still fail completely in some cases, and this can most often already be seen in QC steps. Raw sequencing data will include the sequence with associated quality outputs, which inform us of the likelihood of correct base calling. Based on the quality, analysis of artificial sequence content, duplicated and overrepresented sequences and other similar metrics one can validate the quality of RNA-seq data using a tool such as FastQC [65]. After initial QC, one would use a trimming tool to remove artificial sequences and low quality bases, for example Trimmomatic [66]. The QC should then be repeated to verify that the trimming step has worked as desired.

Read Mapping

One we have done QC on our data we are ready for mapping. Mapping is the process of finding where in the genome each of our reads comes from. In the context of the pig genome, the assumption when mapping is that we have a reference genome available. A reference genome in raw text form is comprised of all the nucleotides, coded as A, T, C and G, sorted in the right order and separated into the chromosomes of that specific reference. As it is a reference, it will be based on a specific source. For example, the original pig genome was based on a female Duroc pig [1]. As high throughput sequencing experiments produce millions of reads, and genomes are billions of base pairs long, when matching a read to the genome, a simple exhaustive search would be too slow for practical purposes, as we would need to do as many comparisons as the length of the genome minus the length of the read for each individual read. Instead, we first index the genome, making searches much more efficient [67, 68]. An example of an indexing technique is a suffix array, which transforms the genome into a sorted list of all the possible suffixes in the genome up to some maximum suffix length, allowing for fast lookup of a sequence [67, 69]. When mapping RNA-seq, one must also take the splice sites into account. As mature mRNA is spliced, many reads will map to two separate exons. If a mapper is not able to handle gaps in the alignment in relation to the reference, it cannot properly map spliced reads. Beyond specific algorithmic techniques for dealing with gaps, mappers also take advantage of the fact that splice sites have a canonical structure, with specific donor and acceptor sequences [67, 70]. Furthermore, mappers can also include known splice sites from annotation in mapping [70]. More on annotation below.

Read quantification and annotation

Once we have mapped our reads, we in principle have all the information we need for analysis. We know that status of our reads – mapped, unmapped or multi-mapped – and we have the exact coordinates for each read. Doing analysis on millions of reads based on the coordinates of each read without further processing would be a daunting task, so instead, we must quantify our mapping in a more summarized form. In this step, we will assume two things: that we have mapped to reference genome and we have available annotation. Genome annotation include all the known genomic elements of a genome and their coordinates. For mammalian genomes, this will typically be genes at the top level, which then can contain further subdivisions, such as introns or exons or individual transcripts. A common format used for annotation is the Ensembl gene annotation system[71]. The gene annotation of the pig was released by Ensembl in 2012 and is currently in version 11.2 as of date of writing. There are two general approaches to quantification – transcript-based quantification, and gene-based quantification. Transcript based assembly methods, such as Cufflinks[72], can use annotation to generate abundance estimates of individual transcript variants. These methods are very computationally intensive and show low concordance, and thus technologies for accurate transcript abundance estimation are not mature yet [64]. The other quantification strategy is to count the number of reads within each annotated genomic feature one wants to quantify. When working with genes, one can simply count the number of reads that are mapping within the coordinate boundaries of the gene. The final output is then a discrete count for each of the genes in our annotation for each sample, which then can be merged into a count matrix for further analysis. More sophisticated quantification methods are possible within this framework, but they follow the same basic principle. Common methods used to quantify reads include HTseq [73] and bedtools [74].

Differential expression analysis

Differential Expression Analysis (DEA) is the procedure of analyzing the expression of individual genes, and identifying significant changes in expression due to differences in condition, disease or phenotype. Before starting any kind of analysis comparing separate sequencing libraries, one must always start by normalizing the data. This is due to technical variation between libraries, as there will often be differences in library size and /or other technical biases, which we are not interested in modelling. Early normalization techniques performed linear transformation of read counts based on library size or GC content, but these methods do properly model more complex effects between sequencing libraries [75] [64]. More complex normalization methods can be split into two categories, normalization using control or using modelling [76]. Normalization by control relies on adding known sequences in known concentrations in each library before sequencing. These sequences can be designed in sophisticated ways to cover a range of lengths, abundances and GC contents [77]. Normalization by modeling is the more common approach, and is used by the most popular DEA tools, such as Limma, EdgeR and DEseq2 [78-80]. These approaches generally assume that most genes between libraries are not differentially expressed, and that the overall expression levels are similar. For example, in DEseq2 normalization, the ratio of read counts and the geometric mean is calculated for all genes in each library. The final correction is the mean of all these ratios in each library, giving one normalization factor per library. If there are large changes in the libraries, then the modelling based approaches may fail to create comparable libraries. When working with data from farm animals, often one does not expect large changes, so the extra expense of control-based methods may not be worth it.

Once we have appropriately normalized our libraries, we can estimate the actual differential expression (DE) between genes. The most widely used technique for modeling DE is based on the negative binomial distribution. The reason to use NB as a model for count distributions, is that read count data has been observed to be heteroscedastic and over dispersed, which cannot be modelled using a Poisson or binomial distribution [81] [82]. The challenge with using the NB, is that for each gene we must estimate a dispersion parameter. If we have large sample sizes, this is not an issue, but in practice, we often do not have enough samples to reliably estimate dispersion. One way to deal with this issue, is take advantage of the large number of genes available in an RNA-seq experiments, and model the dispersion using the data from multiple genes with similar expression ranges [79, 80]. Once we have appropriately modeled our data, differential expression can be calculated in different ways, such as using a generalized linear model, which is implemented in both DEseq2 and EdgeR.

Pathways and enrichment analysis

Given the large amounts of data produced in modern sequencing experiments and the thousands of genes that can be active in a given cell, interpreting results of transcriptomic experiments is often a challenge. Even the output of a relatively simple DE analysis can give hundreds or even thousands of DE genes in some cases. Thus, any tools that can add functional biological knowledge, both to individual, but also to groups of genes can greatly aid in the interpretability of our data, and give us deeper insight into the molecular functional background of the results in our studies. One important resource in this context is Gene Ontology (GO) [83, 84]. GO was created at the turn of the millennia as a response to the ever-growing molecular biological knowledge. The idea was to create a structured database of standardized functional biological knowledge of genes, designed to be easy to use in computational analysis. There are currently over 45,000 terms and 134,000 relations in the GO database, from over 7 million genes and gene products in over 3200 species. The terms and relations have been carefully constructed over 20 years, based on scientific literature and expert domain knowledge. To give a specific example, we can look at GO:0070125, mitochondrial translation elongation. It is defined as “The successive addition of amino acid residues to a nascent polypeptide chain during protein biosynthesis in a mitochondrion”[84] . The term itself is under the translation elongation (GO:0006414), which is part of translation (GO:0006412), which is part of gene expression (GO:0010467) etc. Given a gene ID, it is the easy to retrieve functional pathway information through databases, such as BioMart [84].

Once pathways have been identified in dataset, based on GO, it is often of interest to perform enrichment analysis, thus statically quantifying if certain biological processes are related to the features the dataset phenotypes. For example, if we perform DE analysis based on FE and then find certain pathways enriched, we then have evidence of the functional biological background of FE. Thus, the general principle of pathway analysis is that there will be some abnormal distribution of pathways for genes which are related to specific biological conditions. There are many different methods for pathway enrichment, that each have particular features, and already back in 2008 Huang et al included a list of 68 tools in their article on the topic [85]. They generally work by calculating enrichment of GO terms using an appropriate background, and then applying some common statistical test, such as a Fisher test, χ^2 test or using the hypergeometric distribution [85, 86]. It is then important to apply multiple testing correction when testing for pathway enrichment, due to the many GO terms a gene set can contain. Choosing a background set is relatively straightforward in a transcriptomic experiment, as we can use the total of set of expressed genes. Examples of popular methods include DAVID [87] and GOrilla [86].

All the concepts mentioned above apply in an analogous fashion to metabolite analysis, just based on different tools and databases. One tool for metabolite pathway analysis is IMPala[88]. Given a set of annotated metabolites, IMPala searches for pathway information using 11 different databases, among them the KEGG[89] and Reactome[71]. Given pathway knowledge, it is then also possible to link metabolites to genes in shared pathways.

Network Analysis

If we want to move beyond DEA and individual gene analysis, the next logical step is to use network-based methods. If we imagine the transcriptome of a cell, there is a continuous process of transcription, which in turn is being regulated by transcription regulators, such a transcription factors, or post-transcriptionally, by miRNAs [90]. Thus, a picture emerges of a widely interacting transcriptome. Furthermore, genes in the same metabolic pathways or feedback systems will naturally have related expression patterns [91, 92]. If we look at metabolites, a similar reasoning applies. Metabolites can be the bi-products of the same processes, form part in related pathways and are known to have network like structure [93-95].

A network is defined by two properties: a set of nodes, and the edges between them. Nodes can then represent genes or metabolites, with the edges indicating the connection between them. There have been several proposed methods for gene network analysis proposed in the past [92, 96, 97], however, here we will focus on one of the most popular and widely used, Weighted correlation network analysis (WGCNA) [98].

WGCNA

The final goal in WGCNA is to identify biologically meaningful modules, which can be analyzed via pathway analysis and /or correlated with phenotypes of interest. The first step is to create a network based on gene similarities. Perhaps the simplest and most obvious metric for relation between quantitative measurements is correlation. In particular Pearson correlation [99] is widely used, which is the linear correlation of two variables, ranging from -1 (perfect negative linear correlation) to 1 (perfect positive linear correlation),. The first goal is to create an adjacency matrix based on all pairwise correlations between all gene pairs. In WGCNA, instead of using pure correlations for the edges between two genes, gene i and gene j, the correlation values are raised to a power β, as seen in *Figure 4*. The β value is chosen based on a scale-free network topology criterion. A scale free network is a network where the connections follow a power law, which means that the fraction of nodes $P(k)$ that have k connections is $P(k){\sim}k^{-\gamma}$ for large k. It has been observed that real networks, such as communication and, most importantly, biological networks are observed to have a scale free topology. The optimal fit for β is then the value that maximizes the R^2 value between the linear regression of $log_{10}(P(k))$ and $log_{10}(k)$. In general, WGCNA accepts a network as scale-free if $R^2 > 0.8$

$$a_{ij} = \left|cor(x_i, x_j)\right|^{\beta}$$

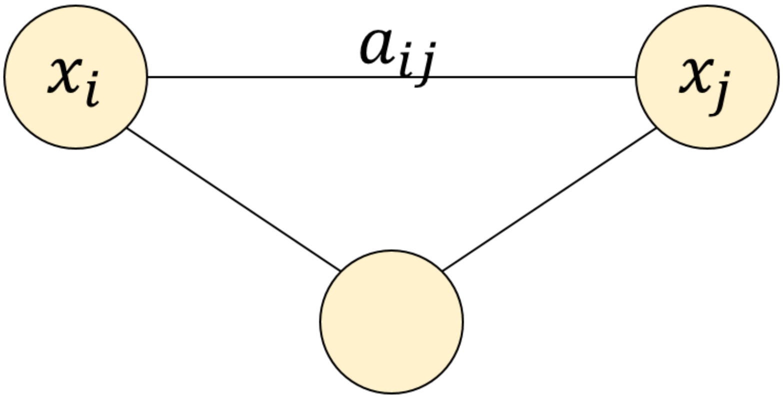


Figure 4 Schematic WGCNA network, with genes and associated edges.

Once a scale-free adjacency matrix has been created, the next step is to identify similarity between nodes for a clustering into modules. To do this, WGCNA uses the topological overlap between nodes, which gives us a topological overlap matrix (TOM). This measure reflects how interconnected two nodes are, and has been found to be biologically meaningful. The TOM is calculated as follows:

$$\omega_{ij} = \frac{\mathrm{l}_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}}$$

With $l_{ij} = \sum_u a_{iu}a_{uj}$ and $k_i = \sum_u a_{iu}$. It is possible to show, that if a_{ij} is between 0 and 1, which it will be if it is based on correlation, that ω_{ij} will also be between 0 and 1. ω_{ij} is a similarity measure, and WGCNA uses $1 - \omega_{ij}$ for the clustering, thus transforming it into a dissimilarity. The TOM dissimilarities are then hierarchically clustered, and a cutoff is selected in the resulting dendrogram. The branches based on this cutoff will be the modules. There is no exact way of defining cutoff thresholds, but WGCNA includes default suggestions[98]. After modules have been generated, there are many analysis possibilities, including but not limited to: correlating module eigenvalues with phenotypes, module pathway analysis and hub gene analysis. All of the above methodology was first presented by Zhang et al in [100].

While WGCNA was designed for gene networks, given the general biological motivation of the method, it is natural to apply it to metabolites. Indeed, there are several studies where WGCNA has been applied in metabolomics analysis [51, 101].

eQTLs

Unraveling the functional genetic background of complex phenotypes is often a challenging endeavor. While many genetic markers have been associated with specific phenotypes and diseases, the functional effect of genetic variation has been more elsuive [102] [103]. One way of analyzing this problem, is to link genetic variance directly to gene expression. This offers a straightforward connection between genetics and function, and aids in functional interpretation, as we can take advantage of gene pathway information. This type of analysis is called expressed quantitative phenotype loci (eQTL) mapping. Thus, an eQTL is a locus, which can explain the genetic variance of the expression of a gene [104]. eQTLs can be separated into two major categories: cis-eQTLs, which are local acting and trans-eQTLs, which are distally located in relation to gene they are acting on. More analysis has been done on cis-eQTls [105].

As cis-eQTLs act locally, they have a more direct functional link to expression of the gene through local genomic context. Furthermore, the number of potential cis-eQTLs only grows linearly with the number of genes, making feasible both computationally and statistically. In contrast, in trans-eQTL analysis, one must test all possible association between each locus included and each gene. This gives rise to computational issues – testing the relation between 10^5 SNPs and genes requires in the order of 10^8 tests. This in turn, give multiple testing problems making it even more difficult to meaningfully detect trans-eQTLs, especially as they are reported to have smaller effect sizes than cis-eQTLS [105, 106]. It is therefore important to have good data strategies when doing trans-eQTL analysis, which could include appropriate filtering of both genetic and expression data using measures such as expression heritability [107], relation to a phenotype of interest [108] or other meaningful measures.

When modelling eQTLs, regardless if it is cis or trans, the actual modelling will look the same. There are essentially two option – additive modelling, were genotypes are transformed into a numeric scale, such as 0 (only reference alleles), 1(heterozygous) and 2 (only non-reference alleles) and fit as a continuous variable, or a factor model, were each genotype is treated as separate factor level. The factor level requires more parameters to be estimated, but is more flexible as it can fit a wider range of scenarios, such as dominant or recessive models [109]. One useful tool for eQTL analysis is Matrix eQTL [110]. Matrix eQTL is able to do both additive and a factor-based model, and due to both computational efficiency and heuristics is very time efficient. This allows it to perform cis and trans eQTL analysis on modern datasets.

Paper A - Metabolomic networks and pathways associated with feed efficiency and related-traits in Duroc and Landrace pigs

Motivation

Measuring FE related phenotypes in pigs is costly. Therefore, if non-invasive techniques could identify predictive FE biomarkers, this method could be applied as an early screening tool in breeding programs, thus saving resources. One possible source for non-invasive biomarkers is the blood metabolome. Beyond practical applications, there is also a scientific interest in identifying links between FE and the blood-metabolome, namely, to strengthen our understanding of the functional background of FE. Blood is an ever-changing tissue, as it is responsible for transporting a wide range of processes, from sugar to immune response. It has also been shown that the blood metabolome is affected directly by environmental factors [111]. If we combine this with the very high growth rate of pigs, one can easily hypothesize that early metabolite screening would not be effective, as the blood, metabolome may be too transient. Therefore, we also wanted to demonstrate that the blood metabolome has enough temporal stability and biological impact to be a useful predictive biomarker.

Data

The results of this paper were based on 109 pigs, including 59 Duroc and 50 Landrace. Blood samples were collected from each pig at two time points, one at the start of the FE testing phase and a weight close to 28 kg, and again 45 days later. Non-targeted metabolomics analysis was performed on blood plasma using LC-MS. Based on this, 729 metabolites were identified in the data.

Methods

To show that the metabolites were meaningful overall and somewhat stable over time, we performed PCA of all metabolites, and fitted a linear model between the two sampling time points for each metabolite. The log-normalized metabolites where adjusted by extracting the residuals from a mixed linear-model, to correct for fixed effects and the random pen effect. Adjusted metabolites were then analyzed using a simple linear model, for daily gain (DG) from birth to end of testing, early daily gain (EDG) from birth start of test phase, testing daily gain (TDG) during test phase, RFI and feed efficiency (FE*). FE* in the context of this paper will be the ratio between growth and feed intake, in contrast to FE, which will still denote feed efficiency in general. Models were performed separately on breed, time point one/two (TP1/TP2), and a joint time point model. The Kolmogorov-Smirnov test (KS-test) was used to assess the p-value distribution of each model in comparison to a theoretical uniform distribution. WGCNA was used to create metabolite modules based on the Spearman correlation of adjusted metabolite concentrations. To associate modules with phenotypes, a linear model between the module's eigenvalues and our phenotypes was applied. Metabolites with p-value < 0.05 in the linear modelling, and from modules with a phenotype-module correlation > 0.2 and p-value < 0.1 were selected for further pathway and gene-metabolite analysis. Pathway enrichment analysis was done using IMPala. Gene-metabolite networks were generated using Metscape.

Results

Based on the visualization of the first two principle components from the PCA, we saw a clear division into our four sampling groups – Duroc vs Landrace and TP1 vs TP2. The linear relationship between metabolites showed that over half of all metabolites had significant (p-value < 0.05) linear relationship between TP1 and TP2. KS-test revealed a left-skewed p-value distribution for most of the models across metabolites in FE and RFI, with a maximum p-value of 0.25 across all 10 models from these two phenotypes. Similar results were also found for the daily gain phenotypes, albeit even more significant. As for individual metabolites, based on an FCR < 0.05, we found 1 metabolite for FE*, 0 for RFI, 9 for EDG, 21 for DG and 37 for TDG. The most significant results were in Duroc TP2 and TDG. Based on the WGCNA analysis several modules correlated with our phenotypes, with the top two modules both being in Duroc, one associated with FE* (cor = 0.34) and the other with TDG (cor = 0.4). The pathway enrichment and gene-metabolite network analysis revealed many pathways, genes and hub metabolites in the four different sampling groups, with highlights being cholesterol sulfate, which was involved in FE*, TDG and RFI in Duroc TP2 and RFI in Landrace TP1 and 2-Aminoadipate, which was related to TDG in Duroc TP2 and EDG and DG in Landrace TP2.

Conclusions

We demonstrated that the blood metabolome is generally biologically relevant, and that it shows stability over time, despite the quick morphological changes happening in pigs during the sampling period. On an overall level, we found an significant association between FE phenotypes and the blood metabolome, and even more significant for growth phenotypes. In particular, we observed that metabolite profiles of pigs were associated with phenotypes that had not been measured yet, as the Duroc/Landrace TP1 KS-tests for FE*, RFI and TDG had a p-value < 0.1 for every category except Landrace TP1 – FE, which had a p-value of 0.25. For the metabolites individually, there were only limited results, except for TDG and DG. Based on the linear modelling, metabolite networks and gene-metabolite networks, we identified several metabolites that could be potential phenotypic biomarkers. Overall, our results showed promise for the application of metabolites in pigs for FE, and in particular, in growth phenotypes, while also revealing the challenge of analyzing FE in general.

Paper B - Genome regulation and gene interaction networks inferred from muscle transcriptome underlying feed efficiency in Pigs

Motivation

Muscle is the most important tissue in pig production, as it harbors most of the economic value of the carcass. Indeed, pig farmers in Denmark are payed by the estimated amount of meat in a carcass. Beyond this, muscle is a major metabolic tissue with a large role in energy metabolism as a whole. The overall goal with improving FE, is also mainly directed towards the improvement in lean weight of animals, which represent the muscle tissue. Thus, there are strong motivations to analyze the relation between FE and the muscle transcriptome, to shed light on the molecular differences between more or less efficient animals. Naturally, previous studies exist analyzing the pig transcriptome for FE. Our study offered several novelties in relation to the past findings. As FE is a complex trait expressed through multiple factors, the individual effects are expected to be weak, thus one should aim to maximize the data available. Our study had the highest number of samples reported in this type of study. Previous studies had focused on extreme outliers or used divergently bred populations. We used pigs from a real breeding, population with a range of efficiency values. In the real world, pigs are not divergently selected for FE, and there are no low FE selected pigs. Therefore, our study design was more realistic and thus, applicable. Finally, two breeds were included in our analysis, instead of only one. This can be seen as a weakness, as it introduces an extra variable into the modeling, but it also means that any results we might find could be more broadly applied.

The ultimate goals of the analysis were to gain new insight into the molecular biological background of FE in muscle tissue and identify genes or pathways that could act as biomarkers for FE.

Data

This study was based on a subgroup of 41 pigs from our main pig group (described in Paper A), including 13 Duroc and 28 Landrace pigs. Muscle tissue samples were collected immediately post slaughter, and RNA-Seq was performed on each sample.

Methods

In Paper B, we applied DEA, gene expression interaction analysis, gene network analysis, pathway enrichment analysis and comparative functional analysis. DEA was done to identify DE genes, using three different methods, Limma, EdgeR and DESeq2. Due to the overall anti-conservative distribution of our p-values in relation to FCR, we selected a group of genes based on the comparison between our empirical p-value distribution and uniformly distributed p-values for pathway analysis. All pathway analysis was done using GOrilla, utilizing the full set of expressed genes in our samples as background. These same genes were also used in a gene-gene interaction analysis to identify possible pairwise interactions among the top FCR genes, and the most interacting genes overall. WGCNA analysis based on all genes was done, clustering the genes into modules. Module eigenvalues were then correlated with our traits, and the genes in top correlating modules were subjected to pathway enrichment analysis. Finally, top DE genes for FCR and breed were analyzed for enrichment in differentially expressed genes in human muscle transcriptome pre- and post-exercise from three separate studies.

Results

In the DEA, we identified 14 DE genes with FDR < 0.1: two genes in Landrace, nine in Duroc and four in a common DEA for both breeds. The highlights included two mitochondrial genes (MRPS11 and MTRM1), and TRIM63, which has been found to be a biomarker for exercise induced muscle damage [112]. Based on the overall p-value distribution of DEA analysis, we selected genes in the common analysis for pathway analysis, using the overlap between all three DE methods, based on each individual analysis' divergence from uniformity in the p-value distribution. Pathway enrichment analysis from the DEA genes revealed five enriched pathways, with 4 being related to mitochondrial gene ontologies.

Based on our pairwise gene interaction analysis, we saw that the interaction p-value of individual genes were left skewed in relation to bootstrapped sets of p-values, as 193 genes out of 853 had more anti-conservative p-values than the most anti-conservative bootstrapped results. Given the context of the analysis, we chose a conservative heuristic approach and reported the top 20 genes based on their overall interaction. This included several interesting genes, such as transcription regulators, lipid metabolism genes and mitochondrial genes. Interestingly, the most interacting gene according to our analysis was ETV1, which is a transcription factor whose activity is mediated by androgen and involved in overall growth [113].

The WGCNA gene network analysis revealed two modules that were correlated with FCR (correlation ≥ 0.4). Pathway analysis of the first module revealed high enrichment of mitochondrial genes, grouped into three overall ontologies: translation elongation, electron transport chain and hydrogen ion transmembrane transport. Furthermore, in this module, seven of the top 10 most connected genes were all from the same gene group, the NADH ubiquinone oxidoreductase group (NDUF). The other significant module included 3744 genes, and many enriched pathways, which were summarized in the DNA repair ontology. This included ontology related to DNA, RNA and amino/nucleic acid processing and metabolism. Combining this with the size of the module, the interpretation could be that the module is related to generic growth and maintenance processes in the cells. As the module was positively correlated with FCR, one can then speculate the higher expenditure on these processes lead to worse efficiency.

Finally, to improve our understanding of the functional nature of FE in muscle, we hypothesized that genes differentially expressed between our two breeds and FCR would be enriched in transcriptomic analysis of muscle pre- and post-exercise. We saw a consistent pattern of enrichment of both breed and FCR associated genes across three different human studies.

Conclusions

We showed that while individual gene relations to FCR may be weak, we could still identify interesting results if we rely on overall distributional effects. We identified interesting genes based in DEA, gene-gene interaction analysis and from the most connected genes in FCR related modules. Based on this, we were able to confirm previous finding of the relation between FE and mitochondrial in a novel context and applying novel methodologies for FCR. We identified the possible involvement of the DNA repair pathway group as relating to FCR. We established a novel link between FCR and exercise induced changes in muscle. Based on our analysis we propose that mitochondrial genes and the NDUF group in particular could be used as FE biomarkers.

Paper C - eQTL and pathway enrichment analysis on FCR and mitochondrial genes of Danish performance tested pigs

Motivation

The analysis of the genetic background of complex traits is often challenging. In the context of animal science traits, this can be particularly difficult, as it prohibitively expensive to do functional molecular studies based on FE on large animals, such as typical farm animals. eQTL analysis, combined with selecting genes linked to FCR is one strategy for connecting genetics to pathways without needing overly complex experimental designs. This can then lead to the identification of possible genetic biomarkers for FCR. One further motivation for this study was to tackle the issue of sample size versus the complexity of trans-eQTL. How can one validate trans-eQTLs, given that there are so many tests involved, combined with their postulated weak effects? We hypothesized that low p-value trans-eQTLs would be enriched for genes that are highly interactive in general, and that would specifically interact with genomic context.

Data

The data in this paper came from a subset of 38 pigs from the 42 pigs included in paper B. These pigs were genotyped using the GGP Porcine HD array (GeneSeek, Scotland, UK), including 68,516 SNPs on 18 autosomes and both sex chromosomes.

Methods

The first important step was to filter the input data in conservative, but non-arbitrary way. For this, we started by selecting the 853 genes we had identified in paper B from the enrichment in the DEA. We also included all genes with a mitochondrial ontology, as these were identified to be associated with FE not only in our study, but in several other studies in multiple species. After expression-based QC, the final list included 1425 genes. The genotype data was filtered based on two criteria: a minor allele frequency > 0.3 for each genotype, and an LD filtering grouping neighboring SNPs linearly across the genome into clusters if the $R^2 > 0.9$ between all possible pairs in each cluster. We then performed both additive and factor based eQTL analysis using Matrix eQTL. We observed similar results as in paper B – a relatively limited set of individual results post multiple testing correction, but a quite significant enrichment of low p-values. Thus, we selected the top 28147 top trans-eQTLs based on the enrichment of p-values below 0.01, which corresponded roughly to the top 10% trans-eQTLs with p-value < 0.01, for targeted pathway enrichment. The targeted pathway enrichment was specifically based on ontologies which were related to expression regulation and DNA binding.

Results

We identified 15 eQTLs (14 trans, 1 cis) with a FDR < 0.1. Beyond these, we also presented and did qualitative analysis of the top 10 eQTLs in each model category (additive/factor based and cis/trans). Several of the eQTL associated genes were differentially expressed among genetically divergent pig breeds, or other contexts indicating possible genetic regulation potential. In our targeted pathway analysis, transcription factors, and negative regulation of expression were highly significantly enriched, in comparison to both the input set and the overall expressed genes in the tissue. Furthermore, both DNA binding and DNA-binding transcription factor activity genes were also highly significantly enriched.

Conclusion

We identified a set of potential eQTLs , using both statistical and qualitative evidence. We also presented novel strategy for validation of trans-eQTLs, based on in-silico biological validation. The enriched pathways we found are not only interesting as a result in our pig population, but present a novel way of dealing with and analyzing the computational and statistically difficult topic of trans-eQTLs.

Discussion, conclusion and perspectives

Paper A

To the best of our knowledge, this was the first time a study was published analyzing FE and DG traits using metabolomics in pig. The work in this paper showed the potential of metabolomics for FE and even more so in DG, but also identified the challenges of the analysis of FE, given its complex multifactorial nature. This challenge was a common theme in all publications, thus we will discuss it further in the overall perspectives.

What do other studies tell us of metabolomics in pigs or possibly other livestock species? The only other FE metabolomics full publications we were able to identify were all in in cattle [51, 114-116]. Given the large morphological differences between pig and cattle in general, and more specifically, their completely different digestive systems, comparing specific metabolites from these studies may not be a good strategy. As we can safely assume that FE is just as multifaceted in cattle, the studies can give us an idea of the overall power of metabolomics in an FE context. Across all four studies, they report 1, 4 and 8 metabolites with significantly different concentrations across FE groups, with the last study [114] not reporting any specific metabolites due to their analysis methods. Thus, we see similar picture as in our study, with only a limited set of significant metabolites. We do not know if they also found an overrepresentation of low p-values, as the overall distribution of results is not reported. If we look at studies involving production pigs, regardless of what phenotypes were studies, Goldansaz et al. report 18 pig metabolite studies in blood plasma [117]. These were identified using text mining techniques, and included somewhat relaxed criteria for what a metabolomics study is, namely a minum of 8 metabolites. In the litterature it is hard to find a good comparison to our study. There are several studies looking at the impact of different type of diets on the metabolome, but all our pigs ate exactly the same feed [118-120]. Another major category is pig animal model studies for various health conditions [121-124]. As the pigs in these studies are subject to drastic experimental conditions, it does not relate to the stable and identical environment for our pigs. There were also more general studies without a treatment group, such as one on analyzing sex dimorphism in the metabolome [125] and one on breed differences in the metabolome. In the breed study, they find 5 metabolites to be different between breed. The breed was not the focus in our study, but we did see clear separation based on PCA, and although not in the article, there were significant breed difference in the metabolome. Overall, the studies above do not offer much comparison, except for possible methodological strategies.

How should we then view paper A? The main usefulness of the paper, is as a pilot study. It offers us the first glimpse into the power of the metabolome for FE and growth phenotypes. How should we continue from this? First, we should discuss what could be improved within the framework of what was already done. While we performed many layers of analysis, including metabolite, pathway and gene information, in the end, we did lack some clarity in the overall results. This does not mean the results are invalid or done improperly. However, due to the various selection criteria and layers of networks, it became a bit unclear what the exact robustness or power of the results was. As several metabolites were quite significant, especially for TDG, these could have been highlighted more. In the pathway enrichment, perhaps we should have selected metabolites on a simple criterion based on the linear modelling, and then from the top modules from WGCNA. Other improvements could be the refinement of the statistical analysis. A more thorough analysis of how metabolites are distributed, what an appropriate model is for metabolite data and how to make the analysis robust could be interesting for further analysis. Here a good start point could be many of the studies mentioned above.

The further analysis and applications of this topic has great potential, and many possibilities. Looking from a molecular perspective, one would like to add more layers of data, such as genomic and transcriptomic data. Thus, we could really take advantage of the pathway and gene-metabolite networking techniques already applied. If we think from a computational perspective, given the complex nature of the data, it seems that the application of machine learning, unsupervised learning methods and some feature selection could be interesting in identifying the predictive power of the metabolites. Performing feature selection in relation to a trait of interest can also add knowledge to be used in pathway and gene-metabolite analysis. The goal with more computational analysis would be to identify the predictive power of metabolites at an early age in relation to the growth and FE. If the goal is to use metabolomics in a practical production context, there are several interesting strategies. Before including any metabolite data in a breeding context, one must ensure they are heritable. Thus, a study demonstrating the heritability of metabolites associated with traits of interest could be interesting, as twin studies have shown a wide range of heritabilities in metabolites [126]. One could also simply include raw metabolite data in breeding, in a similar way, as genomics have been used metabolomics breeding. This is not a realistic goal at moment from a practical perspective and would require more research showing it is worth doing, including all phenotypes one would like to include in a breeding program. In general, all of the above-mentioned ideas would benefit from more data. Perhaps using the initial results, one could devise a more targeted set of metabolites for further analysis, and thus make it more cost effective to expand. The target of this analysis was FE, but one should not forget the importance of daily gain, as it had the second highest improvement value in Danish pigs. Here the results were quite promising in Duroc, indicating the further analysis of potentially early screening metabolites could be fruitful. Beyond genetic factors, it may also be the case that the metabolome is representing overall health, thus the screening could be useful regardless of metabolite heritability, if the metabolites are predictive for pigs that will perform poorly in testing due to underlying issues.

Paper B

In paper B, the muscle transcriptome of pigs was analyzed in relation to FCR, with the novelty of using a continuous FCR as the phenotype, in contrast to using low and high FCR/RFI groups based on divergent selection or highly extreme value selection.

We went through the literature and looked at all previous studies in the paper, so we will just paraphrase the main points here based on the four papers identified in the literature, with some additional comments [127-130]. In general, it was found that most studies were lacking in power, used weak statistical thresholds and used divergence selection for FE. The last point it key in the novelty of our study, as in real production pigs there is no selection for poor FE. The most common results, which we also found, was the relation between mitochondria and FE in muscle. Given the commonality of the mitochondrial relation, perhaps what is needed is a deeper analysis of this result. What is the causal structure, do more efficient animals have higher mitochondrial activity or more mitochondria to begin with, or is it a side effect of efficiency? Increased mitochondrial activity under dietary restriction has been reported, indicating that a link to metabolic efficiency when resources are scarce [131]. Furthermore it has been reported that mitochondrial activity is involved in myoblast differentiation [132], and mitochondrial activity has been linked to protein synthesis through the mTOR pathway [133]. In a small study from 2009 in cattle [134], higher activity, and not higher number of mitochondrial is reported in low RFI animals. All of these studies indicate that mitochondria really do play a role in FE. The other molecular finding we had, the association of DNA repair with FCR, was a novel finding. In our analysis, DNA repair came up as an umbrella term for general nucleic acid metabolism and process. One can argue that there is a biological interpretation of our results, as higher DNA repair was associated with higher FCR, and thus less efficiency. This resonates well with the idea of the effect of biological maintenance on efficiency. In the literature, there are only extremely vague links our observed effect, with a study reporting change in genes regulating DNA repair under feed restriction [135] and one candidate gene related to DNA repair in a QTL region for RFI in Nelore cattle [136]. We should in general be careful with this result, especially as it came from a module in our network analysis that contained roughly one third of the genes. This can make use think that it is simply a cluster for genes that do not fit other places. Thus, more careful analysis and new studies may be needed to confirm this.

The other major finding in paper B was a relation between exercise and FE. This was an interesting novel finding demonstrating the power of modern bioinformatics. As we are able to access data from many studies, and convert genes between species, we can test many interesting hypotheses. Here the idea was simple, after exercise, a muscle grows. As efficient muscle growth is key for FE we hypothesized there could be such a link. Importantly we do not imply that pigs should do exercise, but that pig muscles are in an exercised state without any exercise. As with most of the results, this need further investigation, but also serves to show that there are many strategies one can use to link trait, genes and function. As there is a wealth of both human and animal studies with public data, many results can likely be found based solely on *in silico* approaches.

What is the perspective of paper B? While we used more samples than previous studies, it is clear, that with FE phenotypes, more is better. Furthermore, given the complex nature of the trait, samples from more tissues are also needed, such as liver, and possibly brain, as we have casually observed that Duroc boars, which are the most efficient on average, is generally are less active, thus having behavioral differences. The study of the transcriptome for FE will continue to be difficult, as it is costly to sample a large number of animals and test them for FE. Even if we can collaborate with a commercial breeding test station, the tissue of the best performing animals will not be available. One other avenue we did not pursue was the analysis of the DG phenotypes, which showed promise in the metabolomics paper. The reasons for this was that our overall aims and goals were focused on FE, as it is the more important and difficult phenotype to study. Applying transcriptomic data in a practical breeding context is also a challenge. Instead, one could think of a more targeted approach, based on the mitochondrial connection, as there are techniques for measuring mitochondrial activity. If we can confirm the heritability of the mitochondrial effects, and demonstrate more convincingly the relation to FCR, this could be used as a tool for screening which pigs should do the more thorough testing phase, if non-invasive methods can be commercialized and applied.

Paper C

In the final paper, our aim was to identify potential genetic regulation of genes with potential to be associated with FCR. It was the first eQTL study applied to FCR to the best of our knowledge.

Comparing eQTL studies, as mentioned in paper C, is not a very straightforward thing to do. What is a bit easier, it to compare methodologies, overall strategies and applications? In our case, we performed a quite straightforward and streamlined analysis based on conservative cutoffs, and gene selections. Other studies, with more samples, have included more genetic components, such as heritability calculation of expression levels and pre-GWAS on SNPs [137]. Given our starting point, these strategies were not appropriate, so instead we chose to develop an *in silico* testable hypothesis that could show that the result were meaningful, but also develop new ideas for how to view trans-eQTLS. While the idea of trans-eQTLs being associated with transcription factor is not necessarily novel, directly relating trans-eQTLs with genomic context as we did certainly seem to be. This can have consequences for future trans-eQTL work. First, it would be interesting to see if these effects apply in other datasets. As our observed enrichment were quite strong, there is at least reason to believe so, beyond the theoretical biological considerations of the paper. Given the statistical challenge of genome wide trans-eQTLs, if our results are robust, techniques for conditioning results based on gene type, or on empirical interaction data, could assist in the analysis. In the paper, we also presented individual possible cis and trans-eQTLs. While results were not overly strong, the evidence from the pathway enrichment analysis did lead more credence to these results, which in some cases was backed up by qualitative evidence.

Applying these results as is in a pig production context is straightforward. If one identified eQTLs with strong evidence of relation to both FCR/RFI and with the gene associated having the same relation, one could consider using weighted BLUP, thus giving different weight to different SNPs. This has been shown to work in the literature [138, 139], but at our current knowledge level it is not clear if it would be applicable in pig production. From a basic research stance, it would be interesting to analyze the same pathways we chose at in other datasets. It would also be interesting to use the genomic context as a predictive measure. For example, by identifying DNA binding genes which are targeted by trans-eQTL, and try and predict their associated binding sites, and how genetic variation in both gene and target modulate function.

Overall Perspectives

In the work in this thesis, the common thread was the difficulty of FE as a trait whether as FCR or RFI. Due to the multifaceted nature of the trait, and the relative subtlety of the phenotype, large sample sizes are needed for the identification of significant individual effects. This shifted our focus to methods that could encompass multiple results at once, such as network analysis and pathway enrichment. We often observed that the p-values for our FE related effects were acting as if they were generated using a loaded die rolling an excess of sixes. Thus, there often was a very significant enrichment of low p-values, but they seemed to be lower bounded, thus disappearing after multiple testing correction. This is both encouraging and challenging, as it shows we were able to find general effects, but makes it difficult to be very specific, except for a limited set of results. All three papers in this thesis did identify some potential interesting individual genes/metabolties/SNPs, but in the same vein, all three papers indicate the need of more data. The papers are also inviting for the possibility of the integration of the data presented, as it was the same group of pigs in all papers. Given more time, we could also have applied the ideas in paper B and C to DG related traits, although this was not the focus of the overall project. If we view this thesis from as a basic research point of view, there were several novel strategies and methods applied in novel contexts, which have contributed to the understanding of FE, but also offer potential in general transcriptomic, metabolomic and eQTL analysis. If we view the results from a pig production point of view, we have shown that there is potential in both transcriptomic, genetic and metabolic to act as biomarkers for production traits, but the practical application requires more work.

During the thesis, new knowledge was acquired, which could offer avenues for improving furthering the research presented here. Thus, both more advanced computational, statistical, and given enough data, genetic methods could be applied in the future, using the knowledge that we have developed a basic framework of understanding of FE in a metabolic, transcriptomic and genetics. Beyond the idea of advancing the methodologies, the importance of large sample sizes and good study design cannot be overstated. Given the highly reliable conditions in the Danish breeding system, it does offer possibilities for designing interesting studies building on the studies performed here.

Acknowledgments

I would like to thank Haja Kadarmideen for giving me the chance to work on this project. You offered valuable feedback on science general direction, put up with my somewhat inconsistent nature, and pushed me at the right time when necessary, while allowing me to bring my ideas to the project.

I would also like to thank Claus Ekstrøm, for the feedback and interactions we had. You always had good and useful insights for the project.

Great thanks to everyone, both past and present member of the group. In particular I would like to thank, but in no particular order:

Suraya Salleh, for being of invaluable help with the collection of samples in the countryside and slaughterhouses of Jutland. I thoroughly enjoyed our trips and long days finishing in the lab!

Priyanka Banerjee, thank your huge help with the metabolomics paper. Your help was vital in a time when we needed to push thing. Your hard working mentality and drive was always inspirational.

Gianluca Mazzoni, thank you for the great scientific discussion, it was always great to hear your opinion on all scientific topics, and non-scientific. Your contributions at A-vej will also definitely be remembered.

Markus Drag, for always-enjoyable office discussions, whatever the topic, and for the music!

Wellison Deniz, thank you for your help with the metabolomics paper, and for being a living consultable lexicon on scientific publications

Also, a big thanks to all my colleagues at DTU, it was greatly entertaining to have DTU as a workplace, wherever I was.

Thanks to my mother and sister, who supported me and were always happy to take care of Thalia when needed.

A big thanks to my daughter Thalia, for inspiring me and always being ready with a smile for me. While I cannot understand everything you say yet, I am sure you were trying to help me with my thesis.

Finally, I would like to give a thanks from the heart to my love Effie. You always encouraged me to improve and do better no matter what and were able to be honest to me when I needed it.

References

1. Groenen, M.A.M., et al., *Analyses of pig genomes provide insight into porcine demography and evolution.* Nature, 2012. **491**(7424): p. 393-398.

2. Chen, F.X., E.R. Smith, and A. Shilatifard, *Born to run: control of transcription elongation by RNA polymerase II.* Nature Reviews Molecular Cell Biology, 2018. **19**(7): p. 464-478.

3. Crick, F., *Split genes and RNA splicing.* Science, 1979. **204**(4390): p. 264-271.

4. Pandya-Jones, A. and D.L. Black, *Co-transcriptional splicing of constitutive and alternative exons.* RNA, 2009. **15**(10): p. 1896-1908.

5. Shahryari, A., et al., *Long non-coding RNA SOX2OT: expression signature, splicing patterns, and emerging roles in pluripotency and tumorigenesis.* Frontiers in Genetics, 2015. **6**(196).

6. Gonzalez, I., et al., *A lncRNA regulates alternative splicing via establishment of a splicing-specific chromatin signature.* Nature Structural & Molecular Biology, 2015. **22**(5): p. 370-376.

7. Necsulea, A., et al., *The evolution of lncRNA repertoires and expression patterns in tetrapods.* Nature, 2014. **505**(7485): p. 635-640.

8. Pain, V.M., *Initiation of Protein Synthesis in Eukaryotic Cells.* European Journal of Biochemistry, 1996. **236**(3): p. 747-771.

9. Polderman, T.J.C., et al., *Meta-analysis of the heritability of human traits based on fifty years of twin studies.* Nature Genetics, 2015. **47**(7): p. 702-709.

10. Lush, J.L., *HERITABILITY OF QUANTITATIVE CHARACTERS IN FARM ANIMALS.* Hereditas, 1949. **35**(S1): p. 356-375.

11. Reddy, V.R. and F. Jabeen, *Narrow Sense Heritability, Correlation and Path Analysis in Maize (Zea Mays L.).* Sabrao Journal of Breeding and Genetics, 2016. **48**(2): p. 120-126.

12. Sachidanandam, R., et al., *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.* Nature, 2001. **409**(6822): p. 928-33.

13. Weber, J.L., et al., *Human diallelic insertion/deletion polymorphisms.* Am J Hum Genet, 2002. **71**(4): p. 854-62.

14. Iafrate, A.J., et al., *Detection of large-scale variation in the human genome.* Nat Genet, 2004. **36**(9): p. 949-51.

15. Altshuler, D.M., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-+.

16. Reich, D.E., et al., *Linkage disequilibrium in the human genome.* Nature, 2001. **411**(6834): p. 199-204.

17. VanLiere, J.M. and N.A. Rosenberg, *Mathematical properties of the r2 measure of linkage disequilibrium.* Theoretical Population Biology, 2008. **74**(1): p. 130-137.

18. Christensen, O.F. and M.S. Lund, *Genomic prediction when some animals are not genotyped.* Genetics Selection Evolution, 2010. **42**(1): p. 2.

19. Calus, M.P.L., *Genomic breeding value prediction: methods and procedures.* Animal, 2009. **4**(2): p. 157-164.

20. Jing, L., et al., *Transcriptome analysis of mRNA and miRNA in skeletal muscle indicates an important network for differential Residual Feed Intake in pigs.* Scientific Reports, 2015. **5**.

21. Gilbert, H., et al., *Review: divergent selection for residual feed intake in the growing pig.* Animal, 2017. **11**(9): p. 1427-1439.

22. Koch, R.M., Swiger, L. A., Chambers, D. J. & Gregory K. E, *Efficiency of feed use in beef cattle.* Journal of Animal Science, 1963. **22**(2): p. 486-494.

23. Do, D.N., et al., *Genome-wide association and pathway analysis of feed efficiency in pigs reveal candidate genes and pathways for residual feed intake.* Front Genet, 2014. **5**: p. 307.

24. Yi, Z., et al., *Feed conversion ratio, residual feed intake and cholecystokinin type A receptor gene polymorphisms are associated with feed intake and average daily gain in a Chinese local chicken population.* J Anim Sci Biotechnol, 2018. **9**: p. 50.

25. Nkrumah, J.D., et al., *Genetic and phenotypic relationships of feed intake and measures of efficiency with growth and carcass merit of beef cattle1.* Journal of Animal Science, 2007. **85**(10): p. 2711-2720.

26. Gunsett, F.C., *Merit of Utilizing the Heritability of a Ratio to Predict the Genetic Change of a Ratio.* Journal of Animal Science, 1987. **65**(4): p. 936-942.

27. Hoque, M.A., et al., *Genetic parameters for measures of residual feed intake and growth traits in seven generations of Duroc pigs.* Livestock Science, 2009. **121**(1): p. 45-49.

28. Do, D.N., et al., *Genetic parameters for different measures of feed efficiency and related traits in boars of three pig breeds.* J Anim Sci, 2013. **91**(9): p. 4069-79.

29. Patience, J.F., M.C. Rossoni-Serão, and N.A. Gutiérrez, *A review of feed efficiency in swine: biology and application.* Journal of Animal Science and Biotechnology, 2015. **6**(1): p. 33.

30. PW Knap, W.L., *Pig Breeding for Improved Feed Efficiency*. Feed efficiency in swine, ed. P. JF. 2012, Wageningen: Wageningen Academic Press.

31. Milgen, J.V., et al., *Major determinants of fasting heat production and energetic cost of activity in growing pigs of different body weight and breed/castration combination.* British Journal of Nutrition, 2007. **79**(6): p. 509-517.

32. Quiniou, N., et al., *Modelling heat production and energy balance in group-housed growing pigs exposed to low or high ambient temperatures.* British Journal of Nutrition, 2007. **85**(1): p. 97-106.

33. Romanyukha, A.A., S.G. Rudnev, and I.A. Sidorov, *Energy cost of infection burden: An approach to understanding the dynamics of host–pathogen interactions.* Journal of Theoretical Biology, 2006. **241**(1): p. 1-13.

34. Dritz, S., *Influence of Health on Feed Efficiency*. Feed efficiency in swine, ed. P. JF. 2012, Wageningen: Wageningen Academic Press.

35. Ding, R., et al., *Genetic Architecture of Feeding Behavior and Feed Efficiency in a Duroc Pig Population.* Front Genet, 2018. **9**: p. 220.

36. Goldansaz, S.A., et al., *Livestock metabolomics and the livestock metabolome: A systematic review.* PLoS One, 2017. **12**(5): p. e0177675.

37. Goodacre, R., et al., *Metabolomics by numbers: acquiring and understanding global metabolite data.* Trends in Biotechnology, 2004. **22**(5): p. 245-252.

38. McMurry, J., *Organic chemistry: with biological applications(2nd ed.)*. 2011, Belmont, CA: Brooks/Cole.

39. Dubois, F., et al., *A comparison between ion-to-photon and microchannel plate detectors.* Rapid Communications in Mass Spectrometry, 1999. **13**(9): p. 786-791.

40. *Mass Spectrometry*, in *Kirk‐Othmer Encyclopedia of Chemical Technology*.

41. da Silva, R.R., P.C. Dorrestein, and R.A. Quinn, *Illuminating the dark matter in metabolomics.* Proceedings of the National Academy of Sciences, 2015. **112**(41): p. 12549-12550.

42. Wishart, D.S., et al., *HMDB 3.0--The Human Metabolome Database in 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D801-7.

43. Dias, D.A., et al., *Current and Future Perspectives on the Structural Identification of Small Molecules in Biological Systems.* Metabolites, 2016. **6**(4).

44. Jeffryes, J.G., et al., *MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics.* J Cheminform, 2015. **7**: p. 44.

45. Gao, J., L.B. Ellis, and L.P. Wackett, *The University of Minnesota Biocatalysis/Biodegradation Database: improving public access.* Nucleic Acids Res, 2010. **38**(Database issue): p. D488-91.

46. Wishart, D.S., et al., *HMDB 4.0: the human metabolome database for 2018.* Nucleic Acids Res, 2018. **46**(D1): p. D608-D617.

47. Duggan, G.E., et al., *Metabolomic response to exercise training in lean and diet-induced obese mice.* J Appl Physiol (1985), 2011. **110**(5): p. 1311-8.

48. Jones, D.P., Y. Park, and T.R. Ziegler, *Nutritional metabolomics: progress in addressing complexity in diet and health.* Annu Rev Nutr, 2012. **32**: p. 183-202.

49. May, D.H., et al., *Metabolomic profiling of urine: response to a randomised, controlled feeding study of select fruits and vegetables, and application to an observational study.* British Journal of Nutrition, 2013. **110**(10): p. 1760-1770.

50. Karisa, B.K., et al., *Plasma metabolites associated with residual feed intake and other productivity performance traits in beef cattle.* Livestock Science, 2014. **165**: p. 200-211.

51. Novais, F.J., et al., *Identification of a metabolomic signature associated with feed efficiency in beef cattle.* BMC Genomics, 2019. **20**(1): p. 8.

52. Chapinal, N., et al., *The association of serum metabolites in the transition period with milk production and early-lactation reproductive performance.* J Dairy Sci, 2012. **95**(3): p. 1301-9.

53. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.

54. Best, Myron G., et al., *RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics.* Cancer Cell, 2015. **28**(5): p. 666-676.

55. Smith, S., L. Bernatchez, and L.B. Beheregaray, *RNA-seq analysis reveals extensive transcriptional plasticity to temperature stress in a freshwater fish species.* BMC Genomics, 2013. **14**(1): p. 375.

56. Morin, R. and M.P. Team, *The status, quality, and expansion of the NIH full-length cDNA project: The Mammalian Gene Collection (MGC) (vol 14, pg 2121, 2006).* Genome Research, 2006. **16**(6): p. 804-804.

57. Boguski, M.S., C.M. Tolstoshev, and D.E. Bassett, *Gene Discovery in Dbest.* Science, 1994. **265**(5181): p. 1993-1994.

58. Royce, T.E., J.S. Rozowsky, and M.B. Gerstein, *Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification.* Nucleic Acids Research, 2007. **35**(15).

59. Okoniewski, M.J. and C.J. Miller, *Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations.* Bmc Bioinformatics, 2006. **7**.

60. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.

61. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

62. van Dijk, E.L., et al., *Ten years of next-generation sequencing technology.* Trends in Genetics, 2014. **30**(9): p. 418-426.

63. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nat Rev Genet, 2016. **17**(6): p. 333-51.

64. Stark, R., M. Grzelak, and J. Hadfield, *RNA sequencing: the teenage years.* Nat Rev Genet, 2019. **20**(11): p. 631-656.

65. Andrews, S. *FastQC.* 2010; Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

66. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.

67. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

68. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature Methods, 2012. **9**(4): p. 357-359.

69. Kärkkäinen, J. and P. Sanders. *Simple Linear Work Suffix Array Construction.* 2003. Berlin, Heidelberg: Springer Berlin Heidelberg.

70. Burset, M., I.A. Seledtsov, and V.V. Solovyev, *Analysis of canonical and non-canonical splice sites in mammalian genomes.* Nucleic Acids Research, 2000. **28**(21): p. 4364-4375.

71. Aken, B.L., et al., *The Ensembl gene annotation system.* Database, 2016. **2016**.

72. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nature Protocols, 2012. **7**(3): p. 562-578.

73. Anders, S., P.T. Pyl, and W. Huber, *HTSeq—a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2014. **31**(2): p. 166-169.

74. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-842.

75. Wagner, G.P., K. Kin, and V.J. Lynch, *Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples.* Theory Biosci, 2012. **131**(4): p. 281-5.

76. Evans, C., J. Hardin, and D.M. Stoebel, *Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions.* Brief Bioinform, 2018. **19**(5): p. 776-792.

77. Chen, K., et al., *The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses.* Mol Cell Biol, 2015. **36**(5): p. 662-7.

78. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.

79. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.

80. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

81. Robinson, M.D. and G.K. Smyth, *Small-sample estimation of negative binomial dispersion, with applications to SAGE data.* Biostatistics, 2007. **9**(2): p. 321-332.

82. Froussios, K., et al., *How well do RNA-Seq differential gene expression tools perform in a complex eukaryote? A case study in Arabidopsis thaliana.* Bioinformatics, 2019. **35**(18): p. 3372-3377.

83. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nature Genetics, 2000. **25**(1): p. 25-29.

84. The Gene Ontology Consortium, *The Gene Ontology Resource: 20 years and still GOing strong.* Nucleic Acids Research, 2018. **47**(D1): p. D330-D338.

85. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature Protocols, 2009. **4**(1): p. 44-57.

86. Eden, E., et al., *GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.* BMC Bioinformatics, 2009. **10**: p. 48.

87. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.

88. Kamburov, A., et al., *Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA.* Bioinformatics, 2011. **27**(20): p. 2917-8.

89. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.

90. Chen, K. and N. Rajewsky, *The evolution of gene regulation by transcription factors and microRNAs.* Nature Reviews Genetics, 2007. **8**(2): p. 93-103.

91. Laurendeau, I., et al., *Gene Expression Profiling of the Hedgehog Signaling Pathway in Human Meningiomas.* Molecular Medicine, 2010. **16**(7): p. 262-270.

92. Harris, S.L. and A.J. Levine, *The p53 pathway: positive and negative feedback loops.* Oncogene, 2005. **24**(17): p. 2899-2908.

93. Chen, Y., M.J. Smanski, and B. Shen, *Improvement of secondary metabolite production in Streptomyces by manipulating pathway regulation.* Applied Microbiology and Biotechnology, 2010. **86**(1): p. 19-25.

94. Orawski, A.T. and W.H. Simmons, *Degradation of bradykinin and its metabolites by rat brain synaptic membranes.* Peptides, 1989. **10**(5): p. 1063-1073.

95. Hochberg, U., et al., *Metabolite profiling and network analysis reveal coordinated changes in grapevine water stress response.* BMC Plant Biology, 2013. **13**(1): p. 184.

96. Carey, V.J., et al., *Network structures and algorithms in Bioconductor.* Bioinformatics, 2005. **21**(1): p. 135-6.

97. Henegar, C., K. Clement, and J.D. Zucker, *Unsupervised multiple-instance learning for functional profiling of genomic data.* Machine Learning: Ecml 2006, Proceedings, 2006. **4212**: p. 186-197.

98. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**(1): p. 559.

99. Bravais, A., *Analyse mathématique sur les probabilités des erreurs de situation d'un point*. Mémoires présentés par divers savants à l'Académie de l'Institut de France. 1844, Paris: Imprimerie Royale.

100. Zhang, B. and S. Horvath, *A general framework for weighted gene co-expression network analysis.* Stat Appl Genet Mol Biol, 2005. **4**: p. Article17.

101. DiLeo, M.V., et al., *Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome.* PLoS One, 2011. **6**(10): p. e26683.

102. Visscher, P.M., et al., *10 Years of GWAS Discovery: Biology, Function, and Translation.* The American Journal of Human Genetics, 2017. **101**(1): p. 5-22.

103. Gallagher, M.D. and A.S. Chen-Plotkin, *The Post-GWAS Era: From Association to Function.* Am J Hum Genet, 2018. **102**(5): p. 717-730.

104. Nica, A.C. and E.T. Dermitzakis, *Expression quantitative trait loci: present and future.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1620): p. 20120362.

105. Bryois, J., et al., *Cis and trans effects of human genomic variants on gene expression.* PLoS genetics, 2014. **10**(7): p. e1004461-e1004461.

106. Grundberg, E., et al., *Mapping cis- and trans-regulatory effects across multiple tissues in twins.* Nature Genetics, 2012. **44**(10): p. 1084-1089.

107. Liaubet, L., et al., *Genetic variability of transcript abundance in pig peri-mortem skeletal muscle: eQTL localized genes involved in stress response, cell death, muscle disorders and metabolism.* BMC Genomics, 2011. **12**: p. 548.

108. González-Prendes, R., R. Quintanilla, and M. Amills, *Investigating the genetic regulation of the expression of 63 lipid metabolism genes in the pig skeletal muscle.* Animal Genetics, 2017. **48**(5): p. 606-610.

109. Zhao, F., et al., *Genetic model.* Journal of cellular and molecular medicine, 2016. **20**(4): p. 765-765.

110. Shabalin, A.A., *Matrix eQTL: ultra fast eQTL analysis via large matrix operations.* Bioinformatics, 2012. **28**(10): p. 1353-8.

111. Bartel, J., et al., *The Human Blood Metabolome-Transcriptome Interface.* PLoS genetics, 2015. **11**(6): p. e1005274-e1005274.

112. Baumert, P., et al., *TRIM63 (MuRF-1) gene polymorphism is associated with biomarkers of exercise-induced muscle damage.* Physiological Genomics, 2018. **50**(3): p. 142-143.

113. Cai, C., et al., *ETV1 Is a Novel Androgen Receptor-Regulated Gene that Mediates Prostate Cancer Cell Invasion.* Molecular Endocrinology, 2007. **21**(8): p. 1835-1846.

114. Widmann, P., et al., *Systems biology analysis merging phenotype, metabolomic and genomic data identifies Non-SMC Condensin I Complex, Subunit G (NCAPG) and cellular maintenance processes as major contributors to genetic variability in bovine feed efficiency.* PloS one, 2015. **10**(4): p. e0124574-e0124574.

115. Clemmons, B.A., et al., *Rumen fluid metabolomics of beef steers differing in feed efficiency.* Metabolomics, 2020. **16**(2): p. 23.

116. Clemmons, B.A., et al., *Serum metabolites associated with feed efficiency in black angus steers.* Metabolomics, 2017. **13**(12): p. 147.

117. Goldansaz, S.A., et al., *Livestock metabolomics and the livestock metabolome: A systematic review.* PloS one, 2017. **12**(5): p. e0177675-e0177675.

118. Nørskov, N.P., et al., *Multicompartmental Nontargeted LC–MS Metabolomics: Explorative Study on the Metabolic Responses of Rye Fiber versus Refined Wheat Fiber Intake in Plasma and Urine of Hypercholesterolemic Pigs.* Journal of Proteome Research, 2013. **12**(6): p. 2818-2832.

119. Jégou, M., et al., *NMR-based metabolomics highlights differences in plasma metabolites in pigs exhibiting diet-induced differences in adiposity.* European Journal of Nutrition, 2016. **55**(3): p. 1189-1199.

120. Soumeh, E.A., et al., *Nontargeted LC–MS Metabolomics Approach for Metabolic Profiling of Plasma and Urine from Pigs Fed Branched Chain Amino Acids for Maximum Growth Performance.* Journal of Proteome Research, 2016. **15**(12): p. 4195-4207.

121. Lin, G., et al., *Metabolomic Analysis Reveals Differences in Umbilical Vein Plasma Metabolites between Normal and Growth-Restricted Fetal Pigs during Late Gestation.* The Journal of Nutrition, 2012. **142**(6): p. 990-998.

122. Sachse, D., et al., *The Role of Plasma and Urine Metabolomics in Identifying New Biomarkers in Severe Newborn Asphyxia: A Study of Asphyxiated Newborn Pigs following Cardiopulmonary Resuscitation.* PloS one, 2016. **11**(8): p. e0161123-e0161123.

123. Jiang, P., et al., *Progressive Changes in the Plasma Metabolome during Malnutrition in Juvenile Pigs.* Journal of Proteome Research, 2016. **15**(2): p. 447-456.

124. Polakof, S., et al., *Postprandial metabolic events in mini-pigs: new insights from a combined approach using plasma metabolomics, tissue gene expression, and enzyme activity.* Metabolomics, 2015. **11**(4): p. 964-979.

125. Bovo, S., et al., *Deconstructing the pig sex metabolome: Targeted metabolomics in heavy pigs revealed sexual dimorphisms in plasma biomarkers and metabolic pathways1.* Journal of Animal Science, 2015. **93**(12): p. 5681-5693.

126. Yet, I., et al., *Genetic Influences on Metabolite Levels: A Comparison across Metabolomic Platforms.* Plos One, 2016. **11**(4).

127. Horodyska, J., et al., *RNA-seq of muscle from pigs divergent in feed efficiency and product quality identifies differences in immune response, growth, and macronutrient and connective tissue metabolism.* BMC Genomics, 2018. **19**(1): p. 791.

128. Gondret, F., et al., *A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs.* BMC Genomics, 2017. **18**(1): p. 244.

129. Jing, L., et al., *Transcriptome analysis of mRNA and miRNA in skeletal muscle indicates an important network for differential Residual Feed Intake in pigs.* Sci Rep, 2015. **5**: p. 11953.

130. Vincent, A., et al., *Divergent selection for residual feed intake affects the transcriptomic and proteomic profiles of pig skeletal muscle.* J Anim Sci, 2015. **93**(6): p. 2745-58.

131. Zid, B.M., et al., *4E-BP Extends Lifespan upon Dietary Restriction by Enhancing Mitochondrial Activity in Drosophila.* Cell, 2009. **139**(1): p. 149-160.

132. Rochard, P., et al., *Mitochondrial activity is involved in the regulation of myoblast differentiation through myogenin expression and activity of myogenic factors.* Journal of Biological Chemistry, 2000. **275**(4): p. 2733-2744.

133. Morita, M., et al., *mTOR coordinates protein synthesis, mitochondrial activity and proliferation.* Cell Cycle, 2015. **14**(4): p. 473-480.

134. Bottje, W.G. and G.E. Carstens, *Association of mitochondrial function and feed efficiency in poultry and livestock species1.* Journal of Animal Science, 2009. **87**(suppl_14): p. E48-E63.

135. Connor, E.E., et al., *Enhanced mitochondrial complex gene function and reduced liver size may mediate improved feed efficiency of beef cattle during compensatory growth.* Functional & Integrative Genomics, 2010. **10**(1): p. 39-51.

136. de Oliveira, P.S., et al., *Identification of genomic regions associated with feed efficiency in Nelore cattle.* BMC Genetics, 2014. **15**(1): p. 100.

137. Velez-Irizarry, D., et al., *Genetic control of longissimus dorsi muscle gene expression variation and joint analysis with phenotypic quantitative trait loci in pigs.* BMC Genomics, 2019. **20**(1): p. 3.

138. Teissier, M., H. Larroque, and C. Robert-Granié, *Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: a quantitative trait influenced by a major gene.* Genetics Selection Evolution, 2018. **50**(1): p. 31.

139. Zhang, X., et al., *Weighting Strategies for Single-Step Genomic BLUP: An Iterative Approach for Accurate Calculation of GEBV and GWAS.* Frontiers in Genetics, 2016. **7**(151).

SCIENTIFIC REPORTS
natureresearch

OPEN

Metabolomic networks and pathways associated with feed efficiency and related-traits in Duroc and Landrace pigs

Victor Adriano Okstoft Carmelo[1], Priyanka Banerjee[1], Wellison Jarles da Silva Diniz[1,2] & Haja N. Kadarmideen[1*]

Improving feed efficiency (FE) is a major goal of pig breeding, reducing production costs and providing sustainability to the pig industry. Reliable predictors for FE could assist pig producers. We carried out untargeted blood metabolite profiling in uncastrated males from Danbred Duroc (n = 59) and Danbred Landrace (n = 50) pigs at the beginning and end of a FE testing phase to identify biomarkers and biological processes underlying FE and related traits. By applying linear modeling and clustering analyses coupled with WGCNA framework, we identified 102 and 73 relevant metabolites in Duroc and Landrace based on two sampling time points. Among them, choline and pyridoxamine were hub metabolites in Duroc in early testing phase, while, acetoacetate, cholesterol sulfate, xanthine, and deoxyuridine were identified in the end of testing. In Landrace, cholesterol sulfate, thiamine, L-methionine, chenodeoxycholate were identified at early testing phase, while, D-glutamate, pyridoxamine, deoxycytidine, and L-2-aminoadipate were found at the end of testing. Validation of these results in larger populations could establish FE prediction using metabolomics biomarkers. We conclude that it is possible to identify a link between blood metabolite profiles and FE. These results could lead to improved nutrient utilization, reduced production costs, and increased FE.

With the expanding human population and requirement for nutrient-rich food, there is an increasing demand for improvement of meat production, but simultaneously, to decrease the input costs in terms of feed[1]. Thus, feed efficiency (FE) is the most important trait in commercial pig farming[2] as increasing the amount of meat produced per feed is beneficial both economically and environmentally. Thereby, improving FE is beneficial for producers and increases the sustainability of pork meat production. Fortunately, FE is a highly heritable trait in Danish pigs (ranging from 0.34 in Duroc to 0.40 in Landrace), thus suitable for the genetic selection of pigs with high breeding values in breeding programs aimed at improving this economically important phenotype[3].

Since FE cannot be measured directly, feed conversion ratio (FCR) and residual feed intake (RFI) have been used to evaluate the animal efficiency[4]. FCR determines the ratio of feed intake (FI) to output and found to correlate with growth rate and body weight[3,5]. RFI calculates the difference between the actual and expected FI[6] predicted based on production traits such as average daily gain (ADG)[7]. ADG is also considered important in commercial pig production as pigs with higher ADG can achieve a target market weight within a shorter period than those with lower ADG, thereby saving feeding costs[8]. Thus, selection for RFI has proved to be effective in improving the FE in pigs[3,9,10]. Selection for FCR will results in co-selection for other traits, such as body composition and ADG. In contrast, RFI selects for increased metabolic efficiency without the same side effects[11–13]. RFI and FCR are well correlated, with a reported correlation of over 0.7 in the literature[3].

As part of the existing genetic determinants of FE, genome-wide association studies (GWAS) and differential expression (DE) analyses have reported a large number of polymorphism and genes for RFI or FCR in pigs[9,14]. However, despite these efforts, FE is a complex trait with many aspects involved and the functional molecular background is still somewhat elusive[1]. Among the approaches, the metabolomics profile reveals the relationship between animal genetics and physiological phenotypes[15], thereby increasing the fundamental understanding of

[1]Quantitative Genomics, Bioinformatics and Computational Biology, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark. [2]Department of Genetics and Evolution, Federal University of São Carlos, São Carlos, Brazil. *email: hajak@dtu.dk

Breeds*		FE	EDG	TDG	DG	RFI
Duroc 1	P ≤ 0.05**	62 (30)	42 (24)	33 (23)	47 (32)	64 (31)
	KS test	1.00E-05	0.19099	0.02317	0.02625	0
	FDR ≤ 0.05***	1	0	1	0	0
Duroc 2	P ≤ 0.05**	82 (40)	41 (26)	115 (68)	46 (26)	57 (28)
	KS test	0	0.10092	0	0.8561	0.02687
	FDR ≤ 0.05***	0	0	35	0	0
Landrace 1	P ≤ 0.05**	40 (17)	59 (37)	44 (22)	67 (38)	41 (16)
	KS test	0.25416	2.00E-05	0.00079	0	0.08764
	FDR ≤ 0.05***	0	9	0	1	0
Landrace 2	P ≤ 0.05**	54 (35)	37 (21)	77 (44)	73 (46)	53 (36)
	KS test	0	0	0	0	0
	FDR ≤ 0.05***	0	0	0	0	0
Duroc 1,2	P ≤ 0.05**	83 (46)	24 (12)	54 (23)	96 (61)	69 (36)
	KS test	0	0.004	7e-05	0	0
	FDR ≤ 0.05***	0	0	1	20	0
Landrace 1,2	P ≤ 0.05**	59 (35)	76 (40)	73 (37)	69 (34)	46 (29)
	KS test	0.06419	0	0	0	0.16896
	FDR ≤ 0.05***	0	0	0	0	0

Table 1. Overview of metabolites associated with phenotypic traits in Duroc and Landrace at two time points. *Numbers (1, 2) represents time point 1 and 2 respectively; **Number of significant metabolites with p-value ≤ 0.05 (in the parenthesis are the number of annotated metabolites); P = p-value; KS test = Kolmogorov-Smirnov test; ***Number of metabolites with False discovery rate (FDR) ≤ 0.05; FDR = False discovery rate; FE = Feed efficiency; EDG = Early daily gain; TDG = Testing daily gain; DG = Daily gain; RFI = Residual feed intake.

efficiency and selection. Although affected by prandial activity, many metabolic processes underlie the transport of molecules through the blood. Blood is the sole way of absorption of nutrients into the body, and the blood metabolites are useful as a prime candidate for the study of FE in livestock[16]. In this context, it is also generally considered than improvement of RFI is associated with improved efficiency in the utilization of feed[11,12] and thus improved utilization of nutrients.

An effective way to get insights into the interactions at a molecular level involved in complex phenotypes can be done by applying a network-based approach like weighted gene co-expression network analysis (WGCNA)[17]. In the context of metabolomics, the clusters (modules) represent specific metabolic processes or pathways and gives a better understanding of the function, interaction, and common regulatory mechanisms. WGCNA has been widely applied in pigs and several livestock species with fruitful results[18–21]. Therefore, one of the main objectives was to identify key blood metabolites associated with FE and related traits in Danbred Duroc and Danbred Landrace (referred to as Duroc and Landrace, respectively, further in the text). As the Durocs are more FE than the Landrace, the two breeds serve biological contrast in FE. Furthermore, selecting two diverse breeds can help generalize any results obtained versus only focusing on one breed.

Here, we applied an untargeted metabolomics approach for a better understanding of changes at a molecular level associated with nutrient utilization. We test the hypothesis that we are able to associate metabolite concentrations in blood at an early growth stage to predict future growth and FE measurements, and that metabolites profiles in general are associated with growth and efficiency phenotypes. We applied linear regression models to select the top metabolites predictive of FE, combined the results from network-based methods, and conducted a functional enrichment and pathway analyses to provide potential easy-to-screen candidate metabolite biomarkers and metabolic processes modulating FE in pigs.

Results

Descriptive statistics and linear model analysis.

The phenotypic traits summary, including feed consumed (FC), FE, daily gain (DG), and delta weight (DW), for 109 pigs from Duroc and Landrace breed is shown in Supplementary Table S1. Aiming to ascertain the metabolite profiles concerning FE, we collected the blood samples at two time points (start and end of testing phase) from two breeds of pigs, profiled for the metabolite changes. The start phase was labeled as time point 1 (TP1) and the breeds as Duroc 1 and Landrace 1, while the end of the testing phase as time point 2 (TP2), mentioned as Duroc 2 and Landrace 2.

The number of metabolites for each of the breeds at each time point with p-value ≤ 0.05 are provided in Table 1. The molecular mass, retention time, and p-values of these metabolites for each trait in the breeds at different time points and at the combined time points are provided in Supplementary Table S2.

With an initial dataset of 729 metabolites, only those metabolites with relative standard deviation >0.15 were used for each group based on the raw counts. This amounts to 691 and 702 metabolites in Duroc (TP1 and TP2), while 684 and 689 for Landrace (TP1 and TP2), which were subjected for further analysis. To test if the metabolite profile was associated with the most distinct factors such as age and breed, the data were visualized on the first two principal components, colored by time point and breed, as given in Fig. 1. Further, the significance of the

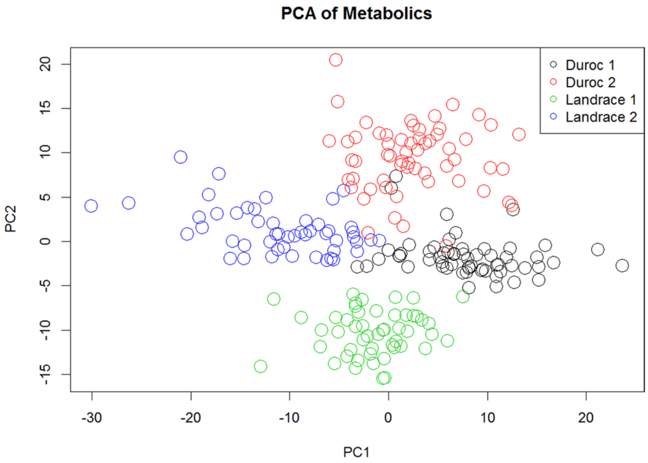


Figure 1. PCA visualization of the two first principal components, colored by breed and sampling. The first component separates the most divergent group – Duroc 2 and Landrace 1. The second component separates Duroc 1 and Landrace 2, the second most divergent group.

linear relationship for each metabolite between the two time points is observed (Supplementary Fig. S1). Over half of the metabolites have a p-value < 0.05 for the linear relationship between the two sampling points, indicating that there is a stability and predictability in the relative metabolite concentrations over time. In Fig. S2 we can see further evidence for this, with a visualization of all the pairwise log metabolite concentrations between the two time points, showing a clear overall relationship.

The significant metabolites at two time points were identified, as given in Table 1. A linear model was fitted to unravel the effect of blood metabolite on the FE phenotypes. The overall significance of divergence from the null hypothesis of no relation between metabolites and phenotypes using the Kolmogorov-Smirnov (KS) test, comparing the observed p-value distributions with the corresponding uniform distribution was tested. This was done to reveal, if there was an overall relation between the metabolites and our phenotypes. Most of the traits have significant metabolite profiles based on the KS test, signifying an overall relation between metabolites and traits. Based on the overall distribution of the KS test p-values, even the highest value of 0.19 in early daily gain (EDG) for Duroc could be significant based on FDR. In Duroc and Landrace, some metabolites were significantly associated with every trait, with the highest number identified in TDG (Table 1). The most significant results for testing daily gain (TDG) (35) and EDG (9) in Duroc 2 and Landrace 1, respectively, after false discovery rate (FDR) correction (Table 1) was identified.

We also did exploratory clustering analysis was done for the metabolites found significant for RFI in Duroc (36) and Landrace (29) (both time points combined) (Table 1, Supplementary Table S2). The heatmap plots in Duroc (Fig. 2) and Landrace (Fig. 3) grouped the metabolites in four specific clusters (Supplementary Table S5) and also the samples separately at TP1 and 2.

Metabolite network analysis.

Since the metabolites interact and/or are a part of the same or related metabolic pathways, a weighted gene network approach using WGCNA[17], that is typically used for gene co-expression analyses, was adapted and implemented for metabolomics data. A signed weighted metabolite network was constructed following the WGCNA pipeline, which identifies modules of functionally related metabolites, summarizes the module based on module eigengene - ME, and relates the MEs with the trait of interest[17]. We constructed the networks separately for both the breeds at two time points to unravel the correlated metabolites with the trait of interest (FE, EDG, TDG, DG, and RFI). Next, we selected the significantly associated modules ($p \leq 0.1$, and correlation ≥ 0.2) that were labeled by color for further analysis. The expression of any FE trait, such as RFI, is dependent on the stage of maturity while for other traits, this correlation is low[22,23]. However, in our study, we observed low to medium correlation for all the traits with respect to the metabolites.

In Duroc (TP1), 144, 131, 335, and 81 metabolites were clustered, respectively, in MEblue, MEbrown, MEturquoise, and MEyellow (Fig. 4A- upper panel). Among the modules, MEbrown was significantly associated with FE and RFI, and MEturquoise with RFI (Fig. 4A – lower panel). From the TP2, 190, 104, 316, and 92 metabolites were clustered in MEblue, MEbrown, MEturquoise, and MEyellow, respectively (Fig. 4B – upper panel). From these modules, significant associations were identified for MEblue (FE, TDG, and RFI), and MEturquoise (TDG) (Fig. 4B – lower panel). In Landrace (TP1), 152 metabolites were clustered in MEblue, 151 in MEbrown, 260 in MEturquoise while 121 in MEyellow (Fig. 4C – upper panel). MEbrown was significantly associated with RFI, while MEturquoise and MEyellow with DG at TP1 in Landrace (Fig. 4C – lower panel). Regarding TP2, 253 metabolites were clustered in MEblue, 142 with MEbrown and 294 with MEturquoise (Fig. 4D – upper panel). Nonetheless, only MEturquoise was associated with EDG and DG (Fig. 4D – lower panel).

The annotated metabolites with $p \leq 0.05$ (Table 1, Supplementary Table S2) and those clustered into the associated modules (Fig. 4, Supplementary Table S3) were subjected to pathway over-representation analysis (Table 2). As the same metabolite in a module can be related to more than one trait, the unique metabolites were screened by taking all the significant modules for all the traits in each breed at each time point (Table 2). Then, we also screened the metabolites for commonality in each breed between the two time points. In Duroc, only a single

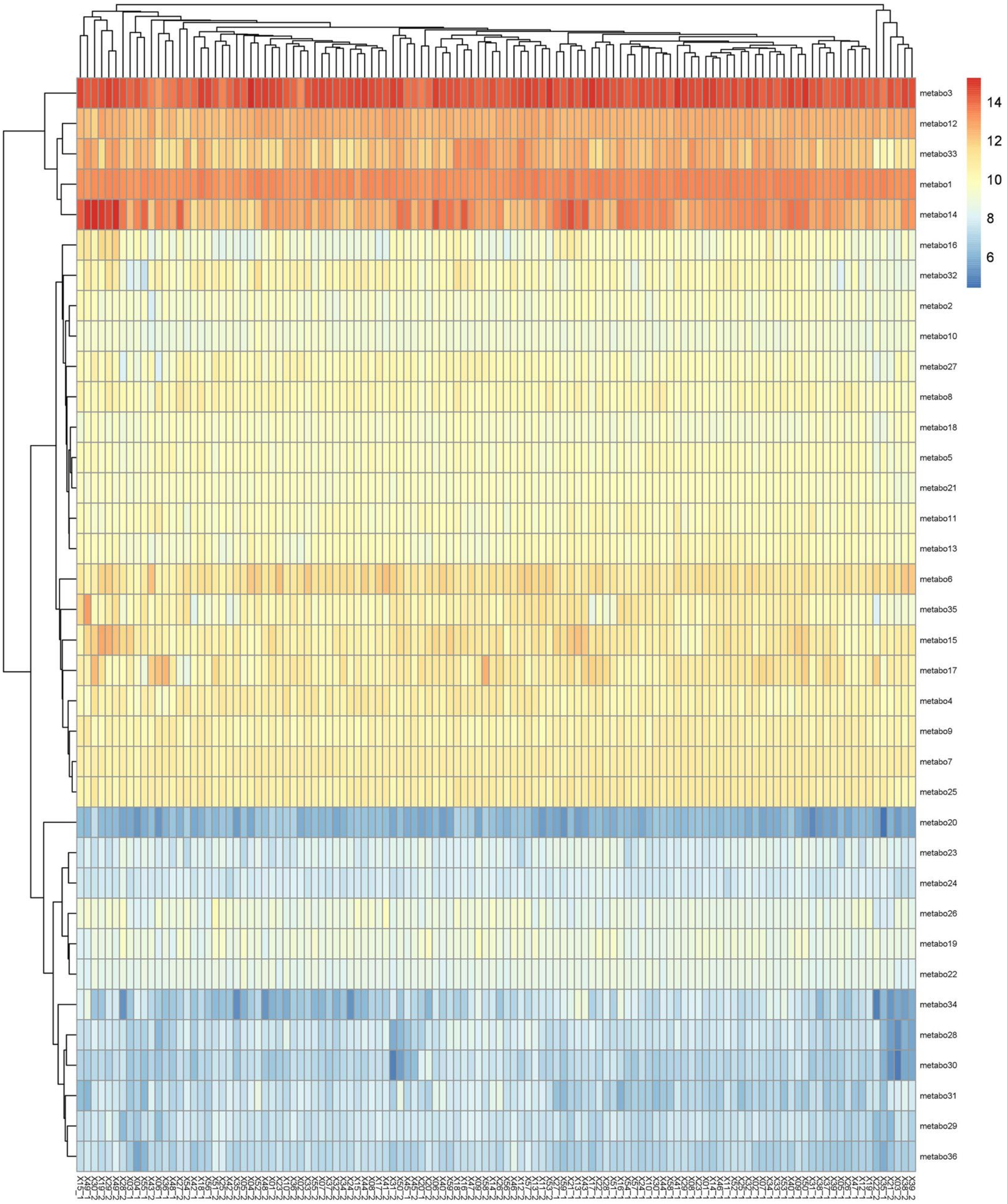


Figure 2. Heatmap constructed using the significant metabolites with RFI in Duroc (time point 1 and 2). The x-axis represents the sample ID at time point 1 and 2 represented as ID_1 and ID_2, respectively; the y-axis represents the metabolites (names of the corresponding metabolites are given in Supplementary Table S5.

metabolite out of 102 was found to be common between the two time points (TP1 and TP2). In Landrace, 36 metabolites were found common in two time points, while 66 metabolites were different in TP1 and TP2.

Pathway over-representation analysis. Exploiting the fact that metabolites are linked through biochemical reactions and thus are partaking in many pathways, we carried out a pathway over-representation analysis based on the integrated molecular level pathway analysis (IMPaLA) software[24]. To reveal the differences at the pathway level, we analyzed the unique metabolites in two different ways. First, by comparing the difference in the metabolites at two time points within the breeds. Second, by comparing the metabolite differences among the breeds (Duroc *vs*. Landrace), taking all the metabolites together irrespective of the time points in each breed (Table 2, Supplementary Table S3). The unique metabolites from two time points were screened, supporting the fact that the FI or FE is affecting the pathways to some extent, thus pointing out the different pathways in TP1 and TP2. The significant over-represented pathways were screened against 7 databases (Kyoto Encyclopedia of Genes and Genomes - KEGG, Edinburgh Human Metabolic Network - EHMN, Reactome, Integrating Network Objects with Hierarchies - INOH, HumanCyc, Biocarta, Pathway Interaction Database - PID) and selected ($p \leq 0.05$) pathways were used for biological interpretations.

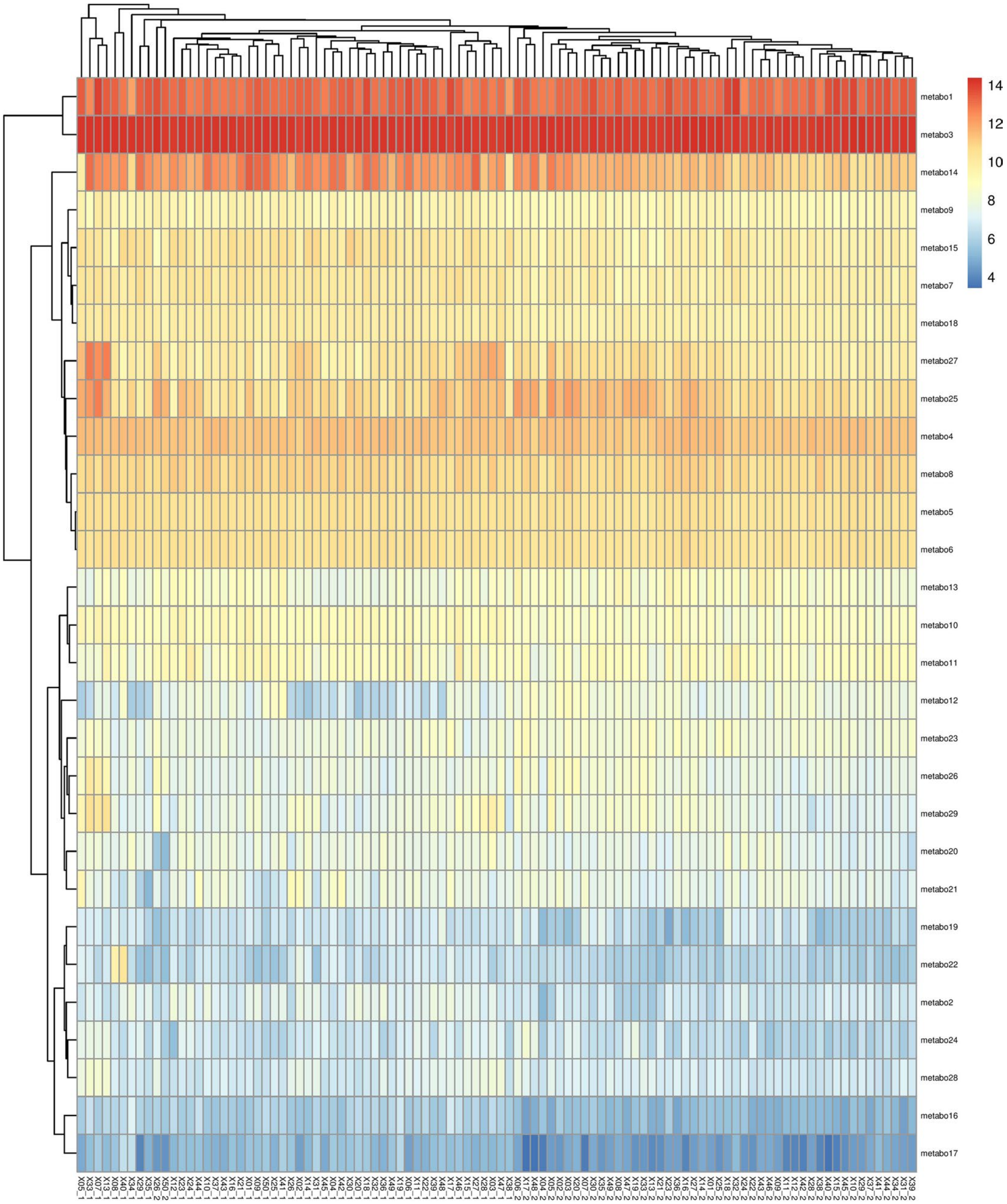


Figure 3. Heatmap constructed using the significant metabolites with RFI in Landrace (time point 1 and 2). The x-axis represents the sample ID at time point 1 and 2 represented as ID_1 and ID_2, respectively; the y-axis represents the metabolites (names of the corresponding metabolites are given in Supplementary Table S5.

In Duroc, 32 metabolites were involved in 49 pathways in TP1 as compared to 35 pathways obtained by 70 unique metabolites in TP2 (Supplementary Table S4). Some of the underlying pathways in TP1 were the metabolism of glycerophospholipid, D-arginine and D-ornithine and choline; mTOR, Arf6, ErbB1, and Arf1 signaling pathways. However, in TP2, synthesis and degradation (Lysine, Valine-Leucine-Isoleucine, pyrimidine deoxyribonucleosides, methionine, glycine betaine, guanosine), bile salts and organic anion SLC transporter and pentose phosphate pathway were identified. Vitamin B6 metabolism was common between TP1 and TP2. Similarly, in Landrace, 36 unique metabolites from TP1 were involved with 20 significantly ($p \leq 0.05$) over-represented pathways, while 37 metabolites were involved with 15 significantly over-represented pathways in TP2. Pathways like digestion of dietary lipid, synthesis of bile salts, valine degradation, valine-leucine-isoleucine biosynthesis were found in TP1. In TP2, the pathways found were degradation of pyrimidine deoxyribonucleosides, methionine, glycine betaine, cysteine biosynthesis, and vitamin B6 metabolism. However, the pathways involved were completely different in TP1 and TP2 in Landrace. This supports the fact that there is an observable difference in the biological level as shown by the difference in metabolites at two time points in both the breeds.

The complete breed analysis, combining both the time points, was also carried out to evaluate the differences in the metabolites and the biological pathways involved, that are specific to the breed. 101 unique metabolites were

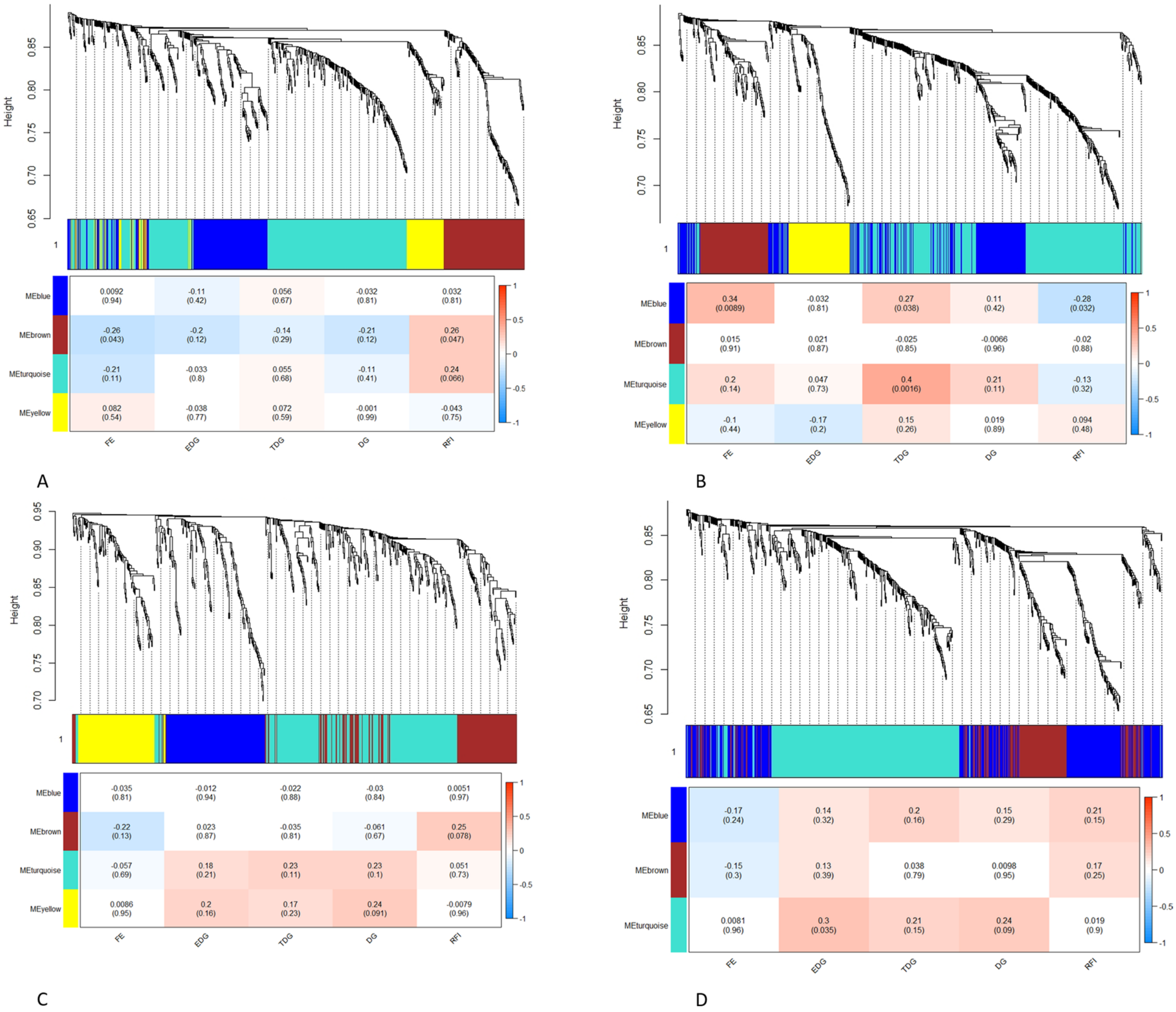


Figure 4. Clustering dendrogram and module-trait correlation plots. The upper panel of each plot (**A**–**D**) represents metabolite-clustering dendrogram obtained by hierarchical clustering of TOM-based dissimilarity with the corresponding module colors indicated by the color row. Each colored row represents color-coded module that contains a group of highly connected metabolites. The lower panel of each plot (a–d) represents the module trait correlation where the x-axis represents feed efficiency trait, and the y-axis represents the modules. Plots (**A**) and (**B**) represents Duroc at time points one and two, respectively, while plots (**C**) and (**D**) represent Landrace at time points one and two. The color-coding in the module-trait correlation plots is based on Spearman's correlation (p-values in parenthesis). Positive and negative correlations are shown in red and blue colors, respectively.

subjected to over-representation pathway analysis leading to their involvement with 50 pathways over-represented at combined time points in Duroc (Supplementary Table S4). Combining both the time points in Landrace, 66 unique metabolites pointed to 10 pathways that were significantly over-represented ($p \leq 0.05$) (Supplementary Table S4). Biological oxidation, Histidine-lysine-phenylalanine-tyrosine-proline-tryptophan catabolism, and methionine salvage were involved with both Duroc and Landrace. All the other pathways were specific to each breed. The pathway differences between the breeds are also given in Supplementary Table S4.

Cluster analysis was carried out for the metabolites significant for RFI in a combined time point (Duroc – 36; Landrace – 29) (Table 1). The differences in the metabolite clustering for two time points in each breed is also observed in the heatmap (Figs. 2 and 3). Pathway analysis of the metabolites clustering together in the heatmaps is given in Supplementary Table S5. In Duroc, 4 significant clusters of 36 metabolites: Cluster 1 (metabo 3–14), cluster 2 (metabo – 16–13), cluster 3 (metabo 6–25), and cluster 4 (metabo 20–36) can be differentiated (Fig. 2). In Landrace, 4 significant clusters of 29 metabolites: cluster 1 (metabo 1 and 3), cluster 2 (metabo – 14–6), cluster 3 (metabo 13–21), and cluster 4 (metabo 19–17) can be differentiated (Fig. 3). The x-axis represented sample clustering of TP1 and TP2 in both the breeds. The annotation of the metabolites as given in the heatmap (y-axis) and their corresponding pathways are given in Supplementary Table S5.

Breed (Time point)*	Number of metabolites**	Module	Trait	Unique metabolites
Duroc (1)	11	Brown	FE	**32**
	9	Brown	RFI	
	21	Turquoise	RFI	
Duroc (2)	13	Blue	FE	**70**
	8	Blue	TDG	
	8	Blue	RFI	
	55	Turquoise	TDG	
Landrace (1)	8	Brown	RFI	**36**
	22	Turquoise	DG	
	6	Yellow	DG	
Landrace (2)	19	Turquoise	EDG	**37**
	36	Turquoise	DG	

Table 2. Significant metabolites for FE traits used for pathway analysis. *Numbers (1, 2) represents time point 1 and 2 respectively. **The metabolites were identified based on the overlapping of the linear model and network association modules. FE = Feed efficiency; EDG = Early daily gain; TDG = Testing daily gain; DG = Daily gain; RFI = Residual feed intake

Network visualization. To visualize and interpret metabolomics data in the context of human metabolic networks, to trace connections between metabolites and genes, and to visualize compound networks, the unique metabolites were cross-referenced with the KEGG database. Only metabolites with specific KEGG IDs were considered for compound-gene network and pathway analysis. The hub metabolites were identified by taking the highly connected metabolites that were associated with more than one gene in the compound gene network (Supplementary Table S6). Hubs are the nodes that are more connected than the average or typical nodes, and consequently are more likely to play crucial biological role.

In Duroc, 63 genes in TP1 pointing to 6 hub metabolites were identified, while in TP2, 79 genes pointing to 14 hub metabolites were identified (Supplementary Table S6). The hub metabolites were specific for each time point. In Landrace, 87 genes underlying 9 hub metabolites in TP1, while 40 genes were pointing to 7 hub metabolites in TP2. 3-Methyl-2-oxobutanoic acid was common hub metabolite in Landrace TP1 and TP2 (Supplementary Table S6). S-(2,2-Dichloro-1-hydroxy)ethyl glutathione was a common hub metabolite identified in Duroc and Landrace TP1, while 3-Methyl-2-oxobutanoic acid and cholesterol sulfate was found to be a common hub between Duroc and Landrace TP2.

A combined time point analysis for the breed identified 20 metabolites for Duroc and 15 for Landrace. Choline, acetoacetate, (R)-Lactate, D-Erythrose 4-phosphate, 3,4-Dihydroxy-L-phenylalanine, Xanthine, Deoxyuridine, phenylacetaldehyde, pyridoxine phosphate, 4-Pyridoxate, Taurolithocholate sulfate, 5-Guanidino-2-oxopentanoate were specific for Duroc while L-Methionine, D-Glutamate, Thiamine, Deoxycytidine, Chenodeoxycholate were specific for Landrace.

Compound-gene network for both the breeds (Fig. 5) along with the putative genes (Supplementary Table S6) underlying the pathways were constructed. In Duroc, 32 metabolites were cross-referenced with the KEGG database thereby identifying 63 genes involved in 5 pathways in TP1 while 70 metabolites were related to 79 genes involved with 11 pathways in TP2. Glycerophospholipid and xenobiotics metabolism was specific pathways for TP1 after compound-gene cross-referencing while metabolism of butanoate, C21-steroid hormone biosynthesis, lysine, phosphatidylinositol phosphate, purine, pyrimidine pathways were specific at TP2. Metabolism of vitamin B6, tyrosine and Glycine-Serine-alanine-threonine was involved in both the time points (Fig. 5). In Landrace, 36 metabolites were related to 87 genes involved in 9 pathways in TP1 while 37 metabolites related to 40 genes involved with 5 pathways. Biosynthesis of androgen and estrogen, bile acid, C-21 steroid hormone. Metabolism of glycerophospholipids, methionine-cysteine, vitamin B1 and xenobiotics were specific for TP1 while metabolism of lysine, pyrimidine, vitamin B6 were specific for TP2. Glycine-Serine-alanine-threonine metabolism and valine-leucine-isoleucine degradation were common pathways between TP1 and TP2 in Landrace after cross-referencing of the metabolites. Interestingly, C21-steroid hormone biosynthesis and metabolism, glycerophospholipid and xenobiotics metabolism were identified only in TP1 in both the breeds and not present in TP2 while lysine, pyrimidine and vitamin B6 metabolism was identified only in TP2 in both Duroc and Landrace (Fig. 5).

Discussion

Improving FE greatly reduces the feed expense and increases the profit for the producers. However, it difficult to measure as it involves the accurate recording of dry matter intake and other features[25]. Therefore, any reliable predictors of FE that can be easily measured and used in selecting animals would be helpful for pig producers. There are many genetic/genomic studies on pig FE in Danish pig breeds[9,26]. However, this is the first study to relate FE with metabolomics to identify metabolomic markers or signatures in Danish pigs.

In our study, using a high throughput UPLC/MS system, we analyzed metabolite concentration in blood collected before and after the FI testing period to search for a metabolomics signature with respect to the FE and other related traits in Danish production pigs at two time points. A clear clustering of sampling time and breed,

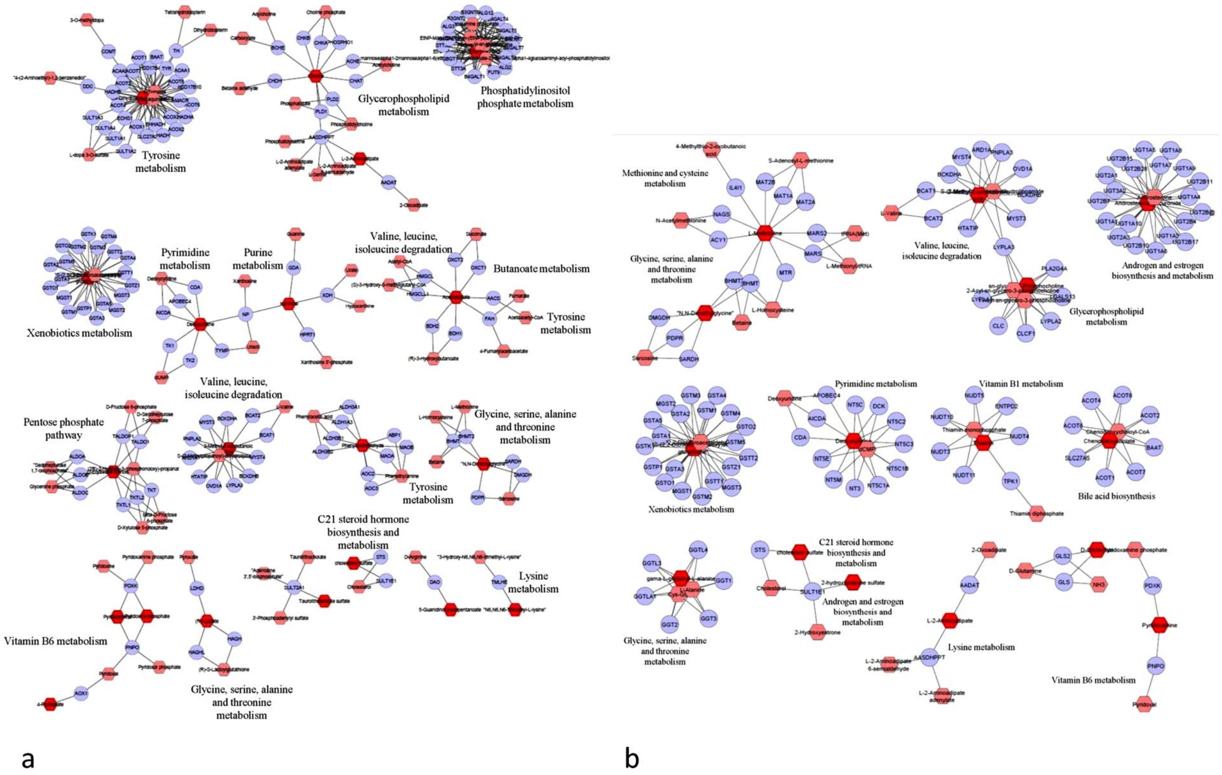

Tyrosine metabolism
Glycerophospholipid metabolism
Phosphatidylinositol phosphate metabolism
Xenobiotics metabolism
Pyrimidine metabolism
Purine metabolism
Valine, leucine, isoleucine degradation
Butanoate metabolism
Tyrosine metabolism
Valine, leucine, isoleucine degradation
Pentose phosphate pathway
Glycine, serine, alanine and threonine metabolism
Tyrosine metabolism
C21 steroid hormone biosynthesis and metabolism
Lysine metabolism
Vitamin B6 metabolism
Glycine, serine, alanine and threonine metabolism
a
Methionine and cysteine metabolism
Glycine, serine, alanine and threonine metabolism
Valine, leucine, isoleucine degradation
Androgen and estrogen biosynthesis and metabolism
Glycerophospholipid metabolism
Pyrimidine metabolism
Vitamin B1 metabolism
Xenobiotics metabolism
Bile acid biosynthesis
C21 steroid hormone biosynthesis and metabolism
Androgen and estrogen biosynthesis and metabolism
Glycine, serine, alanine and threonine metabolism
Lysine metabolism
Vitamin B6 metabolism
b

Figure 5. Compound-gene network for (**a**) Duroc (**b**) Landrace. The network is constructed using 101 metabolites underlying 17 different pathways in Duroc, while 28 metabolites are underlying 6 different pathways in Landrace.

among Duroc (TP2) and Landrace (TP1) (component 1) and Duroc (TP1) and Landrace (TP2) (component 2) gathered the samples according to their breeds and time points in four different groups, and supports the hypothesis of change in the metabolite profiles of the samples according to the breeds and time points. This also shows that the metabolite concentrations are not random and do have meaningful biological information.

We carried out an exploratory analysis by applying untargeted metabolomics, linear and network analyses, and pathway over-representation to unravel the effect of metabolites on FE phenotypes. A stronger association of metabolites with FE was expected at TP2; it is based on data recorded at the second sampling point. We do however believe that any metabolites found in TP1 would be more valuable for selection as this would allow for early screening of the pigs, leading to less wasted resources. As we also do find that the metabolites have a linear association between the time points, we do believe there is a potential for early screening using blood metabolites. Although the relation between the two time points and the lack of significant metabolites at TP1 may seem contradictory, it can likely be explained by several factors. The metabolite concentration in TP1 do not explain all the variation in TP2. If we combine this with the fact that FE is a multifaceted phenotype, which is not strongly controlled by a single factor, and in general is a somewhat subtle phenotype, it is easy to imagine that despite the connection between TP1 and TP2 we do not find the same results in both time points. Thus more data, and possibly a multiple-metabolite model may be needed for successful application of early screening.

From the KS test, we can observe that for most traits, the p-values are not uniformly distributed, with the highest p-value being 0.19. This means that if we apply FDR correction, all traits seem to have an overall relation to our traits, meaning even the borderline significant results are likely to be showing an underlying true effect. This establishes that metabolite profiles are a relevant source of information for our phenotypes of interest. Beyond the relevance of the metabolites for our phenotypes, we also established that for a large proportion of metabolites, the concentrations are linearly related over time. This shows us that despite variation over time, metabolites profiles show a level of temporal stability and predictability in our data.

From the heatmap clustering analysis of the top metabolites based on p-values in Duroc and Landrace separately for RFI, we observed that the samples clustered at two time points in both the breeds. A clear demarcation is observed while clustering the metabolites. In Duroc, the clusters identified were involved with the metabolism of phenylalanine, vitamin B6, arginine and ornithine, digestion of dietary lipids. Regarding Landrace, the clusters identified were found to be involved with biosynthesis of arginine-proline metabolism, bile secretion, and lysine degradation.

We applied a well-known gene co-expression network approach – WGCNA[17] to analyze the metabolomics data in this study. From the network analysis, we found several modules associated with FE, TDG, and RFI in both the breeds at different time points, pointing towards the common pathways influencing these traits. The change in the metabolites found at different time points supports the fact that there are changes in metabolomic levels related to

FE, TDG, and RFI between Duroc and Landrace breeds. We also constructed a compound-gene network for the significant unique metabolites in Duroc and Landrace to identify the pathways after cross-referencing with the KEGG pathways specific for humans and identifying genes underlying these pathways.

Based on the over-representation pathway analysis, we identified some key pathways in two time points in each breed (Supplementary Table S4). We also created a compound-gene network by applying Metscape 3.1.3. The compound-gene network in Duroc pointed towards 13 specific pathways underlying metabolism (butanoate, glycerophospholipid, glycine-serine-alanine-threonine, lysine, purine, and pyrimidine, vitamin B6, tyrosine, C21-steroid hormone, phosphatidylinositol phosphate, and xenobiotics), valine-leucine-isoleucine degradation and pentose phosphate pathway (Fig. 5a). In Landrace, 12 pathways were identified with some of them overlapping with Duroc, while androgen and estrogen biosynthesis, bile acid biosynthesis, methionine, and cysteine metabolism and Vitamin B1 metabolism specific to Landrace (Fig. 5b).

Among all the key metabolites identified in Duroc TP1, we identified choline (C00114), which is involved in glycerophospholipid metabolism and glycine-serine-alanine-threonine metabolism. Choline was found to be a hub metabolite involved with both FE and RFI in Duroc TP1 (Fig. 5, Supplementary Table S3). Choline is an essential nutrient for normal animal growth and performance and has been used as a supplement in the animal diets. Being an essential component of the cell wall and fat metabolism, choline is found to enhance FE and weight gain in ruminants[27]. Furthermore, Choline is a methyl donor taking part in DNA methylation, and is a vital process control the correct expression of genes thus ensuring proper cell development and growth[28].

The hub metabolite pyridoxamine (C00534), was found to be significant in Duroc TP1 (RFI) and Landrace TP2 (EDG, DG), which was identified for Vitamin B6 metabolism. Pyridoxamine phosphate plays an essential role in the interaction of amino acid, carbohydrate, fatty acid metabolism, and TCA cycle. Studies reported the relationship of B6 in tryptophan metabolism of weanling piglets but were unable to detect an effect on the oxidation of the tryptophan pathway and suggested that B6 may stimulate another pathway in tryptophan metabolism[29]. Metabolic shifts in lipid and carbohydrate utilization in high FE animals were reported[14]. They also reported reduced hepatic usage of fatty acid in high FE animals with a molecular alteration in lipid metabolism. A complementary analysis pointed out increased circulating triglycerides accompanied by a lower concentration of saturated and polyunsaturated fatty acids in the liver of high FE pigs[14].

We identified acetoacetate (C00164) to be the most significant for pathways underlying metabolism of butanoate, tyrosine, and valine-leucine-isoleucine degradation in Duroc TP2. Since butanoate is a metabolite of gut flora and involved with energy metabolism, butanoate metabolism may be activated under the conditions of cellular stress[30]. Oxidative stress reprograms lipid metabolism increasing the mitochondrial fatty acid oxidation[31]. Butanoate metabolism was also found to be enriched for differentially expressed genes in Nelore cattle muscle for RFI[32]. In our study, we found acetoacetate as the hub metabolite responsible for butanoate metabolism related to TDG in Duroc TP2 (Supplementary Table S3). However, in the study reported by Akbar[33], the subcutaneous administration of acetoacetate did not affect the FI. Acetoacetate was also found responsible for tyrosine metabolism as it affects *FAH*, an enzyme that catalyzes the last step of tyrosine metabolism. Metatranscriptomic studies revealed the tyrosine pathway to be differentially expressed in rumen microbiome of beef cattle[34]. Acetoacetate was also found as the hub (Fig. 5) for valine, leucine, and isoleucine pathway. We identified 3-methyl-2-oxobutanoic acid specific for this pathway also found to be associated with Duroc TP2. The three amino acids in the pathway are essential and act as a building block for tissue protein synthesis[35].

Deoxyuridine (C00526) and xanthine (C00385) were found to be involved with purine-pyrimidine metabolism in Duroc TP2. Deoxyuridine was associated with FE, while xanthine was associated with TDG in Duroc TP2 (Supplementary Table S3). Both the metabolites are involved in pyrimidine metabolism and are part of the cecal content of digestive segments involved with direct or indirect synthesis or utilization of compounds by the gut microbiota[36]. These metabolites were also reported to be affecting the digestive efficiency in chickens[37]. Previous studies have shown an increase in the concentration of xanthine by increased FI[38,39]. The degradation of rumen fluid into xanthine, hypoxanthine, and uracil by the action of bacterial nucleic acids (DNA, RNA) was reported previously[40]. The decrease in the rumen pH in dairy cows fed with high-grain diets changes the microbiota composition due to their intolerance towards low pH[41].

We identified cholesterol sulfate (C18043) associated with FE, TDG, and RFI in Duroc TP2, whereas with RFI in Landrace TP1 (Supplementary Table S3). The relationship among FI behavior, cholesterol, and triglyceride plasma levels in pigs was reported by Rauw *et al.*[42], wherein a strong co-relation between FI and cholesterol levels was established. However, these authors reported a weak correlation between RFI and cholesterol levels that were completely insignificant after correcting for the FC. The cholesterol pathways were also found to be consistent with the study involved in the regulation of FE in cattle (Holstein and Jersey) as reported by Salleh *et al.*[43].

Pathways such as lysine metabolism are affected by the metabolite 2-Aminoadipate (C00956) and were related to TDG in Duroc TP2, EDG and DG in Landrace TP2, respectively (Supplementary Table S3). Lysine is a limiting amino acid, and its deficiency impairs the animal's immunity and growth performance[44]. Yin *et al.*[45] suggested that the dietary supplementation with lysine influences intestinal absorption and metabolism of amino acids. Lysine restriction inhibits intestinal lysine transport and promotes FI associated with gut microbiome in piglets[45].

Functional annotation revealed some pathways involved with the metabolism and digestive gland secretion during feeding over-represented among the unique hubs and their role in FE, EDG, TDG, DG, and RFI in pigs. Based on the potential role of these metabolites in the metabolism of carbohydrate (butanoate), lipid (steroid, glycerophospholipid, pentose phosphate pathway, bile acid), amino acid (Gly-Ser-Ala-Thr, Lysine, Methionine-cysteine, tryptophan, tyrosine, valine-leucine-isoleucine), nucleotide metabolism (purine, pyrimidine), metabolism of cofactors and vitamins (B3, B6), and metabolism of xenobiotics, their involvement in the feeding behavior and FE traits are conceivable.

The genes identified from the compound-gene network were checked against the previously identified QTLs obtained from Animal Genome PigQTL database, where all previous research on QTLs is curated. Among the 198 genes identified for both Duroc and Landrace from both the time points, 9 genes were previously reported as candidate genes in the QTL database with varied traits (Supplementary Table S6). *NT5E* associated to Deoxycytidine in Landrace TP2 was identified as a candidate gene for RFI in the QTL database[46]. *HSD17B4*, which was associated to 3,4-Dihydroxy-L-phenylalanine in Duroc TP1, was identified as a candidate gene for carcass weight, backfat at tenth rib, and drip loss in Berkshire pigs[47]. Previous studies show the relation of FE with pork quality. Some studies reported that animals with low RFI have less back fat[48–50], less water holding capacity[48] and impaired sensory quality[50]. However, in some other studies, no difference was observed in the pork quality from low RFI pigs and controls with respect to drip loss, but a correlation between RFI and sensory traits related to reduced intramuscular lipid was observed[51]. A candidate gene, *MAOA*, associated to phenylacetaldehyde, was identified in our study that has been reported previously for intramuscular fat, ADG, and loin muscle[52]. Previous studies in Duroc reported high genetic variability due to moderate to high heritabilities for RFI, growth and carcass traits. An increase in the loin eye area was reported with decreased RFI, backfat and intramuscular fat content in Duroc pigs[53]. *NUDT3* was related to thiamin in Landrace TP1 and *PLD2* related to choline in Duroc TP1 in our study was also found to be a candidate gene for loin muscle area and loin muscle depth in pigs[54,55]. The metabolites and the genes identified are consistent with FE related traits. Further studies are warranted to evaluate the repeatability of our results in other pig population.

Conclusions

Our integrated approach using data annotation, linear model association, weighted metabolite network analysis, and pathway over-representation analysis indicated potential targets for biological processes related to FE. The significant metabolites affecting the pathways points out the role of the metabolites concerning to FE and related traits. Overall, we observe several trends in the results. We are able to identify relevant biological relation between our traits and metabolite profiles, but also differences in breed and time points. In contrast, we also see that there is some linear predictability in the metabolites between time points. As the pigs are entering and undergoing a very rapid growth and maturation rate between samplings, it is natural to expect that the underlying metabolite profiles and networks are changing, despite elements of stability in metabolite profiles. This means that strategies for applying metabolite information into a real life farming appear to be complex and require good understanding of the relations and changes in metabolite profiles and time, and the identification of not only key metabolites, but also key time points. Validation of these results in a cohort with more animals and time points would help to establish a framework for future FE prediction using metabolomics biomarker profiles that could be practical to use in large populations other than genomic profiling. More data would also make it possible to model the complex relations in metabolite profiles over time more accurately. Further understanding of the mechanisms driving these trends will result in improved nutrient utilization, reduction in production costs, and increased FE in pigs. To best of our knowledge, this is the first study to report metabolomics profiles related to FE and related traits in Danish pigs.

Methods

Ethical approval.

The blood sampling and experiment were approved and carried out in accordance with the Ministry of Environment and Food of Denmark, Animal Experiments Inspectorate under the license number (tilladelsesnummer) 2016-15-0201-01123, and C-permit granted to the principal investigator/senior author (HNK).

Study design and phenotypes.

The pigs used in this experiment were housed at the pig testing station "Bøgildgård" operated by SEGES within Landbrug and Fødevarer (L&F: Danish Agriculture and Food Council). Pigs were *ad libitum* fed and free water supply. The authors of this study were not responsible for animal husbandry, diet, and care as the testing station is a facility within the Danish breeding program, run by SEGES.

Blood samples were collected at a boar testing station Bøgildgård, owned by SEGES. The pigs were purebred uncastrated males from Danbred Duroc (n = 59) and Danbred Landrace breeds (n = 50), amounting to a total of 109 pigs. The initial bodyweight of the pigs before the testing period was approximately 7 kg, followed by a 5-week acclimatization phase. The pig diet consisted of a feed mixture with the main ingredients being: 39% barley, 27% wheat, 14% soybean meal and 6% oats. For the phenotypic traits, the weight of FC in kg and FE for each pig in the testing phase was measured beginning with an initial weight of around 28 kg for each pig. Bodyweight measured at two time points, the beginning and end of the test, were available from standard test procedure of the testing station and their difference was referred to as delta weight (DW). FE was calculated as the ratio between DW and FC. The testing phase ranged from 41 to 70 days based on the viability of each pig. The DG was calculated for three time phases – birth to testing (EDG), testing start to end (TDG), and birth to testing end (DG). RFI was computed as the difference between the observed daily feed intake (DFI) and the predicted feed intake (pDFI)[56]. All pigs consumed the same feed until the test end.

For the study of metabolites, approximately 5 mL of blood was collected from jugular venipuncture from each pig into tubes containing ethylenediaminetetraacetic acid (EDTA) and immediately placed on ice. Samples were collected at two time points, one at the start of this test phase (approximately 28 kg weight) and the second after 45 days referred to as TP1 and TP2 in the further study. The pigs were sampled at the same time of the day and same day of the week to insure the most comparable sampling. Pigs were in non-fasted state. For the separation of the blood plasma, samples were centrifuged at 3000 g for 10 minutes at 4 °C, and plasma was stored at −80 °C.

Non-targeted metabolomics analysis.

The plasma samples extracted from each pig were subjected to metabolomics analysis. The samples were processed by MS-Omics (http://www.msomics.com/; Denmark), and the analysis was carried out using a UPLC system (UPLC Acquity, Waters) coupled with time of flight mass spectrometer (Xevo G2 Tof Waters). An electrospray ionization interface was used as an ionization source. The analysis was performed in negative and positive ionization mode. The UPLC was performed using a slightly modified version of the protocol described by Catalin *et al.* (UPLC/MS Monitoring of Water-Soluble Vitamin Bs in Cell Culture Media in Minutes, Water Application note 2011, 720004042en).

Raw files were processed using MZmine 2[57]. The mass detection was ascertained, keeping the noise level at 1E2 (negative mode) and 1E3 (positive mode). The chromatogram building was achieved using a minimum time of 0.05 min, a minimum height of 1E3 (positive mode) and 4E2 (negative mode), and *m*/*z* tolerance of 0.01 (5ppm). The local minimum search deconvolution algorithm was used with a baseline cutoff, the minimum peak height of 2E3 (positive mode) and 5E2 (negative mode), and peak duration range of 0.04–5.0 min (positive mode) and 0.05–5.0 min (negative mode). Chromatograms were deisotoped with *m*/*z* tolerance of 0.01 (or 5 ppm) and an RT tolerance of 0.2 minutes for positive and 0.5 minutes for negative modes, respectively. Peak alignment was performed with (*m*/*z* tolerance at 0.01 (or 5 ppm). The peak list was eventually gap-filled with the peak finder module (intensity tolerance at 50% and *m*/*z* tolerance at 0.01 (or 5 ppm).

The identification of the metabolites was performed using both peak retention times (compared against authentic standards included in the analytical sequence) and accurate mass (with an acceptable deviation of 0.005 Da). As a standard quality control, samples with blank >3 were not included. The relative standard deviation between QC samples was kept less than 60, the correlation between the dilution of QC and response was >0.8.

The data were aligned and normalized using total ion intensity. The metabolites were identified by comparison with the online Human Metabolome Database (HMDB)[58] using exact *m*/*z* values and retention time. The metabolites that did not correspond to HMDB were left unannotated. These compounds were annotated based on a library search of the masses in the HMDB with a mass uncertainty of 0.005 Da or 5 ppm. The search in HMDB assumes that all ions originated from the $[M+H]^+$ or $[M+Na]^+$ (in positive ionization) or $[M-H]^-$ (in negative ionization) ions.

Metabolite-trait association analyses. The metabolite data were log-normalized before fitting the linear model. For each group of metabolites, only those with relative standard deviation >0.15 were used, based on the raw counts. The log-normalized metabolite concentration (m_{ijk}) was adjusted for fixed and random effects as follows.

$$m_{ijk} = B_i + S_j + P_k + \varepsilon_{ijk} \tag{1}$$

where,

m_{ijk}: is the relative concentration of each metabolite;
B_i: is the fixed effect for the breed;
S_j: is the batch effect;
P_k: is the random effect from the pen;
ε_{ijk}: is the random residual effect associated with each observation.

For each adjusted metabolite, denoted as $\widehat{m_{ijk}}$,(where, $\widehat{m_{ijk}} = (m_{ijk} - \hat{B}_i + \hat{S}_j + \hat{P}_k)$ from Eq. 1), the linear association with the pig phenotypes was estimated based on the following model:

$$y_{ij} = \widehat{m_{ijk}} + A_j$$

Where,

y_{ij}: is the phenotype (FE, EDG, TDG, DG, RFI) for each animal;
$\widehat{m_{ijk}}$: are the adjusted metabolites based on the Eq. (1);
A_j: is the covariate for animal's sampling age in days;

We did not include the sampling age with our other fixed and random effects when correcting our metabolites, as the sampling age is correlated with our phenotypes. This is because the slower-growing pigs have a higher sampling age as it takes long time for them to reach the testing phase. Adjusting for sampling age *a priori* would thus create biases[59]. Thus, we included sampling age as a covariate in the final models associating corrected metabolites with our phenotypes.

Many models were used, so instead of looking into the specific results of each metabolite in each model, we initially tested the significance of the model based on all metabolites. This was done by using the Kolmogorov-Smirnov test to compare the resulting p-value distribution with the uniform distribution for the parameter of interest in each batch. Cluster analysis and heatmap of significantly different metabolites were generated using the 'pheatmap' package in R (v1.0.12).

Metabolite network analysis. Network analysis was performed using Weighted Gene Co-expression Network Analysis (WGCNA) R package version 1.66[17]. The WGCNA methods have been successfully applied to gene expression data from microarrays[60] and RNA sequencing platforms in animal sciences[18]. Recently this methodology was applied on genome-wide genotype data as well[61]. Hereby, we extended this methodology to build networks using metabolomics data. The methodology, in summary, involved the Spearman correlation between all adjusted metabolite concentrations followed by the transformation of the correlation matrix into an adjacency matrix (AM) by fitting a power coefficient beta (β). The β was chosen by testing the coefficient between 12 and 22 and selecting the one that maximizes the scale-free topology based on the scale-free R^2 value >0.8. From the scaled correlation, the Topological Overlap Measure (TOM), representing the connection between metabolites was calculated. Based on the TOM and applying the *dynamicTreeCut* algorithm, modules of connected metabolites were generated. In each module, the eigengene values of the module metabolites were calculated.

A linear model was fitted among the eigengene values of metabolite modules and the phenotypes to assess the module-phenotype relationship. Further, we intersected the metabolites identified based on the linear association with those from the modules significantly associated with phenotypes. Metabolites with p-values ≤ 0.05 and those clustered into the modules with a phenotypic correlation ≥ 0.2 and $p \leq 0.1$ were selected for subsequent analysis.

Pathway analysis and network visualization.

Over-representation analysis was performed using IMPaLA[24] to identify metabolites underlying pathways meaningful to FE related-traits. IMPaLA takes into account the pathways from 11 public databases, including Reactome[62] and KEGG pathway[63]. Over-represented biological pathways were taken as significant with $p \leq 0.05$.

The visualization of metabolomic data was done in the context of human metabolic networks using Metscape v 3.1.3[64], a Cytoscape plugin. Based on that, we identified the connections between metabolites and the putative genes underlying the pathways in a compound-gene network approach. The key metabolites that were found to be involved in the main pathways were referred to as hub metabolites. The hub metabolites (compound IDs) from each significant pathway were selected and visualized using Cytoscape. A schematic representation of the methodology is given in Supplementary Fig. S3.

Data availability

The datasets generated and/or analyzed during the current study are publicly available upon acceptance of the paper at Metabolights database https://www.ebi.ac.uk/metabolights/MTBLS1384 with accession ID: MTBLS1384. https://doi.org/10.1093/nar/gks1004. PubMed PMID: 2310955.

Received: 21 May 2019; Accepted: 23 December 2019;
Published online: 14 January 2020

References

1. Patience, J. F., Rossoni-Serao, M. C. & Gutierrez, N. A. A review of feed efficiency in swine: biology and application. *J. Anim. Sci. Biotechnol.* **6**(1), 33 (2015).
2. Dekkers, J. C. M. & Gilbert, H. Genetic and biological aspect of residual feed intake in pigs. *9th World Congr Genet Appl to Livest Prod*. 1–8 (2010).
3. Do, D. N., Strathe, A. B., Jensen, J., Mark, T. & Kadarmideen, H. N. Genetic parameters for different measures of feed efficiency and related traits in boars of three pig breeds. *J. Anim. Sci.* **91**(9), 4069–79 (2013).
4. Lu, D. *et al*. The relationship between different measures of feed efficiency and feeding behavior traits in Duroc pigs. *J. Anim. Sci.* **95**(8), 3370–3380 (2017).
5. Fan, B. *et al*. Identification of genetic markers associated with residual feed intake and meat quality traits in the pig. *Meat Sci.* **84**, 645–650 (2010).
6. Koch, R. M., Swiger, L. A., Chambers, D. & Gregory, K. E. Efficiency of Feed Use in Beef Cattle. *J. Anim. Sci.* **22**, 486–494 (1963).
7. Do, D. N., Strathe, A. B., Ostersen, T., Pant, S. D. & Kadarmideen, H. N. Genome-wide association and pathway analysis of feed efficiency in pigs reveal candidate genes and pathways for residual feed intake. *Front. Genet.* **5**, 307 (2014).
8. Quan, J. *et al*. Genome-wide association study reveals genetic loci and candidate genes for average daily gain in Duroc pigs. *Asian-Australasian J. Anim. Sci.* **31**, 480–488 (2018).
9. Gilbert, H. *et al*. Genetic parameters for residual feed intake in growing pigs, with emphasis on genetic relationships with carcass and meat quality traits. *J. Anim. Sci.* **85**, 3182–3188 (2007).
10. Saintilan, R. *et al*. Genetics of residual feed intake in growing pigs: Relationships with production traits, and nitrogen and phosphorus excretion traits. *J. Anim. Sci.* **91**, 2542–2554 (2013).
11. Gilbert, H. *et al*. Review: Divergent selection for residual feed intake in the growing pig. *Animal* **11**, 1427–1439 (2017).
12. Yi, Z. *et al*. Feed conversion ratio, residual feed intake and cholecystokinin type A receptor gene polymorphisms are associated with feed intake and average daily gain in a Chinese local chicken population. *J. Anim. Sci. Biotechnol.* **9**, 1–9 (2018).
13. Nkrumah, J. D. *et al*. Genetic and phenotypic relationships of feed intake and measures of efficiency with growth and carcass merit of beef cattle. *J. Anim. Sci.* **85**, 2711–2720 (2007).
14. Reyer, H. *et al*. Strategies towards Improved Feed Efficiency in Pigs Comprise Molecular Shifts in Hepatic Lipid and Carbohydrate Metabolism. *Int. J. Mol. Sci.* **18**, 1674 (2017).
15. Fontanesi, L. Metabolomics and livestock genomics: Insights into a phenotyping frontier and its applications in animal breeding. *Anim. Front.* **6**, 73–79 (2016).
16. Cônsolo, N. R. B. *et al*. Associations of Blood Analysis with Feed Efficiency and Developmental Stage in Grass-Fed Beef Heifers. *Anim.* **8**, 133 (2018).
17. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
18. Kogelman, L. J. A. *et al*. Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue in a porcine model. *BMC Med. Genomics* **7**, 57 (2014).
19. Drag, M., Skinkyté-Juskiené, R., Do, D. N., Kogelman, L. J. A. & Kadarmideen, H. N. Differential expression and co-expression gene networks reveal candidate biomarkers of boar taint in non-castrated pigs. *Sci. Rep.* **7**, 12205 (2017).
20. Novais, F. J. *et al*. Identification of a metabolomic signature associated with feed efficiency in beef cattle. *BMC Genomics* **20**, 8 (2019).
21. Diniz, W. J. S. *et al*. Detection of Co-expressed Pathway Modules Associated with Mineral Concentration and Meat Quality in Nelore Cattle. *Front Genet*., https://doi.org/10.3389/fgene.2019.00210 (2019).
22. Hoque, M. A. & Suzuki, K. Genetics of residual feed intake in cattle and pigs: A review. *Asian Australasian J. Anim. Sci.* **22**, 747–755 (2009).
23. Archer, J. A., Reverter, A., Herd, R. M., Johnston, D. J. & Arthur, P. F. Genetic variation in feed intake and efficiency of mature beef cows and relationships with postweaning measurements. *Proc 7th Wld Congr Genet Appl Livest Prod* **31**, 221–224 (2002).
24. Kamburov, A., Cavill, R., Ebbels, T. M. D., Herwig, R. & Keun, H. C. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* **27**, 2917–2918 (2011).
25. Nielsen, M. K. *et al*. Review: Life-cycle, total-industry genetic improvement of feed efficiency in beef cattle: Blueprint for the Beef Improvement Federation. *Prof. Anim. Sci.* **29**, 559–565 (2013).
26. Zhang, C. *et al*. Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants. *Genet. Sel. Evol.* **50**, 14 (2018).
27. Gralak, M. A., Lesniewska, V., Puchala, R., Barej, W. & Dymnicki, E. The effect of betaine and rumen undegradable choline on growth rate and feed efficiency in calves. *J. Anim. Feed Sci.* **7**, 229–233 (1998).
28. Niculescu, M. D. & Zeisel, S. H. Diet, Methyl Donors and DNA Methylation: Interactions between Dietary Folate, Methionine and Choline. *J. Nutr.* **132**, 2333S–2335S (2002).
29. Matte, J. J., Ponter, A. A. & Sève, B. Effects of chronic parenteral pyridoxine and acute enteric tryptophan on pyridoxine status, glycemia and insulinemia stimulated by enteric glucose in weanling piglets. *Can. J. Anim. Sci.* **77**, 663–668 (1997).

30. Sauer, J., Richter, K. K. & Pool-Zobel, B. L. Physiological concentrations of butyrate favorably modulate genes of oxidative and metabolic stress in primary human colon cells. *J. Nutr. Biochem.* **18**, 736–745 (2007).
31. Douglas, D. N. *et al*. Oxidative Stress Attenuates Lipid Synthesis and Increases Mitochondrial Fatty Acid Oxidation in Hepatoma Cells Infected with Hepatitis C Virus. *J. Biol. Chem.* **291**, 1974–1990 (2016).
32. Tizioto, P. C. *et al*. Gene expression differences in Longissimus muscle of Nelore steers genetically divergent for residual feed intake. *Sci. Rep.* **6**, 39493 (2016).
33. Physical, A. N. and Metabolic Constraints on Feed Intake in Ruminants: A Systematic Model. *Adv. Dairy Res.* **06**, 2 (2018).
34. Li, F. & Guan, L. L. Metatranscriptomic Profiling Reveals Linkages between the Active Rumen Microbiome and Feed Efficiency in Beef Cattle. *Appl. Environ. Microbiol.* **83**, 1–16 (2017).
35. Zhang, S., Zeng, X., Ren, M., Mao, X. & Qiao, S. Novel metabolic and physiological functions of branched chain amino acids: A review. *J. Anim. Sci. Biotechnol.* **8**, 10 (2017).
36. Nicholson, J. K. *et al*. Host-gut microbiota metabolic interactions. *Science* **336**, 1262–1267 (2012).
37. Beauclercq, S. *et al*. Relationships between digestive efficiency and metabolomic profiles of serum and intestinal contents in chickens. *Sci. Rep.* **8**, 6678 (2018).
38. Khafipour, E. *et al*. Effects of grain feeding on microbiota in the digestive tract of cattle. *Anim. Front.* **6**, 13–19 (2016).
39. Saleem, F. *et al*. A metabolomics approach to uncover the effects of grain diets on rumen health in dairy cows. *J. Dairy Sci.* **95**, 6606–6623 (2012).
40. McAllan, A. B. & Smith, R. H. Degradation of nucleic acids in the rumen. *Br. J. Nutr.* **29**, 331–345 (1973).
41. Slyter, L. L. Influence of acidosis on rumen function. *J. Anim. Sci.* **43**, 910–929 (1976).
42. Rauw, W. M. *et al*. Behaviour influences cholesterol plasma levels in a pig model. *Animal* **1**, 865–871 (2007).
43. Salleh, M. S. *et al*. RNA-Seq transcriptomics and pathway analyses reveal potential regulatory genes and molecular mechanisms in high- and low-residual feed intake in Nordic dairy cattle. *BMC Genomics* **18**, 258 (2017).
44. Liao, S. F., Wang, T. & Regmi, N. Lysine nutrition in swine and the related monogastric animals: muscle protein biosynthesis and beyond. *SpringerPlus* **4**, 147 (2015).
45. Yin, J. *et al*. Lysine Restriction Affects Feed Intake and Amino Acid Metabolism via Gut Microbiome in Piglets. *Cell. Physiol. Biochem.* **44**, 1749–1761 (2017).
46. Do, D. N. *et al*. Genome-wide association and systems genetic analyses of residual feed intake, daily feed consumption, backfat and weight gain in pigs. *BMC Genet.* **15**, 27 (2014).
47. Jo, J. L. *et al*. Association between a non-synonymous HSD17B4 single nucleotide polymorphism and meat-quality traits in Berkshire pigs. *Genet. Mol. Res.* **15**, 1–11 (2016).
48. Lefaucheur, L. *et al*. Muscle characteristics and meat quality traits are affected by divergent selection on residual feed intake in pigs. *J. Anim. Sci.* **89**, 996–1010 (2011).
49. Cai, W., Casey, D. S. & Dekkers, J. C. M. Selection response and genetic parameters for residual feed intake in Yorkshire swine. *J. Anim. Sci.* **86**, 287–298 (2008).
50. Faure, J. *et al*. Consequences of divergent selection for residual feed intake in pigs on muscle energy metabolism and meat quality. *Meat Sci.* **93**, 37–45 (2013).
51. Smith, R. M. *et al*. Effects of selection for decreased residual feed intake on composition and quality of fresh pork. *J. Anim. Sci.* **89**, 192–200 (2011).
52. Terenina, E. *et al*. Association study of molecular polymorphisms in candidate genes related to stress responses with production and meat quality traits in pigs. *Domest. Anim. Endocrinol.* **44**, 81–97 (2013).
53. Hoque, M. A., Suzuki, K., Kadowaki, H., Shibata, T. & Oikawa, T. Genetic parameters for feed efficiency traits and their relationships with growth and carcass traits in Duroc pigs. *J. Anim. Breed. Genet.* **124**, 108–116 (2007).
54. Zhuang, Z. *et al*. Meta-analysis of genome-wide association studies for loin muscle area and loin muscle depth in two Duroc pig populations. *PLoS One* **14**, 1–21 (2019).
55. Choi, I., Bates, R. O., Raney, N. E., Steibel, J. P. & Ernst, C. W. Evaluation of QTL for carcass merit and meat quality traits in a US commercial Duroc population. *Meat Sci.* **92**, 132–138 (2012).
56. Nguyen, N. H. & McPhee, C. P. Selection for growth rate in pigs on restricted feeding. Genetic parameters and correlated responses in residual feed intake. In: *Association for the Advancement of Animal Breeding and Genetics*. 14th Conference, Queensland, NZ, 2001, Queenstown, NZ (2001).
57. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics.* **11**, 395 (2010).
58. Wishart, D. S. *et al*. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
59. Freckleton, R. P. On the misuse of residuals in ecology: regression of residuals vs the analysis of multiple regression. *J Anim Ecol.* **71**, 542–5 (2002).
60. Kadarmideen, H. N., Watson-Haigh, N. S. & Andronicos, N. M. Systems biology of ovine intestinal parasite resistance: disease gene modules and biomarkers. *Mol. Biosyst.* **7**, 235–246 (2011).
61. Carmelo, V. A. O., Kogelman, L. J. A., Madsen, M. B. & Kadarmideen, H. N. WISH-R- a fast and efficient tool for construction of epistatic networks for complex traits and diseases. *BMC Bioinformatics* **19**, 277 (2018).
62. Fabregat, A. *et al*. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
63. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
64. Karnovsky, A. *et al*. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics* **28**, 373–380 (2012).

Acknowledgements

This study was funded by the Independent Research Fund Denmark (DFF) – Technology and Production (FTP) grant (grant number: 4184-00268B). VAOC received partial Ph.D. stipends from the University of Copenhagen and Technical University of Denmark. Authors thank SEGES – Pig Research Centre (VSP) Denmark for access to blood samples and phenotype datasets used in this study as well as farm staff and group members who assisted in the sampling experiments.

Author contributions

H.N.K. conceived and designed this "FeedOMICS" project, obtained funding as the main applicant. V.C. and H.N.K. designed the blood sampling experiments, phenotype data collection, metabolite profiling, and biostatistical/bioinformatics analyses. V.C., P.B., and W.J.S.D. carried out biostatistical and bioinformatic data analysis. All authors collaborated in the interpretation of results, discussion and write up of the manuscript. All authors have read, reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-57182-4.

Correspondence and requests for materials should be addressed to H.N.K.

Genome regulation and gene interaction networks inferred from muscle transcriptome underlying feed efficiency in Pigs

Victor AO. Carmelo[1] and Haja N. Kadarmideen[1]*

[1]Quantitative Genomics, Bioinformatics and Computational Biology Group, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Building 324, 2800, Kongens Lyngby, Denmark

***Correspondence:**
Haja N. Kadarmideen

hajak@dtu.dk

Keywords: muscle transcriptome, pigs, feed efficiency, gene networks, candidate biomarkers

Abstract

Improvement of feed efficiency (FE) is key for sustainability and cost reduction in pig production. Our aim was to characterize the muscle transcriptomic profiles in Danbred Duroc (Duroc) and Danbred Landrace (Landrace), in relation to FE for identifying potential biomarkers. RNA-seq data was analyzed employing differential gene expression methods, gene-gene interaction and network analysis, including pathway and functional analysis. We compared the results with genome regulation in human exercise data. In the differential expression analysis, 13 genes were differentially expressed, including: *MRPS11, MTRF1,* TRIM63, MGAT4A, KLH30. Based on a novel gene selection method, the divergent count, we performed pathway enrichment analysis. We found 5 significantly enriched pathways related to feed conversion ratio (FCR). These pathways were mainly mitochondrial, and summarized in the mitochondrial translation elongation (MTR) pathway. In the gene interaction analysis, highlights include the mitochondrial genes: PPIF, MRPL35, NDUFS4and the fat metabolism and obesity genes: *AACS, SMPDL3B, CTNNBL1, NDUFS4* and *LIMD2*. In the network analysis, we identified two modules significantly correlated with FCR. Pathway enrichment of modules identified MTR, electron transport chain and DNA repair as enriched pathways. In the network analysis, the mitochondrial gene group *NDUF* was a key hub group, showing potential as biomarkers. Comparing with human transcriptomic exercise studies, genes related to exercise displayed enrichment in our FCR related genes. We conclude that mitochondrial activity is a driver for FCR in muscle tissue, and mitochondrial genes could be potential biomarkers for FCR in pigs. We hypothesize that increased FE mimics processes triggered in exercised muscle.

Introduction

In commercial pig production, the cost of feed is the highest individual economic factor (Jing, Hou et al. 2015, Gilbert, Billon et al. 2017). Furthermore, reduction in feed consumption per unit growth is beneficial for the environment, which is a key factor in being able to maintain sustainable and resource efficient production. In this context, there have been continuous efforts to increase feed utilization efficiency in pigs through selective breeding. In the Danish Production pig population, breeding is done at a core central facility where potential breeding sires are tested for FCR through accurate individual measurements of feed intake and growth. Danish production pigs are crossbreds, with the maternal line being Landrace x Danbred Yorkshire, and the paternal line being Durocs The Durocs are well-known for being heavily selected for growth and efficiency, while the two other breeds have had more heavy selection on litter size or piglet survival related traits.

Feed efficiency can be defined in several ways, with the main ones being Residual Feed Intake RFI(Koch 1963) and FCR. FCR is the ratio between feed consumed and growth, while RFI is based on the residual between predicted feed intake and actual feed intake given growth. In general, it is reported that selection for low FCR will result in co- selection for related traits, namely growth rate and body composition (Nkrumah, Basarab et al. 2007, Gilbert, Billon et al. 2017, Yi, Li et al. 2018). In contrast, selection for RFI is more directly focused on metabolic efficiency irrespective of daily gain and growth (Nkrumah, Basarab et al. 2007, Gilbert, Billon et al. 2017, Yi, Li et al. 2018). In general, RFI and FCR are strongly correlated, with a correlation above 0.7 and both show low to medium heritability(Do, Strathe et al. 2013). In general, FCR is simpler to calculate, as RFI calculation is dependent on individual population and production factors (Hoque, Kadowaki et al. 2009, Do, Strathe et al. 2013). However, in pig production, the side effects of FCR selection and simplicity are desired traits, thus perhaps explaining why the pig population in Denmark and in general pig production, FCR has been the main efficiency phenotype used for selection (Gilbert, Billon et al. 2017). One can also hypothesize that FCR is more easily translatable between breeds/populations, as it is a simple dimensionless ratio, which has a simple and generally comparable interpretation. In contrast, it is more difficult to easily compare RFI values across different populations or breeds. In regards to the biological and/or genetic background of FCR in pigs, the results remain somewhat elusive(Ding, Yang et al. 2018), thus inviting for further analysis on the topic.

The key tissue in pig production is muscle, as pig carcasses are valued according to lean meat content. Skeletal muscle is a key organ in carbohydrate and lipid metabolism and plays a large part in the storage of energy from feed (Turner, Cooney et al. 2014, Morales, Bucarey et al. 2017), especially as lean growth has been one of the main goals of pig breeding programs. Increased efficiency has also been positively associated with various meat quality parameters (Czernichow, Thomas et al. 2010, Lefaucheur, Lebret et al. 2011, Smith, Gabler et al. 2011, Faure, Lefaucheur et al. 2013, Horodyska, Oster et al. 2018), showing that improved FE can have multiple positive outcomes. There are only a few studies analyzing muscle tissue transcriptome pf pigs in a FE context(Jing, Hou et al. 2015, Vincent, Louveau et al. 2015, Gondret, Vincent et al. 2017, Horodyska, Wimmers et al. 2018), and thus our knowledge of the muscle transcriptomic background of FE is somewhat limited. In general, the studies available have relied on small samples sizes, weak statistical thresholds and categorical division of lines divergently selected for FE. This means that more studies are still needed to uncover the true underlying transcriptomic background of FE in muscle tissue.

Here, in our study, we aim to characterize the transcriptomic profiles and link them to FE traits measured in Duroc and Landrace, purebred pigs, by fitting FE as a continuous trait over a full spectrum of efficiency, from high to low. Furthermore, the pigs selected for the study all came out of the potential breeding sire population, with no pigs negatively selected for FE, thus better representing real world breeding scenarios than using negative FE selection. We analyzed the muscle transcriptome based on several layers of statistical-bioinformatics analyses: differential expression (DE), gene-gene interaction and network analysis, which was followed up by pathway and functional analysis. The rationale behind the approach was to reveal potential biomarkers that are functionally important and are predictive of FE in pigs. Dealing with complex yet subtle phenotypes can be a challenging, as the signal to noise ratio can be high, and it may be impractical or costly to collect large sample sizes. Therefore, we also suggest a novel method for selecting features based on overall p-value distributions, the divergent count.

To gain more insight on the molecular and functional background of FE, we also hypothesized, that the mechanism between differences in the muscle transcriptome of breeds with different efficiency could be similar to the differences between a rested and an exercised muscle, We adapted a translational genomics approach to investigate this, comparing human data with our data.

Materials and Methods

Sampling and Sequencing

In total, 41 purebred male uncastrated pigs where sampled for this study from two breeds, with 13 Danbred Durocand 28 Danbred Landrace pigs. All pigs were raised at a commercial breeding station at Bøgildgard owned by the pig research Centre of the Danish Agriculture and Food Council (SEGES). The pigs where raised from ~7kg until ~100kg at the breeding station. During this time, all feed intake was measured starting at 28kg and for a period of 40-70 days based on the viability of each pig. All pigs were routinely weighed several times, including at testing start and end for calculation of FCR. FCR was calculated by dividing the growth in the testing period with the feed consumption. Residual Feed Intake (RFI) was also estimated based on the residuals of the following model, from Do et al(Do, Strathe et al. 2013):

$$DFI_{ij} = \mu + DWG_i + \beta_j$$

Where DFI is daily feed intake and DWG is daily weight gain in the period, and β is the batch effect. RFI was calculated separately for each breed, and based on data from a larger population (Duroc *n*=59 and Landrace n=50).

Muscle tissue samples from the psoas major muscle were extracted immediately post slaughter and preserved in RNAlater (Ambion, Austin, Texas). Sample were kept at -25 C, as per protocol, until sequencing

Sequencing

Sequencing was done on BGISEQ-500 platform using the PE100 (pair end, 100bp length) with Oligo dT library prep at BGI Genomics.

QC, Mapping and Read Quantification

Reads were trimmed and adapters removed using Trimmomatic (Bolger, Lohse et al. 2014) version 0.39 with default setting for paired end reads. The QC on the data was done both pre- and post-trimming using FastQC v0.11.9(. The reads were mapped using STAR aligner(Dobin, Davis et al. 2013) version 2.7.1a using default parameters with a genome index based on sus scrofa version 11.1 and using ensemble annotation *sus scrofa* 11.1 version 96 for splice site reference. Default parameters were used for mapping except for the addition of read quantification during mapping using the --quantMode GeneCounts setting. All statistic for the reads can be found in supplementary data 1.

Differential Expression Analysis

To analyze the relationship between FCR and gene expression, we applied the following overall model, and implemented it using several different methods:

$$y_{ijklm} = \mu + \beta_{1_i}(FCR) + \beta_{2_j}(RIN) + \beta_{2_k}(age) + BR_l + BA_m + \epsilon \qquad (1)$$

$y = normalized\ read\ counts$

$\beta_1 = regression\ coefficient\ \ of\ feed\ conversion\ rate$

$\beta_2 = regression\ coefficient\ \ of\ RIN\ (RNA\ Integrity\ value)$

$\beta_3 = regression\ coefficient\ \ of\ Slaugter\ Age(days)$

$BR = effect\ size\ of\ Breed$

$BA = effect\ size\ of\ Batch$

RNA integrity value (RIN) should be corrected for, as it affects expression, and the most appropriate way to correct this is to include it in the model(Gallego Romero, Pai et al. 2014). As the samples had different slaughter days, which affected the collection conditions, we also deemed it necessary to correct for this via the batch effect. Finally, we correct for Breed and age at slaughter, as these are biological factors, which can cause differences in expression.

We used the following 3 methods for the DEA: Limma (Ritchie, Phipson et al. 2015), edgeR (Robinson, McCarthy et al. 2010) and Deseq2(Love, Huber et al. 2014). This was done to increase the robustness of our analysis, as our phenotype of interest is expected to have a subtle effect on the transcriptome due to the complex nature of FE. In addition, we also fit the model for each breed separately using Deseq2, just removing the Breed as a covariate.

Deseq2

We used Deseq2 version 1.22.2. In the Deseq2 analysis, the counts were filtered a priori requiring a minimum of 5 reads for each sample, resulting in a total of 10765 out of 25880 genes being included in the DE analysis in the common breed analysis, and 10687 and 11107 in Landrace and Duroc respectively. As the overall read counts were very similar across experiments (see supplementary data 1), it was deemed sufficient to filter pre normalizing. We then used the default analysis method based on our specified model.

Limma

We used Limma version 3.38.3. For the Limma analysis, the counts were filtered based on the edgeR *filterByExpr*function and normalized using *calcNormFactors* from the same package, as suggested in the limma manual. This resulted in the inclusion of 11146 genes in the analysis. To fit the model we used the *eBayes* method in conjunction with our specified model.

EdgeR

We used edgeR 3.24.3. We used the same normalization and filtering as in the Limma analysis, thus including the same number of genes. We used the *glmQLfit* function and *glmQLTest* to implement our model.

While we used to different set sizes in the analysis, this does not affect the results significantly, as the genes omitted in the Deseq2 analysis are all lowly expressed. Furthermore, in our further analysis we elected to use the smaller and more conservative Deseq2 set to become our reference set for selections and analysis. **Gene Pathway Analysis**

Gene selection

To select a robust set of genes for a gene enrichment analysis when we have non-conservative p-value but only a limited number of genes with a FDR below 0.05, we applied the following strategy:

- Identify the overrepresentation of (low) p-values in comparison to a uniform p-value distribution in our data. We will call this the divergent count.
- Select the top N genes by p-value, where N is the estimated divergent count
- Among the top N genes, select those that are found in all three methods.

To find the divergent count D, we find the interval with the maximum positive divergence between our observed empirical p-values and the same number of uniformly distributed p-values. It is calculated as follows:

$$(1)\ d_i = \left(\sum_{n}^{i=1} x_i \begin{cases} 0\, for\ x_i \geq \frac{i}{n} \\ 1\, for\ x_i < \frac{i}{n} \end{cases} \right) - i$$

$$(2)\ D = \max\{d_1, d_2 \dots d_n\}$$

Where n is the total number of p-values, x_i is the i'th observed p-value in increasing order. Here i is both the index for x and the expected number of p-values between 0 and $\frac{i}{n}$ given a uniform distribution. D is the final divergent count, which is the maximum over all possible values of d..

GOrilla

To perform gene enrichment in GOrilla (Eden, Lipson et al. 2007, Eden, Navon et al. 2009), we translated our *sus scrofra* ensemble gene IDs into human ensemble gene IDs. The background set of genes used in GOrilla was the set of genes from the Deseq2 analysis. We used default settings. Furthermore, we used the Revigo (Supek, Bosnjak et al. 2011) analysis through GOrilla to generate summaries of our enrichment analysis, using default settings.

Feed Efficiency measure

In this study, we elected to use weight gain/feed intake as our FCR measure. It fit the data better than RFI, and FCR is the metric used in the breeding program of our pigs.

Pairwise Gene interaction Analysis

To continue our analysis of the top set of genes identified using the divergent counts in our DE analysis, we decided to apply a pairwise interaction model. First, we adjust the expression based on any factors and covariates that may affect expression for each gene. These factors are the same as in the general DE analysis, giving rise to the following linear model:

$$y_{jklm} = \mu + \beta_{1_j}(RIN) + \beta_{2_k}(age) + BR_l + BA_m + \epsilon \qquad (2)$$

$$y = normalized\ read\ counts$$

$$\beta_1 = regression\ coefficient\ \ of\ RIN\ (RNA\ Integrity\ value)$$

$$\beta_2 = r\ egression\ coefficient\ \ of\ Slaugter\ Age(days)$$

$$BR = Breed$$

$$BA = Batch$$

We then centered and scaled the residuals and then run a model for all pairwise gene interaction in our gene set. The reason we scaled and centered is that this leads to a more flexible and interpretable model regardless of the type of interaction. The interaction model was as follows:

$$y_i = \mu + \beta_1 x_{1_j} + \beta_2 x_{2_k} + \beta_3 \left(x_{1_j} \times x_{2_k}\right) + \epsilon \qquad (3)$$

$$y = FCR\ values$$

$$\beta_1 = regression\ coefficient\ \ of\ residual\ expression\ of\ gene\ 1$$

$$\beta_2 = regression\ coefficient\ \ of\ residual\ expression\ of\ gene\ 1$$

$$\beta_3 = regression\ coefficient\ \ of\ the\ interaction\ between\ gene\ 1\ and\ gene\ 2$$

$$x_{1_j} = residual\ expression\ of\ gene\ 1$$

$$x_{2_k} = residual\ expression\ of\ gene\ 2$$

$$\left(x_{1_j} \times x_{2_k}\right) = \text{product of the two residual expression values}$$

The next step was then to identify significant interactions. As the number of interaction in a dataset grows exponentially to the square of the input space, it is often difficult to detect effects based on classical multiple testing correction methods such as Bonferroni or FDR. This is especially true when dealing with complex phenotypes, as we generally do not expect to find individual large effects. Due to this, instead of focusing on individual results, for each gene, we calculated the divergent count, to assess the divergence of each genes distribution of interaction p-values. We then bootstrapped with replacement samples of 853 p-values from our empirical p-values 10^5 times, calculating the divergent count each time, giving us a bootstrapped distribution of divergent counts, to compare with our empirical distribution

Network analysis

To perform network analysis we used WGCNA(Langfelder and Horvath 2008). First, we filtered the read counts to only include genes with a minimum of 5 un-normalized reads, as was done for the Deseq2 analysis. We then created a correlation matrix based on all pairwise correlation in the data. The correlation matrix was based on un-normalized values as the correlation metric is based of comparison of the samples with themselves, thus it is not affected by the covariates. We then fit the ß parameter for the scaling of the network to create a scale free topology(Zhang and Horvath 2005). The scaled correlation matrix was used as an adjacency matrix that was used to generate the

Topological Overlap Measures (TOM), which represents the final calculation of the relation between genes.

The TOM values of the genes where clustered using the *dynamicTreeCut* function from the dynamicTreeCut cut package with default setting, resulting in a number of module which are arbitrarily named based on colors.

The eigenvalue of each module was then calculated based on the normalized read counts and RIN adjusted count. We did these corrections in this step to remove the technical effects of library size differences and RIN from the eigenvalues, as we did not want technical effects to affect the eigenvalues.. The counts were normalized based on the *calcNormFactors* function from the edgeR package. After this, the counts were adjusted for RIN by fitting the following linear model: $expression = \mu + \text{RIN} + \epsilon$ for all genes, and extracting the residual expression values. Highly correlating models where merged using the *mergeCloseModules* function using a default cut-off. We then calculated the Pearson correlation between corrected and normalized module eigenvalues and our traits and covariates. Pathway analysis was performed on the genes of highly correlated modules, with GOrilla and ReviGO as seen above. Finally, we also identified the top hub genes in relevant modules. This was done based on calculating the intramodular connectivity using the *intramodularConnectivity* function with default settings. We then selected the top hub genes base on the kWithin measure, which represents the connectivity within modules.

Comparison to human exercise data

To test the hypothesis that differences in the muscle tissue transcriptome of Duroc and Landrace and/or FCR related genes mimic differences in rested and exercised muscle tissue, we compared our results with three human data sets(Murton, Billeter et al. 2014, Devarshi, Jones et al. 2018, Popov, Makhnovskii et al. 2019). For each data set, we performed the following:

1. Select the genes differentially expressed between breeds, based on the edgeR analysis
2. For FCR, use the 853 genes from divergent count set
3. Find the same set of genes in the human data – the breed/FCR matching genes. Genes are matched using the biomart R package, based on retrieving the external_gene_name of our sus scrofa ensemble gene identifiers.
4. Separate the human data into two parts – the breed matching set and the background set

5. Using a Fisher Exact test, compare the number of differentially expressed genes for the exercised vs rested muscle in the background set vs the breed matching set.
6. The steps for the breed were also applied to our divergent count set for FCR.

The reason edgeR was used in this part of the analysis, was because it was more flexible to fit to the publicly available data, allowing to compare our results to the other studies. As each dataset was formatted and analyzed differently, we had to process them individually. In the data set from Devarshi et al(dataset 1)(Devarshi, Jones et al. 2018), we chose to use the lean pre exercise vs lean post exercise group as our comparison, and significance was based on the reported cuffdiff analysis. For the set of Murton et al(dataset 2)(Murton, Billeter et al. 2014), we pooled all control vs exercise samples and analyzed them using Limma as the data was microarray data, using the same Limma pipeline as mentioned above in our FE analysis. As the results were weaker in Murton et al, we chose to use P <0.05 as a cutoff for the Fisher exact test. For the set from Popov et al(dataset 3)(Popov, Makhnovskii et al. 2019), we grouped all the 4h post exercise results vs all 4h control non-exercised and performed DE analysis using edgeR with no other covariates using the same settings as our FE analysis above, with significance based on the found FDR values.

Results

Differential Expression analysis

In figure 1 we can see the visualization of the PCA analysis of the count data. There is one main point: there is no clear pattern separating the breeds based on the first two components. Based on the lack of separation of the breeds we gain confidence in the application of a common breed analysis. Any of the lower variance components have a lower proportion of the variation explained than the two observed Principal Components, therefore we are confident that no major proportion of the variation is directly driven by breed. We do observe a significant and detectable effect of breed expression level (as seen further down), meaning there are features in our data which *can* separate the breeds.

In figure S1 we can see the distribution of the uncorrected p-values for the Deseq2 analysis in our two breeds in relation to FCR with the corresponding figure for the common analysis in figure 2(right). In total, the Landrace analysis had one gene with an FDR < 0.1, and Duroc had 8, and we found 4 in the common breed analysis. Overall, we only find a limited set of genes associated with FCR. In table 1, we see the overview over the genes that where differentially expressed at the 0.1 FDR level in the common and individual breed analysis from Deseq2. As in previous studies, we find genes related to mitochondria (MRPS11, MTRM1) and glucose a related gene (*MGAT4A)(Ohtsubo, Takamatsu et al. 2005)*. We also find genes that have been associated with meat quality phenotypes in cattle and pig (MTRF1,KLH30) (Jiang, Michal et al. 2009, Chung, Lee et al. 2015, Dos Santos Silva, Fonseca et al. 2019). Perhaps the most interesting result, is that one of the genes in the Duroc analysis, TRIM63, has been associated as a biomarker for differences in response to exercise induced muscle damage(Baumert, G-REX Consortium et al. 2018), which ties into our comparison to human data below.

As the results were somewhat limited, we chose to continue with a different strategy in the joint breed analysis. Based on the results in figure 2, we see that p-values had an overall anti-conservative distribution for FE in the joint analysis, which showed us some promise for further analysis. We chose to calculate the DE using 3 methods, as we wanted to ensure that our results where robust and replicable, knowing that individual methods can vary in output (Seyednasrollah, Laiho et al. 2015). In figure 2 we can see the overview of the distribution of uncorrected p-values for FCR in all 3 methods, showing an anti-conservative distribution regardless of the method. If FCR was unrelated to gene expression in general, we would expect a uniform p-value distribution in our model. We can statistically confirm the likelihood of our observed p-values under the null hypothesis of no relation between expression and FCR using a Kolmogorov-Smirnov test, and in all 3 methods we reject the null hypothesis with (p-value < 10^-16). This leads us to conclude that there is a relation between the muscle tissue expression and FCR. In table 2 we can see the overview over the significance of our covariates in the 3 methods used for DE analysis. The most significant covariate is RIN, highlighting the importance of correcting for the RIN values when analyzing samples acquired in a non-laboratory setting. It has been previously shown that while RIN values do have an impact on expression values, explicitly controlling for this in a modelling framework should appropriately correct the data in most data points(Gallego Romero, Pai et al. 2014). Furthermore, we see that many genes are differentially expressed between the breeds, which is expected, and that age has an impact on expression. To quantify the observed link between expression and FE, we continue with two strategies – analyzing the overall pathway enrichments for the most significant genes and creating gene expression modules based on network analysis of our gene expression profiles.

Enrichments Analysis

The first step in an enrichment analysis is to select a suitable set of genes. The most general strategy is to pick genes that are differentially expressed after multiple testing correction for such a set. In our analysis, we do not have enough of these for a meaningful enrichment analysis, but we are able to demonstrate an overall relation between FCR and gene expression as seen above in figure 2. In our case, we could select genes with an uncorrected p-value below 0.05, but this is somewhat arbitrary distinction(Butler and Jones 2018). Instead, we chose to make an estimation of the number of additional low p-values in comparison to the uniformly distributed p-values, which represents the null hypothesis of no overall relation between FCR and gene expression. We call this value the divergent count. In essence, we are estimating the interval with the maximum positive divergence between our observed p-value frequencies and the same number of uniformly distributed p-values, assuming an approximately monotonely decreasing p-value distribution in our results. This has the advantage of not relying on arbitrary cutoffs but instead being a property of the overall p-value distribution. In figure 3, we can see a schematic representation of the divergent count. In Figure 4 we can see the a Venn diagram showing the overall divergent counts and overlaps for all three methods, with the full overlap set being the final gene set for enrichment analysis. We can see that a majority of the selected genes are identified by all three methods. This gives us confidence in the robustness of the selected set. To identify enriched functional pathways in our dataset, we chose to use is GOrilla(Eden, Navon et al. 2009). In GOrilla it is possible to give a background set to base the analysis on, making it advantageous for expression data, as it allows us only to use genes actually expressed in our data as a background. For the full output of the analysis, see supplementary table 2. Overall, 5 terms were significant post multiple corrections, with 4 out of these being related to mitochondrial ontologies In figure 5 we can see a summarized output of the significant post multiple testing correction GO-terms and groups based on the GOrilla analysis, using Revigo(Supek, Bosnjak et al. 2011). Based on this, the important overall pathway was translation elongation.

Gene Interaction Analysis

Many strategies can be used to take advantage of the interaction or co-expression between genes. We propose to apply modelling of pairwise gene interactions, which explicitly includes the phenotype of choice, which in our case is FE. This can be advantageous when dealing with complex phenotypes, as it may allow us to capture subtle biological variation. We chose to perform the gene interaction analysis based on the set of genes we identified from the divergent counts in our DE analysis. The visualization of the empirical divergent counts and the bootstrapped counts can be found in supplementary figure 2. Based on these results, the maximum bootstrapped divergent count was 83, and we observed 193 genes with a divergent count over 83. This means that many of the genes' p-value distributions are very anticonservative, and not very likely to happen by chance. There is however, the issue of data independence, as the genes' results are not independent from each other. Due to this, and general concern of data size and weak effects we used a conservative qualitative heuristic and focused on the top 20 genes based on our methodology. From the top 20 genes (see supplementary data 3 for the full results), the overall highlights were several transcription regulators: ETV1(an androgen receptor activate gene), LF1 (transcription factor) and KDM4C (transcription activator and growth related gene) (Bray and Kafatos 1991, Cai, Hsieh et al. 2007, Gregory and Cheung 2014); two mitochondrial genes, KMO and MRPS11(Meinke, Kerr et al. 2019),; two genes related to muscular atrophy - GEMIN7 and PLPP7 (Baccon, Pellizzoni et al. 2002, Meinke, Kerr et al. 2019); on gene implicated in heart development BIN1 (Nicot, Toussaint et al. 2007), two lipid metabolism/obesity related genes ACOT11 and GPD1 (ADAMS, CHUI et al. 2001) (Park, Berggren et al. 2006); and finally 3 genes associated with specific traits in pig IL2RG (Immune system in pigs)(Suzuki, Iwamoto et al. 2012), GGPS1 (meat quality) and PPARA (weak association with fat percentage) (Szczerbal, Lin et al. 2007). Interestingly, MRPS11 was also differentially expressed.

Gene Network Analysis

Based on our network analysis, we identified 19 distinct modules after correcting for RIN and merging the modules based on similarity. Based on the DE analysis, we decided not to focus individually on Landrace or Duroc pigs in the network analysis, and thus the network was generated combining both breeds. Looking at the the clustering in figure 6a,, initially one might think that the network is poorly constructed, as the module dendrogram representation is not very clear. In general, we see that some modules look closely clustered based on the dendrogram, such as the red module, while other are more diffuse. We should however realize that the modules themselves are based on N x N matrix, where n is >10.000. Thus, it is not easy to represent the modules properly in lower dimensions. Therefore, we rely on the module eigenvalue trait correlation and pathway analysis of the modules to asses if they are biologically meaningful. In figure 6b we can see the correlation between the eigenvalue of the modules and the traits and covariates we included in the DE analysis. We observe that the RIN correction of the individual genes has removed all the effect of the RIN on the eigenvalues of our modules. Several of the modules are well correlated with the breed and age, with correlation > 0.5, while FCR is mainly correlated with two modules, red and turquoise. The red and turquoise module include 391 and 3744 genes, respectively. Based on these results we performed GO-term analysis on the red and turquoise module. The red module is more correlated to breed and age than FCR, but we know that breed and FCR are correlated, and in our data, age is correlated with FCR (0.5). It should be noted that the age and FCR correlation is caused by the higher FCR pigs in our data exhibiting lower growth rates, thus needing more time to reach the tissue sampling as the slaughter takes place at a target weight of approximately 100 kg. The turquoise module shows highest correlations in FCR. In figure 7 we see the Revigo summary of the GOrilla GO term analysis performed based on the genes in the red (a) and turquoise (b) modules. In both the red and turquoise modules, a large number of GO terms where significantly overrepresented after multiple testing correction (see supplementary data 4 and 5 for the full list for red and turquoise respectively), indicating that the modules do represent specific biological pathways. In the red module, the most significant group of terms where related to mitochondria, which were grouped into three overall groups – translation elongation, electron transport chain and hydrogen ion transmembrane transport. This mirrors our finding from the DE analysis and the gene interaction analysis. As the module has a negative correlation with FCR, it indicates a relation between higher mitochondrial activity and lower FCR, thus higher efficiency. In the turquoise module, there was one large grouping of terms – DNA repair. This category included many GO terms, related to RNA, DNA, animo acid and nucleic acid metabolism and processing. These processes could be seen as generic growth and maintenance processes, and as the module is positively correlated with FCR, we can speculate the higher activity in DNA repair and related processes are increasing energy spend on maintenance, thus lowering efficiency. Due to the size of the module and the processes involved, it seems that the turquoise module is generically associated with overall cell maintenance and growth processes, giving it a somewhat unspecific functionality. In supplementary data 6 we find the top 10 most connected genes in the red and turquoise module. Interestingly, in the red module 7 out of 10 genes belong to the NADH ubiquinone oxidoreductase group (NDUF), with the remaining 3 also being implicated in mitochondrial function. Thus, the mitochondrial genes are both overrepresented in the red module and the most connected within the module. In the turquoise module, the results are unclear, as the most connected genes do not belong to any specific process, but instead cover a range of general processes that are generally important for cell function. This agrees with the general observation based on the size of the module and the overrepresented GO terms.

Human Exercise Data

To test they hypothesis that improvements in efficiency could be linked to a state mimicking exercise, we compared our divergent counts genes for FCR and the genes differentially expressed between breeds with 3 different human exercise datasets [33-35]. The results can be found in table 3. We are comparing if there is a higher proportion of genes that are significant for exercise-mediated changes in our two subsets, breed and FCR related genes, in relation to the non-differentially expressed genes. We see that in all cases there is a higher proportion of significant genes in the breed and FCR set versus the background set, as the odds ratio between the subsets and the background is always below 1. In general, the breed results are more significant than the FCR genes, but they show similar ratios. This is likely because there are roughly 4 times more breed genes, yielding higher statistical power. Given the overall results, it does seem like both FCR and breed related genes are slightly more significant than background for exercise related changes. We also did the pathway enrichment analysis for the genes that where significant in both one of our three human data sets and in the breed, and FCR set respectively. The overall results are found in figure 7a (breed) and 7b(FCR). In the breed, we find that main categories are cellular metal ion homeostasis and anatomical structure development, based on 702 genes. For FCR, only 42 genes overlap with the human significant genes, meaning the results of the enrichment are not as significant, but the main overall group is regulation of transcription from RNA polymerase II promoter.

Discussion

There have been 4 previous studies analyzing the muscle transcriptome in an FE context (Jing, Hou et al. 2015, Vincent, Louveau et al. 2015, Gondret, Vincent et al. 2017, Horodyska, Wimmers et al. 2018). The study by Gondret et al [18] was based on selecting divergent FE lines of Large White pigs for 8 generations, used 24 samples and was based on microarray. They reported a high number of differentially expressed genes in muscle between the low and high RFI groups (2417), but it is not clear from their paper how many probes were included in the statistical analysis and how this may affect multiple testing correction. They also reported that a gene was considered differentially expressed if one probe met the cutoff regardless of multiple probes did not. They reported that mitochondrial electron chain transport, glucose metabolic process and generation of precursor metabolites and energy as significant pathways for RFI.

In the study from Horodyska et al [17], they used 16 pigs, but included 8 pigs of each gender. They used an uncorrected p-value of 0.01 as their threshold,, with no consideration weather this is appropriate given their overall data distribution. They report 272 genes with p-value < 0.01, which is similar to ours of 243, however we have included less genes in our analysis (14497 vs 10563). Overall, we cannot assess their results as very significant.

In Vincent et al [20], they had 16 female Large Whites from divergent RFI lines, their study was microarray based, but they reported their results based on uncorrected p-values in both expression and proteomics. They do however report finding mitochondrial related probes being significant.

Finally, in Jing et al [19], they had a total sample size of only 6 Yorkshire pigs, based on the selection of the most extreme RFI pigs in a set of 236. They reported 645 DE genes, with 99 with FDR lower than 0.05. However, selecting such few samples at the extreme end of FE does raise the question of replication, as the large differences in RFI/FCR they achieved could easily be caused by factors that are not generally applicable. They found that the most significant pathways in their data were mitochondrial activity, glycolysis and myogenesis pathways. Despite the issues presented with the studies, it is notable that mitochondria are reported to be related to FE multiple times.

In our study, we have the highest number of samples reported (41) and we include two breeds, which do not have directly divergent selection for FCR, but with one of the breeds more positively selected for FCR. Having this setup does present advantages and disadvantages. The advantage in relation to the other studies it that the results may generalize better across breeds. The disadvantage is that we may be fitting breed effects instead of phenotypic effects, but we do account for breed in all our analysis. The other main difference is that we have selected pigs with a range of FCR values, and fit FCR as a continuous value. In general fitting a continuous value is more informative, and the fact that we have a range of pigs that are not divergently selected, may make the results more applicable to a real life setting. In pig production there is no low FE selected line to contrast with, so the biological background of FE in a normal breeding population may be more relevant and interesting.

Another general issue is how to deal with statistical issues in analysis of FE. From the various studies presented above it is clear that FE is a somewhat subtle phenotype in muscle tissue, and thus a lot of data is needed make conclusions. Here we try to tackle this issue by not being overly conservative, but still applying multiple testing correction by using and FDR of 0.1 level for individual results in our DE analysis. Furthermore, we generally try to analyze our data by either taking the overall distribution of results and/or combing genes in groups, to avoid relying on individual weak results.

Differential Expression analysis and Pathway Enrichment

We have analyzed the transcriptomic differences and molecular pathways involved in differences in FCR in two different breeds. Based on DE, we identified 14 genes with an FDR value below 0.1. The highlights here were the finding of mitochondrial genes, and TRIM64, which related to exercise induced muscle damage.

Due to the limited results in the DE analysis, we chose to use a novel approach to perform a pathway enrichment analysis. In practice, we wanted to broaden the number of genes for the pathway analysis, but at the same time also select a robust and meaningful set of genes. To make the analysis more robust, we choose to base the pathway analysis on results from 3 DE expression methods. Furthermore, we elected to select genes based on the overall divergence from the null hypothesis of our p-value distribution, as this should represent a set of genes that is likely to be associated with our trait, even the genes are not significant based on individual FDR corrected p-values . To our knowledge, this is a novel way of selecting a group of genes, which we called the divergent count. Looking at the enriched pathways in our dataset selected based on the divergent counts, we find results that are common in the literature in several species beyond the pig studies already mentioned(Connor, Kahl et al. 2010, Bottje, Lassiter et al. 2017), namely differences in mitochondrial pathways related to FE, summarized as mitochondrial translation elongation in our Revigo summary. While this is not a novel result, we did find it in a novel setting, with larger sample size, novel population selection and using a continuous value for FCR. This acts as further evidence to the link of mitochondrial activity and FE, but also as evidence that it may be relevant in real breeding populations, and not only in divergently selected test populations.

Gene Expression Interaction

Our gene expression interaction analysis is a novel way of finding the most important genes, which has not been applied to FE in pigs before. Based on the qualitative analysis of the top 20 genes, the results seem promising. We found several transcription factors, including the most divergent gene (ELF1), which makes sense in regards to gene interaction. The remaining genes also seemed promising, as they included categories one can expect to be related to muscle growth and FCR, such as lipid metabolism and muscle atrophy. Confirming previous results, we also identified two mitochondrial genes among the top 20.

Gene Network Analysis

Our gene network analysis revealed two modules with a correlation > 0.4 with FCR. Based on the GO term analysis enrichment of the red module, we find many enriched GO terms related to mitochondrial processes, confirming our finding in the other analysis, and from other studies. More specifically, the negative correlation between the red module eigenvalue and FCR also shows that higher mitochondrial activity is positively associated with higher efficiency. Based on the top ten hub genes in the red module we confirm this picture, as all ten genes are related to mitochondria, and seven of them are from the NDUF family, which was also found in the gene expression interaction analysis. The turquoise module was the most correlated module(0.49), and furthermore, it was more correlated to FCR than to our other traits. Based on the GO term analysis, we found that the cluster was highly enriched for genes related to DNA repair, which included GO terms relate to RNA, DNA, animo acid and nucleic acid metabolism and processing. To the best of our knowledge, this is the first evidence of these processes being related to FE in general. The only previous link to DNA repair in livestock was a feed restriction study of cattle(Connor, Kahl et al. 2010). The top ten hub genes of this module did not show a clear picture, with the genes involved in a wide range of processes related to general cell maintenance. This indicates that the turquoise module represents general housekeeping functions, rather than very specific pathways. As the module eigenvalue was positively correlated with FCR, we can speculate that more active DNA repair and maintenance processes represent higher maintenance costs, thus reducing efficiency.

Human Exercise

We have established earlier that the gene expression and molecular background of FE is still somewhat elusive. To try and identify what overall mechanisms could be at play, we hypothesized that differences between our two breeds, which have different overall FE, and genes related to FCR, are more likely to be important for processes involved in exercise. The reason we had this hypothesis, is that the pigs are selected for lean growth, and it is possible that this growth stimulus is similar to the effects induced in muscle by exercise. We found a slight confirmation of this hypothesis, as we found similar favorable odds ratio for our hypothesis in all 3 datasets we tested for both FCR and our breed genes. Our pathway enrichment analysis for FCR did not yield any very significant results, as it was only based on 42 genes. The main overall category identified, based on 4 go terms, was regulation of transcription from RNA polymerase II (pol II) promoters. Interestingly, Actin has been associated with the pre-initiation complex necessary for transcription by RNA polymerase II(Hofmann, Stojiljkovic et al. 2004), which could be relevant given the importance of actin in muscle tissue(Tang 2015). There are also links between a poll II subunit and myogenesis (CORBI, PADOVA et al. 2002). Although these results may be relevant, our data here is too weak for solid conclusions.

In regards to the genes overlapping between exercise and breed differences, the results are more statistically robust, as they are based on an overall larger gene set of 702 genes. Here we find two overall groups – cellular metal ion homeostasis and anatomical structure development. For the first term, we know that the transport of ions is generically vital to muscle function (Wolitzky and Fambrough 1986, Mohr, Krustrup et al. 2007). The second overall term, anatomical structure development, is very generic in terms of function, and includes sub-categories that are related to muscle development, such as muscle structure development.

Overall, the results from the Human data analysis represent a novel hypothesis, but requires more analysis and new experiments on pigs to strengthen the link between FE and exercise. One interesting aspect of this analysis is that in theory pigs could be used as a model for lean growth in sedentary conditions, which in the long run could yield interesting therapeutic possibilities applicable to humans.

Conclusion

We have analyzed the muscle transcriptome from Duroc and Landrace,   twp of the main purebred breeding pigs in Denmark. In contrast to previous studies, we did not use any lines divergently selected for FE, and we included a wider range of FE values, which were modelled as a continuous trait, using the highest number of pigs in a study of this type. We identified several individual genes based on DE analysis and gene-gene interaction analysis that are involved in FCR, with many of them having relevant functional backgrounds from previous studies. We applied a novel strategy to select genes for pathway enrichment, the divergent count. Based on enrichment analysis, gene-gene interaction, network analysis and DE we found several interesting candidate biomarkers genes and pathways. We reinforced the knowledge that mitochondrial activity is important FCR, but using a non-divergently FE selected pig population. Based on the findings, we postulate that mitochondrial genes, and in particular genes from NDUF group or MRPS11 could be used as potential biomarkers for FCR in pigs. Furthermore, all our top genes from our interaction analysis also show promise as potential FCR biomarkers. Finally, we find that there is a putative link between genes involved in exercise related changes in human, and FE in pigs

Conflict of Interest

There were no conflicts of interest.

Ethics

As animals were only sampled post conventional slaughter, no ethics approval was needed for the study.

Author Contributions

HNK conceived and designed the project and obtained funding as the main applicant. VAOC and HNK designed the muscle sampling experiments, phenotype data collection and statistical/bioinformatics analyses. VOAC performed the sampling, data processing, data visualization and bioinformatics and statistical analysis. All authors collaborated in the interpretation of results, discussion and write up of the manuscript. All authors have read, reviewed and approved the final manuscript.

Funding

This study was funded by the Independent Research Fund Denmark (DFF) – Technology and Production (FTP) grant (grant number: 4184-00268B). VAOC received partial Ph.D. stipends from the University of Copenhagen and Technical University of Denmark. The Authors thank the SEGES – Pig Research Centre (VSP) Denmark and Danish Crown for access to samples and phenotype datasets.

Acknowledgments

We would like to thank all group members of the Quantitative Genomics, Bioinformatics and Computational Biology Group for participating in discussion when needed and creating a fruitful scientific environment.

References

Muscle as a Secretory Organ. Comprehensive Physiology**:** 1337-1362.

ADAMS, S. H., C. CHUI, S. L. SCHILBACH, X. X. YU, A. D. GODDARD, J. C. GRIMALDI, J. LEE, P. DOWD, S. COLMAN and D. A. LEWIN (2001). "BFIT, a unique acyl-CoA thioesterase induced in thermogenic brown adipose tissue: cloning, organization of the human gene and assessment of a potential link to obesity." Biochemical Journal **360**(1): 135-142.

Baccon, J., L. Pellizzoni, J. Rappsilber, M. Mann and G. Dreyfuss (2002). "Identification and Characterization of Gemin7, a Novel Component of the Survival of Motor Neuron Complex." Journal of Biological Chemistry **277**(35): 31957-31962.

Baumert, P., G-REX Consortium, M. J. Lake, B. Drust, C. E. Stewart and R. M. Erskine (2018). "TRIM63 (MuRF-1) gene polymorphism is associated with biomarkers of exercise-induced muscle damage." Physiological Genomics **50**(3): 142-143.

Bolger, A. M., M. Lohse and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics **30**(15): 2114-2120.

Bottje, W. G., K. Lassiter, A. Piekarski-Welsher, S. Dridi, A. Reverter, N. J. Hudson and B. W. Kong (2017). "Proteogenomics Reveals Enriched Ribosome Assembly and Protein Translation in Pectoralis major of High Feed Efficiency Pedigree Broiler Males." Front Physiol **8**: 306.

Bray, S. J. and F. C. Kafatos (1991). "Developmental function of Elf-1: an essential transcription factor during embryogenesis in Drosophila." Genes & Development **5**(9): 1672-1683.

Butler, J. and P. Jones (2018). "Theoretical and empirical distributions of the p value." Metron-International Journal of Statistics **76**(1): 1-30.

Cai, C., C.-L. Hsieh, J. Omwancha, Z. Zheng, S.-Y. Chen, J.-L. Baert and L. Shemshedini (2007). "ETV1 Is a Novel Androgen Receptor-Regulated Gene that Mediates Prostate Cancer Cell Invasion." Molecular Endocrinology **21**(8): 1835-1846.

Chung, H. Y., K. T. Lee, G. W. Jang, J. G. Choi, J. G. Hong and T. H. Kim (2015). "A genome-wide analysis of the ultimate pH in swine." Genet Mol Res **14**(4): 15668-15682.

Connor, E. E., S. Kahl, T. H. Elsasser, J. S. Parker, R. W. Li, C. P. Van Tassell, R. L. Baldwin and S. M. Barao (2010). "Enhanced mitochondrial complex gene function and reduced liver size may mediate improved feed efficiency of beef cattle during compensatory growth." Functional & Integrative Genomics **10**(1): 39-51.

CORBI, N., M. D. PADOVA, R. D. ANGELIS, T. BRUNO, V. LIBRI, S. IEZZI, A. FLORIDI, M. FANCIULLI and C. PASSANANTI (2002). "The α-like RNA polymerase II core subunit 3 (RPB3) is involved in tissue-specific transcription and muscle differentiation via interaction with the myogenic factor Myogenin." The FASEB Journal **16**(12): 1639-1641.

Czernichow, S., D. Thomas and E. Bruckert (2010). "n-6 Fatty acids and cardiovascular health: a review of the evidence for dietary intake recommendations." Br J Nutr **104**(6): 788-796.

Devarshi, P. P., A. D. Jones, E. M. Taylor and T. M. Henagan (2018). "Effects of Acute Aerobic Exercise on Transcriptomics in Skeletal Muscle of Lean vs Overweight/Obese Men." The FASEB Journal **32**(1_supplement): lb248-lb248.

Ding, R., M. Yang, X. Wang, J. Quan, Z. Zhuang, S. Zhou, S. Li, Z. Xu, E. Zheng, G. Cai, D. Liu, W. Huang, J. Yang and Z. Wu (2018). "Genetic Architecture of Feeding Behavior and Feed Efficiency in a Duroc Pig Population." Front Genet **9**: 220.

Do, D. N., A. B. Strathe, J. Jensen, T. Mark and H. N. Kadarmideen (2013). "Genetic parameters for different measures of feed efficiency and related traits in boars of three pig breeds." J Anim Sci **91**(9): 4069-4079.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics **29**(1): 15-21.

Dos Santos Silva, D. B., L. F. S. Fonseca, D. G. Pinheiro, M. M. M. Muniz, A. F. B. Magalhaes, F. Baldi, J. A. Ferro, L. A. L. Chardulo and L. G. de Albuquerque (2019). "Prediction of hub genes associated with intramuscular fat content in Nelore cattle." BMC Genomics **20**(1): 520.

Eden, E., D. Lipson, S. Yogev and Z. Yakhini (2007). "Discovering motifs in ranked lists of DNA sequences." PLoS Comput Biol **3**(3): e39.

Eden, E., R. Navon, I. Steinfeld, D. Lipson and Z. Yakhini (2009). "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists." BMC Bioinformatics **10**: 48.

Faure, J., L. Lefaucheur, N. Bonhomme, P. Ecolan, K. Meteau, S. M. Coustard, M. Kouba, H. Gilbert and B. Lebret (2013). "Consequences of divergent selection for residual feed intake in pigs on muscle energy metabolism and meat quality." Meat Sci **93**(1): 37-45.

Gallego Romero, I., A. A. Pai, J. Tung and Y. Gilad (2014). "RNA-seq: impact of RNA degradation on transcript quantification." BMC Biol **12**: 42.

Gilbert, H., Y. Billon, L. Brossard, J. Faure, P. Gatellier, F. Gondret, E. Labussiere, B. Lebret, L. Lefaucheur, N. Le Floch, I. Louveau, E. Merlot, M. C. Meunier-Salaun, L. Montagne, P. Mormede, D. Renaudeau, J. Riquet, C. Rogel-Gaillard, J. van Milgen, A. Vincent and J. Noblet (2017). "Review: divergent selection for residual feed intake in the growing pig." Animal **11**(9): 1427-1439.

Gondret, F., A. Vincent, M. Houee-Bigot, A. Siegel, S. Lagarrigue, D. Causeur, H. Gilbert and I. Louveau (2017). "A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs." BMC Genomics **18**(1): 244.

Gregory, B. L. and V. G. Cheung (2014). "Natural variation in the histone demethylase, KDM4C, influences expression levels of specific genes including those that affect cell growth." Genome Research **24**(1): 52-63.

Hofmann, W. A., L. Stojiljkovic, B. Fuchsova, G. M. Vargas, E. Mavrommatis, V. Philimonenko, K. Kysela, J. A. Goodrich, J. L. Lessard, T. J. Hope, P. Hozak and P. de Lanerolle (2004). "Actin is part of pre-initiation complexes and is necessary for transcription by RNA polymerase II." Nature Cell Biology **6**(11): 1094-1101.

Hoque, M. A., H. Kadowaki, T. Shibata, T. Oikawa and K. Suzuki (2009). "Genetic parameters for measures of residual feed intake and growth traits in seven generations of Duroc pigs." Livestock Science **121**(1): 45-49.

Horodyska, J., M. Oster, H. Reyer, A. M. Mullen, P. G. Lawlor, K. Wimmers and R. M. Hamill (2018). "Analysis of meat quality traits and gene expression profiling of pigs divergent in residual feed intake." Meat Sci **137**: 265-274.

Horodyska, J., K. Wimmers, H. Reyer, N. Trakooljul, A. M. Mullen, P. G. Lawlor and R. M. Hamill (2018). "RNA-seq of muscle from pigs divergent in feed efficiency and product quality identifies differences in immune response, growth, and macronutrient and connective tissue metabolism." BMC Genomics **19**(1): 791.

Jiang, Z., J. J. Michal, J. Chen, T. F. Daniels, T. Kunej, M. D. Garcia, C. T. Gaskins, J. R. Busboom, L. J. Alexander, R. W. Wright, Jr. and M. D. Macneil (2009). "Discovery of novel genetic networks associated with 19 economically important traits in beef cattle." Int J Biol Sci **5**(6): 528-542.

Jing, L., Y. Hou, H. Wu, Y. Miao, X. Li, J. Cao, J. M. Brameld, T. Parr and S. Zhao (2015). "Transcriptome analysis of mRNA and miRNA in skeletal muscle indicates an important network for differential Residual Feed Intake in pigs." Sci Rep **5**: 11953.

Jing, L., Y. Hou, H. Wu, Y. X. Miao, X. Y. Li, J. H. Cao, J. M. Brameld, T. Parr and S. H. Zhao (2015). "Transcriptome analysis of mRNA and miRNA in skeletal muscle indicates an important network for differential Residual Feed Intake in pigs." Scientific Reports **5**.

Koch, R. M., Swiger, L. A., Chambers, D. J. & Gregory K. E (1963). "Efficiency of feed use in beef cattle." Journal of Animal Science **22**(2): 486-494.

Langfelder, P. and S. Horvath (2008). "WGCNA: an R package for weighted correlation network analysis." BMC Bioinformatics **9**: 559.

Lefaucheur, L., B. Lebret, P. Ecolan, I. Louveau, M. Damon, A. Prunier, Y. Billon, P. Sellier and H. Gilbert (2011). "Muscle characteristics and meat quality traits are affected by divergent selection on residual feed intake in pigs." J Anim Sci **89**(4): 996-1010.

Love, M. I., W. Huber and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biol **15**(12): 550.

Meinke, P., A. R. W. Kerr, R. Czapiewski, J. I. de las Heras, C. R. Dixon, E. Harris, H. Kölbel, F. Muntoni, U. Schara, V. Straub, B. Schoser, M. Wehnert and E. C. Schirmer (2019). "A multistage sequencing strategy pinpoints novel candidate alleles for Emery-Dreifuss muscular dystrophy and supports gene misregulation as its pathomechanism." EBioMedicine.

Mohr, M., P. Krustrup, J. J. Nielsen, L. Nybo, M. K. Rasmussen, C. Juel and J. Bangsbo (2007). "Effect of two different intense training regimens on skeletal muscle ion transport proteins and fatigue development." American Journal of Physiology-Regulatory, Integrative and Comparative Physiology **292**(4): R1594-R1602.

Morales, P. E., J. L. Bucarey and A. Espinosa (2017). "Muscle Lipid Metabolism: Role of Lipid Droplets and Perilipins." Journal of Diabetes Research **2017**: 10.

Murton, A. J., R. Billeter, F. B. Stephens, S. G. D. Etages, F. Graber, R. J. Hill, K. Marimuthu and P. L. Greenhaff (2014). "Transient transcriptional events in human skeletal muscle at the outset of concentric resistance exercise training." Journal of Applied Physiology **116**(1): 113-125.

Nicot, A.-S., A. Toussaint, V. Tosch, C. Kretz, C. Wallgren-Pettersson, E. Iwarsson, H. Kingston, J.-M. Garnier, V. Biancalana, A. Oldfors, J.-L. Mandel and J. Laporte (2007). "Mutations in amphiphysin 2 (BIN1) disrupt interaction with dynamin 2 and cause autosomal recessive centronuclear myopathy." Nature Genetics **39**(9): 1134-1139.

Nkrumah, J. D., J. A. Basarab, Z. Wang, C. Li, M. A. Price, E. K. Okine, D. H. Crews, Jr. and S. S. Moore (2007). "Genetic and phenotypic relationships of feed intake and measures of efficiency with growth and carcass merit of beef cattle1." Journal of Animal Science **85**(10): 2711-2720.

Ohtsubo, K., S. Takamatsu, M. T. Minowa, A. Yoshida, M. Takeuchi and J. D. Marth (2005). "Dietary and genetic control of glucose transporter 2 glycosylation promotes insulin secretion in suppressing diabetes." Cell **123**(7): 1307-1321.

Park, J.-J., J. R. Berggren, M. W. Hulver, J. A. Houmard and E. P. Hoffman (2006). "GRB14, GPD1, and GDF8 as potential network collaborators in weight loss-induced improvements in insulin action in human skeletal muscle." Physiological Genomics **27**(2): 114-121.

Popov, D. V., P. A. Makhnovskii, E. I. Shagimardanova, G. R. Gazizova, E. A. Lysenko, O. A. Gusev and O. L. Vinogradova (2019). "Contractile activity-specific transcriptome response to acute endurance exercise and training in human skeletal muscle." Am J Physiol Endocrinol Metab **316**(4): E605-E614.

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." Nucleic Acids Res **43**(7): e47.

Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.

Seyednasrollah, F., A. Laiho and L. L. Elo (2015). "Comparison of software packages for detecting differential expression in RNA-seq studies." Brief Bioinform **16**(1): 59-70.

Smith, R. M., N. K. Gabler, J. M. Young, W. Cai, N. J. Boddicker, M. J. Anderson, E. Huff-Lonergan, J. C. Dekkers and S. M. Lonergan (2011). "Effects of selection for decreased residual feed intake on composition and quality of fresh pork." J Anim Sci **89**(1): 192-200.

Supek, F., M. Bosnjak, N. Skunca and T. Smuc (2011). "REVIGO summarizes and visualizes long lists of gene ontology terms." PLoS One **6**(7): e21800.

Suzuki, S., M. Iwamoto, Y. Saito, D. Fuchimoto, S. Sembon, M. Suzuki, S. Mikawa, M. Hashimoto, Y. Aoki, Y. Najima, S. Takagi, N. Suzuki, E. Suzuki, M. Kubo, J. Mimuro, Y. Kashiwakura, S. Madoiwa, Y. Sakata, A. C. F. Perry, F. Ishikawa and A. Onishi (2012). "Il2rg Gene-Targeted Severe Combined Immunodeficiency Pigs." Cell Stem Cell **10**(6): 753-758.

Szczerbal, I., L. Lin, M. Stachowiak, A. Chmurzynska, M. Mackowski, A. Winter, K. Flisikowski, R. Fries and M. Switonski (2007). "Cytogenetic mapping ofDGAT1, PPARA, ADIPOR1 andCREB genes in the pig." Journal of Applied Genetics **48**(1): 73-76.

Tang, D. D. (2015). "Critical role of actin-associated proteins in smooth muscle contraction, cell proliferation, airway hyperresponsiveness and airway remodeling." Respiratory Research **16**(1): 134.

Turner, N., G. J. Cooney, E. W. Kraegen and C. R. Bruce (2014). "Fatty acid metabolism, energy expenditure and insulin resistance in muscle." J Endocrinol **220**(2): T61-79.

Vincent, A., I. Louveau, F. Gondret, C. Trefeu, H. Gilbert and L. Lefaucheur (2015). "Divergent selection for residual feed intake affects the transcriptomic and proteomic profiles of pig skeletal muscle." J Anim Sci **93**(6): 2745-2758.

Wolitzky, B. A. and D. M. Fambrough (1986). "Regulation of the (Na+ + K+)-ATPase in cultured chick skeletal muscle. Modulation of expression by the demand for ion transport." J Biol Chem **261**(21): 9990-9999.

Yi, Z., X. Li, W. Luo, Z. Xu, C. Ji, Y. Zhang, Q. Nie, D. Zhang and X. Zhang (2018). "Feed conversion ratio, residual feed intake and cholecystokinin type A receptor gene polymorphisms are associated with feed intake and average daily gain in a Chinese local chicken population." J Anim Sci Biotechnol **9**: 50.

Zhang, B. and S. Horvath (2005). "A general framework for weighted gene co-expression network analysis." Stat Appl Genet Mol Biol **4**: Article17.

1 Data Availability Statement

The data will be uploaded to GEO and released if the article is accepted

Gene Name	Breed	FDR	Regulation
PNCK	Landrace	0.0007	Down
Patr-A	Landrace	0.08	Down
MTMR11	Duroc	0.07	Up
C3	Duroc	0.02	Down
LCP1	Duroc	0.02	Up
TRIM63	Duroc	0.08	Down
KLHL30	Duroc	0.07	Down
NANOS1	Duroc	0.08	Up
IGHM	Duroc	0.07	Up
ETV5	Duroc	0.02	Down
MTFR1	Both	0.068	Down
MGAT4A	Both	0.098	Down
SLC38A2	Both	0.098	Up
MRPS11	Both	0.067	Up

Table 1 – Overview of genes with a FDR value < 0.1 in all 3 differential expression analysis. There is only a limited amount of genes differentially expressed at 0.1 FDR level for FE. Notably, out of 4 genes in the common breed analysis there are two genes with mitochondrial related Gene Ontologies - MRPS11, MTRM1. MTFR1 has been implicated in eating quality (measures of meat quality post cooking) in cattle(Jiang, Michal et al. 2009) and as a meat PH QTL in pig(Chung, Lee et al. 2015). Also interesting to note that TRIM63 has been suggested as a biomarker for difference in response to exercise-induced muscle damage(Baumert, G-REX Consortium et al. 2018), KLHL30 has been associated with intramuscular fat and muscle metabolism in Nelore Cattle(Dos Santos Silva, Fonseca et al. 2019). MGAT4A has been linked to diabetes and glucose transport (Ohtsubo, Takamatsu et al. 2005).

Trait	***EdgeR***	***Limma***	***Deseq2***

FCR	*4*	*0*	*0*
Breed	*3633*	*3679*	*3428*
RIN	*5572*	*5763*	*5779*
Age	*503*	*189*	*328*

Table 2 – Over view over the number of genes with FDR < 0.1 in the common breed analysis for all 3 methods and each covariate. In general, we have modest amount of DE genes for FE, while our other covariates have a amny significant genes associated with them.

Data	**P-value Breed**	**Odds ratio Breed**	**P-value FCR**	**Odds ratio FCR**
Dataset 1	0.0017	0.79	0,0046	0.71
Dataset 2	0.0012	0.85	0.22	0.9
Dataset 3	0.12	0.84	0.47	0.88

Table 3 – Results of Fisher exact test comparing the number of genes significant for difference in rested and exercised muscle in divergent count genes for genes found in the divergent count for FCR

and breed and the background for each of the 3 human data sets(dataset 1 (Devarshi, Jones et al. 2018),dataset 2 (Murton, Billeter et al. 2014) and dataset 3 (Popov, Makhnovskii et al. 2019)).

Figures

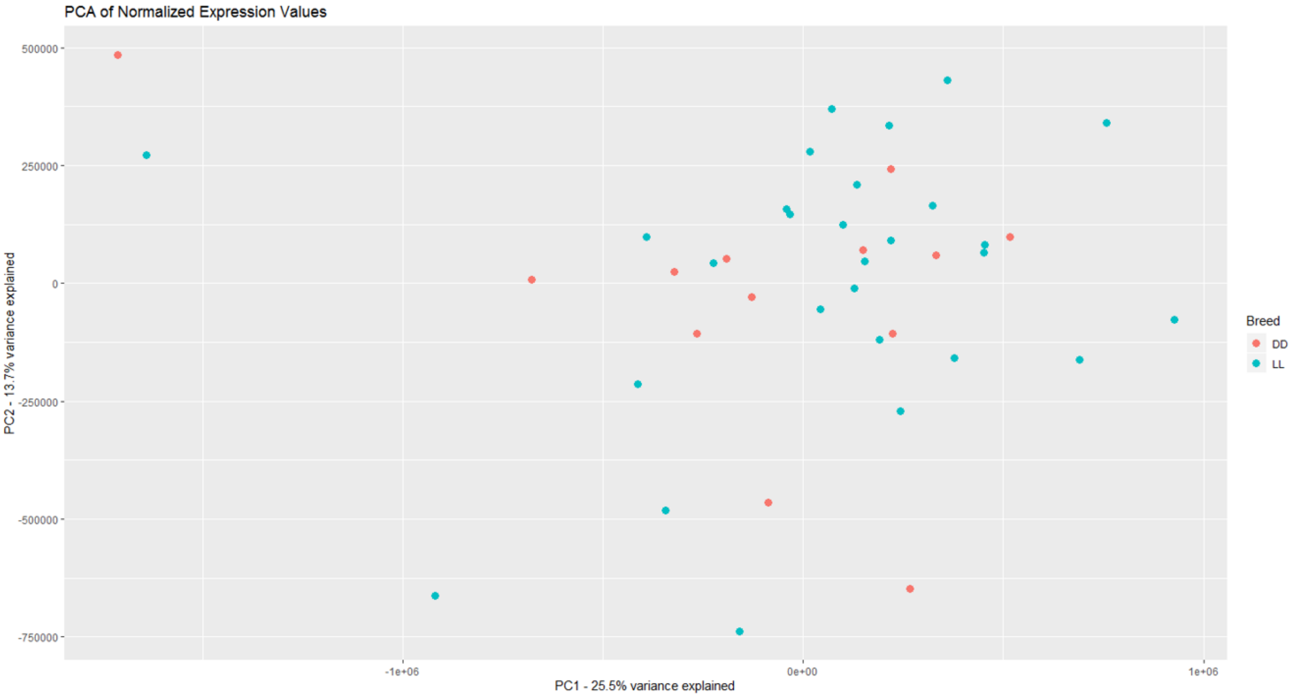
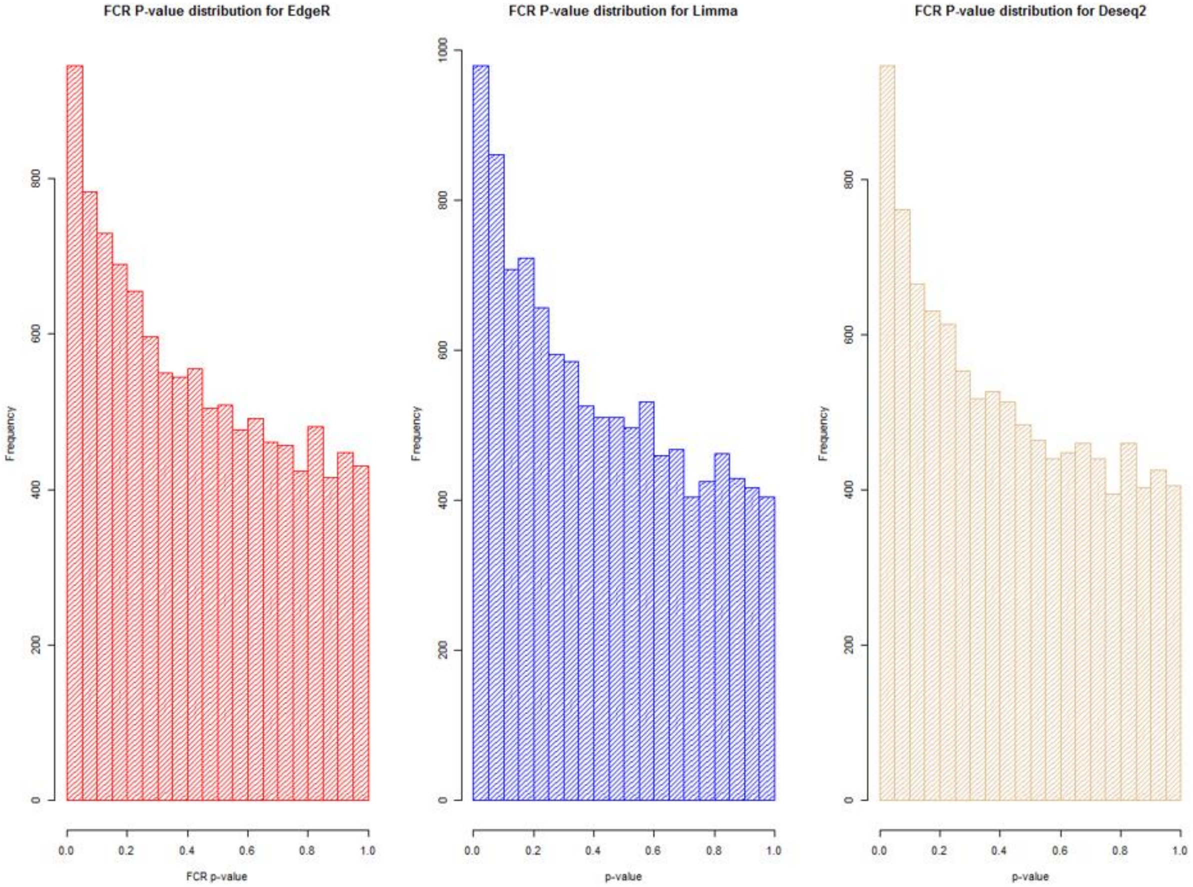


Figure 1 Visualization of the two first principle components in the expression data, with DD being Duroc and LL being Landrace.There is not a clear separation between breeds based on the overall expression, giving credence to a joint breed analysis of the data.


FCR P-value distribution for EdgeR
Frequency
0
200
400
600
800
0.0
0.2
0.4
0.6
0.8
1.0
FCR p-value
FCR P-value distribution for Limma
Frequency
0
200
400
600
800
1000
0.0
0.2
0.4
0.6
0.8
1.0
p-value
FCR P-value distribution for Deseq2
Frequency
0
200
400
600
800
0.0
0.2
0.4
0.6
0.8
1.0
p-value

789 Figure 2 Visualization of the distribution of the p-values testing the relation between FCR and gene expression for all three analysis methods. It is clear in all cases that we observe an anti-conservative distribution, that is, there is an overweight of low p-values.

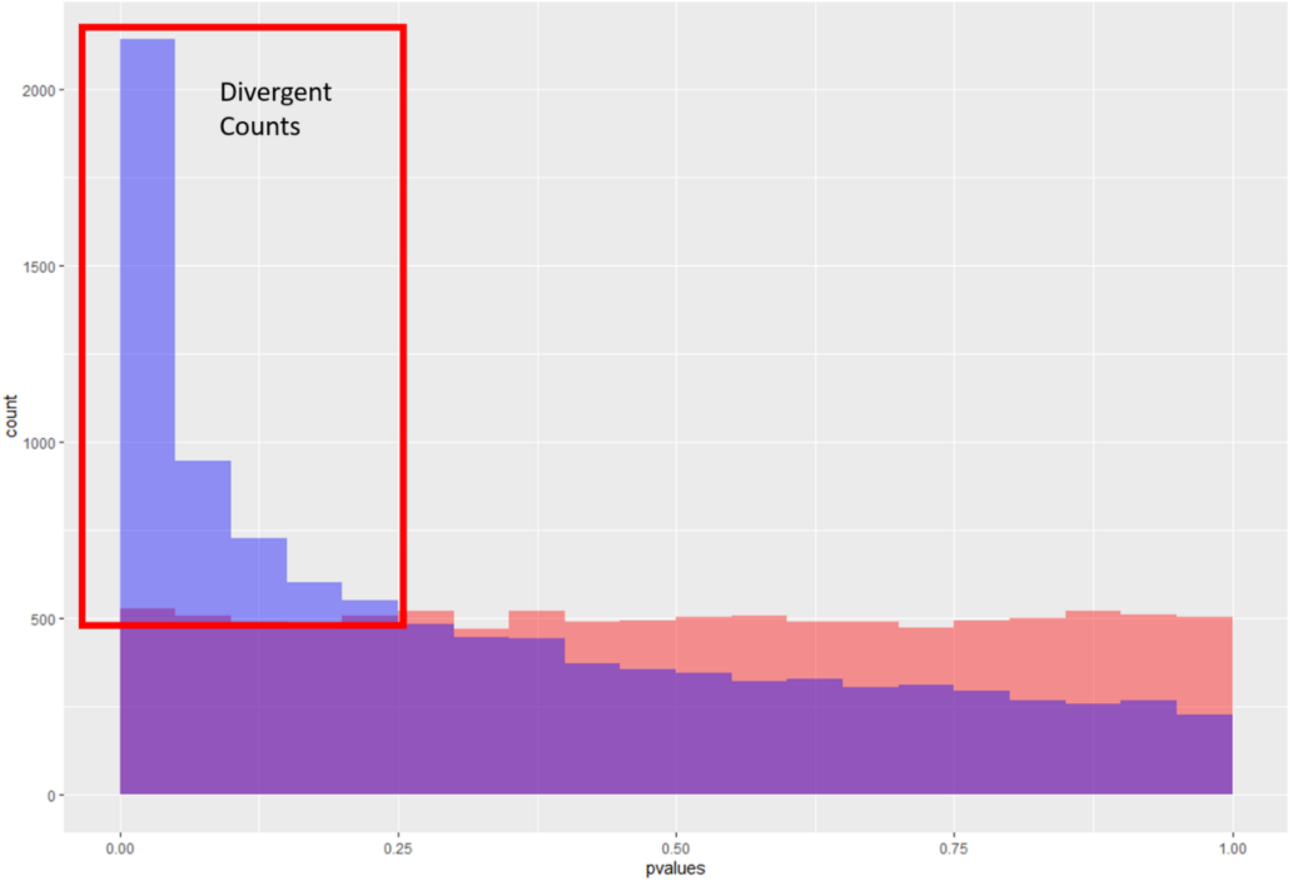


Figure 3 Schematic representation of the divergent counts. Here we see to theoretical p-value distributions, one which is uniform (in red) and one which is anti-conservative (blue). The purple area is where they overlap, and the blue area is the area used to estimate the divergent counts.

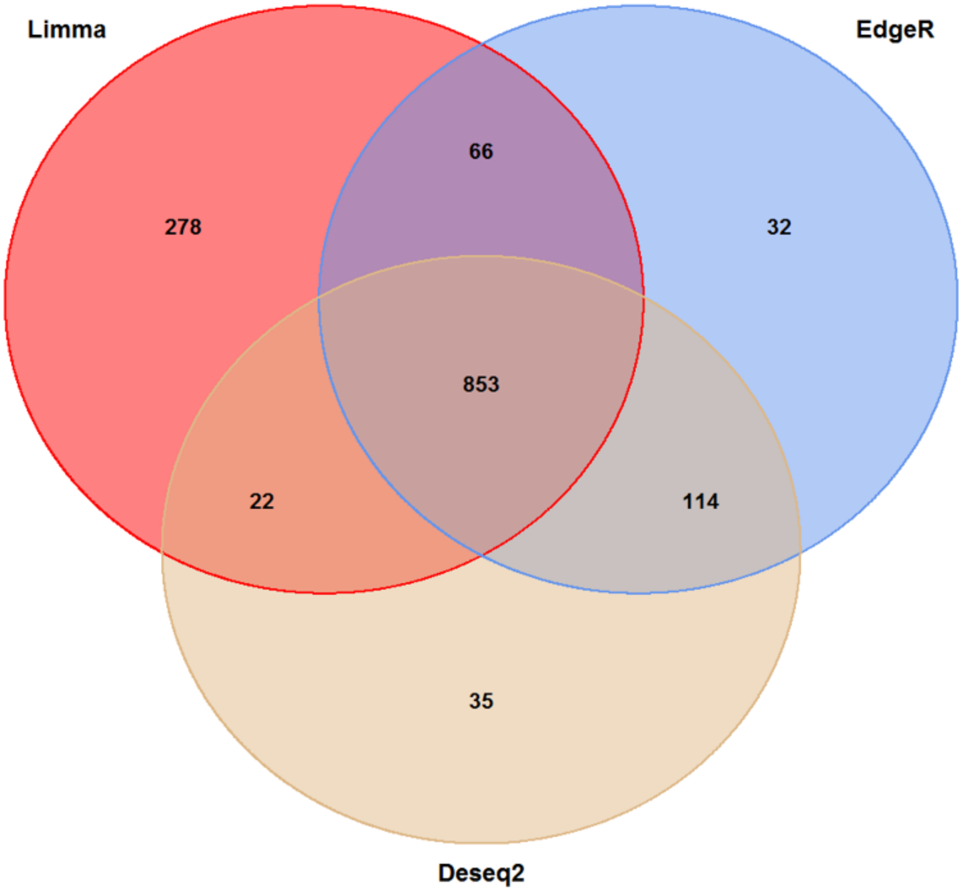


Figure 4 Venn diagram of the overlap in the divergent counts between the three methods. We see here that the Limma is overall less conservative than the two other methods, but in general, the

methods are in high agreement with each other. The final set of genes selected for the enrichment analysis was the 853 triple overlapping set.

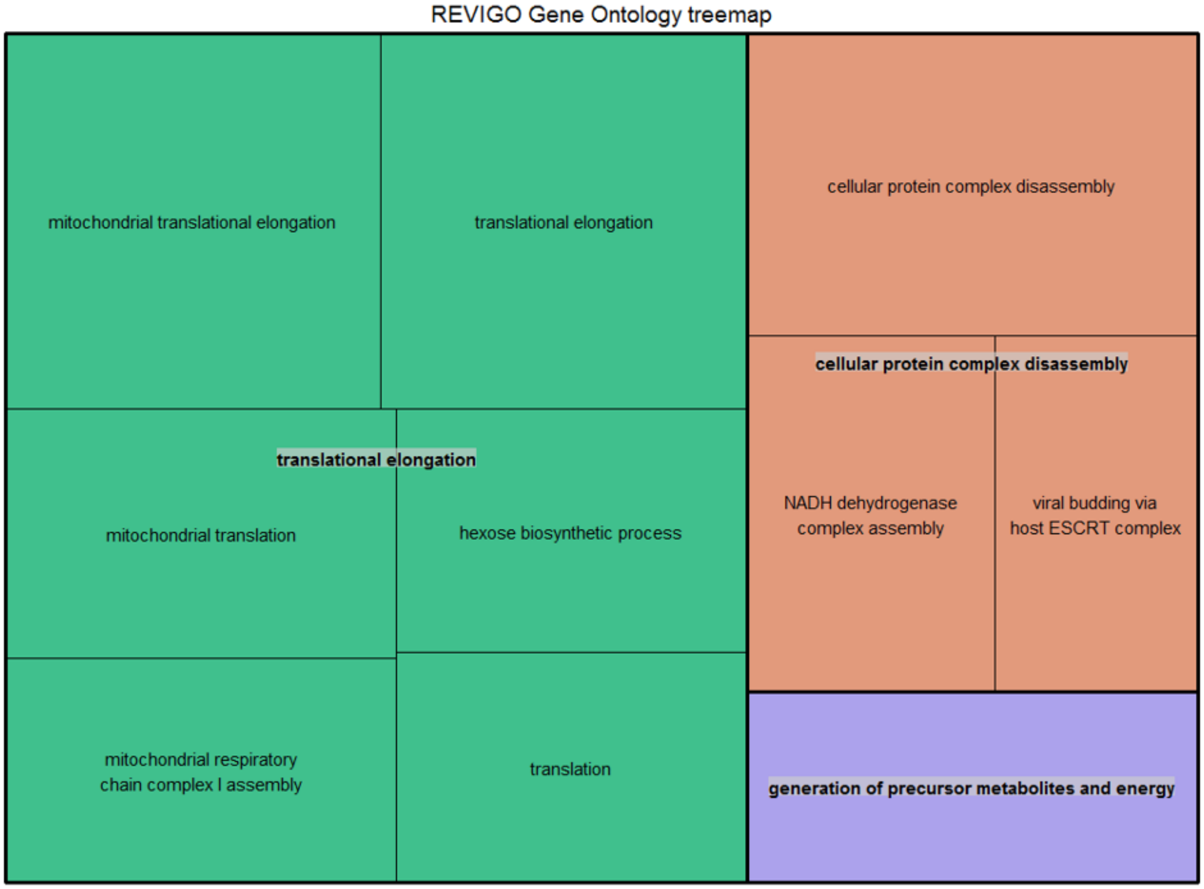


Figure 5 Summarized representation of significant GO- for the genes set generated from the divergent count (853 total genes) overlap based from the DE analysis of FCR. The size of the boxes

is scaled according to the -log10 of the p-value. The most significant individual terms are all in the translation, indicating a link between mitochondrial activity and FE.

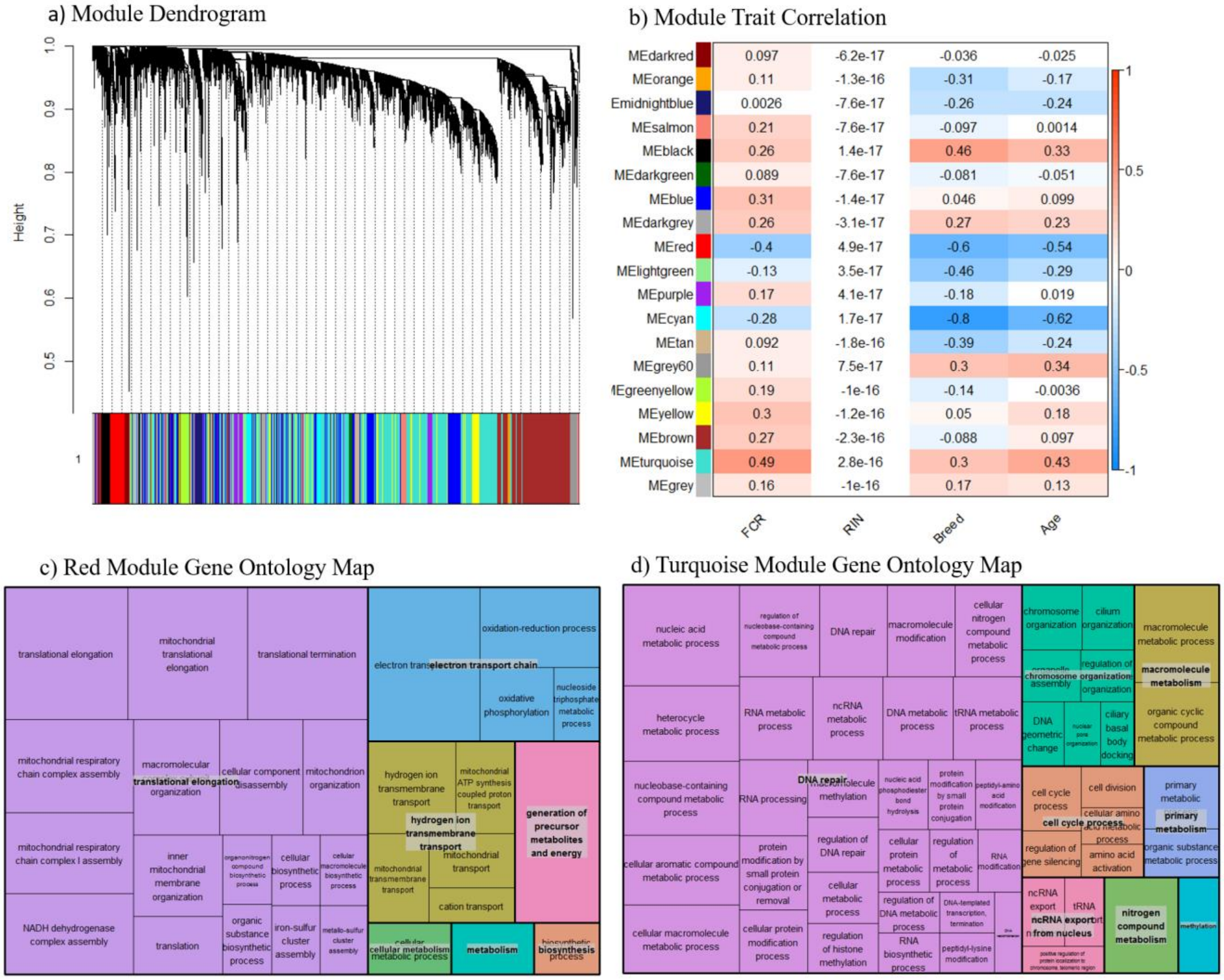


Figure 6 (a) Dendrogram over the module clustering. Looking at the visual clustering not all the modules look equally well defined, but it should be noted that the actual relations in given module cannot be simplified to two dimensions, as the all the relations between the genes exist in N dimentional space, where N is the number of genes. (b) Correlation between module eigenvalue and our traits, including RIN. We see here that the correlation to RIN is essentially 0 in all cases, indicating our linear correction method has worked well. Based on the top two modules **(c)** Summarized representation of significant GO- for genes in the red module of the WGCNA network analysis. The three largets groups are all associated with mitochondria, mirroring the results found in the differential expression analysis and the gene interaction analysis. (b) Summarized representation of significant GO- for genes in the turquoise module of the WGCNA network analysis. The main grouping here is DNA repair, which is not found in our other analysis. This may represent that increased energy expenditure on maintenance processes is reducing FE.

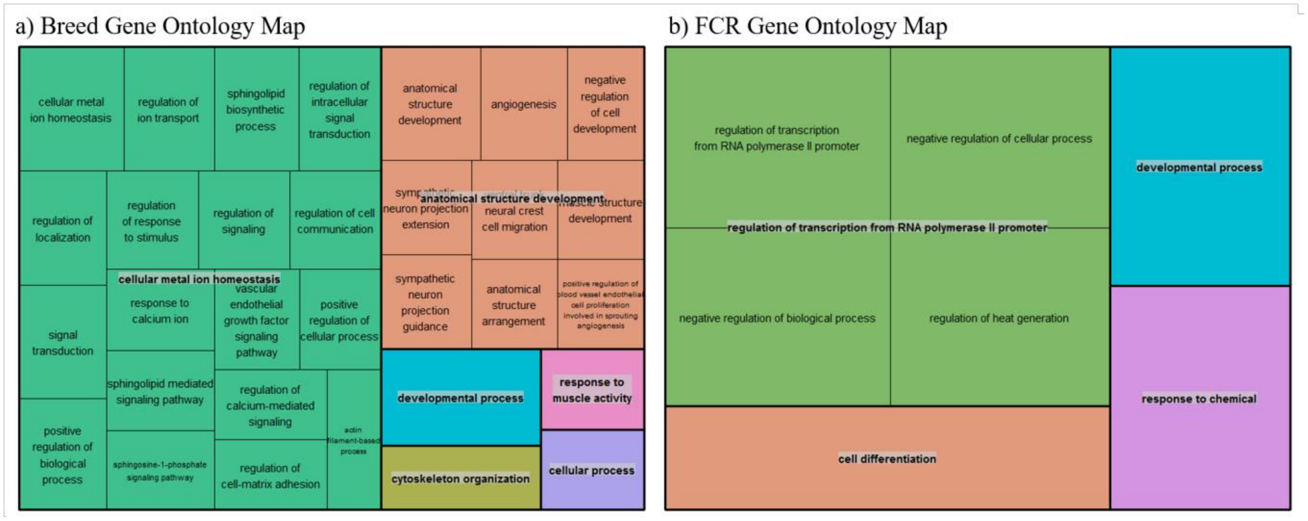


Figure 7 (a) Summarized representation of significant GO- for genes significantly associated with exercise in one of the three human dataset and between the breeds, based on a total of 702 genes. The size of the boxes is scaled according to the -log10 of the p-value. Here we find two overall main categories, cellular metal ion homeostasis and anatomical structure development. (b) Summarized representation of significant GO- for genes significantly associated with exercise in one of the three human dataset and in our divergent set for FCR. The size of the boxes is scaled according to the -log10 of the p-value. Here the main process is regulation of transcription from RNA polymerase. Overall, the categories are not very significant here as it is only based on 42 genes.

Expression QTL and pathway enrichment analysis of feed efficiency and mitochondrial genes in Danish performance tested pigs

Victor AO. Carmelo[1] and Haja N. Kadarmideen[1]*

[1]Quantitative Genomics, Bioinformatics and Computational Biology Group, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Building 324, 2800, Kongens Lyngby, Denmark

***Correspondence:**

Haja N. Kadarmideen

hajak@dtu.dk

Abstract

Feed efficiency (FE) is a key trait in pig production, as it has both economic and environmental impact. FE is a challenging phenotype to study, as it is complex in nature and can be affected by many factors, such as metabolic efficiency and growth, but also activity level. Furthermore, testing for FE is a costly procedure, as it requires specific equipment and monitoring to measure. Therefore, there has been a desire to find functionally relevant genetic variants and biomarkers for FE, not only to assist with improved selection, but also to broaden and deepen our biological understanding of FE. Expression quantitative trait loci (eQTL) are genetic variants that modulates tissue-specific gene expression differences between individuals and thus the downstream gene products and eventually phenotypes. We have done a cis- and trans expressed quantitative trait loci (eQTL) analysis using both a linear and Anova model, in a population of Danbred Durocs (N=11) and Danbred Landrace (N=27). We also used bootstrapping and enrichment analysis to validate and analyze detected eQTLs. We identified 15 eQTLs with FDR < 0.01, affecting several genes found in previous studies of commercial pig breeds. Example include IFI6, PRPF39, TMEM222, CSRNP1, PARK7 and MFF. The bootstrapping results showed statistically significant enrichment of eQTLs with p-value < 0.01 (p-value < $2.2x0^{-16}$) in both cis and trans linear eQTLs. Based on this, enrichment analysis of

top trans-eQTLs was performed, and revealed high enrichment for gene categories and gene ontologies associated with genomic context and expression regulation. This includes transcription factors (p-value=1.0x10^{-13}), DNA-binding (GO:0003677, p-value=8.9x10^{-14}), DNA-binding transcription factor activity (GO:0003700,) nucleus gene (GO:0005634, p-value<2.2x10^{-16}), positive regulation of expression (GO:0010628), negative regulation of expression (GO:0010629, p-value<2.2x10^{-16}).

Introduction

The biological background of complex traits is expressed through molecular processes triggered by a combination of genetics, epigenetics and the environment. While ample genetic markers have been identified for complex traits, the understanding of the functional effect of identified genetic markers is more challenging to identify[1]. Almost per definition, complex traits are controlled by multiple genetic factors [2-4], thus further complicating the biological background. One way of tackling this issues, is to look at direct causal links between genetics and gene expression, thus identifying a direct effect of genetic variation. This allows for a straightforward interpretation of the effect of genetic variation based on pathway and functional knowledge of related genes. This is done through the identification of expressed quantitative loci (eQTL), mapping variants to gene expression. Expression quantitative trait loci (eQTL) are genetic variants that modulates tissue-specific gene expression differences between individuals, thus the downstream gene products and eventually phenotypes. The usage of both the genetic and the transcriptomic layer combined with pathway and phenotype data can be a powerful way of identifying functionally relevant genetic variants and biomarkers for traits of interest. There are, however, several challenges with eQTL analysis. Firstly, if one wanted to map all possible SNP-gene pairs in a modern omics data set, which typically has thousands of expressed genes and at the minimum tens of thousands of SNP, the total amount of tests will be at least in the order of 10^8. This can pose computational challenges, but even worse, multiple testing problems. This is especially relevant as a cursory search of the Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo/) for RNA-seq studies reveals most studies having less than 100 samples. Therefore, it is important to have strategies for these issues when doing eQTL analysis. Example strategies used for filtering the expression data in previous studies include: filtering by estimated heritability of transcripts [5] or using only a limited set of genes[6], and there are many more possibilities.

Feed efficiency (FE) has been known for decades to be an important complex trait in pig breeding. Cost of feed is the largest economic burdens in commercial pig production [7, 8], and lower feed consumption leads to more environmentally friendly production. Improving feed efficiency in livestock is also benefiting reduced greenhouse gas emissions [9]). The two main metrics for feed efficiency are residual feed intake (RFI) [10] and feed conversion rate (the ratio between feed consumed and lean growth) , with the latter being the most used in pig production. Selective breeding has improved FCR in pigs, but this has not led to direct gains in knowledge of the biological drivers of FE in pigs. Even with many studies being done on the subject, the genetic and biological background of FE in pigs is still not well understood [11]. There have been genome-wide association studies (GWAS) and systems genetic studies on RFI phenotypes in Danish pigs that is closely genetically correlated to FCR [12, 13] These studies have revealed important candidate genes for FE via GWAS approach but has not integrated gene expression (transcriptomics) datasets to identify SNPs affecting gene expression in different porcine tissues. The cost and difficulty of measuring FE likely plays into this, in contrast to other traits, such as meat quality or litter size. One concrete example of the usage of FCR is in the Danish pig production where, FCR is improved through a centralized breeding program were potential breeding sires are tested for efficiency via accurate FCR calculations based on measured feed intake and growth.

Muscle is the most important tissue in pig production in regards to production value. Muscle plays a large role in energy metabolism and energy storage[14-16].. As such, there have been multiple studies on the muscle transcriptome in a FE context [7, 17-19]. In comparison, while there are several eQTL studies performed in pig muscle [5, 6, 20, 21], there are none based on FE traits. A connection between FE and mitochondria in muscle has been reported several times, in several species in the littearature [7, 18, 19, 22-24]. In general, it is reported that higher mitochondrial activity is related to increased FE. Given the evidence for mitochondrial effects, identifying genetic regulation of mitochondrial genes could assist in efforts to develop biomarkers for FE.

Here we aimed to perform both cis and trans eQTL analysis on a previously identified set of FCR related genes and mitochondrial genes, in a pig population comprised of Duroc and Landrace purebred pigs. The two-breed analysis serves genetic variation that can aid in the detection of eQTLs, particularly as the Durocs were more heavily selected for FCR. This serves as a targeted approach for our phenotype of interest, but also aids us in reducing the input space we are analyzing in a meaningful way. Furthermore, we hypothesized that genes, which interact with genomic context and/or regulate gene expression are enriched in top trans-eQTLs. If this is true, it is a novel way of validating trans-eQTLs based on functional pathway analysis.

Material and Methods

Sampling and Sequencing

The pigs in this study were the intersection between the pigs genotyped in [25] and Carmelo et al (submitted). All data processing steps follow those two studies, unless otherwise stated. Overall, 38 male uncastrated pigs were included in this study, with 11 purbred Danbred Duroc and 27 purebred Danbred Landarace. The pigs were sent to the commercial breeding station at Bøgildgård, which is owned by the pig research Centre of the Danish Agriculture and Food Council (SEGES) at ~7kg. The pigs were regularly weighed, and feed intake was measured in a test period ranging from 40-70 days from ~28kg of weight. The period of measurement was determined by each pigs commercial viability.

Data selection and filtering

All gene annotation and analysis was done using Sus Scrofa annotation version 11.1.96 from Ensembl.

Gentoype Data and Filtering

The DNA isolation from collected blood and genotyping was performed by to GeneSeek (Neogen company - https://www.neogen.com/uk/). The Genotyping was based on the GGP Porcine HD array (GeneSeek, Scotland, UK), which includes 68,516 SNPs on 18 autosomes and both sex chromosomes. The SNPs were mapped to the sus scrofa genome version 11.1 using the NCBI Genome Remapping from the sus scrofa genome version 10.2. This was done using default settings. To insure that we had a sufficient representation of genotypes for each SNP, we use a MAF (minor allele threshold of 0.3. This removes SNPs that would be underpowered, or that cannot be related to expression changes as due to lack of variation. It also has the advantage of reducing the overall testing space to a more conservative set. This reduced the initial set of SNPs to 27531. The next step performed was to remove groups of SNPs in high Linkage Disequilibrium (LD). To do this, we used the *LD_blocks* function from the *WISH-R* R package[26], which was applied with an R^2 of 0.9. This grouped SNPs linearly across chromosomes into blocks based on a minimum pairwise R^2 value of

0.9 between all SNPs in a block. After this step, 19179 SNPs remained. The genotypes were coded as 0 (homozygote major), 1 (heterozygote) and 2 (homozygote minor) for the eQTL anaylsis.

Expression data, Gene selection and filtering

Muscle tissue samples were extracted from the psoas major muscle immediately post slaughter, and the samples were kept at -25 C in RNA later (Ambion, Austin, Texas). The data was sequenced on the BGISEQ-500 platform using the PE100 (pair end, 100bp length) with Oligo dT library prepapration at BGI Genomics. The reads were trimmed using Trimmomatic [27] version 0.39, with the default setting for paired end reads. Data QC was performed pre- and post-trimming using FastQC v0.11.9. Mapping was done with STAR aligner[28] version 2.7.1a, with default parameters and genome and annotation 11.1 version 96. Beside default parameters, the --quantMode GeneCounts setting was used for read quantification. Our main interest was to investigate genes that could be related to FCR. We therefore based our set of genes on the methods in Carmelo et. al (submitted). In brief, Differential Expression analysis (DEA) was performed using three different DE methods (Limma, EdgeR, Deseq2)[29-31] with FCR as the phenotypes of interest. We then calculated the divergence between our observed p-value distribution for FCR and the uniform distribution for each method, enabling us to select a list of genes that are related FCR. This was motivated by the fact that we had a large overrepresentation of low p-values in the DEA, meaning the distribution was anti-conservative. This resulted in a set of 853 genes. As mitochondrial genes have been implicated in FE in muscle in both our previous study and in several studies in multiple species[7, 18, 19, 22-24], we also selected all genes with a mitochondrial gene ontology (gene ontology id GO:0005739) and included them in the analysis. All genes were filtered to have a minimum of 5 reads in at least 11 samples, as 11 was the size of the Duroc group. Testing revealed that genes with a single expression outlier could result in likely false positives. Therefore, all genes with a single gene with a Z-score above 3 were removed, corresponding to a single observation with normalized expression further than 3 standard deviations from the mean. This resulted in a final gene set of 1425 genes.

eQTL Analysis

Calculation of eQTLs

All of the eQTL analysis was performed using R version 3.5.3. Gene expression was normalized using the *calcNormFactors* from the R package edgeR version 3.34.3. We performed eQTL analysis using the R package MatrixEQTL version 2.3[32]. We added the following covariates in the model: RNA integrity values (RIN), breed, batch and age (days). Given that the samples were collected in slaughterhouse setting, it was necessary to include RIN in the model, but this should not be an issue if appropriately corrected for [33]. As samples were collected on different days, it was necessary to correct for this using the batch effect. Breed and age have an effect on expression, as seen in our previous study [Carmelo et al], and thus must be corrected for. While the samples come from a selection of 28 different breeders in Denmark, there is still some relationship between some pigs, especially if they came from the same breeder. Therefore, a kinship matrix based on 4 generations of pedigree was added as the error covariance matrix instead of using the default identity matrix. The cis-distance was set to 10^6 bp. The analysis was done using both the *modelANOVA* (Anova) and *modelLINEAR* (linear) options, giving both a factor-based model and a linear model fit.

Statistical Significance

After the model was fit, based on the empirical p-value distribution, pathway enrichment analysis was performed on the top putative eQTLs based on the results from the trans-eQTL linear model. The linear model was chosen over the Anova as the empirical p-value distribution for the Anova had an overweight of low p-values, which means that we should avoid using the overall distribution of p-values for conclusions. In the linear version, we observed that the p-values were nearly uniform with a slight overweight of low p-values. To show the significance of this result, we performed bootstrapping by shuffling the genotype values of each SNP while maintaining the same expression values and covariates. We then calculated the number of random eQTLs with p-value < 0.01, for both the cis- and trans-eQTLs. Assuming the shuffled values are normally distributed, we calculated the probability of observing our empirical number of p-values < 0.01. We also saved the lowest, the 10th lowest and the 100th lowest observed p-value for both trans and cis bootstrapped eQTLs for each iteration. The bootstrapping procedure was done 500 times with both Anova and the linear model.

Orthonormalization

To visualize the expression and genotype values on the scale used by Matrix eQTL, we scaled and centered both the design matrix of the covariates, the expression and the genotypes. After this, we used the *mlr.orthogonalize* function from the MatchLinReg package version 0.7.0 to orthogonilize the expression values and genotypes of each relevant gene and SNP in relation to the covariates, respectively, using *normalize=True* as an option. This procedure was done mimicking the method reported in the Matrix eQTL[32].

QTL region and relation to FCR

To verify if our eQTLs were in know quantitative trait loci (QTL) regions, we first defined a region of 100kb upstream and downstream of each SNP to overlap with. The region size was conservatively defined based on reported haplotype block sizes in commercial pigs[34]. We then checked if the SNP coordinate had any overlaps with FCR QTLs from the Pig QTL database[35]. We also did the same procedure with the target gene, except we did not extend the region beyond the gene boundaries.

Pathway Analysis

We hypothesized that, if trans-eQTLs are not false positives, they should be enriched for functional categories which could relevantly cause distal interactions, in comparison to a the background. Therefore, to analyze our top trans-eQTLs, we calculated the number of additional empirically observed low p-values under 0.01, by substracting the expected numer of p-values < 0.01 ($\mathbf{0.01} \times \boldsymbol{N}$) from the observed. We then tested the enrichment of the genes in our top eQTL group for the following gene categories/ontologies: transcription factors(TF) (based on the AnimalTFDB 3.0 pig transcription factors [36]) , DNA-binding (GO:0003677), DNA-binding transcription factor activity (GO:0003700) nucleus gene (GO:0005634), positive regulation of expresison (GO:0010628), negative regulation of expression (GO:0010629) and membrane gene (GO:0016020). Each category was selected based on a biological hypothesis. All GO terms were retrieved using biomart 2.42.0 with annotation from Sus Scrofa 11.1. 96

Results

In figure 1 we can see the overall p-value distribution for both the linear and the Anova eQTL analysis. The linear model is overall well behaved, with uniform p-values and a small increase of low p-values. In the Anova model, the spike of high p-values may be due to issues with model assumptions in some cases, but as there large number of eQTLs it is not practical to do model diagnostics on each eQTL. This does not mean individual Anova based eQTL cannot be valid, but we should be careful with drawing results based on the overall distribution. The cis-eQTLs have a more uneven overall distribution, but likely due to the lower amount of tests. Looking at the individual results using a threshold for FDR of 0.1, the only analysis that give any significant results was the Anova analysis, yielding 14 significant trans-eQTLs and 1 cis-eQTL. It should be noted, that in our linear analysis, due to the left-skewing of the p-value distribution, all trans-eQTLs with p-value < 0.01 (N=301213) have and FDR value of 0.9 or better. This means that it is very likely that we have real trans-eQTLs, we just lack the power to identify them individually. Given this, and the results from the bootstrapping analysis (see below), we elected to show the top ten 10 eQTLs for each analysis, except the Anova trans, were we selected all with FDR < 0.1. In figure 2 we can see the visualization of the top 6 eQTLs in the linear trans model, ordered from lowest p-value (top left) to highest (bottom right). Given the low p-values reported, the visualization, especially of the first one, does not seem to support the results. The explanation is found in the Matrix eQTL implementation deals with covariates. In Matrix eQTL, all covariates, expression and genotypes are centered and scaled, and then the expression and genotype vectors are both orthogonized in relation to the covariate matrix. Only after this step is the linear relationship between expression and genotype calculated. In figure 3, we can see the same plot in scatter plot form, based on the transformed values. Here we can see a clear linear relationship between the transformed expression values and genotype values.

Bootstrapping

Bootstrapping is a useful tool when dealing with complex data, allowing us to get estimates of the likelihood of our observations without explicit probability calculations. Here, we wanted to show that our spike in low p-values in the linear analysis was statistically unlikely to happen by chance. Based on 500 bootstraps, we estimating the mean and the variance of the number of p-values < 0.01, and compared this with our observed number. Assuming normally distributed counts, which is a fair assumption given our sample size and the scale of the data, the probability of our observed value is essentially 0, which is also visualized in figure 3. In table 2 we can see the comparison of the 1st,10th and 100th p-values in our bootstrapped data with our empirical data. Overall, the real data performs better as we go down in rank. This indicates that the real data has lower bound on significance, but overall the results are not achievable by chance.

Pathway enrichment analysis

As our genes were pre-selected, there is no a-priori reason to perform enrichment analysis. In particular, there is no particular meaning in finding that the cis-eQTLs are enriched for some pathway. The cis-eQTLs are simply tests of correlation between local genomic context and expression, and significance denotes identification of possible genetic expression regulatory mechanisms, not underlying pathways. In contrast, for the trans-eQTL, there are meaningful hypothesis we could state. Why would a gene have significant association to a distal genetic element? We hypothesized that genes that interact with genomic context and/or are expression regulatory would be enriched in the low p-value group in comparison to the overall genes used. This includes genes that directly interact with genomic context, such as DNA binding genes and regulatory genes, such as transcription factors and positive or negative expression regulators. To test our hypothesis, we selected the top 28147 SPN-gene pairs from our linear trans eQTL analysis. This represented our observed surplus of low p-values found when comparing with a uniform p-value distribution for eQTLs with a p-value < 0.01, motivated by our results in from the bootstrapping (figure 2). Traditionally, one might test our hypothesis using a pathway enrichment tool, but given that the eQTL data had a special structure, including repeated entries of the same genes from a smaller background set, it was not suitable for typical methods. Instead, we used a more targeted approach, selecting specific categories we believed tested our hypothesis. In table 1, you can see the result for the enrichment of the top genes compared to the initial background set, using the Fisher test to derive p-values, with selected gene ontologies and categories. The results from the enrichment are quite striking, as we get very significant enrichment for DNA binding genes, transcription factors and DNA binding transcription factor activity. All these categories fit our hypothesis, as they engage directly with distal genomic context. We also tested for nucleus genes, as we expect genes that are active in the nucleus to be more likely to interact with genomic context. Furthermore, we tested for general expression regulation, with the positive and negative expression regulation categories. Intriguingly, positive regulation was slightly depleted or unchanged, while negative expression regulatory genes was the most enriched category. Finally we included membrane genes, as a control category which includes a large number of genes, as we do not believe they have a reason to be enriched, which they are not. As a control of the enrichment, we also compared with all expressed genes in our samples. This aids in the interpretation, and acts as a control, as if there was high divergence in the two comparisons the results might just be an artefact of our methodology. We see similar results comparing with all genes, and due to the large number of genes in both the expressed set and the trans-eQTLs, we get very significant p-values.

Discussion

In this study, we applied Matrix eQTL to a set of genes previously identified as having associations to FCR. We have presented that top results of both cis and trans eQTLs based on both linear association and a factorial genotype mode (Anova). There have been several muscle eQTL studies in pig before [5, 6, 21, 37-41]. However, direct comparison of results is quite challenging, for several reasons. None of the other studies where applied to FCR, as the genes and SNPs selected were generally selected based on the phenotype of interest, this limits the overlap. Furthermore, due to the statistical challenges, many divergent strategies were employed, for example using a pre-GWAS[38], picking a limited set of pathway specific genes[6] or using a limited set of microsatellites[41]. Some studies also included heritability analysis [5, 21]. The studies above include both crossed, purebred and F2 half-sib pig populations. Given all these factors, and the novelty of FCR in an eQTL context, we cannot compare our study very specifically to othersm, and one should view our study as a pilot study for FCR eQTLs. Specifically, we have only a limited number of samples given the genetic context, and thus we view our individual eQTLs with caution, and they should be confirmed in larger population. We do however present novel strategies in an eQTL context, which show promising results, and could be generally applicable to other eQTL studies.

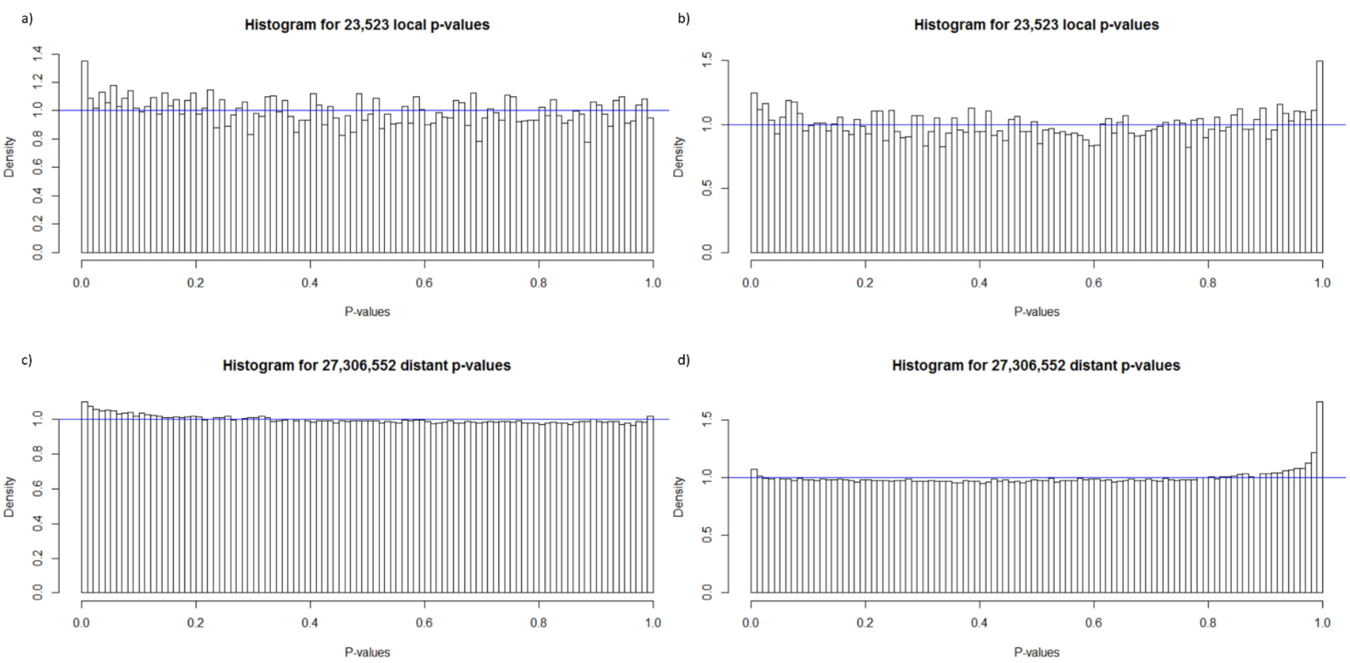
We have included two pure breeds in our analysis, Duroc and Landrace, which in of itself is an unusual choice. Many studies published have inbred lines, but it has been suggested that it would be advantageous to do eQTL analysis on a natural genetically varying population [42], such as two separate breeds. For the input SNPs, we made several choices for maximizing the number of relevant SNPs to include. First, we selected a quite high cutoff of 0.3 MAF. This allows us to have high enough variation at each included SNP, given our low sample size. It has also been shown, that in chip-based data such as ours, the overall structure in the data is robust to different MAF cutoffs[43], thus this should not impart any biases into the results. Finally, we grouped SNPs in high LD (R^2 >0.9) into blocks and used tagging variants to represent blocks. This allowed us to reduce the space further, removing redundant genetic information, thus relaxing our multiple testing thresholds. In regards to our cis-eQTL distance, we chose a 1Mb window, which is on the lower end for pig studies [21], however given our low samples size we wanted to keep the cis analysis as conservative as possible.

In regards to individual eQTLs, one should be careful with over interpreting the results, but instead view the eQTLs as candidates for further study. Based on a qualitative analysis, we do however find several interesting genes among our top eQTL candidates. IFI6, a gene implicated in apoptosis regulation through mitochondrial pathways[44], has been previously related to meat and carcass quality[45]. PRPF39, a pre-mRNA processing gene, has previously been related to a trans-eQTL in Durocs with divergent fatness[41]. DNAJB1, a heatshock gene, was found to be downregulated in lean pigs [46]. TMEM222, a transmembrane protein, was found to be differentially expressed between tissues and genotypes between Korean native pigs and Yorkshires[47]. CSRNP1, the Cysteine And Serine Rich Nuclear Protein 1 gene, was found to be a metabolic response gene in relation to feed intake in Durocs. INTS7, an RNA processing gene, was associated with a SNP significant for meat quality in Chinese pigs[48]. The ACOX3 gene, a fatty acid metabolism gene, had previous cis-eQTLs identified for it[6]. The PARK7 gene, a gene that codes for a protein that protects from oxidative stress[49], does not appear in a pig related context in the literature, but it is found in a known FCR QTL region, as well as the mitochondrial fission factor (MFF). CPT1B, the arnitine palmitoyl transferase 1B gene, was differentially expressed in large whites versus an indigenous high-fat breed[50]. The Potassium Calcium-Activated Channel Subfamily M Alpha 1 gene, KCNMA1, had been previously found to diverge in expression between Large White and Basque pigs[51]. Metaxin (MTX2), a mitochondrial gene, was a candidate gene for red blood cell count in a Duroc x Erhualian population based on a nearby genome wide significant SNP [52]. Synaptotagmin 12 (SYT12) was found in the area with which explained the largest variance in piglets porn in 3520 Durocs[53]. Myosin XIX (MYO19) was a candidate gene for eating behavior traits due to a nearby significant SNP in the same Duroc population our pigs come from[54]. The Uncharacterized Protein C7orf50 has previous cis-eQTLs idenfied in a behavioral context in humans[55]. While this might seem like a mixed group of results, he main takeaway, is that each of the genes mentioned above have appeared in previous contexts that demonstrate genetic regulation and association with traits under selection in commercial pigs, thus giving qualitative evidence that increases the likelihood of our eQTLs being true positives despite the sample size..

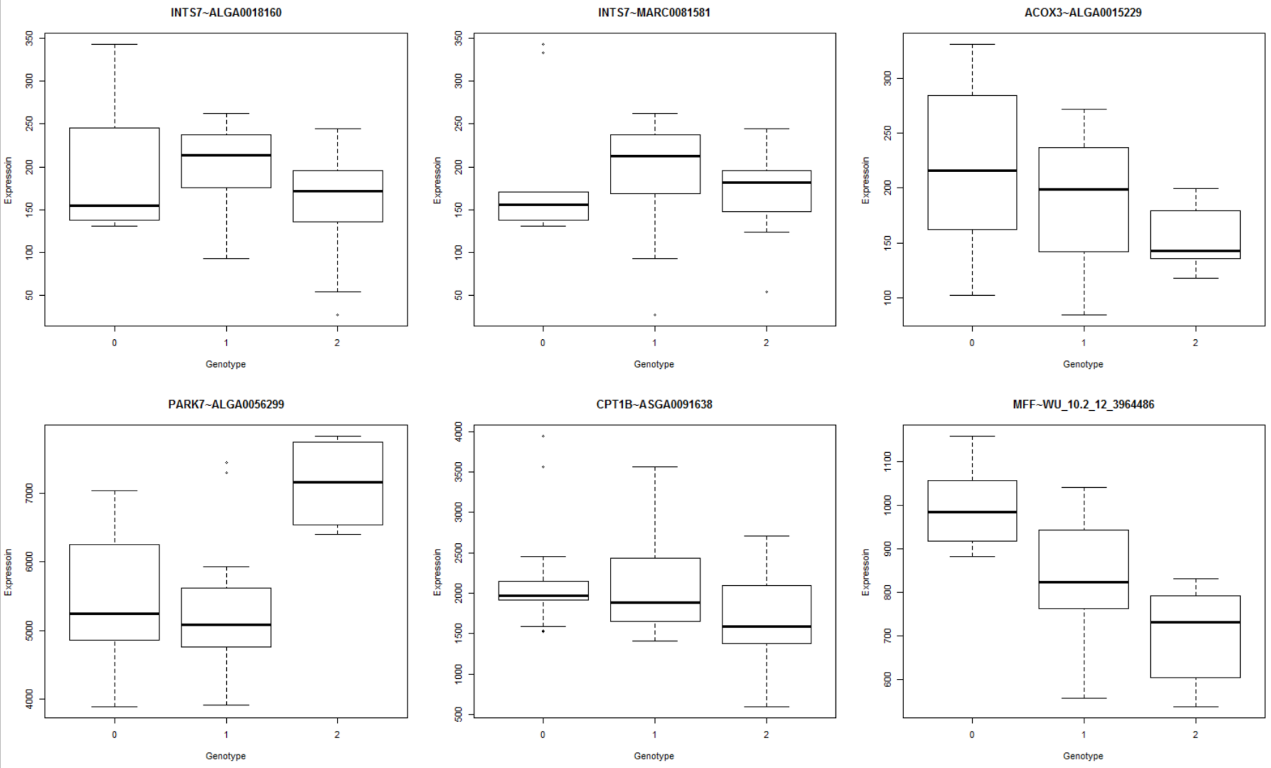
The final and perhaps most interesting result in our analysis stems from the enrichment analysis in the linear trans-eQTL analysis. We had initially hypothesized that we would find enrichment for genes that interact with genomic context and highly interacting genes. The findings, and their significance level, show a strong overrepresentation of DNA-binding genes, DNA-binding with transcription factor activity genes and transcription factors. These results have a quite straightforward interpretation - genes that interact on a genomic level have a higher chance of having trans-eQTL activity. This can be both mediated through direct interactions, but also through indirect effects, such as transcription factor acting on each other, thus mediating their own genetic effect to other genes. The more intriguing result is the contrast between negative and positive gene regulation. Given our sample size and study power, it is difficult to assess individual genes, and thus properly grasp specific interpretation of these results. In general, given the complexity of gene expression regulation, further specific study is needed before we have a proper understanding of the contrast between negative and positive expression, and the rest of the enrichment results. Based on our analysis, we propose that these enrichments could be general effects, and thus can assist us in the validation of true trans-eQTLs. Essentially, identification of such biologically relevant effects can be used as an extra layer of evidence for true positive trans-eQTLs.

Conclusion

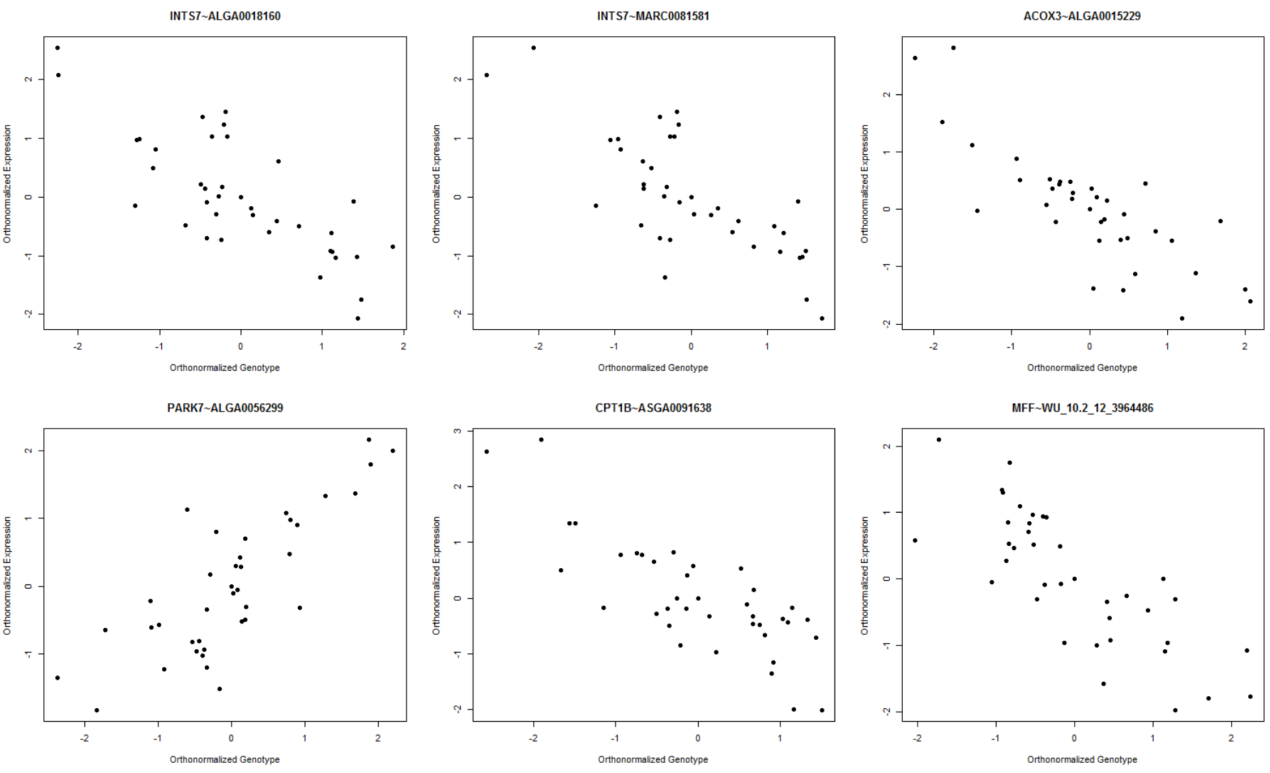
In this study, using a population of purebred Durocs (N=11) and purebred Landarace (N=27) pigs and a set of previously identified FCR genes, we did a cis and trans-eQTL anaylsis based on both linear and Anova models. We identified 15 eQTLs at 0.1 FDR level, reported several more with marginal significance, which all could serve a potential FCR biomarkers. In our linear analysis, we identified a highly statistically significant increase of p-values below 0.01, based on bootstrapping. Based on this, we performed pathway enrichment analysis on the top 28147 linear trans-eQTLs, testing the hypothesis that genes that interact with genomic context, and generally, gene expression regulators, would be enriched in top trans-eQTLs. We identified highly significant enrichment for transcription factors, DNA-binding (GO:0003677), DNA-binding transcription factor activity (GO:0003700) nucleus gene (GO:0005634), negative regulation of expression (GO:0010629) and depletion for positive regulation of expression (GO:0010628).



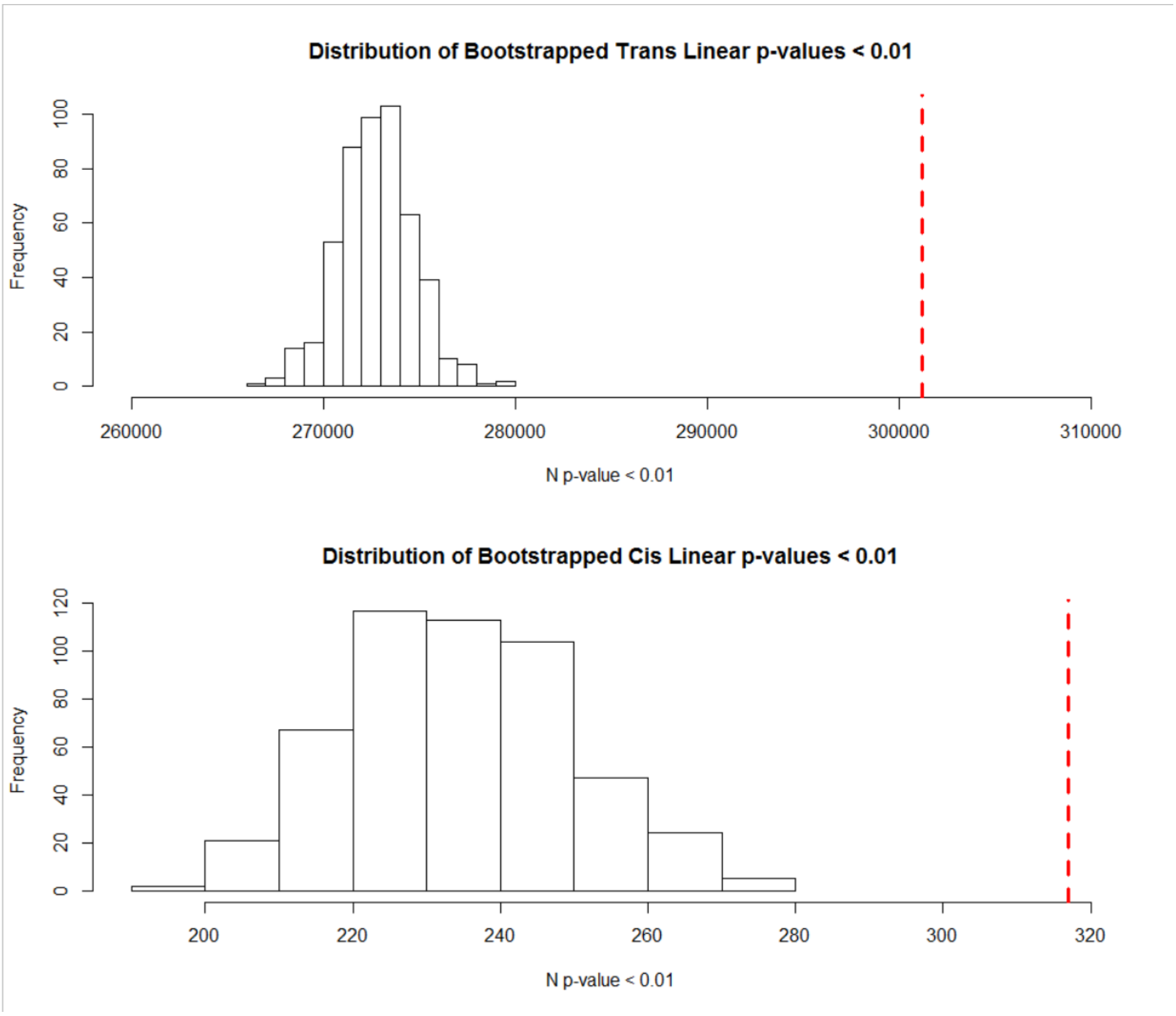
*Figure**Fejl! Ingen tekst med den anførte typografi i dokumentet.** 1. Histograms of the p-value distribution of all cis (a,b) and trans(c,d) eQTL pairs in the linear(a,b) and Anova(c,d) models. Based on the overall distribution, we see a slight anti-conservative trend in the linear p-values in both cis and trans eQTLs.*



*Figure**Fejl! Ingen tekst med den anførte typografi i dokumentet.** 2. Boxplot of the top 6 trans-eQTLs from linear analysis. Comparing with the summary from table 1, it seems unexpected that the top left boxplot is of the most significant eQTL. . Overall the 3rd and the 6th ranked eQTLs look visually more appealing. This because the genotype here are the raw values, and the expression values are only normalized normally, not taking covariates into account.*



*Figure**Fejl! Ingen tekst med den anførte typografi i dokumentet.** 3. Scatter-plot of the orthonormalized expression and genotype values for the top 6 trans-eQTLs in the linear analysis. The linear relationship is quite clear on the transformed values, in comparison to the boxplots of the untransformed values.*



*Figure**Fejl! Ingen tekst med den anførte typografi i dokumentet.** 3. Histograms of the number of p-values below 0.01 in our 500 bootstrapped linear trans and cis-eQTLs analysis. The red dotted line represents the observed values., The likelihood of observing such extreme values by chance is essentially 0 in both cases, if we model the likelihood based of the normal distribution.*

Snp	Gene	P-value	FDR	Chr	Position	Analysis
WU_10.2_7_1320670	SLC20A2	2.36e-11	0.00064	7	1132634	Anova trans
WU_10.2_7_115152142	SLC20A2	1.21e-10	0.00096	7	108750676	Anova trans
ASGA0040859	SLC20A2	1.35e-10	0.00096	9	3330061	Anova trans

WU_10.2_18_2886712	SLC20A2	1.41e-10	0.00096	18	2875724	Anova trans
ASGA0042452	IFI6	7.68e-10	0.0042	9	31552392	Anova trans
WU_10.2_14_148897646	PRPF39	3.89e-09	0.018	14	137024504	Anova trans
ALGA0109564	DNAJB1	1.23e-08	0.048	15	68774343	Anova trans
ALGA0053497	TMEM222	1.40e-08	0.048	9	60196621	Anova trans
WU_10.2_12_33709155	GCAT	2.01e-08	0.058	12	32811156	Anova trans
ASGA0013363	DLC1	2.11e-08	0.058	3	11038140	Anova trans
ASGA0091484	CSRNP1	2.55-08	0.059	4	118914325	Anova trans
ALGA0009614	CSRNP1	2.58e-08	0.059	1	256190584	Anova trans
WU_10.2_13_216907306	POGZ	3.21e-08	0.063	13	207030935	Anova trans
ASGA0083137	DLC1	3.22e-08	0.063	9	138505307	Anova trans
ALGA0018160[1]	INTS7	5.61e-09	0.15	3	27307613	Linear trans
MARC0081581[1]	INTS7	3.25e-08	0.31	3	27346598	Linear trans
ALGA0015229[2]	ACOX3	3.39e-08	0.31	2	116633408	Linear trans
ALGA0056299[2]	PARK7[1]	5.21e-08	0.36	10	1400269	Linear trans
ASGA0091638[2]	CPT1B	8.19e-08	0.38	4	626787	Linear trans
WU_10.2_12_3964486	MFF[1]	8.38e-08	0.38	12	4217210	Linear trans
ALGA0115669[2]	PARK7[1]	9.79e-08	0.38	10	1187360	Linear trans
DRGA0015709[1]	COMTD1	1.22e-07	0.42	16	2090820	Linear trans
ALGA0087901[1]	NSUN2	1.63e-07	0.45	15	129751572	Linear trans
WU_10.2_7_740616	Glycine N-phenylacetyltransferase	1.64e-07	0.45	7	618465	Linear trans
INRA0015708[1]	NES	4.07e-06	0.096	4	94242001	Anova cis
WU_10.2_15_134661069	ABCB6	1.31e-05	0.11	15	121481724	Anova cis
WU_10.2_6_27531636	NUDT21	1.36e-05	0.11	6	30064483	Anova cis
MARC0009689	HDHD5	3.01e-05	0.18	5	69473204	Anova cis
WU_10.2_15_150992806	RAMP1	6.83e-05	0.32	15	136516507	Anova cis
MARC0112128	KCNMA1	0.00010	0.41	14	79922641	Anova cis
WU_10.2_15_91334711	MTX2	0.00022	0.55	15	81867240	Anova cis
WU_10.2_2_4374745	SYT12	0.00022	0.55	2	5445193	Anova cis
ALGA0019808[1]	MEIS1	0.00026	0.55	3	76307479	Anova cis
WU_10.2_3_18580686	STX4	0.00027	0.55	3	18010007	Anova cis
ASGA0054417	MYO19	0.00018	0.63	12	38196853	Linear cis
WU_10.2_X_128169493	RBMX	0.00018	0.63	X	112221790	Linear cis

WU_10.2_12_39624033	MYO19	0.00026	0.63	12	37981199	Linear cis
WU_10.2_3_183721	C7orf50	0.00031	0.63	3	335933	Linear cis
ALGA0108896[1]	CRYM	0.00035	0.63	3	24920076	Linear cis
ALGA0061099	MRPS31	0.00035	0.63	11	16202962	Linear cis
ALGA0061107	MRPS31	0.00035	0.63	11	16236530	Linear cis
ASGA0030240	NSUN4	0.00042	0.63	6	165835717	Linear cis
WU_10.2_14_153092095	ECHS1	0.00047	0.63	14	141129811	Linear cis
WU_10.2_14_153836231	ECHS1	0.00047	0.63	14	141357898	Linear cis

Table 1. Overview over top cis and trans eQTls in all 4 four sub-analyses. [1]Genes or Snps in known FCR qtl regions. [2]Snps with p-value < 0.05 for linear association with FCR

Model	Min P-value	10th P-value	100th P-value
Anova Cis	0.13	0.126	0
Anova Trans	0.044	0.062	0
Linear Cis	0.984	0.688	0
Linear Trans	0.148	0.024	0.018

Table 2. Probability of observing a lower p-value the lowest, 10th lowest p-value and 100th lowest p-values in our bootstrapping. In general, in relation to our random eQTLs we perform better except in the linear cis analysis, but not very significantly. It is interesting to note that by the 100th p-value all the analysis outperform random data. This indicates that we do have true, but perhaps weak effects, and that it sets an upper bound on the power we have to find individual eQTLs post FDR correction.

Category	N hits in P<0.01	Odds Ratio	P-value	Odds ratio expressed genes	P-value expressed genes
Transcription Factor	3145	2.27	$1.0x10^{-13}$	1.40	$<2.2x10^{-16}$
DNA binding	3394	2.20	$8.9x10^{-14}$	1.73	$<2.2x10^{-16}$

DNA-binding transcription factor activity	2721	3.36	$<2.2x10^{-16}$	2.36	$<2.2x10^{-16}$
Positive regulation of expresison	346	0.67	0.07	0.67	$8.9x10^{-6}$
Negative regulation of expresison	1887	4.34	$<2.2x10^{-16}$	5.39	$<2.2x10^{-16}$
Nucleus gene	8811	1.33	$2.4x10^{-6}$	1.18	$1.8x10^{-12}$
Membrane gene	7707	1.06	0.30	1.15	$1.8x10^{-8}$

Table 3. Enrichment analysis based on the linear trans-eQTLs with p-value < 0.01, based on the Fisher exact test. The enrichment was compared with the original input set of 1425 genes, and to the set of expressed genes in our muscle samples for additional comparison.

References:

1. Gallagher, M.D. and A.S. Chen-Plotkin, The Post-GWAS Era: From Association to Function. Am J Hum Genet, 2018. **102**(5): p. 717-730.
2. Wood, A.R., Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet, 2014. **46**(11): p. 1173-86.
3. Hirschhorn, J.N., Genetic approaches to studying common diseases and complex traits. Pediatr Res, 2005. **57**(5 Pt 2): p. 74R-77R.
4. Johnson, G.C. and J.A. Todd, Strategies in complex disease mapping. Curr Opin Genet Dev, 2000. **10**(3): p. 330-4.
5. Liaubet, L., et al., Genetic variability of transcript abundance in pig peri-mortem skeletal muscle: eQTL localized genes involved in stress response, cell death, muscle disorders and metabolism. BMC Genomics, 2011. **12**: p. 548.
6. González-Prendes, R., R. Quintanilla, and M. Amills, Investigating the genetic regulation of the expression of 63 lipid metabolism genes in the pig skeletal muscle. Animal Genetics, 2017. **48**(5): p. 606-610.
7. Jing, L., et al., Transcriptome analysis of mRNA and miRNA in skeletal muscle indicates an important network for differential Residual Feed Intake in pigs. Scientific Reports, 2015. **5**.
8. Gilbert, H., et al., Review: divergent selection for residual feed intake in the growing pig. Animal, 2017. **11**(9): p. 1427-1439.
9. Basarab, J.A., et al., Reducing GHG emissions through genetic improvement for feed efficiency: effects on economically important traits and enteric methane production. Animal : an international journal of animal bioscience, 2013. **7 Suppl 2**(Suppl 2): p. 303-315.
10. Koch, R.M., Swiger, L. A., Chambers, D. J. & Gregory K. E, Efficiency of feed use in beef cattle. Journal of Animal Science, 1963. **22**(2): p. 486-494.
11. Ding, R., et al., Genetic Architecture of Feeding Behavior and Feed Efficiency in a Duroc Pig Population. Front Genet, 2018. **9**: p. 220.
12. Do, D.N., et al., Genome-wide association and pathway analysis of feed efficiency in pigs reveal candidate genes and pathways for residual feed intake. Front Genet, 2014. **5**: p. 307.
13. Do, D.N., et al., Genome-wide association and pathway analysis of feed efficiency in pigs reveal candidate genes and pathways for residual feed intake. Frontiers in Genetics, 2014. **5**(307).
14. Morales, P.E., J.L. Bucarey, and A. Espinosa, Muscle Lipid Metabolism: Role of Lipid Droplets and Perilipins. Journal of Diabetes Research, 2017. **2017**: p. 10.

15. Muscle as a Secretory Organ, in Comprehensive Physiology. p. 1337-1362.
16. Turner, N., et al., Fatty acid metabolism, energy expenditure and insulin resistance in muscle. J Endocrinol, 2014. **220**(2): p. T61-79.
17. Horodyska, J., et al., RNA-seq of muscle from pigs divergent in feed efficiency and product quality identifies differences in immune response, growth, and macronutrient and connective tissue metabolism. BMC Genomics, 2018. **19**(1): p. 791.
18. Gondret, F., et al., A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs. BMC Genomics, 2017. **18**(1): p. 244.
19. Vincent, A., et al., Divergent selection for residual feed intake affects the transcriptomic and proteomic profiles of pig skeletal muscle. J Anim Sci, 2015. **93**(6): p. 2745-58.
20. Ponsuksili, S., et al., Discovery of candidate genes for muscle traits based on GWAS supported by eQTL-analysis. Int J Biol Sci, 2014. **10**(3): p. 327-37.
21. Velez-Irizarry, D., et al., Genetic control of longissimus dorsi muscle gene expression variation and joint analysis with phenotypic quantitative trait loci in pigs. BMC Genomics, 2019. **20**(1): p. 3.
22. Connor, E.E., et al., Enhanced mitochondrial complex gene function and reduced liver size may mediate improved feed efficiency of beef cattle during compensatory growth. Functional & Integrative Genomics, 2010. **10**(1): p. 39-51.
23. Bottje, W.G., et al., Proteogenomics Reveals Enriched Ribosome Assembly and Protein Translation in Pectoralis major of High Feed Efficiency Pedigree Broiler Males. Front Physiol, 2017. **8**: p. 306.
24. Eya, J.C., M.F. Ashame, and C.F. Pomeroy, Association of mitochondrial function with feed efficiency in rainbow trout: Diets and family effects. Aquaculture, 2011. **321**(1): p. 71-84.
25. Priyanka Banerjee, V.A.O.C., Haja N. Kadarmideen, Genome-wide epistatic interaction networks affecting feed efficiency in Duroc and Landrace pigs. Front. Genet, 2020.
26. Carmelo, V.A.O., et al., WISH-R– a fast and efficient tool for construction of epistatic networks for complex traits and diseases. BMC Bioinformatics, 2018. **19**(1): p. 277.
27. Bolger, A.M., M. Lohse, and B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014. **30**(15): p. 2114-20.
28. Dobin, A., et al., STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013. **29**(1): p. 15-21.
29. Ritchie, M.E., et al., limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res, 2015. **43**(7): p. e47.
30. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010. **26**(1): p. 139-40.
31. Love, M.I., W. Huber, and S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol, 2014. **15**(12): p. 550.
32. Shabalin, A.A., Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics, 2012. **28**(10): p. 1353-8.
33. Gallego Romero, I., et al., RNA-seq: impact of RNA degradation on transcript quantification. BMC Biol, 2014. **12**: p. 42.
34. Veroneze, R., et al., Linkage disequilibrium and haplotype block structure in six commercial pig lines. Journal of Animal Science, 2013. **91**(8): p. 3493-3501.
35. Hu, Z.-L., C.A. Park, and J.M. Reecy, Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. Nucleic Acids Research, 2018. **47**(D1): p. D701-D710.
36. Hu, H., et al., AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Research, 2018. **47**(D1): p. D33-D38.
37. Ponsuksili, S., et al., Discovery of candidate genes for muscle traits based on GWAS supported by eQTL-analysis. International journal of biological sciences, 2014. **10**(3): p. 327-337.
38. Steibel, J.P., et al., Genome-wide linkage analysis of global gene expression in loin muscle tissue identifies candidate genes in pigs. PloS one, 2011. **6**(2): p. e16766-e16766.
39. Chen, C., et al., A genome-wide investigation of expression characteristics of natural antisense transcripts in liver and muscle samples of pigs. PloS one, 2012. **7**(12): p. e52433-e52433.
40. Perry, K.R., et al., P3030 Identification of expression quantitative trait loci for longissimus muscle microrna expression profiles in the Michigan State University Duroc × Pietrain pig resource population. Journal of Animal Science, 2016. **94**(suppl_4): p. 67-67.
41. Canovas, A., et al., Segregation of regulatory polymorphisms with effects on the gluteus medius transcriptome in a purebred pig population. PLoS One, 2012. **7**(4): p. e35583.
42. Gilad, Y., S.A. Rifkin, and J.K. Pritchard, Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet, 2008. **24**(8): p. 408-15.
43. Linck, E. and C.J. Battey, Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. Molecular Ecology Resources, 2019. **19**(3): p. 639-647.
44. Qi, Y., et al., IFI6 Inhibits Apoptosis via Mitochondrial-Dependent Pathway in Dengue Virus 2 Infected Vascular Endothelial Cells. PLOS ONE, 2015. **10**(8): p. e0132743.
45. Kayan, A., et al., Investigation on interferon alpha-inducible protein 6 (IFI6) gene as a candidate for meat and carcass quality in pig. Meat Science, 2011. **88**(4): p. 755-760.
46. Zambonelli, P., et al., Transcriptional profiling of subcutaneous adipose tissue in Italian Large White pigs divergent for backfat thickness. Animal Genetics, 2016. **47**(3): p. 306-323.

47. *Li, X., et al., Quantitative gene expression analysis on chromosome 6 between Korean native pigs and Yorkshire breeds for fat deposition. Genes & Genomics, 2010.* ***32****(4): p. 385-393.*

48. *Liu, X., et al., Genome-wide association analyses for meat quality traits in Chinese Erhualian pigs and a Western Duroc × (Landrace × Yorkshire) commercial population. Genetics Selection Evolution, 2015.* ***47****(1): p. 44.*

49. *Zhang, Y., et al., Elevated expression of DJ-1 (encoded by the human PARK7 gene) protects neuronal cells from sevoflurane-induced neurotoxicity. Cell Stress and Chaperones, 2018.* ***23****(5): p. 967-974.*

50. *Gao, Y., et al., Detection of differentially expressed genes in the longissimus dorsi of Northeastern Indigenous and Large White pigs. Genet Mol Res, 2011.* ***10****(2): p. 779-91.*

51. *Damon, M., et al., Comparison of Muscle Transcriptome between Pigs with Divergent Meat Quality Phenotypes Identifies Genes Related to Muscle Metabolism and Structure. Plos One, 2012.* ***7****(3).*

52. *Nan, J.-h., et al., Genetic parameter estimation and genome-wide association study (GWAS) of red blood cell count at three stages in a Duroc×Erhualian pig population. Journal of Integrative Agriculture, 2020.* ***19****(3): p. 793-799.*

53. *Stafuzza, N.B., et al., A genome-wide single nucleotide polymorphism and copy number variation analysis for number of piglets born alive. BMC Genomics, 2019.* ***20****(1): p. 321.*

54. *Do, D.N., et al., Genome-wide association study reveals genetic architecture of eating behavior in pigs and its implications for humans obesity by comparative mapping. PloS one, 2013.* ***8****(8): p. e71509-e71509.*

55. *Liao, C., et al., Multi-tissue probabilistic fine-mapping of transcriptome-wide association study identifies cis-regulated genes for miserableness. bioRxiv, 2019: p. 682633.*