Institut für Tierwissenschaften, Abt. Tierzucht und Tierhaltung
der Rheinischen Friedrich–Wilhelms–Universität Bonn

Application of knowledge discovery and data mining methods in livestock genomics for hypothesis generation and identification of biomarker candidates influencing meat quality traits in pigs

Inaugural - Dissertation

zur

Erlangung des Grades

Doktor der Agrarwissenschaft

der

Landwirtschaftlichen Fakultät

der

Rheinischen Friedrich–Wilhelms–Universität

zu Bonn

von

Sudeep Sahadevan

aus

Bharananganam, Kerala, India

Referent :	Prof. Dr. Karl Schellander
Koreferent :	Prof. Dr. Martin Hofmann-Apitius
Tag der mündlichen Prüfung :	28 November 2014
Erscheinungsjahr :	2014

"If a man will begin with certainties, he shall end in doubts; but if he will be content to begin with doubts, he shall end in certainties."

Francis Bacon

Application of knowledge discovery and data mining methods in livestock genomics for hypothesis generation and identification of biomarker candidates influencing meat quality traits in pigs

Recent advancements in genomics and genome profiling technologies have lead to an increase in the amount of data available in livestock genomics. Yet, most of the studies done in livestock genomics have been following a reductionist approach and very few studies have either followed data mining or knowledge discovery concepts or made use of the wealth of information available in the public domain to gain new knowledge. The goals of this thesis were: (i) the adoption of existing analysis strategies or the development of novel approaches in livestock genomics for integrative data analysis following the principles of data mining and knowledge discovery and (ii) demonstrating the application of such approaches in livestockgenomics for hypothesis generation and biomarker discovery. A pig meat quality trait termed androstenone measurement in backfat was selected as the target phenotype for the experiments.

Two experiments were performed as a part of this thesis. The first one followed a knowledge driven approach merging high-throughput expression data with metabolic interaction network. Based on the results from this experiment, several novel biomarker candidates and a hypothesis regarding different mechanisms regulating androstenone synthesis in porcine testis samples with divergent androstenone measurements in back fat were proposed. The model proposed that the elevated levels of androstenone synthesis in sample population could be due to the combined effect of cAMP/PKA signaling, elevated levels of fatty acid metabolism and anti lipid peroxidation activity of members of glutathione metabolic pathway. The second experiment followed a data driven approach and integrated gene expression data from multiple porcine populations to identify similarities in gene expression patterns related to hepatic androstenone metabolism. The results indicated that one of the low androstenone phenotype specific co-expression cluster was functionally enriched in pathways related to androgen and androstenone metabolism and that the members of this cluster exhibited weak co-expression in high androstenone phenotype. Based on the results from this experiment, this co-expression cluster was proposed as a signature cluster for hepatic androstenone metabolism in boars with low androstenone content in back fat. The results from these experiments indicate that integrative analysis approaches following data mining and knowledge discovery concepts can be used for the generation of new knowledge from existing data in livestock genomics. But, limited data availability in livestock genomics is a hindrance to the extensive use such analysis methods in livestock genomics field for gaining new knowledge.

In conclusion, this study was aimed at demonstrating the capabilities of data mining and knowledge discovery methods and integrative analysis approaches to generate new knowledge in livestock genomics using existing datasets. The results from the experiments hint the possibilities of further exploring such methods for knowledge generation in this field. Although the application of such methods is limited in livestock genomics due to data availability issues at present, the increase in data availability due to evolving high throughput technologies and decrease in data generation costs would aid in the wide spread use of such methods in livestock genomics in the coming future.

Einsatz von Methoden der Datengewinnung und Wissensentdeckung in der Nutztiergenomforschung zur Hypothesengenerierung und Identifizierung von Kandidaten-Biomarkern die ein Fleischqualitätsmerkmal beim Schwein beeinflussen

Neuste Entwicklungen im Bereich der Genomik und in den Technologien für das Genom Profiling führten zum Anstieg der verfügbareren Datenmengen des Nutztiergenoms. Jedoch folgten die meisten Studien in der Nutztiergenomforschung dem reduktionistischen Ansatz und nur wenige Studien den Methoden der Datengewinnung und Wissensentdeckung oder nutzten bestehende Informationen aus der öffentlichen Domain, um neue Erkenntnisse zu gewinnen. Die Ziele dieser Dissertation waren: (i) bestehende Analysestrategien aufzunehmen oder neue Methoden in der Nutztiergenomforschung für die integrative Datenanalyse zu entwickeln. Dabei kamen Methoden der Datengewinnung und der Wissensentdeckung zum Einsatz. Und (ii) dadurch die Anwendung dieser Ansätze in der Nutztiergenomforschung zur Hypothesengenerierung und zur Entdeckung von Biomarkern zu veranschaulichen. Für die vorliegenden Experimente diente als Ziel-Phänotyp ein Schweinefleischqualitätsmerkmal, welches durch die Messungen von Androstenon im Rückenfett gekennzeichnet ist.

Zwei Versuche werden in der Dissertation abgehandelt. Das erste Experiment folgte einem wissensgesteuerten Ansatz und brachte high-throughput Expressionsdaten mit metabolischen Interaktionsnetzwerken in Verbindung. Basierend auf diesen Versuchsansatz konnten verschiedene neuartige Kandidaten-Biomarker identifiziert und Hypothesen gebildet werden die mit Mechanismen der Androstenonsynthese in Hodenproben vom Schwein mit divergenten Androstenongehalten aus dem Rückenfett in Verbindung stehen. Für die Stichprobe mit erhöhten Androstenonsyntheselevel konnte mittels dieses Models ein kombinierter Effekt aus dem cAMP/PKA Signalweg sowie einem erhöhten Level des Fettsäuremetabolismus und Antilipid-Peroxidationsaktivität als Teile des Glutathion Stoffwechselwegs aufgedeckt werden. Das zweite Experiment folgte einem Daten-basierenden Ansatz und integrierte Genexpressionsdaten von multiplen Schweinepopulationen, mit dem Ziel Ähnlichkeiten in Genexpressionsmustern bezogen auf den Lebermetabolismus von Androstenon zu identifizieren. Die Ergebnisse ergaben, dass der Phänotyp niedriger Androstenongehalt spezifische Co-Expressions-Cluster aufwiesen die funktionell mit Pathways, die in Verbindung mit dem Androgen und Androstenon Metabolismus stehen, angereichert sind. Diese Clustermitglieder wiesen im Gegenzug schwache Co-Expressionen zu dem Phänotyp hoher Androstenongehalt auf. Basierend auf diesen Ergebnissen konnte das ermittelte Co-Expressions-Cluster als ein Signatur-Cluster für den hepatischen Androstenenmetabolismus von Ebern mit niedrigem Androstenongehalt im Rückenfett dargestellt werden. Die Ergebnisse beider Versuche zeigten, dass integrative Analysemethoden, die der Datengewinnung und der Wissensentdeckung folgen, für die Gewinnung neuer Erkenntnisse aus bereits vorhandenen Daten in der Nutztiergenomforschung benutzt werden können. Allerdings, machte es die begrenzte Datenverfügbarkeit in der Nutztiergenomik hinderlich solche Analysemethoden im Bereich der Nutztiergenomforschung extensive zu Nutzung um neues Wissen zu gewinnen.

Abschließend war das Ziel der Studie die Möglichkeiten der Methoden der Datengewinnung und

der Wissensentdeckung sowie die der integrativen Analysemethoden, als Verfahren zur Gewinnung von neuem Wissen in der Nutztiergenomforschung aus bereits vorhandenen Daten, darzustellen. Die Ergebnisse dieser Experimente verweisen auf die Möglichkeiten weiter an diesen Methoden zur Weiterentwicklungen in diesen Bereichen, zu forschen. Obwohl der Einsatz solcher Methoden in der Nutztiergenomforschung, aufgrund der zurzeit begrenzt verfügbaren Daten limitiert ist, unterstützen die sich durch entwickelnden high-throughput Technologien entstehende Daten und die sinkenden Datengenerierungskosten die weit verbreitete Nutzung dieser Methoden in der Nutztiergenomforschung in der Zukunft.

Contents

List of Figures

List of Tables

1. Introduction

The conventional method of breeding livestock animals for favorable traits involves visual evaluation of animals and keeping records of performance characteristics based on pedigree and phenotype of the animals. In the genomic and post genomic era, advanced genetic and genomic technologies have also been used to determine various aspects of the genotype of animals (Holloway and Morris, 2008). The advantage of using genomic selection over conventional methods is that the animals can be selected at a young age for traits such as fertility, disease resistance and feed conversion rates, which are expensive and laborious to measure (Hayes et al., 2013). The use of genetic and genomic studies in veterinary sciences have been increasing steadily (Figure 1.1). If the number of abstracts indexed in Pubmed is taken as an indicator of the number of studies published, it can bee seen from the figure that the number of genetics or genomics related studies in animal sciences have been growing annually. At present, breeding practices in involves a combination of conventional breeding methods and advanced genetic technologies to refine and understand the genetics of favorable characters in livestock species (Holloway and Morris, 2008). Thus, the livestock genomics research field primarily involves identifying and studying the genetic machinery behind various traits of economical importance in livestock animals in an effort to improve these traits. Following the advancements in human biology and genetics, livestock genomics also adopted high throughput technologies such as microarray expression profiling, SNP chips for Genome wide association studies (GWAS) and Next generation sequencing (NGS) to study the genetics of farm animals.

With the advancements in whole genome profiling technologies, there has been an increase in the quantity of data available in livestock genomics. As per the current statistics (in early 2014), for *B. taurus* (cattle) there are 6,769 datasets in GEO database (GEO Datasets B. taurus, 2014) and (microarray and other high throughput data) and 765 (SRA Datasets B. taurus, 2014) SRA experiments (NGS data). In case of *S. scrofa*, there are 8,848 GEO datasets (GEO Datasets S. scrofa, 2014) and 1,966 SRA experiments (SRA Datasets S. scrofa, 2014) publicly available. In addition to these large publicly available datasets, there are improvements in gene function and pathway annotations for livestock species. According to the current statistics, there are 20,045 bovine gene products and 19,749 porcine gene products annotated[1] in the Gene Ontology annotation project (Hill et al., 2000). Additionally, in KEGG database (Kanehisa and Goto, 2000) for bovine and porcine genomes there are 279 pathways annotated per genome [2,3]. Although there

[1] http://www.geneontology.org/GO.current.annotations.shtml last accessed March 6, 2014
[2] http://www.kegg.jp/kegg-bin/search_pathway_text?map=bta&mode=1 last accessed March 6, 2014
[3] http://www.kegg.jp/kegg-bin/search_pathway_text?map=ssc&mode=1 last accessed March 6, 2014

is an increase in the number of publicly available datasets for livestock genomics, it has to be taken into consideration that these numbers are still small in comparison to the data available for human, mouse and other model organism species. Even this limited amount of publicly available data can be investigated to learn new patterns and to extract new knowledge.

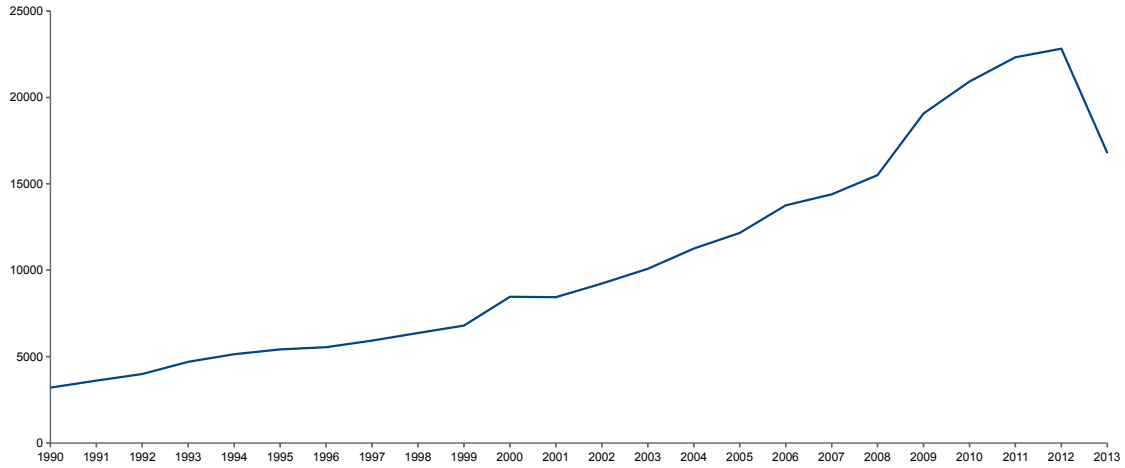

Figure 1.1: Growth of genetics and genomic studies in animal sciences. Figure shows the number of abstracts indexed in Pubmed per year from 1990-2013 on genetics and genomics studies in veterinary sciences. Search query used in Pubmed: "(genomic OR genetic) AND veterinary [sb]". The reason for the sudden drop in 2013 could be that a large number of studies from 2013 are still left to be indexed, as this Pubmed query was performed in early 2014.

Majority of the (high throughput) studies in livestock genomics have been focused on identifying and explaining the differential expression of genes/association of Single Nucleotide Polymorphisms (SNPs) in large scale expression matrices or in GWAS experiments. Very few studies in this field have made use of the wealth of information available in various public databases to study the genetics behind favorable traits in livestock genomics. The data analysis approaches in livestock genomics have mostly been following a reductionist approach, analyzing various components of the cellular system individually for biomarker identification. However, human medicine and development have been following integrative analysis approaches to understand the genetics behind a variety of diseases and phenotypes.

Integrative analysis in molecular biology refers to merging multiple datasets or data resources in order to study a phenotype, identify biomarkers and generate hypothesis for further evaluation. The design philosophy behind such analysis method is that a phenotype or a disease is seldom the consequence of a change in a single effector gene or gene product, but rather the result of a multitude of changes in a complex interaction network (Loscalzo and Barabasi, 2011). The usual end result of such methods are diagnostic pathways or subnetworks. In human development and medicine, these diagnostic pathways and diseases subnetworks are demonstrated to enhance the prediction accuracy of disease states and to be more reproducible than single biomarkers (Chuang et al., 2010). In essence, integrative analysis approaches are used to understand the effects of different large scale zones of the biological system, rather than focusing on the individual components. Systems biology is an interdisciplinary branch of biomedical research that mainly targets the complex biological interactions within a biological system using various holistic data

analysis approaches. These approaches primarily deal with 'omics' data at the level of mRNAs, proteins and metabolites. Rigorous integration of heterogeneous data is a prime requirement in systems biology to achieve comprehensive, quantitative and predictive understanding using mathematical modeling (Sauer et al., 2007).

Two computational theoretic concepts that are often discussed in association with systems biology and integrative analysis approach are data mining and knowledge discovery. Data mining refers to the application of algorithms to extract specific patterns from data. Knowledge discovery is a concept used to highlight that knowledge is the end product of a data driven discovery process (Fayyad et al., 1996a). The key difference between a data mining approach and a knowledge discovery process is that the latter also describes the background steps involved, such as data selection, data preparation, data cleaning, incorporating additional prior knowledge and result interpretation (Fayyad et al., 1996a). In a broad sense, it can be said that the concepts of data mining and knowledge discovery are the underlying themes in integrative analysis approaches and systems biology. In addition to the aforesaid concepts, two additional analysis concepts that are often discussed along with integrative analysis approaches are knowledge driven and data driven approaches. As the name suggests, knowledge driven approaches involves integrating prior knowledge with datasets to gain new knowledge. On the other hand, data driven approaches integrate large volumes of data to identify patterns and to gain new knowledge from the data itself.

As discussed before, there have been very few attempts in livestock genomics either to make use of the publicly available data or to make use of data mining and knowledge discovery methods in order to identify candidate biomarkers or to generate hypothesis on the cellular mechanisms involved in the manifestation of economically important phenotypes in livestock genomics. The primary challenge in this case is that the majority of data mining and knowledge discovery analysis pipelines or integrative analysis workflows were mainly developed with model organism species in mind and to make use of the large volumes of data available for model organism species. In livestock genomics however, far less data is publicly available and therefore the bulk of algorithms and workflows developed may not be useful. Nevertheless, data available in livestock genomics can still be used for knowledge discovery purposes.

Taking the limitations of data availability in livestock into consideration, the major goals of this thesis were defined as:

(i) Adopt existing data analysis approaches or generate new analysis strategies for integrative data analysis in livestock genomics using principles of data mining and knowledge discovery.

(ii) Demonstrate the application of integrative analysis approaches in livestock genomics by using these analysis approaches for hypothesis generation and biomarker discovery on existing data from an economically important phenotype.

For achieving these goals, androstenone content in porcine backfat was chosen as a target analysis trait. The accumulation of androstenone in porcine adipose tissues is one of the primary reasons for a meat quality trait known as boar taint. Boar taint is often described as an off odor or

off taste often noticeable from meat products derived from non castrated boars, primarily due to a lipophilic sex steroid known as androstenone (Bonneau, 1982). Androstenone is mainly synthesized in testis and metabolized in liver (James Squires, 2010). Surgical castration of piglets is one of the most widely practiced method to reduce androstenone by reducing or limiting the synthesis of androstenone (Haugen et al., 2012). But, on grounds of animal welfare, European Union has mandated the abolishment of piglet castration without anesthesia by 2018 (Mörlein et al., 2012). A limitation with the current studies to understand androstenone metabolism is that none of the studies tried to visualize the mechanism of androstenone biosynthesis or metabolism as the result of multifaceted cellular mechanisms and tried only to explain the biological processes and pathways in androstenone biosynthesis and metabolism in terms of individual QTLs, SNPs or candidate genes.

Two experiments were devised in thesis to demonstrate the use of data mining and knowledge discovery driven integrative analysis in livestock genomics in the light of the current economic importance given to androstenone genomics in porcine. The first knowledge driven experiment dealt with the gene interactions and metabolic processes involved in the synthesis of androstenone in testis and made use of the existing knowledge on gene interaction networks associated with steroid hormones biosynthesis. A restriction of this approach in terms of studying androstenone biosynthesis is that none of the major pathway databases contain data on metabolic reaction steps or gene interactions involved in androstenone biosynthesis. As a work around to this limitation, androstenone biosynthesis is treated as an offshoot of steroid hormone (testosterone) synthesis pathway in testis under the assumption that the pathways and interaction events that affect steroid hormone biosynthesis could also affect androstenone biosynthesis. The existing knowledge on hepatic androstenone metabolism is limited to a handful of candidate biomarkers and hence it was not possible to follow a knowledge driven experimental setup in the second experiment. Additionally, since liver is the end point for the metabolism of a large number of compounds, it may not be possible to pinpoint biomarkers based on analysis of a single sample population. Hence, in the second experiment, a data driven experiment combining expression data from three porcine sample populations were followed to understand population/breed similarity in the gene expression patterns related to androstenone metabolism.

The rest of this thesis is structured into four different chapters: Chapter 2 “Literature Review” gives an overview on current state of the art in livestock genomics research, data analysis approaches and integrative analysis approaches. Chapter 3 “Material and Methods” describes the materials and experimental methodology followed in this thesis, Chapter 4 “Results and Discussion” describes and discusses the results from the experiments and this thesis is concluded in the final Chapter 5 “Conclusion”.

2. Literature review

The origins of modern livestock genomics can be traced back to a series of conferences in the early 1990s where strategies and collaborations were developed to maximize the resources available to animal genetics during that period (Womack, 2005). Major research areas in livestock genomics study the genetics behind animal growth, nutrition, milk production, meat production and reproduction related traits in an effort to improve these traits. Genome sequencing efforts in livestock genomics began with the release of the first draft of chicken (*G. gallus*) genome in March 2004 and that of the cattle (*B. taurus*) genome in September 2004 (Fadiel et al., 2005). Quantitative genetics technologies used in livestock genomics also progressed from the use of restriction fragment length polymorphism (RFLP) towards making use of linkage disequilibrium (LD) for the construction of linkage maps, quantitative trait loci (QTL) detection and finally towards marker assisted selection (MAS), a concept of establishing association between various genetic markers and phenotypic trait of interest (Hu et al., 2011). Molecular genetics approaches used in livestock genomics also evolved from the identification of biomarkers to the sequencing of expressed sequencing tags (ESTs) and identification of individual sequence polymorphisms to the use of high throughput genome technologies such as microarrays, SNP chips and finally to use of Next Generation Sequencing (NGS) technologies for sequencing whole genomes.

2.1 Major areas of research in livestock genomics

Genomic selection of economically important traits is the underlying theme for majority of the research topics in livestock genomics. Some of the major research areas, development and success stories in this field are detailed in this section.

In dairy cattle, progeny testing based genomic selection have been performed for improving milk production (Pryce and Daetwyler, 2012; Schaeffer, 2006). It has been demonstrated in Irish cattle population that genomic selection has improved the genetic change for milk production and fertility (Wickham, 2012). According to the data from 2010, reliabilities for predicted transmitting ability (PTA) for milk production ranged from 74-81% in young Holstein bulls (Wiggans et al., 2011). In addition to progeny testing, genomic selection for traits such as feed conversion ratios, body weight gain and dry matter intake (DMI) in dairy cattle have also been subjected to active research (de Haas et al., 2012; Pryce et al., 2012). According to Pryce and Daetwyler (2012), the reliabilities of upto 60% in genetic gain is achievable in dairy cattle using genomic selection (Pryce and Daetwyler, 2012). However, in beef cattle, the adoption of genomic selection technologies has been slower in comparison to dairy cattle due to the low to moderate breeding values of beef

cattle traits such as reproduction, carcass traits, meat quality and feed efficiency (Hayes et al., 2013; Mujibi et al., 2011; Saatchi et al., 2011; Weber et al., 2012). Hayes et al. (2013) pointed out that the low breeding values for economically important traits in beef cattle might be due to the small number of reference population for beef cattle and the large number of important beef cattle breeds, unlike dairy cattle (Hayes et al., 2013). Nevertheless, using a set of hypothetical marker panels, it was predicted that DNA testing could increase the selection response in beef cattle between 29 - 158% (Van Eenennaam et al., 2011). To understand the disease resistance and tolerance traits related to protozoan parasite infection, functional genomics studies are being conducted in *B. taurus* and *B. indicus* cattle species (Glass et al., 2012). Further research have also been conducted on the genomics of various reproductive traits and issues related to in vivo and in vitro culture conditions for cattle embryos (Gad et al., 2012; Humblot et al., 2010). Since published literature can directly reflect the trends in research field, a MeSH[1] term (Rogers, 1963) analysis was done with the search query "(cattle OR cow OR bovine OR B. taurus) AND economic AND traits" to identify and understand the published trends in studies related to economic traits in cattle. Figure 2.1 is a word cloud of MeSH terms based on Pubmed abstracts returned for the search query. This figure hints that major economic traits that are actively researched and published in bovine genomics are dairying, lactation, milk, pregnancy, meat, body weight and fertility related traits.

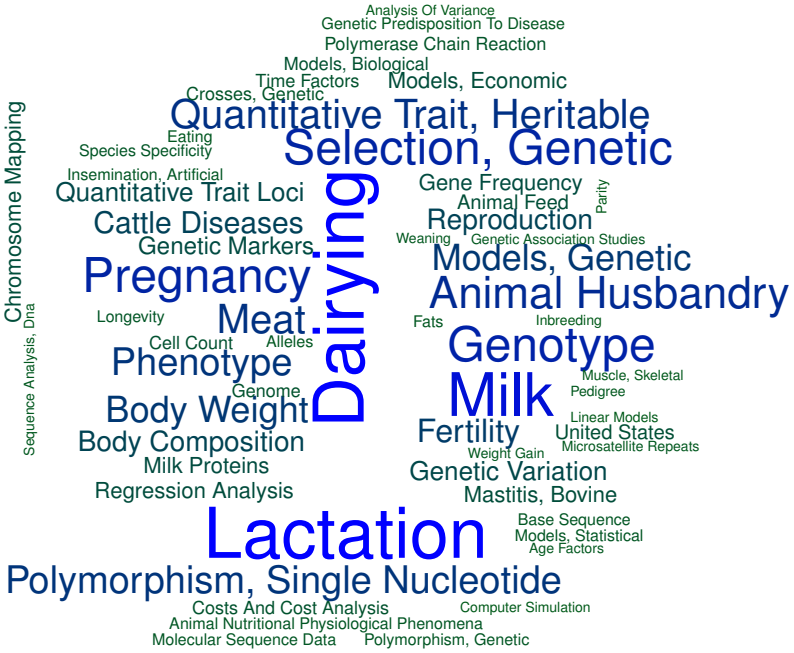

Figure 2.1: Bovine economic traits MeSH cloud. The figure is generated using MeSH clouds retrieved from LigerCat (Sarkar et al., 2009) and R package wordcloud (Fellows, 2013). The MeSH terms removed from the wordcloud representation are: breeding, cattle, male and female. The font size of the terms in the figure directly reflects the frequency of occurrence of these MeSH terms in the set of abstracts returned for search query.

Genomics of a number of economically important traits in pigs has also been major research topic in livestock genomics. Feed conversion rates and daily gain in pure bred porcine population have actively been researched (Ostersen et al., 2011). In case of contribution of maternal trait to

[1]http://www.nlm.nih.gov/mesh/ last accessed March 18, 2014

total genetic genetic gain, it was shown that genotyping and selection of female pigs increased the genetic gain upto 55% in comparison with conventional breeding methods (Lillehammer et al., 2013). Additional investigation has also been done to understand the cellular mechanisms behind porcine meat quality traits such as water holding capacity, driploss, intra muscular fat and androstenone content in backfat (Brunner et al., 2012; Gunawan et al., 2013; Ma et al., 2013). Substantial amount of work has also been devoted to reveal the genetics behind immunity related traits in various porcine breeds. Based on the investigation of a number of immunity related genes in porcine, Flori et al. (2011) called for a more sustainable production system, where animal health can be improved by slight trade-offs in performance characteristics (Flori et al., 2011). To understand the traits related to innate immunity levels in pig, mapping of quantitative trait loci related to innate immunity levels in pigs have also been conducted (Uddin et al., 2011). A MeSH cloud analysis using the query "(pig OR porcine OR swine OR S. scrofa) AND economic AND traits" indicate that economic traits of active research in porcine genomic community are meat, body composition, reproduction, litter size, muscle and body weight related traits, with primary importance given to meat related traits (Figure 2.2).

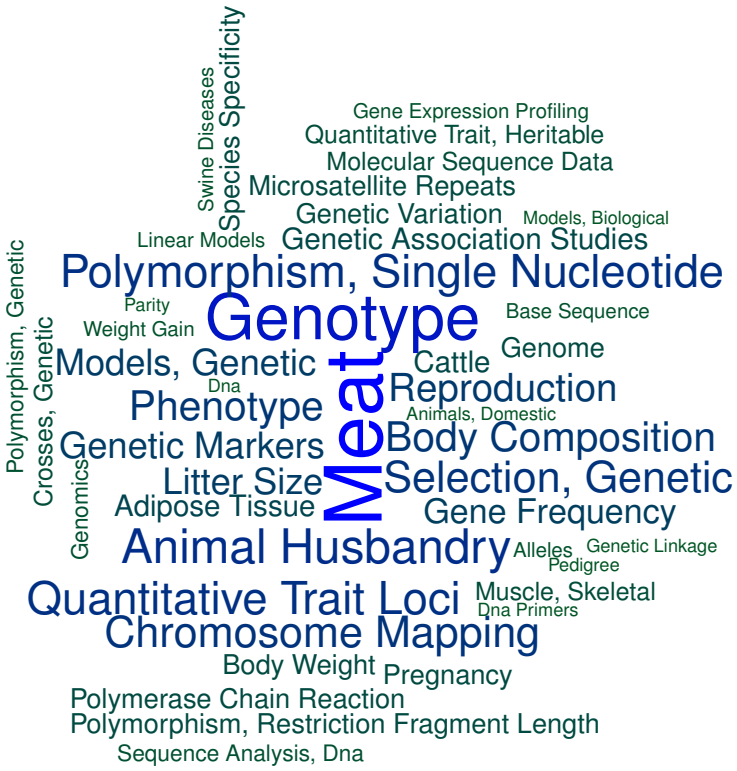

Figure 2.2: Porcine economic traits MeSH cloud. The figure is generated using MeSH clouds retrieved from LigerCat (Sarkar et al., 2009) and R package wordcloud (Fellows, 2013). The MeSH terms removed from the wordcloud representation are: breeding, swine, male, female and Sus scrofa. The font size of the terms in the figure directly reflects the frequency of occurrence of these MeSH terms in the set of abstracts returned for search query.

In addition to cattle and pig, the genomics of other economically important livestock species such as sheep, poultry and horse are also under active study to improve the economically important traits. In dairy sheep, genomics of lactation related traits such as milk yield, fat content and somatic cell scores are being investigated (Duchemin et al., 2012). Furthermore, genotypes related to meat and wool related traits in sheeps were also researched (Daetwyler et al., 2010). As a

result of this, it was shown that the estimated genomic values of wool traits such as fleece weight and fiber diameter are higher than 60% (Daetwyler et al., 2012). In poultry, quantitative traits related to feed conversion rates in chicken were also investigated (González-Recio et al., 2009). SNP markers for resistance to *Salmonella* carrier-state in commercial egg laying chicken lines were also studied to check *Salmonella* propagation and hence reduce food safety concerns (Calenge et al., 2011). Researchers have also scrutinized the genomics of a number performance related traits in various horse breeds. A genome wide analysis examined SNP markers associated with aesthetics and performance related traits in a number of non-thoroughbred horse breeds (Petersen et al., 2013). In thoroughbred horses, a genome wide scan revealed a number of genetic markers related to performance and exercise related traits (Gu et al., 2009).

To future proof livestock species for the challenges in the coming years, researchers in livestock genomics have been investigating a number of various traits in addition to economically important ones. About 250 - 500 liters of methane gas per day are generated by ruminant livestock (Johnson and Johnson, 1995). Methane, one of the green house gases is a major contributor to global warming. Genomic studies to select cattle population with a potential to reduce enteric emissions of methane and increase feed efficiency has been initiated (Basarab et al., 2013; de Haas et al., 2011). To compensate for the major climatic changes in the upcoming decades, researchers have also identified genomic markers for high milk production under climate change scenarios (Hayes et al., 2009). Based on the literature citations above, it can be concluded that although major consideration in livestock genomics is given to genomic selection for economically important traits, researchers are also examining various other genetic aspects related to animal welfare, health and adapting livestock species for new challenges in the future.

2.2 Data resources and analysis approaches in livestock genomics

2.2.1 Data resources

Similar to model organism genomics, major sources of data in livestock genomics are the standard biological databases. Ensembl database[2] holds genome assemblies of livestock species such as cattle, chicken, duck, horse, pig, sheep and turkey[3]. In addition to assembled genomes in Ensembl databases, NCBI databases[4] have large volumes of nucleotide, protein and gene annotation data related to livestock genomics. Moreover, the amount of data available for livestock species in public databases have been on the rise. This growth of publicly available livestock genomic data can be illustrated using an example. Figure 2.3 shows the growth in number of gene annotations available in NCBI Entrez gene database[5] for livestock species over a timespan of 10 years. As the figure shows, there has been an increase in the number of gene annotations available for livestock species and also the number of livestock species for which gene annotation information is available. With the advent of high-throughput technologies in genomics, the amount of publicly available gene expression data for livestock genomics species have also been on the rise. Table 2.1 shows

[2]http://www.ensembl.org/index.html last accessed March 13, 2014
[3]http://www.ensembl.org/info/about/species.html last accessed March 13, 2014
[4]http://www.ncbi.nlm.nih.gov/guide/all/ last accessed March 13, 2014
[5]http://www.ncbi.nlm.nih.gov/gene/ last accessed March 13, 2014

the statistics of publicly available genomic, proteomic, functional annotations and expression data for three livestock species: cattle, pig and chicken.

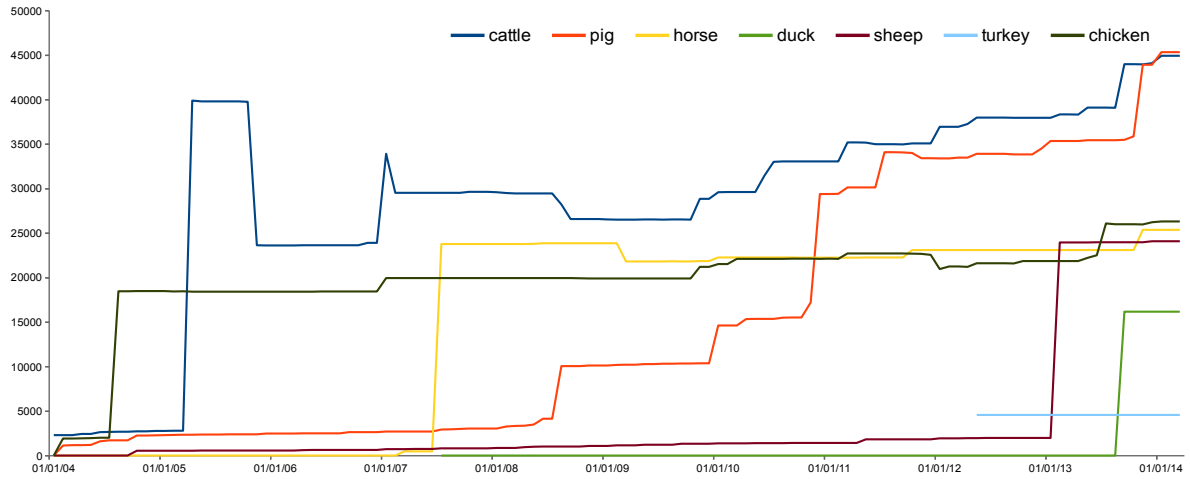

Figure 2.3: Number of gene annotations available in NCBI Entrez gene database for major livestock species. Figure shows the growth in number of gene annotations over a period of 10 years from 2004 to 2014. The statistics include all the gene annotation information, including those of genes withdrawn from major genome release. Data collected in March 2014.

Along with the traditional set of public databases, livestock genomics community also maintains a set of custom databases to store livestock specific data. Chief among them is the animalgenome[6] repository maintained by the National Animal Genome Research Program (NAGRP) of the U.S Department of Agriculture (USDA). Animalgenome acts as a repository for livestock specific databases, genome maps and other resources. At present, this repository stores genomic data from cattle, chicken, pig, horse and various fish species. This repository also stores custom animal genome annotation tracks including cattle, chicken, horse, pig, sheep and fish species[7] and hosts a BioMart server for livestock species[8]. Quantitative trait loci (QTL) information related to various favorable traits in animals is a characteristic feature in livestock genomics and to store and query through these QTL related information, Animal QTLdb[9] (Hu et al., 2013b) has been developed. This database collects all the publicly available QTL data, copy number variations (CNVs) and association data either from published literature or from laboratory reports subjected to publication and collects more than 50 parameters for a single QTL. The linkage map associated with QTLs can display QTL distances in either centiMorgans (cM) or corresponding physical locations in base pairs (bp) (Hu et al., 2013b). Table 2.1 contains the number of various QTLs and related traits deposited in Animal QTLdb for the livestock species cattle, pig and chicken. Along the lines of Animal QTLdb, another QTL database, Bovine QTL Viewer[10] was developed to store QTL information related to economically important traits such as weight gain, milk fat content and intramuscular fat in bovine (Polineni et al., 2006). This database is based on data from other databases such as INRA BOVMAP[11] and USDA-MARC (Kappes et al., 1997) and

[6]http://www.animalgenome.org/ last accessed March 13, 2014
[7]http://www.animalgenome.org/gbrowse/ last accessed March 14, 2014
[8]http://www.animalgenome.org:8181/ last accessed March 14, 2014
[9]http://www.animalgenome.org/cgi-bin/QTLdb/index last accessed March 14, 2014
[10]http://genomes.sapac.edu.au/bovineqtl/home.php last accessed April 8, 2014
[11]http://locus.jouy.inra.fr/cgi-bin/bovmap/intro2.pl last accessed April 2, 2014

mainly consists of an integrated QTL databases and a QTL viewer to display QTLs based on chromosomal position (Polineni et al., 2006). The QTL traits are divided into categories including behavior linear characteristics, body conformation general characteristics, body conformation linear characteristics, carcass quality, mastitis, milk fat, milk protein, milk yield, parasite load, parasite resistance, pigmentation and red blood cell mass. A web based tool AnnotQTL [12] (Lecerf et al., 2011) was developed to assist researchers to characterize and select candidate genes from a given QTL region. AnnotQTL is designed to work with data from livestock species including cattle, pig, chicken, horse and dog integrating data from external databases including gene annotation from biological databases, Gene Ontology annotations and SNPs along with QTL data (Lecerf et al., 2011). SNPchiMP[13] (Nicolazzi et al., 2014) is an open access database designed to manage and resolve the ambiguities in SNP co-ordinate mappings between reference genome and various SNP chips. Currently, this database is designed to work only with bovine genome and integrates data from dbSNP builds 136 and 137 along with Illumina SNP chip data and Affymetrix chip data (Nicolazzi et al., 2014).

A trait correlation database, CorrDB[14] (Hu et al., 2013a) has also been developed to store and search various publicly available genotype-phenotype correlation data. As per the current statistics, the database holds 3,635 correlation data points on 276 economically important traits related to milk production, meat production, growth and health in cattle. To provide a repository for quantitative trait loci related to dairy cattle, a QTL database[15] was created for cattle dairy production traits (Khatkar et al., 2004). The dairy production related traits stored in this database are: milk yield, milk composition (protein yield, protein %, fat yield, fat %), and somatic cell score (SCS) [16]. AgBase[17] (McCarthy et al., 2006) is a curated public resource for the functional analysis of various agriculture animal and plant genomes. AgBase uses controlled vocabularies from the Gene Ontology project and allows the users to search the database using plain text queries, perform sequence similarity searches, taxonomy and Gene Ontology based searches. ANEXdb[18] animal expression database was developed to account for inadequate direct gene/transcript annotations available for livestock species. ANEXdb integrates a microarray expression database ExpressDB and EST annotation database AnnotDB. ExpressDB hosts Affymetrix and two color microarray data and AnnotDB contains porcine ESTs from Iowa Porcine Assembly (IPA) (Couture et al., 2009). Following the footsteps of OMIM®[19] (Online Mendelian Inheritance in Man), a database of human diseases with a known genetic component, OMIA[20] (Online Mendelian Inheritance in Animals) has been developed to archive genetic data on various inherited disorders, single locus traits and genes in animals. At present, this database contains information on 214 animal species including livestock animals. Table 2.1 gives figures on various

[12]http://annotqtl.genouest.org/ last accessed April 2, 2014

[13]http://bioinformatics.tecnoparco.org/SNPchimp/home/ last accessed April 2, 2014

[14]http://www.animalgenome.org/cgi-bin/CorrDB/index last accessed March 14, 2014

[15]http://firefly.vetsci.usyd.edu.au/reprogen/QTL_Map/ last accessed April 4, 2014

[16]http://firefly.vetsci.usyd.edu.au/CMS/reprogen/QTL_Map/index.php?Page=Project+Description last accessed April 4, 2014

[17]http://agbase.msstate.edu/index.html last accessed March 14, 2014

[18]http://www.animalgenome.org/anexdb/index.php last accessed March 28, 2014

[19]http://www.ncbi.nlm.nih.gov/omim last accessed March 14, 2014

[20]http://omia.angis.org.au/home/ last accessed March 14, 2014

traits and disorders available in OMIA database for cattle, pig and chicken. ReCGiP[21] (Yang et al., 2010) is a database of candidate genes related to pig reproduction. The candidate genes in this database falls into six major porcine reproductive traits such as spermatogenesis, oogenesis, fertilization, preimplantation development, embryo implantation and placental development (Yang et al., 2010). The candidate genes in this database are literature derived using named entity recognition (NER) approach. In addition to candidate genes, gene co-occurrence network based on co-mentions in articles, Gene Ontology annotations, OMIM (human) and KEGG pathway mappings related to candidate genes can also be retrieved from this database (Yang et al., 2010). A genome-wide analysis was conducted to understand the patterns of transcript expression in pig (Freeman et al., 2012). A custom Affymetrix array was used to profile the transcriptome expressions and this genome wide expression atlas was generated based on expression data from 62 cell/tissue types. The results from this study are made publicly available[22] as a genome wide expression atlas and can be used for the functional annotation of uncharacterized genes based on cluster assignment of transcripts (Freeman et al., 2012).

ArkDB[23] is a public repository currently hosted by the Roslin Institute[24] for genome mapping data mainly from livestock species along with other animal species. ArkDB hosts chromosomal, linkage, cytogenetic and radiation hybrid maps for species such as cattle, chicken, pig, sheep, duck, horse and various fish species. Similar to human HapMap project, bovine and porcine HapMap projects analyzed the genome wide patterns in variations in cattle and pig genomes (Gibbs et al., 2009; Megens et al., 2010). ChickVD, a chicken sequence variation database was also created to facilitate functional and evolutionary studies in avian genetics (Wang et al., 2005). Similar to Encyclopedia Of DNA Elements[25] (ENCODE) (The ENCODE Project Consortium, 2004) human genome project to identify all functional elements of the human genome, AgEncode[26] project has been initiated to study functional elements in genomes of food animals including ruminants, swine, poultry and various fish species. Moreover, various protein - protein interaction databases also contain protein interactions from livestock species. Data statistics for cattle, pig and chicken protein interactions in databases IntAct and BioGRID interaction databases are given in Table 2.1.

In essence, conventional biological databases and several dedicated livestock genomics databases store biological, genomic and phenotypic data related to farm animal genomics and various production traits. To facilitate consistent and unambiguous communication between livestock genomics researchers and data repositories and to deal with the standardization issues related to livestock genomics data, Animal Trait Ontology for Livestock [27] (ATOL) was developed (Golik et al., 2012). The major domains of ATOL are: welfare trait, growth and meat production trait, mammary gland and milk production trait, egg trait, nutrition trait, fatty liver trait and

[21] http://klab.sjtu.edu.cn/MDpigs/index.html last accessed April 8, 2014
[22] http://www.macrophages.com/pig-atlas last accessed April 8, 2014
[23] http://www.thearkdb.org/arkdb/ last accessed March 28, 2014
[24] http://www.roslin.ed.ac.uk/ last accessed March 28, 2014
[25] http://www.nature.com/encode/#/threads last accessed April 2, 2014
[26] http://www.livestockgentec.com/media-and-outreach/conference/2-uncategorised/203-eu-us-animal-biotechnology-working-group-agencode-workshop last accessed April 2, 2014
[27] http://www.atol-ontology.com/index.php/en/ last accessed April 2, 2014

reproduction trait [28]. The livestock species represented in ATOL include cattle, sheep, trout, rabbit, chicken, turkey and pig along with two model species mouse and zebrafish.

Table 2.1: Statistics for publicly available data in three major livestock species: cattle, chicken and pig, data as of March 2014. Data statistics for human is given for comparison purposes.

Data type	***B. taurus*** (cattle)	***S. scrofa*** (pig)	***G. gallus*** (chicken)	***H. sapiens***
Nucleotides	225,600	536,811	118,112	10,472,013
Nucleotide ESTs	1,559,498	1,669,349	600,434	8,704,884
Proteins (NCBI)	123,472	56,949	52,265	851,871
Proteins (SwissProt)	5,984	1,413	2,257	20,266
SNPs	22,055,952	28,665,189	9,415,942	73,362,051
microRNAs (miRBase)	783	326	996	2,578
Functional annotations				
GO gene product annotations	20,045	19,749	13,106	44,900
KEGG pathway annotations	279	279	162	284
Expression datasets				
GEO Datasets	6,769	8,848	4,653	648,818
SRA Experiments	765	1,966	364	143,402
ArrayExpress experiments	20	27	56	2,394
Animal QTLdb data				
QTLs	8,305	9,862	3,919	NA
Traits	467	653	297	NA
OMIA data				**OMIM data**
Total traits/disorders	452	227	209	22,304
Mendelian traits/disorders	189	49	127	1,707
Mendelian trait/disorder with key mutation known	92	23	39	1,856
Protein - protein interaction data				
IntAct database	1,562	119	715	144,630
BioGRID database	309	57	6	147,806

2.2.2 Analysis approaches in livestock genomics

The data analysis approaches in livestock genomics have mostly followed the genetic technologies used for data generation. The current analysis approaches used in livestock genomics can be broadly classified into a three major groups: (i) statistical modeling of traits (ii) biomarker analysis and (iii) mathematical and computational modeling.

2.2.2.1 Statistical modeling of traits

Statistical modeling of traits is primarily used to model the the effects of various biomarker candidates either for genomic selection or for the estimation of breeding values. In these studies, biomarkers from either the analysis of individual biomarker candidates or from high-throughput studies are used. In general, statistical models and selection theory in animal breeding follows the

[28] http://www.atol-ontology.com/index.php/en/les-ontologies-en/visualisation-en last accessed April 2, 2014

infinitesimal genetic model of quantitative genetics, where it is assumed that a trait is affected by a large number of biomarkers with very small and additive effects (Dekkers, 2012). Genomic selection is defined as a marker assisted selection method in which genetic markers covering the whole genome are used and the markers are assumed to be in Linkage Disequilibrium (LD) with QTL to minimize the number of estimated effects per QTL (Goddard and Hayes, 2007). Genomic breeding values (GEBVs) are calculated as the sum of the effects of various biomarkers or the effects of these biomarkers across the whole genome and tries to capture the QTLs contributing to that trait (VanRaden et al., 2009). The effects of such biomarkers are first inferred in large populations with phenotype information and subsequently, only the effects from biomarkers are used to compute GEBV. These GEBV estimations have been shown to increase the accuracy of genetic merit (VanRaden et al., 2009). According to Goddard and Hayes (2007) the three major steps involved in the statistical analysis to estimate GEBV are:

(i) assessing QTLs through various markers

(ii) estimating the effect of QTLs on genotypes and

(iii) summation of QTL effects for candidate selection and GEBV estimation (Goddard and Hayes, 2007).

To estimate breeding value on selection of candidates, linear mixed model methodology have been used in livestock breeding programs (Dekkers, 2012). To predict the effect of SNPs in genomic estimated breeding values (GEBVs) a method called BLUP (best linear unbiased prediction) is used. In this linear modeling approach SNP effects are modeled as zero mean non random variables with a common effect variance and it is assumed that these variables are independently and identically distributed (Meuwissen et al., 2001). A number of genome wide association studies (GWAS) published in livestock genomics used linear mixed models to estimate genomic breeding values based on SNP genotype and related traits. Data from Illumina BovineHD Genotyping BeadChip assay and phenotypic traits were analyzed using a linear mixed model approach to assess the effect of SNPs in estimated growth related breeding values in bovine (Utsunomiya et al., 2013). Similarly, another study also used linear mixed model to estimate the effect of SNPs in GEBVs related to production traits in cattle (Guo et al., 2012). In a related approach, conventional pedigree based relationship matrix in BLUP models are substituted with genomic relationship matrix (GRM) defining additive covariance between animals derived from high density SNP genotyping technologies, giving rise to a method known as Genomic Best Linear Unbiased Prediction (GBLUP) (Dekkers, 2012). In addition to linear models, Bayesian hierarchical models are also used in the estimation of breeding values. There are two levels of data modeling in these Bayesian approaches: first, at the level of data and second at the level of variances at chromosome segments (Meuwissen et al., 2001). Bayesian least absolute shrinkage and selection operator (Bayesian LASSO) method was also used to fit marker effects to a regression model. In this approach Bayesian LASSO is used to generate a regression model in which effects of various markers, predictors and other covariates are considered jointly (de los Campos et al., 2009). In addition to the methods described here, several additional methods were also developed

to estimate GEBV (Hayashi and Iwata, 2010; Meuwissen et al., 2009; Shepherd et al., 2010; Sun et al., 2012; Yi and Banerjee, 2009). To serve as a benchmark dataset to compare genomic prediction methods, a pig dataset termed PIC dataset has been made available (Cleveland et al., 2012). PIC dataset was generated by a pig genus company called PIC[29] and comprises of data from a population of 3,534 pigs. The dataset contains high density genotypes generated on Illumina PorcineSNP60 chip and five purebred traits with heritabilities ranging from 0.07 to 0.62 (Cleveland et al., 2012).

2.2.2.2 Biomarker analysis

Identification and investigation of single or multiple candidate biomarkers related to a phenotypic trait have long been practiced in livestock genomics. The biomarkers could be genes, proteins, associated polymorphisms, metabolomes or QTLs related to a phenotypic trait. In livestock genomics, investigation of biomarkers can be categorized into (i) candidate biomarker analysis and (ii) high-throughput studies. In candidate biomarker analysis, the activity or effect of a biomarker under one phenotypic case is compared against the other to understand the role of/effect of the biomarker in the phenotype. For example, Islam et al. (2013) studied the age related expression of porcine T helper related cytokines by comparing the expression of candidate biomarker genes such as IL-2, IL-4, IFN $-\gamma$ and IL-10 in pigs under various age groups (Islam et al., 2013).

Following the footsteps of human genomics and medicine, livestock genomics also began using high-throughput technologies such as microarray, SNP chips and NGS technologies to understand the genetics elemental to various phenotypic traits. The choice of the high-throughput platform used depend upon the nature of the investigation, species, model system, tissue or cell type under investigation and the economics (Smith and Rosa, 2007). As per the current statistics in GEO database, there are 130 high-throughput platforms for bovine, 89 platforms for porcine and 12 platforms for chicken. Since there was no comprehensive information on the high-throughput data analysis approaches used in livestock genomics, the material and method section from a random collection of 50 full text articles in livestock high-throughput studies (random corpus) were manually analyzed. Figure 2.4 gives an overview on the major data analysis approaches used in livestock genomics. Additional details, such as the species used, high-throughput platform, analysis approaches and Pubmed identifiers (Pmids) are given in Appendix Table 1.

Figure 2.4 indicates that the analysis of differentially expressed genes/transcripts is one of the major themes in livestock high-throughput data analysis. The term 'differential expression analysis' is used to indicate a broad range of statistical approaches from standard R/Bioconductor packages for microarray/RNA-seq expression data analysis to Student's t-test, Wilcoxon ranksum test and other statistical tests used to compute the difference in gene/transcript expression values in two or more phenotypes. A detailed table giving the frequency of each analysis approach mention in the random corpus is given in Appendix Table 2. Although some of this statistical methods are individually listed in Appendix Table 1, the broad classification 'differential expression analysis' was necessary since a number of articles in the random corpus did not detail the methods used to identify the differentially expressed genes. In addition to statistical tests for group comparison

[29]http://www.genusplc.com/about/pic.aspx last accessed April 8, 2014

such as ANOVA, Student's t-test, Fisher's exact test, Chi-squared test and Mann Whitney U test, dimension reduction technique PCA (Principal Component Analysis) and clustering methods such as hierarchical clustering, k-means clustering and other analysis methods including interaction network analysis and correlation network analysis were also used in high-throughput studies in livestock genomics (Figure 2.4, Appendix Table 1).

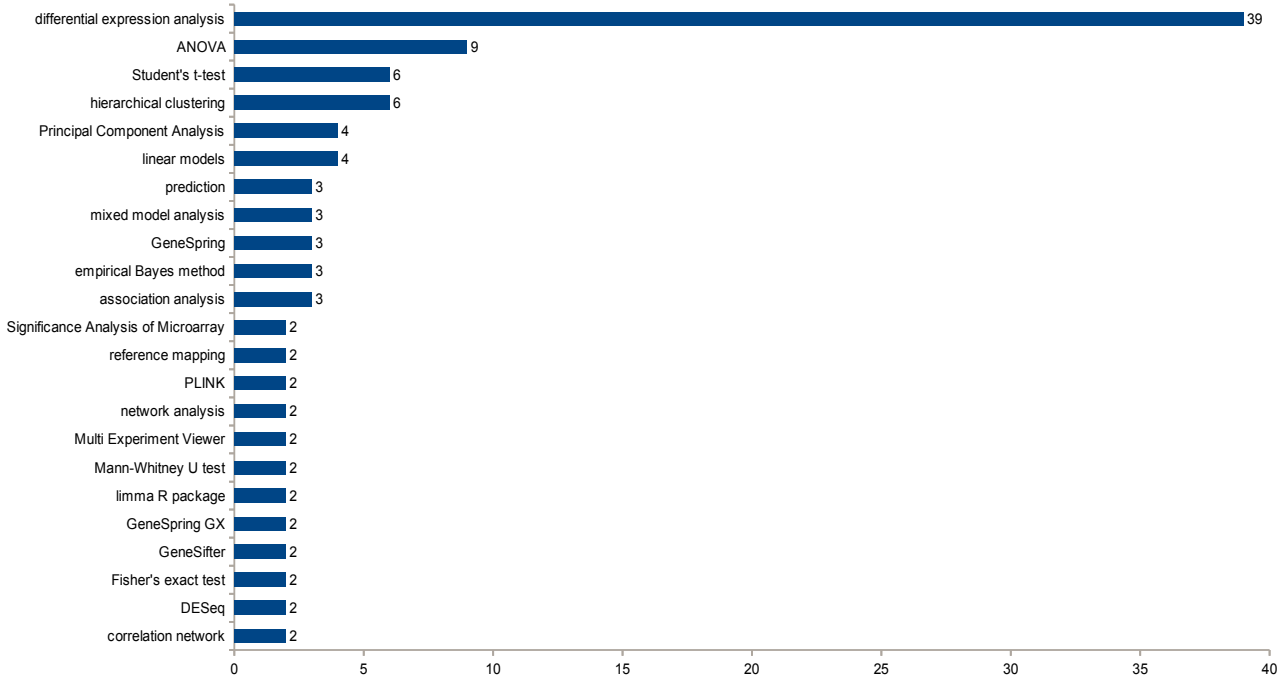

Figure 2.4: Major analysis approaches and methodologies used in high-throughput studies in livestock genomics. Data from manual analysis of the material and method section of 50 full text articles. The figures on the barplot indicate the frequency of the analysis approach/concept mention in the corpus.

A number of studies in random corpus used de novo assembly, reference mapping and prediction methods to identify novel microRNAs in livestock species (Appendix Table 1). In a meta analysis study, te Pas et al. (2012) used publicly available microarray data to identify the common differentially expressed genes in a number of chicken salmonella experiments (te Pas et al., 2012). In this study, the authors normalized the expression datasets using R limma package and meta analysis was carried out using metaMA (Marot et al., 2009) R package and compared the list of differentially expressed genes (DEGs) common in all the experiments, unique to individual experiments and unique to the combined study and identified that a number of host metabolic pathways and functions were similar in different chicken lines when infected with divergent Salmonella serovars (te Pas et al., 2012). For ranking candidate genes associated with quantitative traits and diseases in livestock species, Jiang et al. (2012) implemented a network based gene prioritization method (Jiang et al., 2012). In this method, using a set of genes derived from text mining, genome wide expression profiling, ortholog mapping and network based prioritization approach, a relevancy score was calculated and was finally aggregated with the phenotypic data and using this analysis approach, a number of candidate genes for bovine mastitis were prioritized (Jiang et al., 2012). In an additional study, a partial correlation and information theory approach was used to infer gene correlation networks and co-expression clusters in bovine

skeletal muscle and adipose tissue based on gene expression data from 822 genes in 9 experiments and 47 conditions (Reverter et al., 2006). In a yet another study, Pearson correlation based weighted gene coexpression analysis (WGCNA) (Langfelder and Horvath, 2008) was used to derive gene co-expression clusters for beef marbling using data from multiple publicly available microarray datasets (Lim et al., 2014).

The analysis approaches in material and methods section the random corpus indicate that biomarker analysis approaches in livestock genomics mainly follows the classical methods to identify candidate biomarkers such as DEGs and associated SNPs. Although meta analysis approaches, interaction network analysis and literature mining approaches are also being used, the number of experiments utilizing these approaches are minuscule in comparison to conventional methods.

2.2.2.3 Mathematical and computational modeling

Besides statistical analysis for trait selection and biomarker analysis, mathematical and computational modeling approaches have also been used in livestock genomics. Doeschl-Wilson (2011) argues that in case of livestock host pathogen interactions, biomarkers alone cannot predict the most disease prone or infected animals with 100% accuracy and that mathematical host pathogen interaction models would be able to describe the root biological process related to disease mechanisms and how these processes change over time (Doeschl-Wilson, 2011). The mathematical models developed for studying host pathogen interactions can be divided into three categories:

(i) The first category consists of the mathematical models describing infection patterns and immune system dynamics within a host. These models are used to aggregate data from multiple studies into a comprehensive framework (Doeschl-Wilson, 2011).

(ii) The second category of models accounts for the underlying relationship between immunological pathways and biological processes related to survival or production. It is assumed in these models that when resources are scarce, trade-offs can occur between continuing survival/production related biological process and triggering an immune response (Doeschl-Wilson, 2011).

(iii) The final category of mathematical models for host pathogen interactions addresses the co-evolution between various livestock hosts and pathogens and tries to understand how the control mechanisms involved affect the genetics of hosts and pathogens (Doeschl-Wilson, 2011).

Although these models are grouped into three, it is possible that there are overlaps in the analysis methods used in these models. A schematic representation of the three different groups of mathematical models used in host pathogen interaction modeling is given in Figure 2.5. The mathematical methodologies used in host pathogen interaction studies can be differential equation systems, stochastic mechanistic models, cellular automata and agent based models or

bioinformatics and systems biology algorithms (Doeschl-Wilson, 2011). In addition to modeling host pathogen interactions, a systems biology based mathematical model was used to study the effects of multiple perturbations on bovine estrous cycle to identify the biological processes involved in the development of cystic ovaries (Boer et al., 2012).

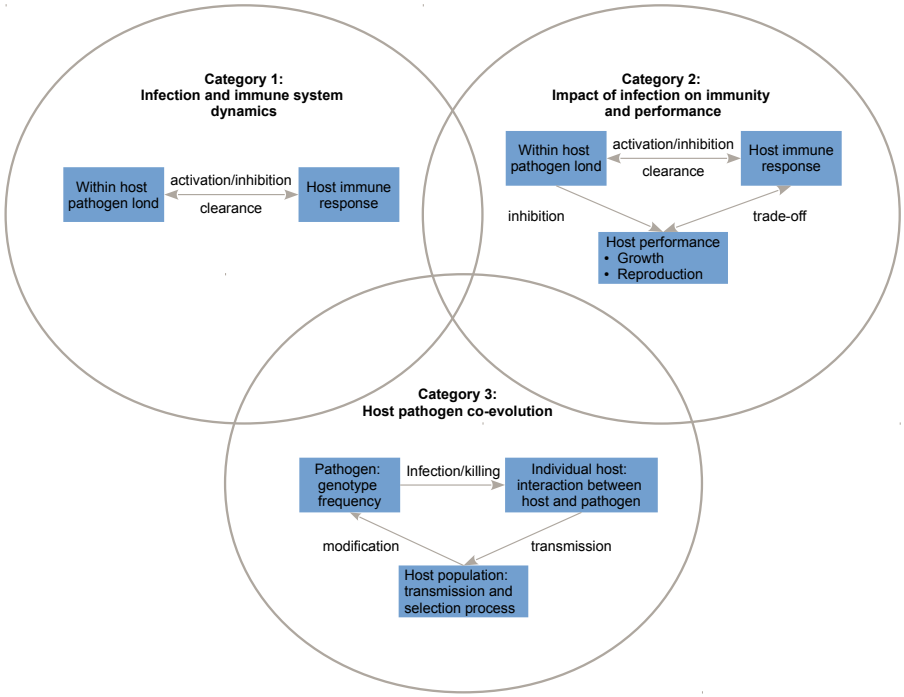

Figure 2.5: Three groups of mathematical models used to study host pathogen interaction models in livestock. Figure adopted from Doeschl-Wilson (2011).

In short, a large variety of diverse analytical approaches are used in livestock genomics to understand the relationship between behavioral patterns of biomarkers and phenotype under investigation. The trends in livestock genomics data analysis approaches hints that although a large number of analysis approaches are being used, conventional biomarker analysis and statistical modeling of economically important traits take the prime spots.

2.3 Androstenone and boar taint genomics

Boar taint is often described as an unpleasant smell or taste noticeable from meat products derived from un-castrated male pigs (Bonneau, 1982). Regulating boar taint is important to the pork industry since it was shown that the odor of boar taint causing compounds are likely to be detected by consumers (Bonneau et al., 1992). A major reason for boar taint is the accumulation of androstenone, a lipophilic sex steroid in adipose tissues of pigs. Androstenone is a male sex pheromone synthesized mainly in testis and metabolized in liver (Bonneau et al., 1992). The accumulation of androstenone in adipose tissues can be the result of either a high rate of testicular synthesis of androstenone or/and a low rate of hepatic degradation (Robic et al., 2008). One of the widely practiced methods to reduce boar taint is the surgical castration of piglets to limit the synthesis of androstenone (Haugen et al., 2012). But, representatives of European farmers, meat industry, retailers, scientists, veterinarians and animal welfare NGOs have issued a declaration

to end surgical castration of piglets without using anesthesia in European union by January 1, 2018 [30] thus creating a need to develop non surgical methods to limit androstenone content in porcine adipose tissues and hence reduce boar taint. The two proposed non surgical methods to reduce boar taint are: (i) the use of chemicals or drugs to reduce boar taint (Dunshea et al., 2001) and (ii) breeding for favorable characteristics to reduce boar taint (Frieden et al., 2011). In this regard, it should be noted that the European Food Safety Authority (EFSA) has already expressed concerns over consumer perception of meats from animals treated with chemicals and drugs to reduce boar taint (Spoolder et al., 2011).

To develop non surgical methods to reduce androstenone, it is necessary to understand the genetic mechanisms involved in the synthesis and degradation of androstenone. The enzyme cytochrome P450 11A catalyzes the cleavage of cholesterol to pregnenolone, the precursor molecule for androgen synthesis in testis (Robic et al., 2008). The synthesis of androstenone (5 α-androst-16-en-3-one) from pregnenolone in testis is catalyzed by the enzymes cytochrome P450C17 (CYP17A1) and enzymes of andien-β synthetase system such as cytochrome b5 (CYB5) along with other reductases (James Squires, 2010; Robic et al., 2008). In the final step of androstenone synthesis, the Δ4 double bond in 4,16-androstadien-3-one is reduced by the enzyme 5 α reductase (James Squires, 2010). A schematic representation of major steroid substrates and enzmes involved in androstenone synthesis is given in Figure 2.6. 3α-androstenol and 3β-androstenol are the final metabolites of androstenone in both testis and liver. In liver, androstenone under go Phase II conjugation reactions to form glucuronide conjugates and sulfoconjugates (James Squires, 2010).

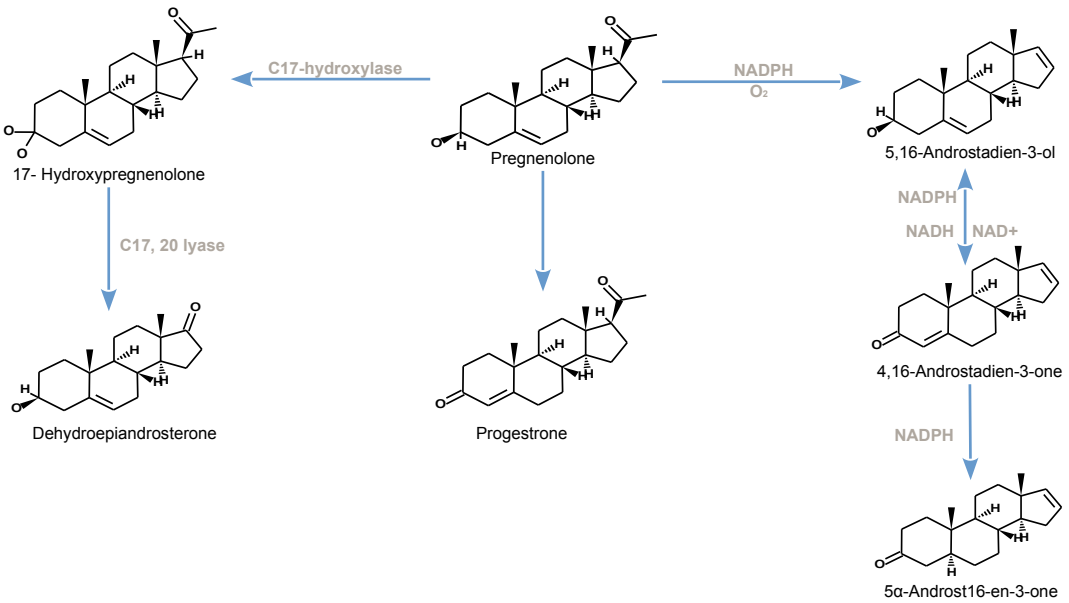

Figure 2.6: Schematic representation of androstenone biosynthesis in porcine testis. Figure adopted from James Squires (2010).

The rate of synthesis of androstenone is slow for young piglets and increases steadily during puberty (Andresen, 1976). A number of studies have already tried to understand the cellular mechanisms involved in androstenone metabolism. High expression of CYB5 was pointed out

[30] http://ec.europa.eu/food/animal/welfare/farm/initiatives_en.htm last accessed March 19, 2014.

as one of the reasons for the overproduction of 16-androstene steroids in testis (Davis and Squires, 1999). The sulfoconjugation of 16-androstene steroids in porcine testis and liver is mainly catalyzed by hydroxysteroid sulfotransferase enzyme (SULT2A1) (Sinclair et al., 2005). QTLs related to androstenone levels in pigs have also been under investigation. Androstenone related QTLs were identified on chromosomes 2, 4, 6, 7 and 9 in a cross between Large White and Meishan pigs (Lee et al., 2005). According to another study performed on an experimental cross between Large White and Meishan pig breeds, suggestive QTLs for fat androstenone were identified on pig chromosomes 3, 4 and 10 (Quintanilla et al., 2003). Additional QTL studies were carried out on Large White × Meishan cross (Boulliou-Robic et al., 2011), Norwegian Landrace and Duroc breeds (Grindflek et al., 2011) and Duroc, Landrace, and Yorkshire breeds (Gregersen et al., 2012). High-throughput microarray gene expression studies have been performed to understand the difference in gene expression profiles in testis tissues of pigs with extreme high and low levels of androstenone (Leung et al., 2010; Moe et al., 2007b). Additionally, transcriptome profiles of a number of candidate genes in testis tissues of pigs with large difference in androstenone measurements has also been investigated (Grindflek et al., 2010). A GWAS experiment performed on pure bred animals from a composite Duroc sire-line identified candidate SNPs associated with androstenone trait on porcine chromosomes 1 and 6 (Duijvesteijn et al., 2010). An in-house study using data from RNA-seq technology has also been performed to identify candidate biomarkers for varying levels of androstenone in porcine testes samples (Gunawan et al., 2013).

In comparison to the number of studies done to understand testicular androstenone synthesis, fewer studies have been carried out to understand the hepatic androstenone metabolism. In liver, breed differences in the expression of androstenone metabolizing enzymes 3β-HSD and SULT2B1 have been reported in Norwegian Landrace and Duroc pigs (Moe et al., 2007a). Nicolau-Solano et al. (2006) asserted that the liver specific regulation of 3β-HSD expression could explain the low rate of hepatic androstenone metabolism based on a study conducted on 13 Large White and Meishan pigs (Nicolau-Solano et al., 2006). Another study also pointed out the relevance of 3β-HSD enzyme in hepatic androstenone metabolism based on the investigation in Large White and Meishan breeds (Doran et al., 2004). Experiments performed on Yorkshire pigs lead to the conclusion that the enzyme hydroxysteroid sulfotransferase (HST) could be responsible for the sulfoconjugation of all 16-androstene steroids including androstenone in liver (Sinclair et al., 2005). A microarray study performed on two pig breeds, Norwegian Landrace and Duroc have identified a number of candidate genes responsible for hepatic androstenone metabolism in both breeds and by studying the gene expression profiles in two breeds, the authors also tried to identify the breed differences in hepatic androstenone metabolism (Moe et al., 2008). The in-house RNA-seq experiment conducted on a sample population of Duroc × F_2 also identified a number of candidate genes and polymorphisms that might be responsible for hepatic androstenone metabolism (Gunawan et al., 2013).

2.4 Data mining and Knowledge discovery

Data mining is the process of examining volumes of data in multiple contexts to abstract the data into useful information (Palace, 1996). The five major components of data mining are: extraction and transformation of data, data storage and management, data access provisions, data analysis and data/result presentation (Palace, 1996). There are two major categories of data mining tasks: descriptive and predictive (Han and Kamber, 2011). Descriptive data mining is used to identify the general properties of the data where as predictive data mining is used to infer trends from data and to generate predictions (Han and Kamber, 2011). The relationships identified in data mining applications between data points can be divided into four major types (Palace, 1996):

(i) Classes: grouping of data into multiple classes. In a biomedical scenario, presence/expression of certain specific biomarkers in tissue samples can be used to classify individuals as either healthy or diseased.

(ii) Clusters: data points are grouped according to the relationship with other data points. In life sciences, data clustering can be used to identify groups of biomarkers with similar expression profiles.

(iii) Associations: data mining technologies can be used to identify associative patterns or rules among data points and is not equivalent to Genome wide association analysis in a genomics context. In genomics context, association mining is related with using expression profiles of genes for phenotype disease classification (Creighton and Hanash, 2003).

(iv) Sequential patterns: data mining applications can be used either to identify or to predict patterns and trends in data. In biomedical realm, an example usage is the time series analysis of expression patterns or prediction of changes in cellular interaction patterns during disease progression.

Knowledge discovery, also referred as Knowledge Discovery in Databases (KDD) is a concept that is discussed along with data mining. Knowledge discovery is defined as the process of identifying potentially useful, innovative, credible and ultimately understandable patterns of data (Fayyad et al., 1996c). Data mining is one of the many steps in a knowledge discovery process and at a basic level, knowledge discovery primarily deals with the development of methods and techniques to process and make sense of the data (Cios et al., 2007; Fayyad et al., 1996b). The basic steps in a knowledge discovery process are: developing an understanding of the application domain, creating a target data set, data cleansing and preprocessing, data reduction and projection, choosing data mining task, choosing data mining algorithm, data mining, interpreting the mined patterns and consolidating the knowledge discovered (Fayyad et al., 1996b). Figure 2.7 gives a schematic representation of the major steps involved in knowledge discovery process. A key difference between knowledge discovery process and data mining is that the term knowledge discovery is used to denote the entire process of discovering useful knowledge from data where as data mining is the application of algorithms to identify specific patterns from the data. Data

mining is an inherent part of knowledge discovery, but knowledge discovery emphasizes the fact that knowledge is the end product of a data driven discovery (Fayyad et al., 1996b).

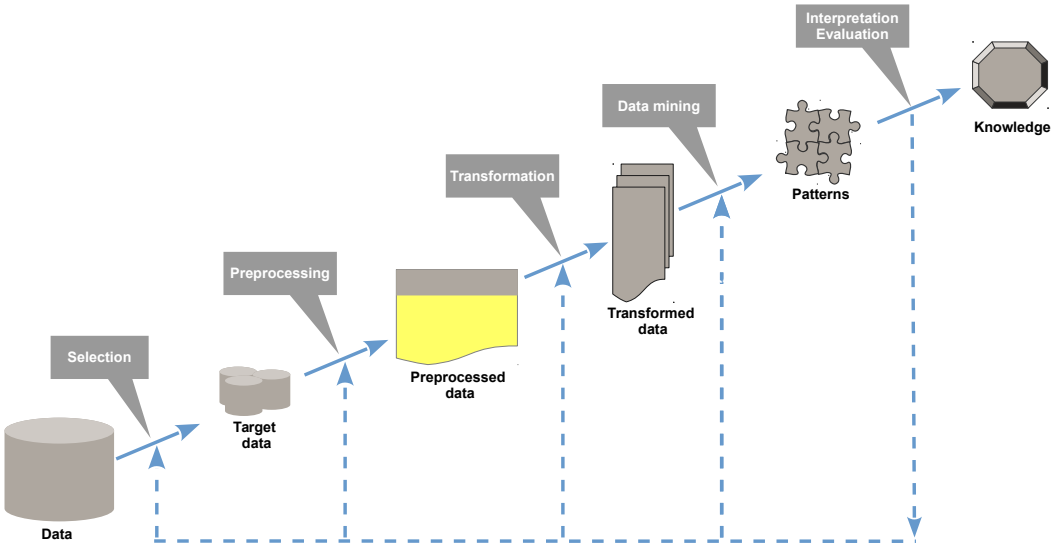

Figure 2.7: Schematic representation of knowledge discovery process. Figure adopted from Fayyad et al. (1996b).

In biology, data mining and knowledge discovery methods are used in a wide variety of applications. However, prior to the application of data mining and knowledge discovery concepts in biology, a number of biological data related constraints have to be taken into account (Brusic and Zeleznikow, 1999). These constraints are detailed below:

Complexity of biological data: All biological data including expression measurements and interaction data are derived from complex biological systems. Currently, the data structures in use fail to encode the underlying hierarchical and interconnected biological processes, but are assumed to be a part of the background knowledge. In this situation, understanding the context of data generation is a prerequisite for the correct selection of data analysis methods (Brusic and Zeleznikow, 1999).

Fuzziness of biological data: A number of experimental methods are used in biological sciences to generate quantitative results. It can happen quite often that the results from a number of different experiments on the same phenotype are partially overlapping, but not fully. Even replicating the same experimental setup would not necessarily yield identical results since experimental outcomes in biology can vary depending on a number of variables such as differences in temperature or pH, difference in culture media, cells or cell lines and technical variability related to the chemicals and instrument set ups used. These biological and technical variations in biological experiments leads to the overall fuzziness of the data and therefore quantitative measurements in biological data are only approximate measurements. In order to select appropriate analysis method and tools it is crucial to consider this overall fuzziness of biological data (Brusic and Zeleznikow, 1999).

Biases and misconceptions: Data generated in biology are subject to biases either due the inherent properties of the system under consideration or due the presence of related motifs or historical reasons. In biological research, certain fields are analyzed in depth where as some other

fields remain relatively unexplored. Typically, new researches are based on previous results and conclusions. Researchers try to explain biological systems using a set of rules and it can happen that further research will be directed towards the application of these rules. If these defined rules explain only a part of the possible behavior of the system, the rule abiding part of the biological system will be explored in detail where as the rest of the biological system will be under explored. In a similar manner, understanding biological system with limited data can lead to over/under simplification of the system and hence can lead to errors. As a result of these issues, a careful assessment of the data is necessary before setting up analysis pipelines (Brusic and Zeleznikow, 1999).

Effect of noise and errors: The major sources of errors and noises in biological data are experimental setups, technical variability in chemicals and instruments used, differences in data measurement, reporting, annotation and processing techniques. Due to the complexity in biological systems, it is difficult to set an error level for biological experiment. Although it is not possible to eliminate errors or noises from biological data, selection of data analysis can be guided by the estimation of noise levels in data (Brusic and Zeleznikow, 1999).

Some of the data mining and knowledge discovery application fields in biology are: gene expression analysis, protein/RNA structure prediction, phylogenetics, identification of sequence and structural motifs, genomics and proteomics, gene finding, RNAi and microRNA analysis, drug design, modeling of biochemical pathways and text mining in biology (Zaki et al., 2003). According to Nguyen et al. (2013) application of knowledge discovery and data mining models is necessary to extract information and knowledge from biomedical data on complex biological systems and to understand the progression of complex diseases (Nguyen et al., 2013). In conclusion, knowledge discovery and data mining are the background themes in a number of analysis approaches in biology. A brief literature review on the methods and tools are given in section 2.5.1.

2.5 Integrative analysis approaches

In life science context, integrative analysis approaches refers to the integration of results or datasets from a number of experiments or data resources to understand the complex systems in living beings. A major factor fueling integrative data analysis approaches is the technological advancements in profiling various cellular properties on a genome wide scale. Advances in whole genome profiling technologies have lead to an increase in the availability of genomic and proteomic datasets including epigenomic data, transcriptomic data, sequence variation data and interactome data (Hawkins et al., 2010). The primary objective of integrative data analysis approaches is to identify the hidden relationships and infer new knowledge based on various biological systems (Kumar, 2011). This section discusses major sources of biomedical data for integrative data analysis methods followed by major concepts used in integrative data analysis approaches and finally reviews the state of the art methods in integrative analysis approaches.

High throughput technologies have enabled the genome scale mapping of DNA methylation events and covalent modifications (Johnson et al., 2007; Lister et al., 2009; Ren et al., 2000). Histone modifications of a genome can be identified by Chromatin immunoprecipitation methods

(ChIP-chip or ChIP-seq) (Park, 2009) and chromatin structure can be determined by DNase I Hypersensitivity Site technologies such as DHS-Seq or DNase-Seq and DHS-chip (Boyle et al., 2008). The growth in transcriptomic data was initially due to the use of microarray chips for profiling the transcriptome abundance under various phenotypic conditions. Although microarray chips have given way for Next generation sequencing technologies such as RNA-seq recently, a large volume of publicly available transcriptome profiles were generated using microarray technologies. In addition to estimating transcriptome abundance, RNA-seq can also detect non coding mRNAs and gene fusion events (Maher et al., 2009). The major aim of sequence variation study is to link a genetic variant to a phenotype. The growth in sequence variation data can be attributed to the use of SNP genotyping arrays and more recently, to the surge in the use of NGS technologies (Hawkins et al., 2010). Interaction datasets in life sciences can either refer to genetic or physical interactions in the genome or proteome level. These datasets are mainly generated by means of large scale genome or proteome wide experiments and the major sources of these interaction are biological databases specialized for archiving interaction data. In protein - protein interaction networks, nodes represent proteins where as edges represent the physical interaction between the proteins (Amar and Shamir, 2014). In case of genetic interactions, nodes represent genes and edges represent the response of the organism to knock-out experiments (Amar and Shamir, 2014). Published scientific literatures are yet additional sources of information in the biomedical realm. Various systems have already been developed to identify and extract the various biomedical concepts and the relationship between them in published articles (Krallinger and Valencia, 2005).

Since datasets from all these high throughput technologies explain different sections of a cellular machinery, integrating and analyzing these datasets together will help to reveal the co-ordination between various cellular features such as gene transcripts, polymorphisms, gene and proteomic interactions and epigenetic effects in the fundamental genome mechanisms and in the manifestation of a disease or a phenotype. According to Chen and Hofestädt (2006) integrating information from various metabolic systems and the interactions between them is the key to systems analysis strategy. It is important to gain an understanding of the relationship among genomic, proteomic and pharmacological components of the biomedical system to devise treatment strategies (Chen and Hofestädt, 2006). Figure 2.8 gives a schematic representation of the integrative biomedical systems architecture as proposed by Chen and Hofestädt (2006) for systems analysis strategies. Two major underlying concepts in integrative approaches used in life sciences are knowledge driven analysis and data driven analysis. As the name suggests, knowledge driven approaches refers to the usage of existing knowledge in association with genome wide datasets to reach conclusions. These existing knowledge could be either literature based evidences retrieved from scientific literatures in life sciences or the wealth of knowledge in various biological databases hosting gene/protein interaction information, metabolic networks or various mathematical models developed (Chang et al., 2008) based on existing information. Data driven approaches on the other hand, relies on integrating multiple datasets or data resources so as to identify the common patterns in the data.

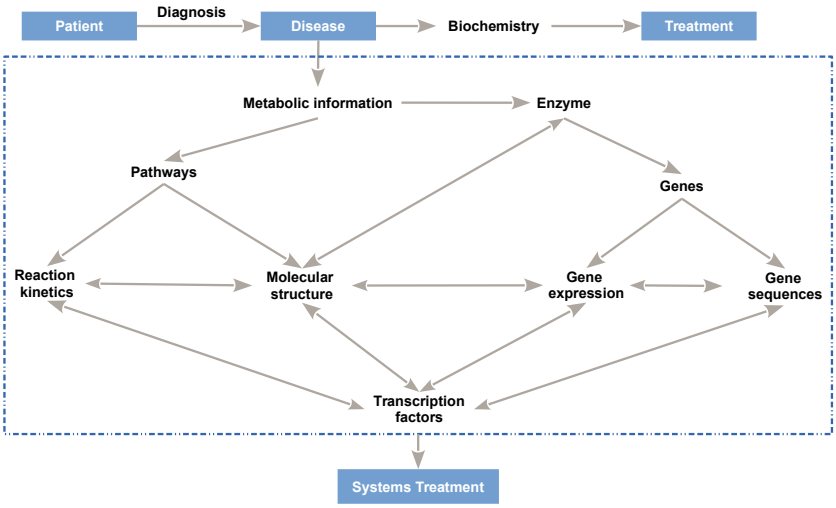

Figure 2.8: Schematic representation of biomedical systems architecture as proposed by Chen and Hofestädt (2006). Figure adopted from (Chen and Hofestädt, 2006)

One of the major challenges in integrative analysis approaches is data integration. Three major data integration approaches in integrative analysis strategies are:

(i) data complexity reduction

(ii) unsupervised integration and

(iii) supervised integration (Hawkins et al., 2010).

Data complexity reduction techniques are mainly performed to reduce the complexity of the experiment datasets used. Since high throughput technologies like microarray and NGS generate thousands or millions of probe reads or short read sequences for a given cell/tissue type, it becomes difficult to account for all these data points and to encode the behavior in a model. An approach to reduce the complexity of such data sets is to abstract the datasets to a number of genomic regions of strong signal and yet another approach would be to perform an intersection analysis on multiple experiments results or datasets (Hawkins et al., 2010). Unsupervised integration methods, the second class of integrative analysis approaches are based on the assumption that relevant patterns occur commonly in data and hence can be identified. A commonly used unsupervised method is clustering. Clustering approaches are employed to identify the common patterns of gene expression, epigenetic states and interactomes. Unsupervised integration methods can often be used to identify correlations among different experiments (Hawkins et al., 2010). Although unsupervised methods can identify the novel patterns in data and generate hypothesis, the disadvantage is that the novel patterns alone cannot advance the knowledge in biomedical sciences. The third integrative analysis approach, supervised integration methods mainly focus on hypothesis testing by incorporating additional datasets or experiments. Supervised integrative analysis approaches begins with a prediction based on an observation and ends with a test for the prediction (Hawkins et al., 2010).

Components of the cellular system carrying out various biological functions are thought to have a

modular organization (Alon, 2003; Hartwell et al., 1999). These modular organization can be in terms of genes, proteins, enzymes, metabolites or a combination of a number of these biological components. A number of integrative analysis approaches in systems biology are devoted to identifying these modules/functions of biological components. According to Mitra et al. (2013) the module discovery methods can be grouped into four broad classes:

(i) identification of active modules

(ii) identification of conserved modules

(iii) identification of differential modules and

(iv) identification of composite modules (Mitra et al., 2013).

Identification of active modules refers to overlaying interaction networks with expression profiles of genes, transcripts, proteins or other biological molecules. The computational integration of expression profiles and interaction networks have been used widely to derive context dependent interaction sub-networks (modules) (Mitra et al., 2013). Identification of conserved modules refers to the identification of components or interactions preserving the core biological functions across different biological conditions or species of animals. By the identification of such conserved modules it would be possible to understand the basic biological processes and to predict the evolutionary basis that leads the generation of such conserved modules (Mitra et al., 2013). The methods classified under the identification of differential modules try to identify the functional modules showing different patterns of interaction under various phenotypic conditions. In these cases, the modules of interest are those modules that are differently present, absent or modified in various phenotypic conditions under investigation (Mitra et al., 2013). In biological systems, protein - protein interaction networks, gene regulatory networks and metabolic interaction networks define different parts of the biological system. Identification of composite modules refers to the identification of modules that jointly form interaction, regulatory or metabolic networks (Mitra et al., 2013). In addition, various analysis strategies can be strung together into multiple integrative module analysis methods leading to increasing complexity and sophistication (Mitra et al., 2013).

2.5.1 Literature review: Integrative analysis approaches

A prime argument for integrative analysis approaches is that by integrating broad spectrum of data it would be possible to interpret the intrinsic mechanisms underlying the genetic machinery of a biological system in a particular phenotypic condition or a disease state. Some of the state of the art approaches used in integrative analysis approaches are detailed in this subsection.

D'Antonio and Ciccarelli (2011) studied the evolutionary change in gene duplicability by comparing gene and network properties of four species of organisms: *E. coli*, *S. cerevisiae*, *D. melanogaster* and *H. sapiens* representing increasing evolutionary complexity. This integrative investigation of genes and protein protein interaction networks in these species revealed that in all the species, ancestral singleton hubs are at the core of the networks and are highly conserved, where as the

genes acquired as a result of progressive evolution encoded less connected and less central proteins (D'Antonio and Ciccarelli, 2011). By linking this multi species integrated interaction network to human cancer biology, the authors revealed that cancer mutations mainly affects either the evolutionary conserved ancestral hub genes, essential for the basic functioning of the cell or a second group of genes that are involved in regulatory processes (D'Antonio and Ciccarelli, 2011). For investigating the genetics behind various metabolic disorders, an analysis approach was developed to merge several biomedical information resources together (Chen and Hofestädt, 2006). The model implemented in this work made use of Petri nets to model and simulate biological systems. In this model, the authors integrated metabolic disorder information from OMIM with metabolic reactions involved in urea cycle from KEGG, ExPASy[31] and BRENDA[32] (Schomburg et al., 2002) and transcription factors from a commercial database called Biobase. Using urea cycle disorder as a case study, the authors were able to describe the mechanisms and pathways involved in urea metabolism, the gene regulatory regions and patterns, the metabolomics and transcriptomics of the urea cycle disorder (Chen and Hofestädt, 2006).

An integrative analysis approach was developed to identify a set of common tumor specific pathways that could differentiate a number of malignant tumor types from healthy samples. This approach integrated high throughput expression datasets from multiple cancer studies and projected the gene status calculated on to a pathway interaction network (Efroni et al., 2007). This study argues that the common pathways identified as a result of this integrative analysis approach is a better predictor for a cancer outcome (Efroni et al., 2007). To generate biologically meaningful results and to improve the statistical power in breast cancer research, an integrative analysis approach was performed combining publicly available expression data from Affymetrix GeneChips and Illumina BeadArrays (Turnbull et al., 2012). After performing quality control and normalization on expression datasets, a linear additive model was used to combine the expression data. Based on the results from the study, the authors concluded that expression data from different microarray platforms could be integrated despite the difference in technology and this integrated analysis strategy could lead to robust analysis with improved statistical power in comparison to analyzing individual datasets (Turnbull et al., 2012). For studying the integration of external signaling pathways with the core transcription network in embryonic stem cells, ChIP-seq data from embryonic stem cells were analyzed in combination with publicly available gene expression experiments (Chen et al., 2008b). Sheng et al. (2011) developed an integrative analysis approach based on Independent Component Analysis (ICA) and a method called 'gene shaving' (Sheng et al., 2011). Gene shaving is a Principal Component Analysis (PCA) based method to identify subsets of genes with consistent gene expression patterns and large variation across multiple conditions (Hastie et al., 2000). ICA methods are used for data analysis and signal processing and are used to find the linear representation of unknown non-Gaussian data. Based on the results from the combined investigation of copy number data from breast cancer cell lines and gene expression profiles from multiple breast cancer datasets, the authors concluded that this integrative data analysis can be used for identifying subsets of genes with similar or

[31]http://www.expasy.org/ last accessed April 7, 2014
[32]http://www.brenda-enzymes.info/ last accessed April 7, 2014

dissimilar expression patterns (Sheng et al., 2011).

Dudley and Butte (2009) used a network paradigm for the integration of inter disease genomic relationships and biofluid proteomes and this framework was further applied for the identification of disease specific protein markers. Based on this approach, authors generated blood plasma biomarker network by integrating genomic profiles from 136 diseases with 1,028 blood plasma proteins (Dudley and Butte, 2009). Additionally, a urine biomarker network was also generated by linking 577 proteins detectable in urine to genomic profiles from 127 diseases (Dudley and Butte, 2009). The analysis of these networks revealed that more than 80% of the protein biomarkers are related to multiple disease conditions (Dudley and Butte, 2009). A prostrate cancer study used an integrative analysis based on graph prototyping to differentiate between 6 prostrate cancer networks and 7 benign networks (Kugler et al., 2011). In this study, the authors generated phenotype specific expression networks from publicly available prostrate cancer datasets and computed graph edit distances based on these networks. It was found that based on the graph distance metric chosen, the distance within the cancer networks are statistically different from the distance between benign and cancer networks (Kugler et al., 2011). For identifying active microRNAs (miRNAs) and their functions related to gastric cancer, Tseng et al. (2011) used an integrative analysis approach merging gene expression profile with protein interaction networks and miRNA expression profile. Through this analysis, the authors demonstrated that an integrated network based approach can be used to determine the nature of miRNA regulated gene expression (Tseng et al., 2011). Additionally, according to the authors, this integrative analysis method also helped in the identification of a number of miRNA regulated protein interaction networks involved in the manifestation of gastric cancer (Tseng et al., 2011).

In another study, predicted human interactome network was analyzed together with cancer genomics data and Gene Ontology information to identify interaction subnetworks activated in cancer (Rhodes et al., 2005). Protein-protein interactions from model organisms *S. cerevisiae*, *C. elegans* and *D. melanogaster* were used to predict orthologus human protein-protein interaction networks. This network was complemented with shared biological functional annotations from Gene Ontology and protein domain information. In this study, correlation coefficients of gene expression values were calculated based on expression data from 65 microarray studies were retrieved from Oncomine Cancer Microarray Database[33] (Rhodes et al., 2004) and a naive Bayes classifier was used to predict high scoring interaction subnetworks (Rhodes et al., 2005). To prove the validity of the prediction model, the authors confirmed the colocalization of two genes predicted in the model and experimentally confirmed two protein protein interactions predicted in the model (Rhodes et al., 2005). In a handful of studies, combined analysis of co-expression clusters and protein-protein interaction networks were used for gene prediction and have been shown to outperform standard clustering algorithms (Amar and Shamir, 2014). An algorithm known as MORPH (module-guided ranking of candidate pathway genes) was used to identify unknown genes in biological pathways (Tzfadia et al., 2012). In this work, 216 microarray expression profiles from *A. thaliana* and 53 *S. lycopersicum* (tomato) expression profiles were

[33]https://www.oncomine.org/resource/login.html last accessed April 7, 2014

utilized along with species specific pathways from PMN[34] and MapMan[35] and protein protein interaction data from PAIR database[36] (Lin et al., 2011). MORPH algorithm uses a number of clustering approaches to separate genes in expression data and protein protein interaction network into modules. In each module, MORPH identifies genes that are already annotated to pathways and computes average expression patterns based on these known genes. In the next step, similarity of the unknown genes in the modules to average expression patterns were calculated and finally, the scores are normalized and merged from all the modules to obtain a ranking for all candidate genes (Tzfadia et al., 2012). An overview of the MORPH algorithm is given in Figure 2.9. Based on results, authors concluded that MORPH prediction results provided valuable candidate genes on specific pathways (Tzfadia et al., 2012).

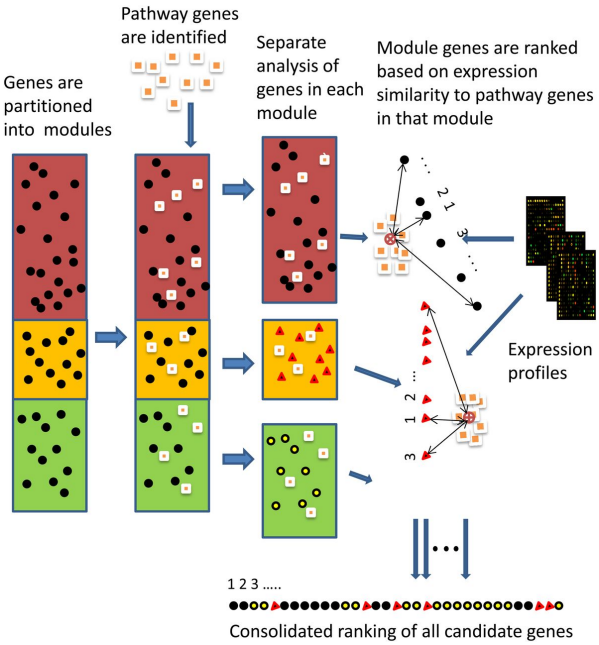

Figure 2.9: An overview of MORPH algorithm. Figure adopted from Tzfadia et al. (2012).

In addition to the analysis methods developed to handle individual problems, various publicly available tools were also developed to facilitate integrative analysis approaches in genomics. Some of the publicly available tools are described in this section. GeneMANIA[37] (Warde-Farley et al., 2010) is a prediction server for generating hypothesis about a given list of genes by incorporating data from a number of interaction databases, metabolic pathways and gene expression data. At present, GeneMANIA includes 1,850 association networks containing 531,122,832 interactions mapped to 187,657 genes from organisms such as yeast (*S. cerevisiae*), worm (*C. elegans*), fly (*D. melanogaster*), mouse (*M. musculus*), plant (*A. thaliana*), human (*H. sapiens*), zebrafish (*D. rerio*) and rat (*R. norvegicus*). GeneMANIA uses two different network weighing methods. For a long list of genes supplied by the user, a basic weighting method is used to learn from the long list of genes and to construct a gene list specific network. For a short list of genes, the algorithm tries

[34]http://www.plantcyc.org/ last accessed April 7, 2014
[35]http://mapman.gabipd.org/web/guest last accessed April 7, 2014
[36]http://www.cls.zju.edu.cn/pair/ last accessed April 7, 2014
[37]http://www.genemania.org/ last accessed April 7, 2014

to retrace the GO co-annotation patterns (Warde-Farley et al., 2010). Integrative multi-species prediction[38] (IMP) is a web server to interpret experimental results in the context of functional predictions and networks (Wong et al., 2012). At present IMP supports seven model organisms including *H. sapiens*, *M. musculus*, *R. novegicus*, *D. melanogaster*, *D. rerio*, *C. elegans* and *S. cerevisiae* (Wong et al., 2012). This web server performs a graphical search on the organism's gene network based on the input gene list. The workflow in IMP uses a Bayesian pipeline to integrate protein-protein interaction data, phylogenetic profiles, expression data, phenotypes and Gene Ontology annotations into a functional relationship network (Wong et al., 2012). This network is used to predict genes and associated phenotypes using an SVM (Support Vector Machine) classifier (Guan et al., 2010). The authors demonstrated the application of IMP by using EVE1 transcription factor in zebrafish as a case study and predicted the functional role of EVE1 in anterior-posterior pattern formation (Wong et al., 2012).

To make use of the large amount of data generated from the Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium, a web analysis server canEvolve[39] was developed (Samur et al., 2013). canEvolve communicates with GEO to retrieve microarray and NGS data and stores normalized expression data. The data analysis framework utilizes R programming language and Bioconductor data analysis packages to perform differential gene and mRNA expression analysis, protein expression, copy number alterations and survival analysis (Samur et al., 2013). In addition to conventional data analysis, canEvolve performs integrative analysis such as GSEA (Gene Set Enrichment Analysis) (Subramanian et al., 2005) and analysis of gene expression profiles together with copy number alterations and miRNA expression profiles. Additionally, the server also stores protein-protein interaction data from STRING (Szklarczyk et al., 2011) database, gene target transcription factor information from TRANSFAC (Matys et al., 2006) and miRNA-target information from PICTAR (Chen and Rajewsky, 2006; Samur et al., 2013). Another web platform, i-cisTarget [40] was developed to enable the analysis of regulatory features in *D. melanogaster* genome (Herrmann et al., 2012). The main analysis problems addressed here are the identification of enriched regulatory features in a set of co-expressed genes or related genomic loci and using these enriched regulatory features to predict *cis*-regulatory modules (CRMs) and infer regulatory networks (Herrmann et al., 2012). Based on benchmark and validation tests in 15 co-expressed datasets, 21 ChIP datasets and 628 curated gene sets, the authors concluded that the analysis approach used in i-cisTarget leads to the identification of meaningful regulatory features (Herrmann et al., 2012).

Based on the literature citations mentioned above, it is clear that a large number of application specific analysis methods and general purpose tools developed make use of the integrative analysis strategies for discovering new knowledge based on existing datasets or databases. However, most of the analysis pipeline developed for integrative data analysis in biomedical sciences are developed to make use of the large volumes of data available in human or other model organisms. As the figures in Table 2.1 (section 2.2.1) shows, the number of genomic or proteomic

[38] http://imp.princeton.edu/ last accessed April 7, 2014
[39] http://www.canevolve.org/ last accessed April 7, 2014
[40] https://gbiomed.kuleuven.be/apps/lcb/i-cisTarget/ last accessed April 7, 2014

annotations and publicly available gene expression data in livestock genomics comparatively small in comparison to data availability in model organisms or human. The major challenging factor in adopting the aforementioned analysis methods to livestock genomics are: (i) the lack of adequate genomic, proteomic and functional annotations in most of the livestock species and (ii) insufficient publicly available high-throughput/expression datasets for integrative analysis and modeling approaches. This challenge is further confounded by the lack of standardized datasets for application development in a livestock genomics scenario, where as a number of standardized datasets are available in human and model organism genomics for the development of species specific or even phenotype and disease specific analysis strategies. The literatures presented in this brief review of integrative analysis approaches is an indication of the broad application potentials of integrative analysis approaches in biomedical field. As detailed, the data sets utilized in these approaches range from publicly available expression profiles to interaction data, functional annotations and data from published articles.

3. Materials and Methods

This chapter describes the materials and various methodologies used in this thesis. The first section materials (see section 3.1) describes the datasets and various algorithms and softwares used for data analysis. Data analysis approaches and mining techniques used in this thesis are explained in the section methods (see section 3.2)

3.1 Materials

3.1.1 Data

This section describes gene expression data sets, interaction networks and database mappings used as data inputs for various analysis methods in thesis.

3.1.1.1 RNA-seq gene expression data

The RNA-seq expression data used in this thesis is from a previous in-house experiment (Gunawan et al., 2013) conducted in order to understand the genetic mechanism behind androstenone metabolism. In the original study, testis and liver tissue samples were harvested from 10 boars which in turn were selected from a pool of 100 boars with an average androstenone value of $1.36 \pm 0.45 \mu g/g$. This pool of 100 animals was a commercial population of Duroc $\times$ F_2 cross breeds. In this pool of animals, boars with a fat androsteone level of $0.5 \mu g/g$ or less were defined as low androstenone (LA) animals and boars with a fat androstenone concentration of $1.00 \mu g/g$ or more were defined as high androstenone (HA) animals. Among the selected 10 boars, 5 animals with an extreme high androstenone measurement of $2.48 \pm 0.56 \mu g/g$ were selected as high androstenone (HA) sample population and 5 animals with an extreme low androstenone measurement of $0.24 \pm 0.06 \mu g/g$ were selected as low androstenone (LA) sample population (Gunawan et al., 2013). Analyzing the ancestry details of these animals revealed that among these selected 10 animals, two sets of 3 animals each: 1 LA and 2 HA animals in the first set and 2 LA and 1 HA animals in the second set were half siblings. SMART cDNA library construction kit (Clontech, USA) was used for library preparations. Sequencing was done externally on an Illumina HiSeq 2000 system by GATC Biotech AG[1] (Konstanz, Germany) (Gunawan et al., 2013). This data is publicly available in GEO database under the accession id: GSE44171[2].

[1] http://www.gatc-biotech.com/en/index.html last accessed July 10, 2014
[2] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44171 last accessed July 10, 2014

3.1.1.2 Microarray data

In addition to the in-house RNA-seq data, a microarray dataset was also used in this thesis. This dataset is publicly available in GEO database (accession id: GSE11073[3]) and was generated to study the hepatic gene expression differences in hepatic androstenone metabolism of two pig breeds, Duroc and Norwegian Landrace (Moe et al., 2008). A total of 58 animals per breed was used in this experiment. For HA Duroc animals the average androstenone level was 11.57 ± 3.2 ppm and for LA Duroc animals, the average androstenone level was 0.37 ± 0.17 ppm (Moe et al., 2008). In case of Norwegian Landrace animals, average measurement of androstenone in HA animals was 5.95 ± 2.04 ppm where as the average androstenone level for LA animals was 0.14 ± 0.04 ppm (Moe et al., 2008). The microarray platform used was a two color cDNA chip. The probes in this chip was designed using cDNA clones from Sino-Danish Pig Genome Sequencing Consortium since the chip was designed before the release of the pig genome assembly (Archibald et al., 2010). The probes were mapped to human gene transcripts from NCBI Refseq database. There were a total of 26,877 PCR products and 867 control features on this chip[4]. In this thesis, this microarray dataset was used to generate multi population co-expression clusters for liver androstenone metabolism (see section 3.2.2.2).

3.1.1.3 KEGG gene interaction networks and pathway mappings

Protein - protein interaction networks and pathway mappings used in this experiment were retrieved from KEGG[5] database (Kanehisa and Goto, 2000) (Release 60.0). Enzyme - enzyme interactions and protein - protein interactions in each KEGG pathway mapped to KEGG gene ids were retrieved from KEGG SOAP based web service using a custom perl script. This retrieved interaction network was comprised of 23,198 edges (interactions) between 3,510 nodes (genes) mapped to 197 pathways. This KEGG interaction network was used in experiment 1 (see section 3.2.2.1).

3.1.1.4 SNP annotations

In this thesis, *Sus scrofa* SNP annotations from dbSNP[6] database (build 137) were used in the variant calling pipeline (see section 3.2.2.1) as the list of known mutations in the *Sus scrofa* genome build. This annotation contained information such as chromosomal id, SNP position on the chromosome (in base pairs), reference and alternate alleles and dbSNP rs ids on 486,585 SNPs in the Variant Calling File (VCF) format.

3.1.2 Algorithms and softwares

This section describes 'off the shelf' data analysis algorithms and softwares used in this thesis. All most all of the softwares and tools described below are freely available for academic use and a vast majority of these are open source projects. A number of tools detailed here are primarily used for (RNA-seq) gene expression data analysis.

[3]http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11073 last accessed July 10, 2014
[4]http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL6173 last accessed July 10, 2014
[5]http://www.genome.jp/kegg/ last accessed July 10, 2014
[6]http://www.ncbi.nlm.nih.gov/SNP/ last accessed July 10, 2014

BEDTools

BEDTools[7] (Quinlan and Hall, 2010) is a suite of softwares to address common genomics tasks such as comparing, manipulating and annotating genomic features in the standard gene annotation formats such as Browser Extensible Data (BED) and General Feature Format (GFF). BEDTools also allows the comparison of sequence alignments in BAM format to gene annotation features in either BED or GFF format. BEDTools can be used for a wide variety of tasks such as coverage analysis, measuring similarity of DNase hypersensitivity measurements, extracting promoter sequences from a genome, identifying regions lacking coverage and calculating GC content. In this thesis, BEDTools suite software `coverageBed` was used to compute the read depth coverage of genes in BAM sequence alignment format in the RNA-seq analysis pipeline (see section 3.2.1.1). `coverageBed` can be used to calculate the depth and coverage of genomic features given in BED or GFF format against a sequence alignment BAM file. This utility can compute the coverage of reads spanning the chromosomal co-ordinates of a gene/transcript and also the fraction of reads that overlaps multiple genomic features[8].

Bowtie

Bowtie[9] is a memory efficient ultra fast aligner for mapping short DNA sequences to large genomes, written in c++ (Langmead et al., 2009). To index reference genomes, Bowtie uses a combination of Burrows-Wheeler transform (BWT) (Burrows and Wheeler, 1994) and the FM index (Ferragina and Manzini, 2000, 2001). Burrows-Wheeler transform is primarily used in the field of data compression, but index built using BWT allows indexing and searching of large texts in a memory efficient manner. FM index is an indexing algorithm utilizing BWT and allows for substring queries on the index. For the purposes of genome alignment, developers of Bowtie extended FM index functionality to allow for sequencing errors and genetic variations (Langmead et al., 2009). Benchmark comparisons done against other next generation sequencing aligners such as Maq[10] (Li et al., 2008a) and SOAP[11] (Li et al., 2008b) showed that with very small sacrifices in sensitivity (number of reads aligned) Bowtie alignments ran considerably faster than Maq or SOAP alignment (Langmead et al., 2009). The results from an independent benchmark study on 9 NGS aligners showed that Bowtie maintained one of the best performances in a number of evaluation criteria (Hatem et al., 2013). In this thesis, Bowtie was used as the aligner in TopHat suite to map RNA-seq raw reads to porcine reference sequences (see section 3.2.1.1).

Consensus clustering

Consensus clustering is a clustering approach in which a number of different clustering solutions from the same dataset is used to find a single clustering result. The clustering outputs from a number of clustering solutions are not deterministic. For a number of graph clustering algorithms (Blondel et al., 2008; Clauset et al., 2004; Lancichinetti et al., 2011; Raghavan et al., 2007), even

[7] http://bedtools.readthedocs.org/en/latest/index.html last accessed March 12, 2014
[8] http://bedtools.readthedocs.org/en/latest/content/tools/coverage.html last accessed March 12, 2014
[9] http://bowtie-bio.sourceforge.net/index.shtml last accessed October 9, 2013
[10] http://maq.sourceforge.net/ last accessed October 9, 2013
[11] http://soap.genomics.org.cn/soapaligner.html last accessed October 9, 2013

if all the parameters supplied to the algorithm are kept constant, clustering solutions can still vary slightly depending on the random seed (random number) chosen to initiate clustering. The usual solution to this problem is to select a clustering result at random or supplying a random seed as a fixed parameter to the clustering algorithm. In both of these scenarios, the clustering output can either be one of the best solutions from the algorithm or one of the subpar solutions. Consensus clustering methods have been proposed as a solution to this problem. Consensus clustering have been shown to improve the robustness and stability of clustering solutions and are less sensitive to outliers and sample variations (Nguyen and Caruana, 2007; Topchy et al., 2005). A variety of strategies have been proposed for obtaining consensus clusters from a set of different clustering solutions from the same dataset (Goder and Filkov, 2008; Lancichinetti and Fortunato, 2012; Strehl et al., 2002; Topchy et al., 2005).

The greedy solutions proposed by Strehl et al. (2002) and Lancichinetti and Fortunato (2012) uses consensus matrices to identify consensus clusters. A consensus matrix is generated from the co-occurrence of vertices in the set of input clustering solutions. The consensus matrix generated is then subjected to further clustering using the graph clustering algorithm adopted in the first phase leading to a new set of solutions. This procedure is repeated until the generated consensus matrix cannot be clustered anymore (until complete consensus is reached) (Lancichinetti and Fortunato, 2012). A flowchart representation of consensus clustering technique is given in Figure 3.1. In this thesis, the greedy consensus clustering algorithm as proposed by Lancichinetti and Fortunato (2012) was used to generate consensus LA and HA co-expression clusters related to porcine hepatic androstenone metabolism (see section 3.2.2.2).

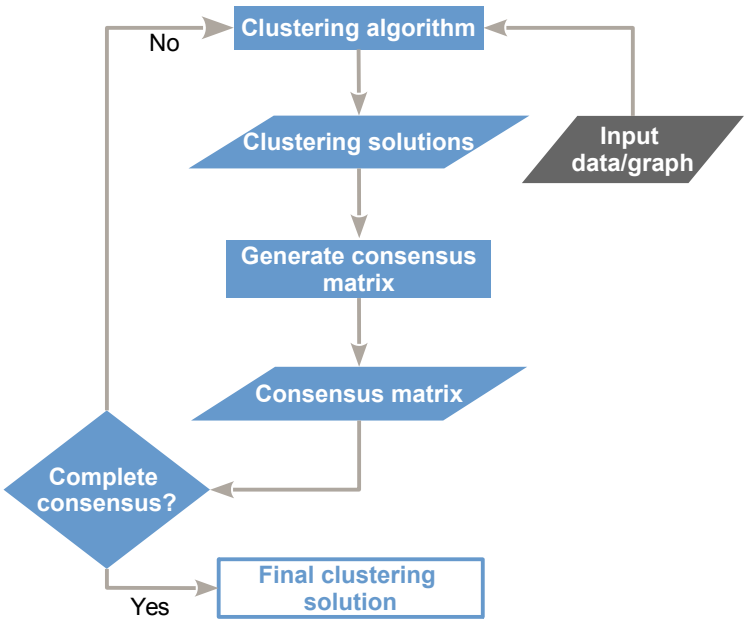

Figure 3.1: Consensus clustering flowchart

Cutadapt

Cutadapt[12] (Martin, 2011) is a tool for removing adapter sequences from high throughput sequencing data in an error tolerant manner. It can be used to trim adapter sequences from the output of a number of sequencing machines either from the 3' end or from the 5' end of the read and can perform gapped alignments with mismatches and indels. One of the features of Cutadapt is that an adapter can be searched and removed multiple times from a read. This feature is helpful in cases where adapters are joined multiple times during the library preparation process (Martin, 2011). Cutadapt was primarily used for raw read quality control and adapter pruning in this thesis (see section 3.2.1.1).

Cytoscape

Cytoscape[13] (Shannon et al., 2003) is an open source platform for the visualization and analysis of complex networks. It is a platform for integrating molecular networks with high throughput expression data and data from other resources into a unified network. The large collection of plugins available for Cytoscape allows the users perform network analysis, clustering and modeling experiments in an input network. In this thesis, Cytoscape was mainly used as a platform for the visualization of the generated interaction networks and clusters.

FastQC

FastQC[14] is a quality control tool aimed at providing initial quality control checks for sequencing data from high throughput experiments. It provides a set of diagnostic plots based on measures such as: per base sequence quality scores, per sequence quality scores, per base GC content and over represented sequences to evaluate the overall quality of the sequencing data. FastQC quality control checks can be done either manually for each sequencing file using the Graphical User Interface or through the batch processing mode using the perl script provided as a part of the FastQC tool. In this thesis, raw read quality evaluation using FastQC was the initial step in RNA-seq data analysis (see section 3.2.1.1).

Gene Ontology semantic similarity

Similarity of gene or gene products can be assessed based on their sequence similarity or functional similarity. Two genes/gene products can be functionally similar if they share the same molecular functions or biological processes. For genes, transcripts and proteins, the primary source of functional annotation is Gene Ontology (Hill et al., 2000). Gene Ontology is divided into three sub ontologies: molecular function (MF), biological processes (BP) and cellular components (CC). These ontologies are directed acyclic graphs (DAGs) and the specificity of the annotation increases from root to the leaves, where the root (parent) nodes describe a generalized concept (either MF, BP or CC) and leaf nodes describe a specialized concept. An illustration of this concept is given in Figure 3.2 using an example. The functional relationship between gene or

[12]http://code.google.com/p/cutadapt/ last accessed October 7, 2013
[13]http://www.cytoscape.org/ last accessed October 9, 2013
[14]http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ last accessed October 7, 2013

gene products can be quantified using GO annotations and GO semantic similarity is one of the methods to estimate these functional relationships.

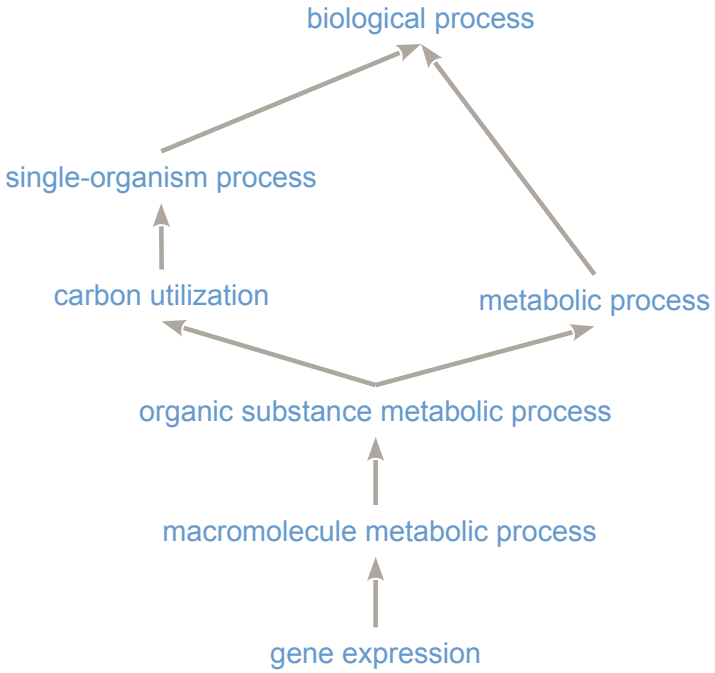

Figure 3.2: GO directed acyclic graph representing the GO term "gene expression" and all the parent nodes of the term.

Algorithms for estimating GO semantic similarity can be classified into two major groups: (i) Information content based methods (Jiang and Conrath, 1997; Lin, 1998; Resnik, 1999; Schlicker et al., 2006) and (ii) Graph based methods (Wang et al., 2007). Information content based methods are based on frequency of a given GO term in a corpus of GO annotations and the most recent common ancestor (MRCA) of two given GO terms. The identification of MRCA for two GO terms is necessary here since GO allows multiple parents for each term and since two GO terms can share the same path by multiple parents. In information content based methods, probability of MRCA is calculated as the ratio of number of occurrences of MRCA or children of MRCA to the total number of GO annotations in the corpus. Information content is calculated as the negative log value of this probability, indicating that the least common MRCA (or MRCA further away from the root nodes) contains greater amount of information (Jiang and Conrath, 1997; Lin, 1998; Resnik, 1999; Schlicker et al., 2006). The graph based Wang method (Wang et al., 2007) postulate that a GO term can be represented as a DAG consisting of the term, the ancestors of the term in GO graph and all the edges in between the terms. Semantic value of this GO term is then defined as the ratio of aggregate contribution of all terms in DAG to the value of the term. The semantic similarity between two GO terms A and B is calculated as the ratio of sum of the semantic values of all ancestor terms of A and B to the sum of semantic values of A and B. The authors argue that the proposed method takes into account the location of the terms in GO graphs and the semantic relationship between the GO terms and ancestor terms (Wang et al., 2007). Wang similarity method is implemented in R Bioconductor package GOSemSim (Yu et al., 2010). BMA (best-match average strategy) is a method to combine the

semantic similarity scores between two genes into one single score (Yu et al., 2010). Given a table of semantic similarity measures, this method calculates the average of maximum similarities on each row and column (Yu et al., 2010) and the similarity values ranges from 0 to 1, 0 being the lowest score possible and 1 being the highest score possible.

In this thesis, GO semantic similarity as proposed by Wang et al. (2007) and implemented in R Bioconductor package GOSemSim was used to estimate the semantic relationship between enriched clusters in experiment 2 (see section 3.2.2.2). The GO semantic similarity scores were then combined into one single score using the BMA strategy discussed above.

Genome Analysis Toolkit(GATK)

Genome Analysis Toolkit (GATK)[15] (McKenna et al., 2010) is a software package that offers a number of tools for analyzing next generation sequencing data. The functionalities offered by GATK are primarily focused on variant discovery and genotyping. Algorithms in GATK software suite including `RealignerTargetCreator`, `IndelRealigner`, `CountCovariates` and `TableRecalibration` were used in a variant calling pipeline. A brief description of these tools are given below.

`RealignerTargetCreator`: generate a list of doubtful alignments that were in need of realignment.

`IndelRealigner`: correct misalignments due to indels (insertion/deletion) in the list of doubtful alignments generated in the first step and generate a new set of alignments in BAM format from the original input files.

`CountCovariates`: used to traverse all the locii that are not in the known list of mutations from the dbSNP database based on the assumption that the mismatches with the reference seen in this step are errors and therefore indicators of poor base quality.

`TableRecalibration`: recalculates base quality scores for reads based on the table generated by `CountCovariates` utility and overwrites the base quality scores for the reads based on empirical observations.

In this thesis, these GATK software suite algorithms were used in association with the Picard utility `MarkDuplicates` as a part of the variant calling pipeline in experiment 1 (see section 3.2.2.1).

Infomap

Infomap[16] (Rosvall et al., 2010) is a graph (network) clustering algorithm based on an information theoretic method called the map equation. The clusters generated by the Infomap algorithm are non overlapping, that is, each node is assigned to one and only one module. Given a graph (network), the algorithm computes the fraction of time each node is visited by a random walker and uses these visit frequencies to search for possible module (cluster) partition spaces. The search results are further refined by a simulated annealing approach (Rosvall and Bergstrom, 2008). Given module partitions, the map equation calculates the average bits per step to describe

[15] http://www.broadinstitute.org/gatk/ last accessed October 14, 2013

[16] http://www.mapequation.org/code last accessed March 4, 2014

an infinite random walker on a network based on these the module partitions. To find the best partitions of the graph, the algorithm tries to minimize the map equation by identifying the clusters of nodes in which the random walker spends a significant period of time before moving on to another cluster of nodes (Rosvall and Bergstrom, 2008).

A benchmark test (Lancichinetti and Fortunato, 2009) compared the performance of a number of a number of graph clustering and community detection algorithms. The algorithms subjected to comparison were : divisive hierarchical algorithm (Girvan and Newman, 2002; Newman and Girvan, 2004), fast greedy modular optimization algorithm (Clauset et al., 2004), exhaustive modular optimization and simulated annealing algorithm (Guimerà et al., 2004), fast modularity optimization algorithm (Blondel et al., 2008), divisive hierarchical algorithm (Radicchi et al., 2004), Cfinder (Palla et al., 2005), Markov Cluster Algorithm (MCL) (van Dongen, 2000), Infomod structural algorithm (Rosvall and Bergstrom, 2007), Infomap algorithm (Rosvall et al., 2010), spectral algorithm (Donetti and Muñoz, 2005), expectation maximization (EM) algorithm (Newman and Leicht, 2007) and Potts model algorithm (Ronhovde and Nussinov, 2009). Two graph clustering benchmarks, GN benchmark (Girvan and Newman, 2002) and LFR benchmark (Lancichinetti et al., 2008) were used in this comparison study. In the LFR benchmark comparison, the clustering performance of algorithms were compared in multiple graph types such as: undirected and unweighted graphs, directed and unweighted graphs, undirected and weighted graphs, undirected unweighted graphs with overlapping communities. Based on the comparisons done using GN and LFR benchmarks, the authors concluded that Infomap algorithm has the best reliable performance in a number of real world scenarios among the tested graph clustering algorithms.

Based on this conclusion by Lancichinetti and Fortunato (2009), in this thesis, Infomap algorithm is used in the second data driven experiment (section 3.2.2.2) to identify gene clusters in low and high androstenone co-expression networks (undirected weighted graphs).

Picard

Picard[17] is a suite of Java based commandline utilities for manipulating sequence alignments in SAM format. Picard supports both SAM text format (SAM) and SAM binary format (BAM). Picard utility `MarkDuplicates` is used to mark duplicate reads in SAM/BAM files. This duplicate marking helps to reduce biases in the variant calling pipeline by flagging the duplicate reads mapped to same part of the reference genome and removing these sequences from further processing in the variant calling pipeline. In this thesis, Picard utility `MarkDuplicates` was used as a part of the variant calling pipeline. An illustration of Picard `MarkDuplicates` run is given in Figure 3.3 (Figure adopted from presentation gatk talks: Mapping and dulicate marking[18]).

[17] http://picard.sourceforge.net/ last accessed October 14, 2013

[18] http://www.broadinstitute.org/gatk//events/2038/GATKwh0-BP-1-Map_and_Dedup.pdf last accessed March 12, 2014

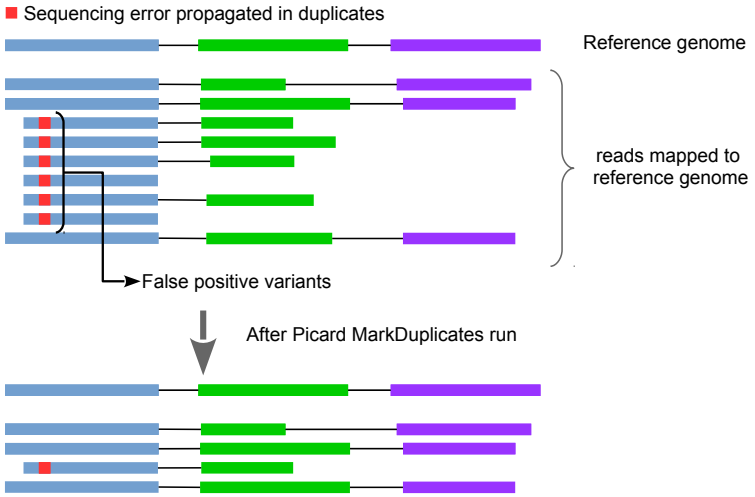

Figure 3.3: An illustration of the effects of Picard `MarkDuplicates` utility run on BAM format files.

SAMStat

SAMStat[19] (Lassmann et al., 2011) is an open source C program for plotting nucleotide over-representation and other statistics from mapped and unmapped reads in NGS SAM and BAM file formats. In this thesis, SAMStat was mainly used to generate the sequence alignment statistics for BAM files generated from mapping testis and liver raw reads to Sscrofa10.2 genome annotation.

SAMTools

SAMTools[20] (Li et al., 2009) is a software suite that provides utilities for manipulating alignments in the Sequence alignment/map (SAM) format. The utilities implemented in SAMTools can be used for post-processing alignments in SAM format. These post-processing functions include sorting, merging, indexing, printing alignments and variant calling. In this thesis, the utilities implemented SAMTools were primarily used for SAM to BAM format conversion and SAMTools `mpileup` function was used for variant calling. The variant calling algorithm `mpileup` in SAMTools is designed under the assumption that all the variants are biallelic. Additionally it is also assumed that at a sequence position, the sequencing and mapping errors of individual reads are independent. Variant calling is done based on a Bayesian inference approach utilizing the prior knowledge that at most of the aligned positions in a sequence, the base pairs from sample reads are homozygous to the reference sequence (Li, 2011).

Seqtk

Seqtk[21] is a lightweight tool for preprocessing sequences in FASTQ and FASTA format. The implemented functionalities include: trimming low quality regions, trimming user specified regions, extracting sequences from a set of user defined regions and converting FASTQ files into FASTA files. Seqtk was also used in the initial quality control phase in RNA-seq data analysis in this thesis (see section 3.2.1.1).

[19] http://samstat.sourceforge.net/ last accessed October 22, 2013
[20] http://samtools.sourceforge.net/ last accessed October 13, 2013
[21] https://github.com/lh3/seqtk last accessed October 8, 2013

SnpEff

SnpEff[22] (Cingolani et al., 2012) is a program for categorizing the effect of polymorphisms in gene annotations. SnpEff can be used to annotate genomic locations and to predict coding effects of a variant in the genome. Using this tool, polymorphisms can be categorized as belonging to intronic, untranslated region, upstream, downstream, splice site, or intergenic regions of the genome. The predicted coding effects of the variants include: synonymous or non-synonymous amino acid replacement, start codon gains or loss, stop codon gains or loss and frame shifts. In this thesis, SnpEff was used to annotate the polymorphisms identified in RNA-seq genome annotations using SAMTools in the variant calling pipeline (see section 3.2.2.1).

Statistical significance of clusters

Statistical significance of a cluster is defined as the probability of finding a cluster in graphs (networks) without any community structure (random graphs) (Lancichinetti et al., 2011). The algorithm developed by Lancichinetti et al. (2010) for estimating the statistical significance of clusters is detailed in this section. The authors have provided the binaries and source codes[23]. In this algorithm, random graphs are generated based on a model known as configuration model (Molloy and Reed, 1995). Given an input graph, configuration model generates random graphs by randomly connecting the nodes while preserving the node degree distribution. In the algorithm proposed by Lancichinetti et al. (2010), given an input graph and a cluster generated from this graph, the stochastic null model is generated in such a way that the original edges between the nodes in the cluster are preserved (within cluster edges) where as the outgoing edges, edges connecting cluster nodes with other non cluster nodes are randomly shuffled (Lancichinetti et al., 2010). This rewiring step also allows multiple edges between two nodes and self loops (Lancichinetti et al., 2010).

Once a null model has been selected, the probability that a node k has i internal nodes in a cluster C can be modeled as a hypergeometric distribution (Lancichinetti et al., 2010). This hypergeometric modeling is followed by a cumulative probability calculation step and provides a means to rank nodes in a cluster based on their link to other nodes in the cluster. The cumulative probabilities are derived from a set of different degrees generated using a bootstrap strategy (Lancichinetti et al., 2011). The statistical significance of a cluster is calculated as the probability of inclusion of the second worst node in the cluster (Lancichinetti et al., 2010). The estimation of statistical significance is followed by a cluster up clean up procedure (Lancichinetti et al., 2011). In this step, nodes outside the cluster are either added to the cluster or nodes inside the cluster are trimmed off from the cluster. This inclusion or trimming depends on the probability (of inclusion of node in a given cluster) calculated in the previous step. If the probability calculated for nodes outside a cluster are statistically significant with respect to a given threshold, the nodes are added to the cluster, where as, if the probability calculated for nodes in the cluster are not statistically significant, these nodes are trimmed off from the final cluster (Lancichinetti et al., 2011).

[22] http://snpeff.sourceforge.net/ last accessed December 10, 2013

[23] https://sites.google.com/site/andrealancichinetti/software last accessed April 11, 2014

In this thesis, this algorithm was used as a part or experiment specific methods in experiment 2 (see section 3.2.2.2) to estimate the statistical significance of the co-expression clusters and to perform the cluster clean up procedure.

TopHat

TopHat[24] (Trapnell et al., 2009) is a splice junction mapper algorithm to align RNA-seq reads to a known genome independent of the existing splice sites defined for the genome. TopHat is implemented in c++ and python and for aligning RNA-seq reads to genome, TopHat uses an 'off the shelf' high-throughput aligner Bowtie. In TopHat read mapping pipeline, for each read, Bowtie reports one or more alignments with a few number of mismatches (default number of mismatches: 2) in the 5' bases of the read. In case of bases at the 3' end of the read, Bowtie allows additional mismatches based on a Phred-quality-weighted Hamming distance threshold. This criteria for allowing mismatches is based on the observation that the 3' end of the reads contain more sequencing errors in comparison to 5' end of the read (Hillier et al., 2008). TopHat allows Bowtie to report more than one alignment per read and suppresses all alignment for reads with more than a predefined (default number of alignments: 10) number of alignments. The developers of TopHat claim that this approach allows the reporting of multireads from genes with multiple copies while suppressing the alignments to low complexity regions of the genome (Trapnell et al., 2009).

In this thesis, TopHat was used as an aligner to map RNA-seq raw reads after the quality control process to *Sus scrofa* genome build Sscrofa10.2 (see section 3.2.1.1).

3.2 Methods

The method section of this thesis is divided into two main subsections, (i) RNA-seq data quality control, mapping and normalization and (ii) Experiment specific methods. The first subsection describes the methods used in gene expression data quality control and normalization and applied in general to the RNA-seq expression data used in both analysis and the second section details the data mining and analysis procedures that are specific to each of the experiments carried out as a part of this thesis.

3.2.1 RNA-seq data quality control, mapping and normalization

3.2.1.1 Data quality control and mapping

In the original study, the raw reads from the RNA-seq data were mapped to NCBI *Sus scrofa* genome build Sscrofa9.2[25] (Gunawan et al., 2013). But in case of the experiments discussed here, the raw reads (in .fastq files) were remapped to a new NCBI *Sus scrofa* genome build released at the time, Sscrofa10.2[26]. The first step in this remapping process was the quality control step. In this step, the quality of the raw read sets (testis and liver) were independently

[24] http://tophat.cbcb.umd.edu/index.shtml last accessed October 9, 2013
[25] http://www.ncbi.nlm.nih.gov/assembly/111518/ last accessed October 14, 2013
[26] http://www.ncbi.nlm.nih.gov/assembly/304498/ last accessed October 14, 2013

assessed using FastQC quality control tool. Over represented PCR primers, bad quality bases (with Phred score <20) and bases with fluctuating GC content were identified in this step (reported by FastQC) and removed from the raw sequencing data using a combination of Cutadapt and Seqtk tools. Cutadapt, as the name suggests, was mainly used for pruning PCR primers from the raw reads. A well known issue with the next generation sequencing Illumina systems is the low quality of the bases at the 3' end of the reads. A recommended procedure in this case is to exclude these bases of the reads from further processing (Minoche et al., 2011), for which Seqtk was primarily used. The selection of threshold cut-off (Phred score >20) was arbitrary and yet this cut-off threshold ensured that only the reads with a base quality score of 99% or more were retained for further analysis. The pruned datasets obtained as a result of this quality control step were aligned to the *Sus scrofa* genome build Sscrofa10.2 using the "splice aware" mapping algorithm TopHat (see section 3.1.2). According to Trapnell et al. (2009), in RNA-seq experiments, the major objective of mapping raw reads to genome are (i) identification of novel transcripts and (ii) abundance estimation of transcripts (Trapnell et al., 2009). In this thesis, mapping raw reads to genome was primarily used for the abundance estimation of transcripts. As explained in the section Algorithms and softwares (section 3.1.2), to compute read depth coverage for each gene and to generate gene read coverage (gene expression) matrices, sequence alignments in the BAM format and gene feature annotations in GFF format were given as inputs for BEDTools `coverageBed` utility. The next step was to filter genes with low read counts testis and liver expression matrices. In both cases (testis and liver), genes with mean read count <25 in HA and LA phenotypes were removed from the raw read count expression matrix before further processing. Table 3.1 shows the number of genes in each gene expression matrix generated from RNA-seq sequence alignment files.

Table 3.1: RNA-seq expression data statistics

Sample tissue type	**Number of genes before pruning**	**Number of genes after pruning**	**HA samples**	**LA samples**
Testis samples	21,340	16,760	5	5
Liver samples	18,427	11,736	5	5

3.2.1.2 Expression data normalization

In RNA-seq experiments, the expression of a gene is measured as the number of reads mapping into a particular genomic interval, unlike the probe intensity values measured in microarray experiments. The measured RNA-seq gene expression values follows a negative binomial distribution (Robinson et al., 2010) in contrast to the normally distributed gene expression values in microarray experiments. A major challenge raised by this difference in data distribution is that the classical linear modeling analysis procedures developed for microarray data mining and analysis assumes the data to be normally distributed and hence cannot be directly applied to RNA-seq expression data. Although various non parametric procedures (distribution free methods) can be used in this context, the initial "trial and error" experiments have shown that the results given by such

analysis procedures were statistically non significant, owing to the small sample size of the data set (number of phenotypes per sample: 5) used in the experiments performed in this thesis and also due to the limited power of non parametric methods to draw significant conclusions from data sets with small sample sizes. Additionally, in the second experiment (see section 3.2.2.2), to combine RNA-seq meta data with microarray meta data it is necessary that expression data from all the experiments follow the same distribution.

Recently, Law et al. (2013) proposed applying normal distribution based microarray like statistical analysis methods to RNA-seq read count data. This proposed model is based around the principle that accurate modeling of the mean-variance relationship intrinsic to the data generating process is essential to design statistically powerful methods (Law et al., 2013). Mean-variance modeling at the observational level (voom) estimates mean-variance relationship in the read count data and computes weights for each observation based on this relationship (Law et al., 2013). In order to overcome the limitations of small sample sizes and non parametric methods to an extend and also following the proposed idea in (Law et al., 2013), the RNA-seq gene expression matrix was normalized and log2 transformed using `voom` function implemented in limma `R` package (Smyth, 2005). Comparison of various normalization and differential expression analysis methods for RNA-seq data have shown that voom normalization combined with limma package to be relatively unaffected by outliers and to perform well under many conditions (Soneson and Delorenzi, 2013). An additional study (Rapaport et al., 2013) concluded that modeling RNA-seq gene count data as log normal distribution with appropriate pseudo counts (limma voom modeling) is a reasonable approximation of the data.

3.2.2 Experiment specific methods

In this thesis, in addition to the common methods described above, different experiment specific analysis were also performed. This section describes these specific analysis procedures and each of the experiment specific analysis sections are subdivided into subsections to describe the different methods followed in each of the experiments.

3.2.2.1 Experiment 1: Pathway based analysis of genes and interactions influencing porcine testis samples from boars with divergent androstenone content in back fat

The major aim of this analysis was to identify and study the dominant metabolic pathways and interactions involved in the maintenance and regulation of steroidogenesis and androstenone biosynthesis in porcine testicular tissues. For this purpose, an integrative knowledge driven approach merging together interaction network and pathway information from KEGG database and gene expression data from RNA-seq experiments was used. But, a current limitation of this approach in terms of studying androstenone metabolism is that none of the major pathway databases contain data on metabolic reaction steps or gene interactions involved in androstenone biosynthesis. As a work around to this limitation, androstenone biosynthesis was considered as an offshoot of steroid hormone (testosterone) synthesis pathway in testis under the assumption that the pathways and interaction events that affect steroid hormone biosynthesis could also affect

androstenone biosynthesis.

This analysis section has three subsections (i) identification of significant interactions, (ii) KEGG pathway enrichment and (iii) variant calling. The methods used in the identification of statistically significant pathway interactions are described in the first subsection, the steps followed in interaction pathway enrichment are detailed in the second subsection and the last subsection describes the gene polymorphism analysis performed.

Identification of significant interactions

The major objective behind this analysis was to identify significant pathway interactions by merging RNA-seq gene expression data and KEGG pathway interaction network (section 3.1.1.3). In this analysis, the gene expression data from only the testis samples in RNA-seq expression data (see section 3.1.1.1) was used. As noted in Table 3.1, the normalized porcine testis expression matrix contained expression measurements of 16,760 genes in 10 samples and as described in section 3.1.1.3, the KEGG gene interaction network contain interactions only for 3,510 genes. Hence, the first step in this analysis procedure was to trim the the testis gene expression data set for genes in the KEGG interaction network. As a result of this trimming, only 2,871 genes in common between the gene expression data set and the KEGG interaction network were retained. The KEGG interaction network was also trimmed down to 2,871 genes and contained 23,198 edges.

In the next analysis phase, Pearson Correlation Co-efficient (PCC) of gene expression values were calculated for both HA and LA testis samples separately and the edges of the trimmed pathway interaction network were weighted with these correlation values. This step gave rise to two different pathway interaction networks: in the first network, the edges were weighted with correlation coefficients derived from LA testis expression data ("LA network") and in the second network, edges were weighted with correlation coefficients derived from HA testis expression data ("HA network"). Both LA and HA networks were comprised of 2,871 nodes and 15,960 edges. In order to identify the interactions that were significantly different between both LA and HA networks, the edge weights (correlation coefficients) of both networks were transformed to z-score using Fisher-r-to-z transformation based on the equation:

$$z = \frac{1}{2} ln \frac{(1+r)}{(1-r)} \ , \text{ where r is the PCC} \tag{3.1}$$

Following the calculation of z-scores for interactions in both networks, the differences between the z-scores were also calculated. For an edge z-score in LA network, the corresponding edge z-score from HA network was retrieved and the difference between the z-scores was calculated as:

$$zscore_{DIFF} \ = \ zscore_{LA} \ - \ zscore_{HA} \tag{3.2}$$

In the following analysis step, in order to identify significant $zscore_{DIFF}$, a two step evaluation criteria based on permutation and random sampling was used (Ripley, 1987). Permutation and random sampling based methods for estimating significance thresholds have already been used in

high throughput studies (Gatti et al., 2009; Zhang et al., 2012). The evaluation criteria used in this step were:

(i) $zscore_{DIFF}$ should be significant at a threshold of empirical p-value <0.05 against a set of zscores generated from randomly sampling the original expression data set.

(ii) At least one of the correlations, either from LA expression set or from HA expression set used to calculate the $zscore_{DIFF}$ must be significant at a threshold of empirical p-value <0.05 against a set of correlations generated from randomly sampling the original expression data set.

For generating the set of zscores used in evaluation criteria (i), a random expression matrix was generated by randomly shuffling and assigning the whole testis gene expression values into two sample groups. The purpose behind random shuffling and assigning of expression values was to break up the original ordering and classification of the expression values and samples as belonging to either HA or LA sample set and generate two complete random expression matrices (expression sets). Pearson correlation coefficients, zscores and zscore differences were calculated on these random expression matrices following the previously described steps and the entire process was repeated 10,000 times to generate a set of random zscore differences ($zscore_{RAND}$) for each interaction. The significance threshold empirical p-value for each $zscore_{DIFF}$ was calculated as:

$$Pval_{Empirical} = \frac{\#\ zscore_{RAND} > zscore_{DIFF}}{N}\ ,\ \text{where N} = 10,000 \tag{3.3}$$

A similar procedure was followed for calculating significance threshold empirical p-value for correlations in evaluation criteria (ii), where empirical p-value was calculated between correlation coefficients from randomly sampled expression data and the original correlation coefficients from LA or HA datasets. Once the significant interaction (correlation) identification was complete, the identified significant interactions were further classified into 8 correlation types such as: HA positive, HA positive significance, HA negative, HA negative significance, LA positive, LA positive significance, LA negative and LA negative significance. The rules used for classification of these correlation types and edge colors and line styles used in visualization of these correlation types are given in Table 3.2. These classification rules were mainly used in the visualization step, and all the interaction networks in this work were visualized using Cytoscape. All the above mentioned analysis procedure were carried out in the statistical computing platform `R` and several custom functions were written in `R` to perform these analysis steps. In this analysis, R package igraph[27](Csardi and Nepusz, 2006) was used for network analysis and manipulation.

[27]http://igraph.sourceforge.net/ last accessed October 14, 2013

Table 3.2: Interaction edge classification rules. Set of rules used for the classification of interactions(correlations) and assigning correlation types, edge color and line styles

Correlation coefficients	**Correlation coefficient coefficient in HA testis samples**	**Correlation coefficient coefficient in LA testis samples**	**Edge color for visualization**	**Edge line style for visualization**
HA positive	positive and significant	negative	red	solid line
HA positive significance	positive and significant	positive	red	dashed line
HA negative	negative and significant	positive	light green	solid line
HA negative significance	negative and significant	positive or negative	light green	dashed line
LA positive	negative	positive and significant	green	solid line
LA positive significance	positive	positive and significant	green	dashed line
LA negative	positive	negative and significant	orange	solid line
LA negative significance	positive or negative	negative and significant	orange	dashed line

KEGG pathway enrichment analysis

Once the identification of significant interactions were completed, the next step in this analysis was the identification of pathways enriched for significant interactions. In this step, rather than performing the conventional gene enrichment analysis, an interaction enrichment analysis was performed following the school of thought that the interactions of a gene reveals more about the functions of that particular gene in a phenotype. A custom function was written in `R` to perform the hypergeometric test to asses the pathways over-represented for significant interactions. The pvalues generated by the R `phyper` function were then corrected for multiple testing using Benjamini–Hochberg procedure. Finally, the pathways with a p-adjusted value of < 0.05 from this analysis were considered as significantly enriched (over-represented) for the identified interactions.

Variant calling

This section describes the analysis methods used in the variant calling pipeline. The variant calling pipeline used utilities and tools implemented in software suites Gatk, Picard and SAMTools function `mpileup`.

The input data used in this pipeline were :

(i) BAM format sequence alignments from TopHat (see section 3.2.1.1)

(ii) Sscrofa10.2 DNA sequences in FASTA format and

(iii) SNP annotations in VCF format (see section 3.1.1.4)

The variant calling pipeline described below was adapted from the GATK guideline on best practices for variant calling[28]. The variant calling pipeline used GATK algorithms and Picard

[28]http://www.broadinstitute.org/gatk/guide/best-practices last accessed October 14, 2013

function `MarkDuplicates` for realigning and re-indexing the bam files and SAMTools function `mpileup` for variant calling. Figure 3.4 shows the workflow followed for variant calling pipeline in this thesis. In the final step in this pipeline, the realigned and recalibrated reads in BAM format are used for variant calling by using the SAMTools function `mpileup`. The initial set of polymorphisms obtained from samtools was further filtered down with the parameters: Root Mean Square (RMS) Phred quality score greater than 20, read depth greater than 50 and SNP quality score greater than 20. Furthermore, all the polymorphisms mapped to intronic positions of genes were excluded from this analysis. The chromosomal position and reference alleles of the final filtered set of polymorphisms were crosschecked against dbSNP database (Build 136) to identify the variants that were already represented in the SNP database. The possible amino acid coding effects of these polymorphisms such as synonymous mutation, non synonymous mutation, start/stop codon gain or loss and genomic positions such as upstream, downstream, in UTR (un-translated region) were predicted using SnpEff software.

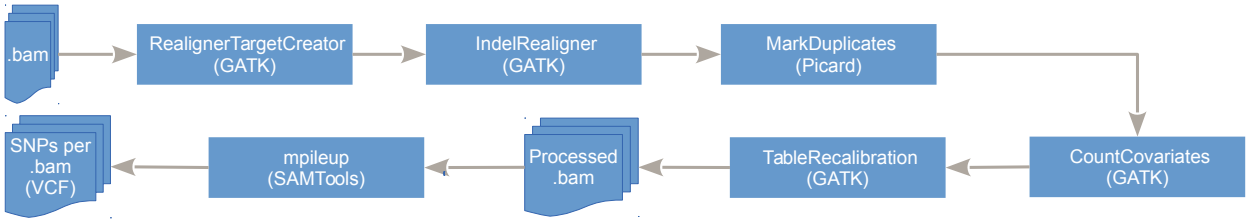

Figure 3.4: Flow chart of variant calling pipeline used in this experiment.

A schematic diagram of the entire workflow used in this experiment is given in Figure 3.5.

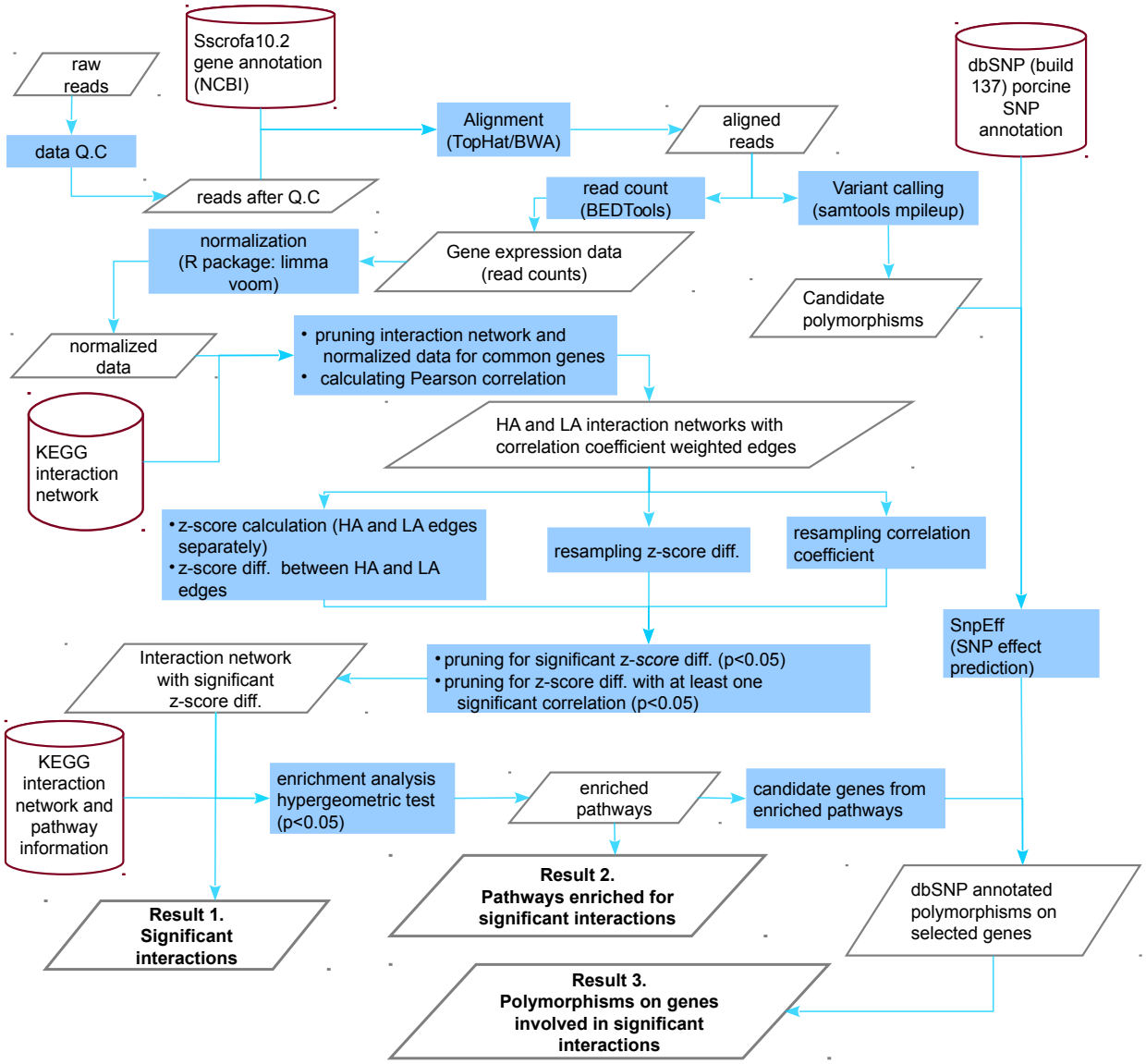

Figure 3.5: Pathway based analysis workflow. Legend: White parallelograms with grey outline: Input/output data and results. White cylinders with red outline: data from external databases. Rectangles with light blue shades: various tools and analysis processes used in this workflow.

3.2.2.2 Experiment 2: Identification of gene co-expression clusters in liver tissues from multiple porcine populations with high and low backfat androstenone phenotype

One of the difficulties in studying porcine androstenone metabolism is the lack of external knowledge on this subject. Although a number of studies (Doran et al., 2004; James Squires, 2010; Moe et al., 2007a, 2008; Nicolau-Solano et al., 2006; Robic et al., 2008; Sinclair et al., 2005) have detailed a number of biomarkers for hepatic androstenone metabolism, majority of the enzymes and pathways involved in the liver metabolism of androstenone are still largely unknown. Since there are sizable discontinuities in external knowledge about hepatic androstenone metabolism, using a knowledge driven approach in this case would not yield fruitful results. A suitable alternative approach in this scenario would be to follow a data driven approach incorporating gene expression data from multiple high throughput experiments in multiple porcine breeds/populations on hepatic androstenone metabolism and to analyze these datasets to identify similarities in gene expression patterns. The advantages of such an analysis procedure would be:

(i) by combining data from multiple populations it would be possible to understand the breed similarities in androstenone metabolism

(ii) since the analysis includes data from multiple populations, the candidate biomarkers can be used fill current gaps in the understanding of androstenone hepatic metabolism and finally,

(iii) the analysis results could be used as a comparison standard to understand breed differences.

Hence, in this experiment, a data driven analysis method combining multiple high-throughput gene expression datasets was followed to identify the common gene expression patters in hepatic androstenone metabolism of three different pig populations. For this purpose, publicly available gene expression datasets on porcine hepatic androstenone metabolism were retrieved and then combined into two datasets based on the androstenone measurements and phenotype assignments in the original studies.

Publicly available high throughput gene expression data from three pig populations, generated in two different experiments were used in this analysis. The number of expression matrices and porcine populations were limited to three here since it was not possible to obtain publicly available expression data on porcine hepatic androstenone metabolism for any other breed during the time of this study. Among the expression data selected, one was from an in-house RNA-seq experiment performed on a sample population of Duroc $\times$ F_2 boars (Gunawan et al., 2013). The liver samples from 5 boars with extreme high levels of androstenone measurement in backfat were categorized as high androstenone animals (HA) and liver samples from 5 boars with extreme low levels of androstenone measurement in backfat were categorized as low androstenone animals (LA). Expression dataset on porcine testicular androstenone metabolism generated as a part of this study was used in the previous experiment (section 3.2.2.1). Additional details of this dataset are described in section 3.1.1.1. The remaining two data sets were from a microarray experiment based on a custom porcine cDNA microarray platform. In this experiment, gene expression

profiling was performed on boar liver samples from two breeds, Duroc and Norwegian Landrace. Expression profiling was performed separately for each breed and both datasets contained 29 HA animals and 29 LA animals each (Moe et al., 2008). This dataset is described in section 3.1.1.2.

Table 3.3 gives an overview of the gene expression datasets used in this experiment.

Table 3.3: Expression dataset details

Dataset	**#Genes**	**#Common genes**	**#LA samples**	**#HA samples**	**Breed**	**GEO dataset id**	**GEO platform id**
DuF2	11,736	7,693	5	5	Duroc × F_2	GSE44171	GPL11429
Duroc	11,186		29	29	Duroc	GSE11073	GPL6173
Landrace					Norwegian Landrace		

Various steps used in RNA-seq data quality control, mapping, generating expression matrix and normalization of expression data are described in section 3.2.1.

Microarray data retrieval and mapping

The next step in this analysis was the retrieval, normalization and mapping of microarray expression data in (Moe et al., 2008) to gene identifiers from Sscrofa10.2 gene build. The data normalization procedure described in the original microarray experiment is as follows: after hybridization and scanning, the mean foreground intensities were log transformed and normalized using print-tip loess normalization procedure in R (R Development Core Team, 2013) limma package (Moe et al., 2008). The duplicate correlation feature in the limma package was used to estimate the correlation between duplicate spots in the array and finally the effect of the normalization procedure was assessed using MA-plots and box plots (Moe et al., 2008). Since the standard procedures of normalization were followed in the original experiment, the normalized expression datasets were retrieved from the corresponding GEO dataset using R package GEOQuery (Davis and Meltzer, 2007).

One of the challenges in analyzing these microarray datasets together with in-house RNA-seq dataset was the mapping between the custom probe ids used in the microarray platform and Entrez gene ids used in RNA-seq expression dataset. The cDNA microarray chip (see Table 3.3) used in the experiment was designed before the release of the pig genome (Archibald et al., 2010) and used cDNA clones from Sino-Danish Pig Genome Sequencing Consortium as probes[29]. The cDNA probes were mapped to human gene transcripts in NCBI Refseq database and used custom probe identifiers. Since these custom designed microarray probes and Entrez gene ids from RNA-seq dataset were not directly compatible, a mapping between the microarray probe identifiers and NCBI Entrez gene identifiers was generated. For this purpose, sequence alignments were performed between the FASTA sequences of these custom probes and Sscrofa10.2 Refseq cDNA sequences mapped to Entrez gene ids. An all-vs-all BLAST was performed using NCBI

[29]http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL6173 last accessed March 4, 2014

standalone BLAST executable (version: 2.2.28+) (Mount, 2007) and reciprocal blast approach was used to remove ambiguous mappings. The Sscrofa10.2 sequence database generated for BLAST-ing consisted of 25,890 cDNA sequences mapped to Entrez gene ids and the microarray probe sequence database was comprised of 26,877 sequences. In this step, mappings were generated between 11,251 microarray cDNA probes and 11,186 Entrez gene ids. In order to avoid the conflicts where multiple cDNA probes were mapped to an Entrez gene id, the expression values from the probe with the largest variance between sample expression values was mapped to the corresponding Entrez gene id and the remaining conflicting probe ids and expression values were discarded from further analysis. Only 7,693 genes were common between the three datasets, and hence the expression values from only these genes were retained in all the datasets for further analysis. In the next step, the expression matrices were regrouped according the phenotype assignment and generated 2 expression datasets: an LA set and an HA set with 3 expression matrices each (Table 3.3).

Generating multi breed co-expression networks

In this study, Pearson correlation coefficient between gene pairs in an expression matrix was used as a measure of co-expression. The principal aim behind this experiment was to generate signature gene co-expression networks by merging metadata from multiple gene expression datasets to study porcine hepatic androstenone metabolism. A number of methods have been developed to merge and analyze datasets across multiple microarray experiments (Chang et al., 2013; Shabalin et al., 2008; Xu et al., 2008). But a recent study (Almeida-de Macedo et al., 2013), argues that rather than merging multiple datasets together, combining statistical results is a better method for estimating correlation coefficients of a gene pair across multiple datasets. This argument is based on the analysis of correlation coefficients between gene pairs by combining 19 heterogeneous microarray datasets into one pool (Almeida-de Macedo et al., 2013).

Stuart et al. (2003) developed a method for computing gene co-expression clusters across microarray datasets from multiple species. In this method, the authors calculated correlation coefficient between gene pairs in each dataset and further computed rank order statistics for each gene pair (Stuart et al., 2003). The rank order statistics for each gene pair (each unique correlation coefficient) was calculated as the ratio of its rank (ordered position) to the total number of gene pairs (unique correlation coefficients). In each dataset, for correlation coefficients $-1 < cor_1 < cor_2 < \cdots < 0 < \cdots < cor_{n-1} < cor_n < 1$, the rank ratios followed the order $0 < r_1 < r_2 < \cdots < r_{n-1} < r_n < 1$.

Under the assumption that the rank ratios follows a uniform distribution $U(0, 1)$ the cumulative density $F(r_n)$ for each rank order (correlation coefficient/gene pair) is :

$$F(r_n) = \begin{cases} 0 & \text{for } r_n \leq 0 \\ r_n & \text{for } 0 < r_n < 1 \\ 1 & \text{for } r_n \geq 1 \end{cases}$$

The rank ratios were calculated for each expression matrix independently. Assuming that the

cumulative density functions of each gene pair (rank ratio) are independent across multiple datasets, the joint cumulative density function (joint c.d.f) of each rank ratio (each gene pair) across multiple species was calculated based on the following equation:

$$P(r_1, r_2, \cdots, r_n) = n! \int_0^{r_1} \int_{r_1}^{r_2} \cdots \int_{s_{n-1}}^{r_n} d_{s1}, d_{s2}, \cdots, d_{sn} \tag{3.4}$$

In this equation n is the number of species in the study and $r_1, r_2, \cdots, r_n$ are the rank order ratios of a gene pair in multiple species (datasets). This method proposed by Stuart et al. (2003) has the advantage that only the statistical results from multiple datasets are combined and thus possibly avoids Simpon's paradox and generating equivocal results as stated by (Almeida-de Macedo et al., 2013).

In this work, the aforesaid approach proposed by Stuart et al. (2003) was adopted to generate the signature co-expression networks related to porcine hepatic androstenone metabolism. As a first step, Pearson correlation coefficients were calculated for gene pairs in all the 6 expression matrices (3 LA and 3 HA expression matrices) separately. Since there were 7,693 (n=7,693) common genes among all the datasets, a total of 29,587,278 unique gene pairs (unique correlation coefficient values) were generated per dataset (total number of unique correlation coefficients (gene pairs) calculated as : $\frac{n\times(n-1)}{2}$). Based initial experiments, it was found that due to this high number of unique correlation coefficient values, using signed values of correlation coefficients for rank order calculation would result in high rank order ratios even for correlation coefficients with a very small positive value. Since these rank ratios were used for computing the joint cdf, even the gene pairs with very small positive correlation coefficients in all the three expression matrices of a dataset would receive a high joint cumulative probability. In order to overcome this issue, the absolute value of correlation coefficients was used to compute the rank order statistics of gene pairs. After calculating the rank order ratios of gene pairs in all the expression matrices, gene pair correlation coefficients and rank order ratios were compiled into either LA or HA set according to the phenotype assignment described in the previous subsection.

In the next step, all the gene expression matrices in LA and HA datasets were pruned for correlation coefficients $\geq +0.50$. This pruning step was aimed at removing all those gene pairs with conflicting directionalities. For example, gene pairs ELAC1, gene id: 100524839 and LAPTM5, gene id: 100624193 with a correlation coefficient of +0.852 in DuF2 expression matrix, -0.371 in Duroc expression matrix and +0.250 in Landrace expression matrix. The correlation coefficients described here are from LA dataset. Additionally, this pruning step had the added advantage that the gene pairs in both LA and HA set with negative correlation coefficients or very low positive correlation coefficients were also removed, thus reducing the number of calculations and hence the computing time needed. After this pruning step, the number of remaining gene pairs in LA and HA sets were 43,480 (from 3,648 genes) and 42,309 (from 2,826 genes) respectively. The joint cumulative probability of rank order ratios for these gene pairs in LA and HA sets were calculated using equation 3.4. The cumulative probabilities generated in this step for gene pairs in LA and HA sets were used as edge weights for the gene pairs and thus two phenotype specific edge weighted co-expression networks were generated: an LA network with 43,480 edges among

3,648 nodes (genes) and an HA network with 42,309 edges and 2,826 nodes (genes). These LA and HA co-expression networks were further used as inputs for graph clustering and community detection.

Identifying statistically significant co-expression clusters

In this experiment, a combination of Infomap clustering algorithm (see section 3.1.2) and consensus clustering technique (see section 3.1.2) was used to generate clusters from LA and HA co-expression networks. All the input parameters, except random seed were kept constant for clustering LA and HA networks and 500 clustering solutions were generated in each iteration (per network). Complete consensus clusters were generated from LA network after 3 iterations where as complete consensus clusters were generated from HA network after only 2 iterations. Figure 3.6 gives an overview of the LA and HA consensus clustering runs and the total number of clusters generated per run for each network.

Although consensus clustering technique can enhance the accuracy and reliability of the resulting clusters, this method still cannot guarantee the significance of a cluster with respect to the input network. Since the initial LA and HA co-expression networks had a large number of nodes (3,648 and 2,826 respectively), it could be possible that some of the clusters generated from these networks are not specific to the phenotype at all, but random collection of nodes either as a result of the large number of nodes in the initial networks or as a result of an artifact in the cluster algorithm. In this work, the aim was to select only the clusters which were not random but specific to the given input network. So, in the next step, a cluster clean up process and assessment of the statistical significance of the clusters was performed by applying the methodology proposed by (Lancichinetti et al., 2010) (for a detailed explanation of this method see section 3.1.2). After this step, clusters with less than 10 nodes and significance score (p-value) ≥ 0.05 were excluded from further analysis.

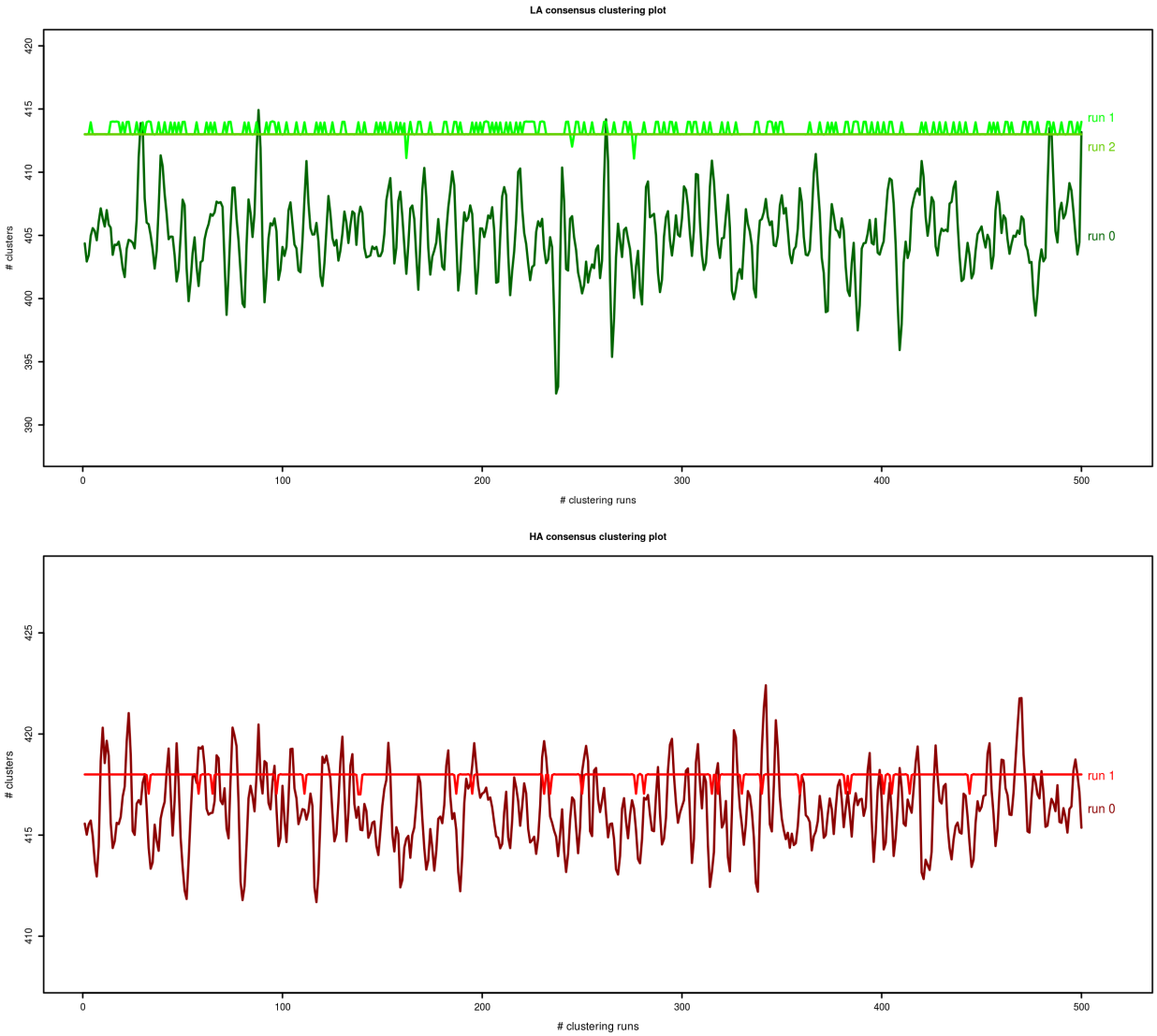

Figure 3.6: LA HA networks consensus clustering. Legend: "run 0" in both graphs indicate first clustering run using LA and HA networks, "run 1" indicates clustering run for the first consensus cluster and "run 2" indicates clustering run for the second consensus cluster.

Enrichment analysis

To identify and describe the biological functions of these significant co-expression networks, Gene Ontology (GO) and KEGG enrichment analysis were performed for each cluster. GO enrichment analysis was limited to the biological process sub tree of the Gene Ontology and was performed using the R package topGO (Alexa and Rahnenfuhrer, 2010). The algorithm used by topGO package takes into account the hierarchical structure of GO graph and transfers annotations from child nodes to parent nodes of the graph for significance testing using Fisher's exact test (Alexa et al., 2006). KEGG enrichment analysis was performed using a custom R script and Fisher's exact test was used for testing the significance of KEGG annotated pathways. In both of these enrichment analyses, only the GO terms/KEGG pathways with significance p-value < 0.05 and with ≥ 5 annotated genes were selected as significantly enriched.

Cluster similarity analysis

Once the significant clusters in LA and HA networks were identified and enrichment analysis was performed, the next step was to calculate the similarity between these significant LA and HA clusters. In this step, the physical and functional similarity between significant LA and HA clusters were calculated. It should be noted that the physical similarity was calculated for all significant LA and HA clusters while functional similarity was calculated only for the clusters with GO enrichment. The reason behind cluster similarity assessment was to understand the physical and functional similarity between the clusters, whether the physical overlap between LA and HA clusters were significant and whether the clusters showed a high degree of functional similarity irrespective of the physical overlap.

Physical similarity

Physical similarity between LA and HA clusters were calculated using a hypergeometric test. For each significant LA cluster, an HA cluster was retrieved and hypergeometric test was performed between the nodes of these clusters to identify the overlap. In this step, only LA - HA similarity was tested since Infomap clustering algorithm generates non overlapping clusters. P-values were generated using the `phyper` function in R environment and the hypergeometric test results were pruned at a significance threshold of p-value < 0.05.

Functional similarity

Functional similarity between LA and HA significant clusters was established by calculating the Gene Ontology semantic similarity (Lord et al., 2003; Schlicker et al., 2006; Wang et al., 2007). In GO enrichment analysis, a number of clusters showed significant enrichment for GO biological process. In this step, the functional similarity only between those clusters showing significant GO enrichment were assessed. For calculating the semantic similarity between GO terms, the Wang method (Wang et al., 2007) as implemented in GOSemSim (Yu et al., 2010) bioconductor package was used (a detailed explanation of GO semantic similarity method is given in section 3.1.2). In this step, semantic similarity was calculated between all enriched LA and HA clusters. For enriched GO terms in each LA or HA cluster, GO terms from another LA or HA cluster was drawn and semantic similarity was calculated between these terms using Wang method and these similarity measurements were combined into a single value using best-match average strategy (BMA) (Yu et al., 2010). These semantic similarity values were termed sim_{CLUS} for future references. The semantic similarity values from this step ranges from 0 to 1, with 0 being the lowest value possible and 1 being the highest possible value.

Although the step mentioned above allows to calculate semantic similarity between two enriched clusters in this analysis, this step does not provide a cut-off threshold to indicate whether the similarity between the two clusters were significant or not. To provide a significant cut-off point for semantic similarity, an empirical approach based on random sampling was utilized. In this

step, all GO biological process annotations for porcine genes were retrieved and sampled two sets of GO terms from these annotations. The number of sampled terms were also kept random and were drawn from the number of GO terms enriched for either LA or HA clusters. Semantic similarity was calculated between these two sets GO terms based on the Wang method. The aim behind this step was to generate a baseline semantic similarity measure from two sets of randomly drawn porcine GO biological process annotations. This whole step was repeated 10,000 times to generate a set of random semantic similarity measures. These random semantic similarity values were termed as sim_{RAND} for further references. Finally, the significance threshold cut-off empirical p-value for each sim_{CLUS} was calculated as:

$$Pval_{Empricial} = \frac{\#\, sim_{RAND} > sim_{CLUS}}{N}, \text{ where N} = 10,000. \tag{3.5}$$

The threshold cut off used here was $Pval_{Empricial} < 0.05$. In the next step, two cluster similarity graphs were generated based on physical similarity assessment and functional similarity assessment. The nodes of these graphs represented LA or HA clusters and edges represented significant similarity measurement (physical or functional) between them. These graphs were visualized using the biological network visualizing platform, Cytoscape.

A schematic diagram of the entire workflow used in this experiment is given in Figure 3.7.

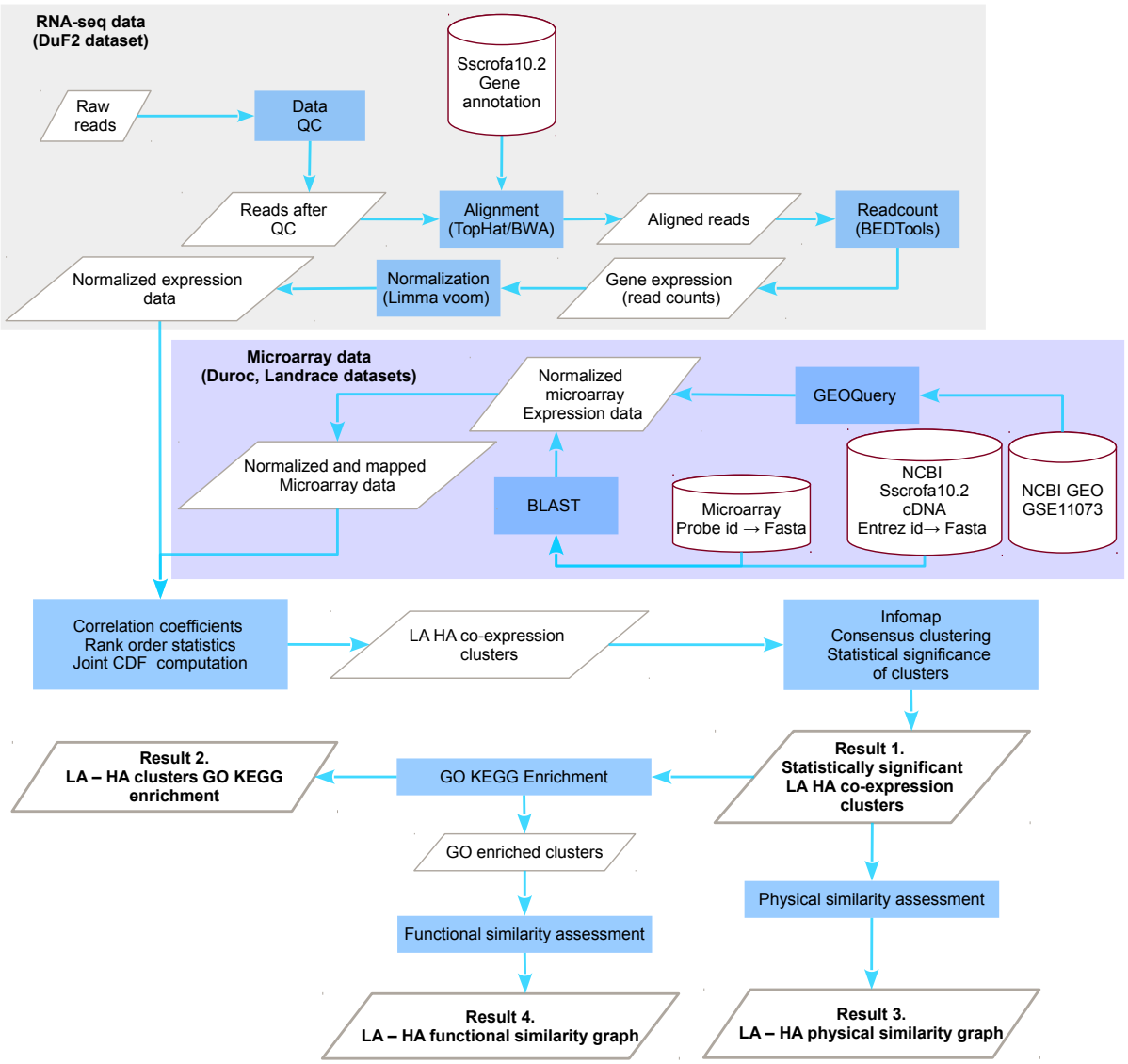

Figure 3.7: Co-expression cluster analysis workflow. *Legend*: White parallelograms with grey outline: Input/output data and results. White cylinders with red outline: data from external databases. Rectangles with light blue shades: various tools and analysis processes used in this workflow.

4. Results and Discussion

This chapter describes the results obtained from the investigation of porcine testis and liver tissue high throughput data through knowledge driven and data driven approaches described in the material and methods chapter. This chapter is divided into two main sections, the first section explains and discusses the results from the knowledge driven integrative analysis performed on gene expression dataset from porcine testis tissues and the second section explains and discusses the results from the data driven analysis performed on expression datasets from porcine liver tissues.

Before detailing the results of the experiments, this subsection gives an overview on data mapping and alignment statistics for the in-house RNA-seq data used in this thesis. Analyzing the alignment of testes sample raw reads to Sscrofa10.2 revealed that $80 \pm 5\%$ of the reads in all the testis samples were aligned to the genome. The alignment statistics of testis samples are given in Table 4.1. Similarly, the alignment of liver raw reads to Sscrofa10.2 gene annotations showed that 74-89% of raw reads in liver were mapped (see Table 4.1).

Previous RNA-seq studies on porcine tissues have shown a wide range of variability in the percentage of reads mapped to the reference genome. It was reported that the up to 44.1% of raw reads from porcine male gonads were mapped to reference genome (Esteve-Codina et al., 2011) and the in-house RNA-seq study reported alignments ranging from 40.8% to 56.63% (Gunawan et al., 2013). In a number of other porcine transcriptomics studies, the percentage of annotated reads varied from 15.6% to 74.9% (Bauer et al., 2010; Chen et al., 2011; Jung et al., 2012; Ramayo-Caldas et al., 2012). This difference in mapping percentages could be the result of a number of factors including GC content, independent cell types, laboratory protocols, primer biases and dinucleotide fragmentation sites (McIntyre et al., 2011). Another factor influencing the percentage of reads mapped to reference genome among multiple studies cited above and the mapping statistics in Table 4.1 could be the choice of reference genome build used for mapping and the quality control parameters used in the initial data quality control phase.

Table 4.1: Testis and Liver samples alignment statistics

Phenotype	sample number	# mapped reads	# unmapped reads	% mapped
LA testis	1	9,370,900	2,591,380	78 %
	2	7,880,721	2,257,957	78 %
	3	22,220,951	4,052,960	85 %
	4	18,367,833	3,254,373	85 %
	5	15,480,913	5,133,215	75 %
HA testis	1	20,814,444	3,609,890	85 %
	2	22,764,774	4,180,618	84 %
	3	9,477,859	2,579,939	79 %
	4	8,344,818	2,422,822	77 %
	5	9,611,381	2728058	78 %
LA liver	1	17,520,764	2,654,624	87 %
	2	25,886,718	3,985,968	87 %
	3	9,811,260	2,148,315	82 %
	4	7,640,379	1,904,670	80 %
	5	18,935,689	24,39,887	89 %
HA liver	1	17,906,455	2,751,493	87 %
	2	7,616,917	1,763,874	81 %
	3	33,133,982	3,993,499	89 %
	4	6,412,629	2,227,964	74 %
	5	7,033,081	2,022,279	78 %

4.1 Pathway based analysis of genes and interactions influencing porcine testis samples from boars with divergent androstenone content in back fat

In this section, the results of the analysis methods described in section 3.2.2.1 are explained and discussed. As described in the method section, the analysis of testis data set was focused around identifying statistically significant gene interactions between HA and LA interaction networks. The results from the identification of significant interactions showed that 1,023 interactions between 826 genes were significant in both HA and LA testis datasets. Analysis of this interaction network revealed that these 1,023 interactions formed into an interaction network and the largest connected component of this network contained 848 edges (interactions) and 563 nodes (genes) (Figure 4.1).

4.1.1 Significant interaction network analysis

Network analysis performed in Cytoscape (section 3.1.2) showed that the significant interaction network had a total of 95 connected components, clustering coefficient value of 0.036, a path length of 8.490 and the average number of neighbors is 2.477. Additional network statistics from this analysis are given in Table 4.2. Analysis of the node degree distribution of the network showed that the network exhibits scale free topology following the power law distribution of node degrees (Figure 4.2), a characteristic nature of interaction networks in biology. Node degree calculations have also revealed that genes such as LOC100623707 (POLR2G), ADCY9, PDE8B,

NUDT2, PDE8B and LOC100620235 (PIK3R1) were some of the highly connected genes in this network. Among the significant interactions in this network, 209 interactions were LA positive, 201 interactions were LA negative, 257 interactions were HA positive and 220 interactions were HA negative (Table 4.2).

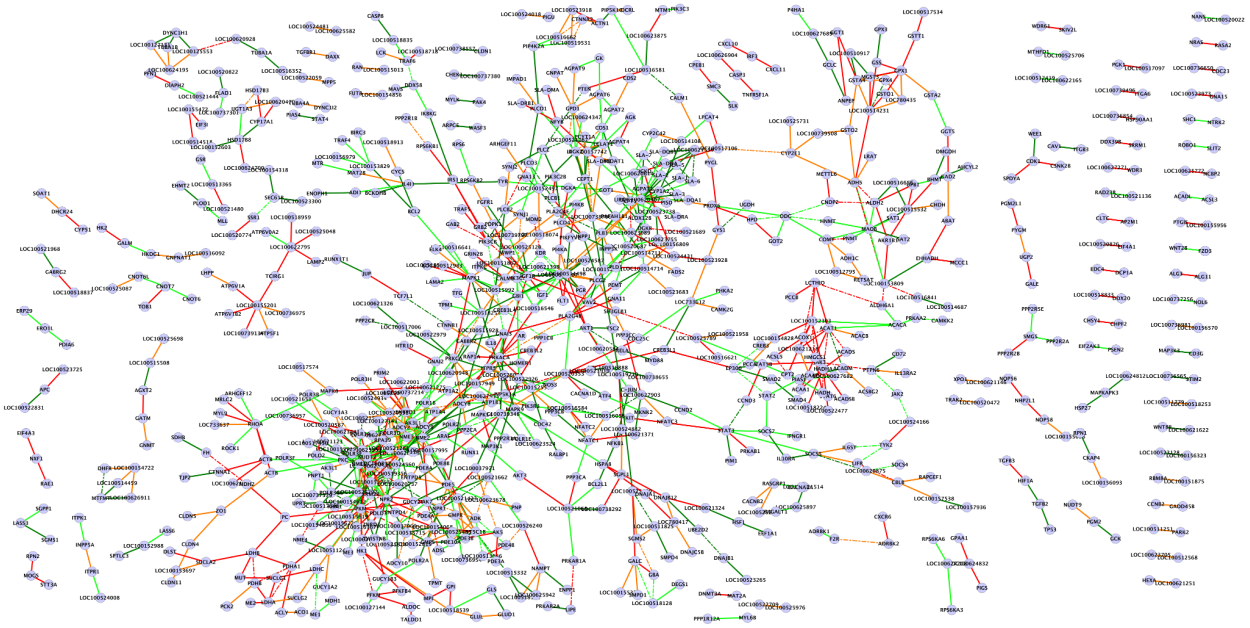

Figure 4.1: Testis HA and LA dataset significant interactions. *Legend*: nodes – genes, edges – interactions with significant z-scores. *Edge legend*: Red solid edges: interactions positive and significant in HA samples, negative in LA samples. Red dashed edges: interactions positive and significant in HA samples, positive in LA samples. Orange solid edges: interactions positive in HA samples, negative and significant in LA samples. Orange dashed edges: interactions negative in HA samples, negative and significant in LA samples. Dark green solid edges: interactions positive and significant in LA samples, negative in HA samples. Dark green dashed edges: interactions positive and significant in LA samples, positive in HA samples. Light green solid edges: interactions positive in LA samples, negative and significant in HA samples. Light green dashed edges: interactions negative in LA samples, negative and significant in HA samples.

Table 4.2: Testis HA LA dataset significant interaction network statistics

Network feature	**Significant interactions network**	**Significant interactions network in enriched pathways**
Number of nodes	826	718
Number of edges	1,023	865
Clustering coefficient	0.036	0.039
Characteristic path length	8.490	8.558
Avg. number of neighbors	2.477	2.409
LA positive interactions	209	173
LA positive significance interactions	35	31
LA negative interactions	201	166
LA negative significance interactions	30	24
HA positive interactions	257	217
HA positive significance interactions	42	39
HA negative interactions	220	189
HA negative significance interactions	29	26

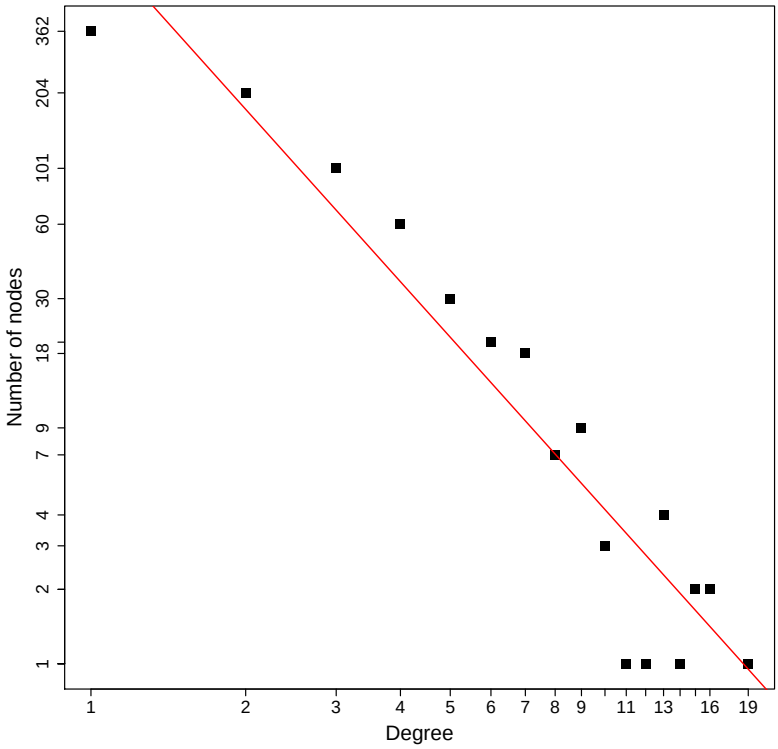

Figure 4.2: Significant interaction network node degree distribution. Power law fit is indicated by the red line in the figure.

4.1.2 Pathway enrichment analysis

The major aim behind pathway enrichment analysis was to annotate significant interactions with metabolic pathways and to identify the key pathways and interactions that might be relevant for porcine testicular steroidogenesis and androstenone synthesis. Pathway enrichment analysis showed that out of 1,023 significant interactions, 865 interactions between 718 genes were enriched in 92 pathways (Table 4.3). Among these enriched pathways, top 5 enriched pathways in terms of the number of interactions were: purine, pyrimidine and glycerophospholipid metabolism pathway, phosphatidylinositol signaling system and Jak-STAT signaling pathway (Table 4.3). Significant interactions in pathways such as synthesis and degradation of ketone bodies, steroid biosynthesis, oxidative phosphorylation, butanoate metabolism, drug metabolism – other enzymes and RNA transport were found only in HA samples where as the interactions in antigen processing and presentation pathway, intestinal immune network for IgA production, autoimmune thyroid disease and allograft rejection pathways were found only in case of LA sample set (Table 4.3).

Table 4.3: KEGG pathway enrichment analysis

pathway id	**pathway name**	**p.adj**	**# total significant interactions**	**# significant interactions in LA samples**	**# significant interactions in HA samples**
ssc04650	Natural killer cell mediated cytotoxicity	0.02	11	3	8
ssc04672	Intestinal immune network for IgA production	0.00	7	7	0
ssc04660	T cell receptor signaling pathway	0.00	13	5	8
ssc04662	B cell receptor signaling pathway	0.00	14	6	8
ssc04612	Antigen processing and presentation	0.00	19	19	0
ssc00360	Phenylalanine metabolism	0.00	6	3	3
ssc04630	Jak-STAT signaling pathway	0.00	33	22	11
ssc00380	Tryptophan metabolism	0.01	5	3	2
ssc04621	NOD-like receptor signaling pathway	0.01	3	2	1
ssc04622	RIG-I-like receptor signaling pathway	0.02	3	2	1
ssc05310	Asthma	0.00	7	7	0
ssc05323	Rheumatoid arthritis	0.00	7	7	0
ssc05322	Systemic lupus erythematosus	0.00	7	7	0
ssc05320	Autoimmune thyroid disease	0.00	16	16	0
ssc00410	beta-Alanine metabolism	0.03	5	1	4
ssc05330	Allograft rejection	0.00	16	16	0
ssc04210	Apoptosis	0.03	7	2	5
ssc05010	Alzheimers disease	0.00	5	2	3
ssc05030	Cocaine addiction	0.03	4	2	2
ssc00280	Valine, leucine and isoleucine degradation	0.00	23	5	18
ssc00270	Cysteine and methionine metabolism	0.02	6	4	2
ssc00260	Glycine, serine and threonine metabolism	0.01	7	2	5
ssc00250	Alanine, aspartate and glutamate metabolism	0.02	5	2	3
ssc00240	Pyrimidine metabolism	0.00	74	37	37
ssc04720	Long-term potentiation	0.03	9	1	8
ssc05416	Viral myocarditis	0.00	17	16	1
ssc05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.00	2	1	1
ssc00340	Histidine metabolism	0.00	4	3	1
ssc00740	Riboflavin metabolism	0.02	2	2	0
ssc00350	Tyrosine metabolism	0.00	11	7	4
ssc00330	Arginine and proline metabolism	0.00	9	7	2

Table 4.3: KEGG pathway enrichment analysis (continued...)

pathway id	**pathway name**	**p.adj**	**# total significant interactions**	**# significant interactions in LA samples**	**# significant interactions in HA samples**
ssc04360	Axon guidance	0.01	10	5	5
ssc04370	VEGF signaling pathway	0.00	19	8	11
ssc03008	Ribosome biogenesis in eukaryotes	0.00	6	1	5
ssc03015	mRNA surveillance pathway	0.01	7	3	4
ssc03013	RNA transport	0.03	9	0	9
ssc03018	RNA degradation	0.02	6	2	4
ssc00760	Nicotinate and nicotinamide metabolism	0.01	8	4	4
ssc04070	Phosphatidylinositol signaling system	0.00	44	24	20
ssc05152	Tuberculosis	0.00	14	5	9
ssc05166	HTLV-I infection	0.01	18	7	11
ssc05160	Hepatitis C	0.01	6	4	2
ssc05134	Legionellosis	0.01	3	2	1
ssc05145	Toxoplasmosis	0.01	7	4	3
ssc00140	Steroid hormone biosynthesis	0.02	8	3	5
ssc00190	Oxidative phosphorylation	0.00	7	0	7
ssc00600	Sphingolipid metabolism	0.00	21	11	10
ssc00230	Purine metabolism	0.00	187	94	93
ssc00620	Pyruvate metabolism	0.00	17	5	12
ssc00650	Butanoate metabolism	0.00	8	0	8
ssc00640	Propanoate metabolism	0.00	13	3	10
ssc04910	Insulin signaling pathway	0.00	16	6	10
ssc04914	Progesterone-mediated oocyte maturation	0.00	11	6	5
ssc00670	One carbon pool by folate	0.00	7	4	3
ssc04012	ErbB signaling pathway	0.02	7	1	6
ssc04010	MAPK signaling pathway	0.02	20	7	13
ssc04962	Vasopressin-regulated water reabsorption	0.03	2	0	2
ssc00480	Glutathione metabolism	0.00	19	6	13
ssc04976	Bile secretion	0.02	7	6	1
ssc04510	Focal adhesion	0.03	21	12	9
ssc04514	Cell adhesion molecules (CAMs)	0.00	17	16	1
ssc00020	Citrate cycle (TCA cycle)	0.00	8	1	7
ssc00010	Glycolysis / Gluconeogenesis	0.00	13	2	11
ssc00071	Fatty acid metabolism	0.00	23	3	20
ssc00072	Synthesis and degradation of ketone bodies	0.00	4	0	4
ssc00983	Drug metabolism - other enzymes	0.02	3	0	3
ssc04530	Tight junction	0.02	17	8	9
ssc00052	Galactose metabolism	0.00	4	3	1

Table 4.3: KEGG pathway enrichment analysis (continued...)

pathway id	**pathway name**	**p.adj**	**# total significant interactions**	**# significant interactions in LA samples**	**# significant interactions in HA samples**
ssc04520	Adherens junction	0.00	12	2	10
ssc00100	Steroid biosynthesis	0.05	2	0	2
ssc05223	Non-small cell lung cancer	0.02	5	1	4
ssc05222	Small cell lung cancer	0.04	4	1	3
ssc05221	Acute myeloid leukemia	0.00	9	4	5
ssc05220	Chronic myeloid leukemia	0.02	6	2	4
ssc05210	Colorectal cancer	0.00	7	4	3
ssc05212	Pancreatic cancer	0.00	7	6	1
ssc05211	Renal cell carcinoma	0.03	4	1	3
ssc05214	Glioma	0.04	8	4	4
ssc05213	Endometrial cancer	0.05	3	1	2
ssc05216	Thyroid cancer	0.01	3	3	0
ssc05215	Prostate cancer	0.00	12	5	7
ssc05218	Melanoma	0.02	8	4	4
ssc04114	Oocyte meiosis	0.00	16	5	11
ssc00520	Amino sugar and nucleotide sugar metabolism	0.00	9	2	7
ssc05200	Pathways in cancer	0.00	32	23	9
ssc04141	Protein processing in endoplasmic reticulum	0.00	23	10	13
ssc04145	Phagosome	0.00	11	5	6
ssc00563	GPI-anchor biosynthesis	0.02	3	0	3
ssc00561	Glycerolipid metabolism	0.00	21	12	9
ssc04150	mTOR signaling pathway	0.00	6	3	3
ssc00565	Ether lipid metabolism	0.00	16	11	5
ssc00564	Glycerophospholipid metabolism	0.00	55	29	26

Although the pathways such as purine, pyrimidine and glycerophospholipid metabolism pathway, phosphatidylinositol signaling system and Jak-STAT signaling pathway were some of the top enriched pathways in this analysis, literature references (Altamirano et al., 2009; Fix et al., 2004; Losel et al., 2003; Ray et al., 2013; Sakata et al., 2000; Sharifi and Mottaghi, 2012) suggest that a number of these pathways were activated by steroid hormones through various signaling pathways and may not have directly influenced steroidogenesis. However, some of the enriched pathways of interest were: steroid hormone biosynthesis pathway, fatty acid metabolism, oxidative phosphorylation, glutathione metabolism and sphingolipid metabolism. These pathways were chosen as pathways of interest since steroid hormone biosynthesis is the major pathway synthesizing testosterone and androstenone and also on account of literature based evidences that the metabolites from glutathione metabolism, sphingolipid metabolism and fatty acid metabolism can influence steroid hormone biosynthesis (Chen et al., 2008a; Hu et al., 2010; Lucki and Sewer, 2010). Based on these enriched pathways and significant interactions, five major assumptions were formalized on the synthesis and maintenance of steroidogenesis and androstenone metabolism in

the porcine testis samples. These assumptions are discussed in the following subsections.

4.1.2.1 Steroid hormone biosynthesis

As expected, steroid hormone biosynthesis pathway is one of the pathways enriched for significant interactions (Table 4.3). In this pathway, five significant interactions (correlations) were positive in HA sample set and three significant interactions were positive in LA sample set (Figure 4.3). One of the interactions positive in HA sample set was the interaction between the genes CYP17A1 and HSD17B3 (Figure 4.3). The enzyme encoded by CYP17A1 gene converts 17 α-Hydroxy progesterone into androstenedione (Boron and Boulpaep, 2005) and the hydroxysteroid dehydrogenase enzyme encoded by HSD17B3 gene catalyzes the conversion of androstenedione to testosterone (Payne and Hardy, 2007). Since CYP17A1 gene product is involved in the initial steps of steroid hormone synthesis, a number of studies have reported this gene as a candidate gene for androstenone biosynthesis (Billen and Squires, 2009; Leung et al., 2010; Moe et al., 2007a,b).

Another HA positive interaction in these results was the interaction between the genes CYP17A1 and LOC100620470 (HSD17B6) (Figure 4.3). In this second interaction involving CY17A1 gene, the interactant LOC100620470 (HSD17B6) encodes 17 β-hydroxysteroid dehydrogenase type 6 enzyme, which catalyzes the conversion of testosterone back to androstenedione (Tindall and Mohler, 2009). Although none of the high throughput studies on androstenone synthesis mention LOC100620470 (HSD17B6) expression in relation with androstenone, this gene is reported to be in an androstenone related QTL region (Grindflek et al., 2011). The third HA positive interaction in steroid hormone biosynthesis pathway was the interaction between the genes LOC100620470 (HSD17B6) and UGT1A3 (Figure 4.3). The enzyme encoded by UGT1A3 gene, a LOC100620470 (HSD17B6) interaction partner catalyzes the glucuronidation of testosterone to testosterone glucuronide (Kuuranne et al., 2003). The fourth HA positive interaction in this pathway was between genes HSD17B8 and LOC100624700 (UGT2C1) (Table 4.3). Among these interaction partners, the former codes for the enzyme hydroxysteroid (17-β) dehydrogenase 8, primarily involved in testosterone inactivation (Hartley et al., 2000) and the latter encodes UDP-glucuronosyltransferase 2C1 enzyme. Although UDP-glucuronosyltransferase 2C1 enzyme is known to catalyze the conjugation of endogenous compounds, its exact function in relation with hydroxysteroid dehydrogenase enzyme remains unclear. The final positive interaction in HA samples was the interaction between genes HSD17B3 and UGT1A3 (Figure 4.3). As described above, the enzyme encoded by HSD17B3 converts androstenedione to testosterone and UGT1A3 gene product catalyzes the glucuronidation of testosterone to testosterone glucuronide. The evidences described here could indicate that both testosterone synthesis and degradation steps were active in HA sample set.

In case of LA sample set, positive interactions were CYP17A1 – HSD17B8 interaction, HSD17B8 – UGT1A3 interaction and HSD17B8 - LOC100152603 (UDP-glucuronosyltransferase) interaction (Table 4.3). As mentioned above, CYP17A1 codes for an enzyme catalyzing 17α-Hydroxy progesterone to androstenedione conversion and the enzyme hydroxysteroid (17-β) dehydrogenase 8 encoded by HSD17B8 gene inactivates testosterone. The remaining interaction partners of

HSD17B8 gene, UGT1A3 and LOC100152603 (UDP-glucuronosyltransferase) primarily catalyzes the conjugation and removal of various endogenous compounds. It should be noted that in all the three interactions positive in LA sample sets, the gene HSD17B8 was one of the interaction partners and the major function of the protein encoded by this gene is testosterone inactivation. These results and evidences could be and indication that in low androstenone animals, testicular testosterone concentrations were primarily affected by a low amount of synthesis coupled with active testosterone inactivation and degradation steps.

A recent study (Lervik et al., 2013) has shown that estimated breeding value of androstenone was positively related to plasma testosterone levels and it was also shown that genetic correlation between androstenone (plasma and fat) and sex steroids were high in pure bred Duroc and Landrace populations (Grindflek et al., 2011). Based on these evidences from published studies and the observation that the enzymes involved in the synthesis of testosterone also catalyzes androstenone synthesis and since both the compounds are derived from pregnenolone (James Squires, 2010), it could be postulated that in HA animals, an active testosterone synthesis could also imply active synthesis of androstenone.

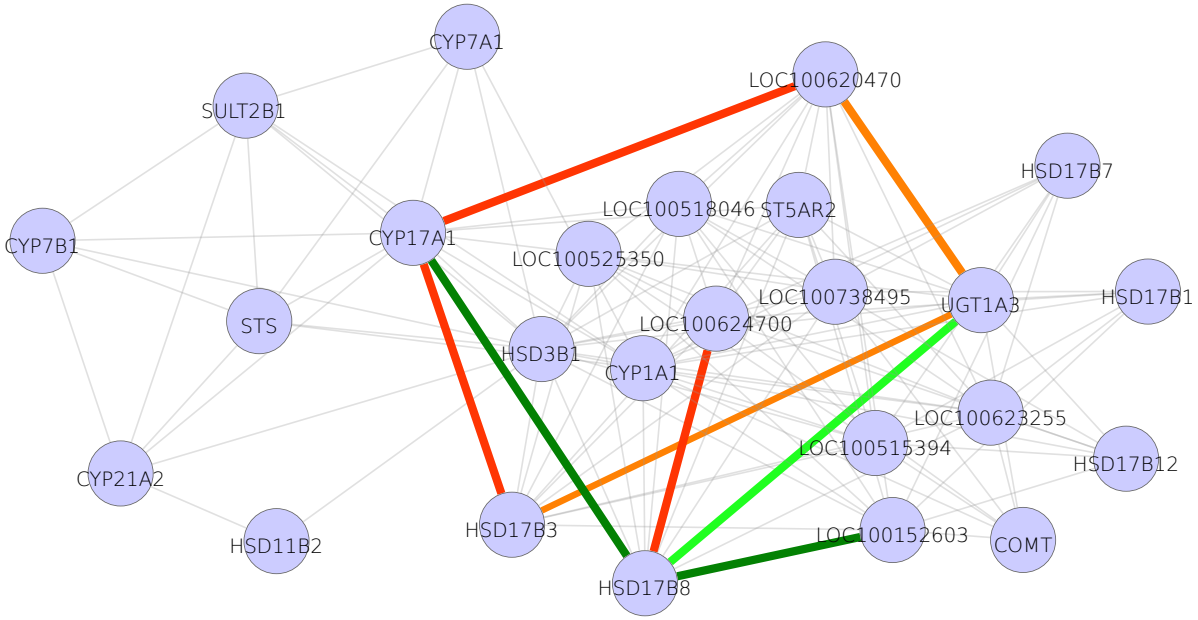

Figure 4.3: Steroid hormone biosynthesis pathway. *Legend*: Red solid edges: interactions positive and significant in HA samples, negative in LA samples. Red dashed edges: interactions positive and significant in HA samples, positive in LA samples. Orange solid edges: interactions positive in HA samples, negative and significant in LA samples. Orange dashed edges: interactions negative in HA samples, negative and significant in LA samples. Dark green solid edges: interactions positive and significant in LA samples, negative in HA samples. Dark green dashed edges: interactions positive and significant in LA samples, positive in HA samples. Light green solid edges: interactions positive in LA samples, negative and significant in HA samples. Light green dashed edges: interactions negative in LA samples, negative and significant in HA samples. Grey edges: Non significant interactions, part of KEGG network data.

4.1.2.2 Glutathione metabolism

Glutathione metabolism was another major metabolic pathway enriched for significant correlations (interactions) in these results (Table 4.3). Literature evidence suggests that the depletion of intracellular gluathione pool significantly decreases testosterone production (Chen et al., 2008a)

and that a decrease in glutathione peroxidase (Gpx) activity affects testosterone synthesis since Gpx activity reduces lipid peroxidation (Chandra et al., 2000). Additionally, it has also been indicated that alterations in glutathione redox cycle might play significant roles in detoxifying mechanisms in testes (Mori et al., 1989).

This analysis identified seven GPX1 gene interactions to be positive in HA sample set (Figure 4.4). GPX1 gene encodes glutathione peroxidase enzyme, primarily involved in the detoxification of hydrogen peroxide. GSTA2, a GPX1 interaction partner in glutathione metabolism pathway exhibits high activity against lipid peroxidation (Fishbein, 2011). GSTA4, another GPX1 interactant metabolizes lipid peroxidation product 4-hydroxynonenal (4-HNE) by conjugating it with glutathione (GSH) (Sharma et al., 2011). GPX1 – GSTA2 interaction (correlation) and GPX1 – GSTA4 interaction (correlation) were positive in HA phenotype, possibly indicating that the combined action of the enzymes encoded by these genes reduced lipid peroxidase activity in HA samples and thus had a positive effect on testicular steroidogenesis.

In this scenario, it should also be taken into account that the majority of reactive oxygen species (ROS), the primary agent in lipid peroxidation is a by-product of mitochondrial oxidative phosphorylation (West et al., 2011). Pathway enrichment analysis and further investigations have shown that oxidative phosphorylation pathway was enriched for significant interactions (Table 4.3) and that a number of interactions (correlations) in oxidative phosphorylation pathway were positive in HA dataset (Table 4.3, Figure 4.5). From these results it could be assumed that in HA samples, an active glutathione metabolism pathway was balancing the negative side effects of an active mitochondrial oxidative phosphorylation, specifically, the peroxidation of lipids triggered by ROS. Interaction evidences also shows the gene GGT1 as an interaction partner for the gene GSTA4 and that the interactions were positive in HA dataset (Figure 4.4). Conversion of glutathione (GSH) into cysteinyl glycine and γ-glutamate catalyzed by GGT1 gene product is an essential step that helps to maintain cellular levels of glutathione and cysteine. GGT1 deficient male mice have been shown to be infertile (Kumar et al., 2000). Although KEGG interaction network includes an interaction between GSTA4 and GGT1, at this point, additional evidence for this interaction could not be found in any published literature.

Based on the evidences stated above, it could be postulated that in HA testis tissues, an active glutathione metabolic pathway resulted in reduced lipid peroxidase activity and thus an increased steroidogenesis and androstenone biosynthesis. In this regard, the genes GPX1 and its interactions partners such as GST family genes GSTA4 and GSTA2 and gene GGT1 in glutathione metabolism as could be further investigated as candidate biomarkers for their involvement in porcine testicular steroid biosynthesis and androstenone biosynthesis. Among the genes involved in significant interactions in this pathway, the gene GSTO1 is previously reported to be differentially expressed in high androstenone (Duroc) boars (Moe et al., 2007b).

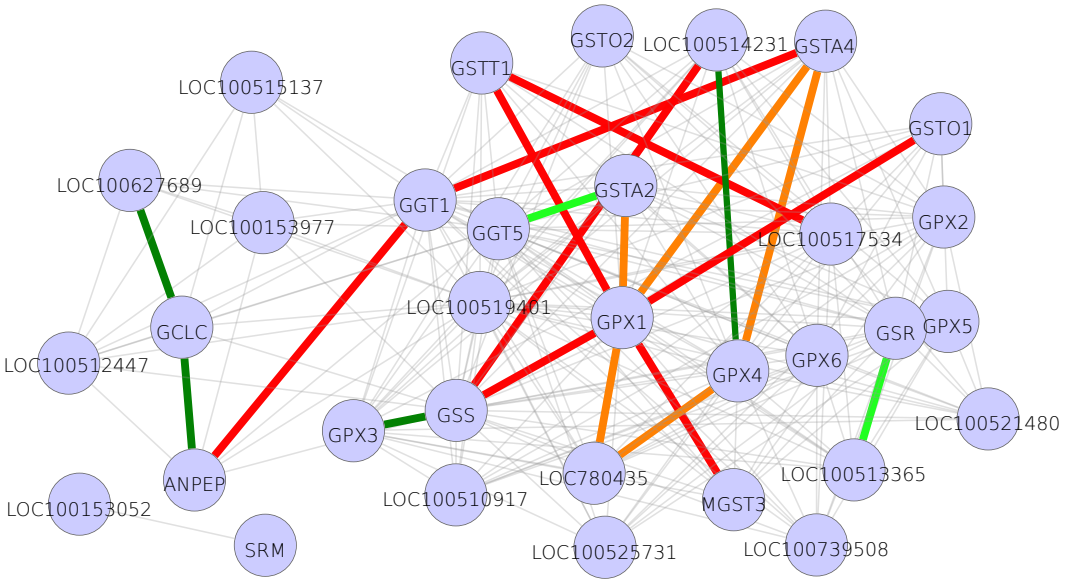

Figure 4.4: Glutathione metabolism. *Legend*: Red solid edges: interactions positive and significant in HA samples, negative in LA samples. Red dashed edges: interactions positive and significant in HA samples, positive in LA samples. Orange solid edges: interactions positive in HA samples, negative and significant in LA samples. Orange dashed edges: interactions negative in HA samples, negative and significant in LA samples. Dark green solid edges: interactions positive and significant in LA samples, negative in HA samples. Dark green dashed edges: interactions positive and significant in LA samples, positive in HA samples. Light green solid edges: interactions positive in LA samples, negative and significant in HA samples. Light green dashed edges: interactions negative in LA samples, negative and significant in HA samples. Grey edges: Non significant interactions, part of KEGG network data.

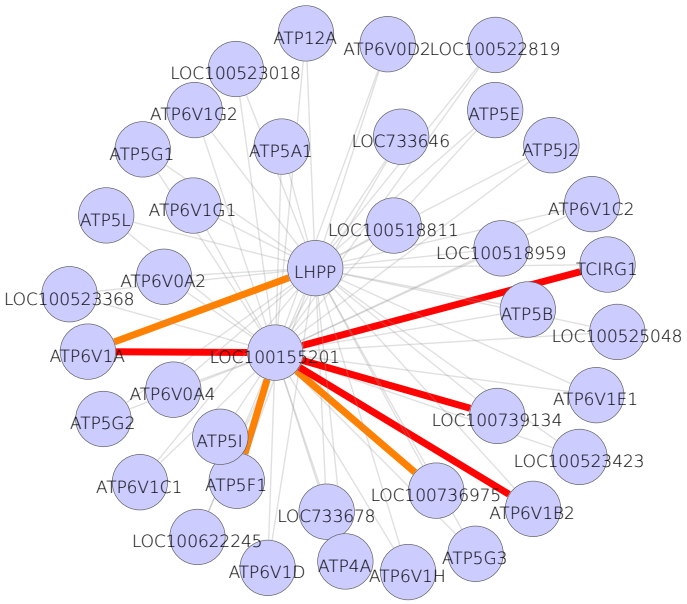

Figure 4.5: Oxidative phosphorylation. *Legend*: Red solid edges: interactions positive and significant in HA samples, negative in LA samples. Red dashed edges: interactions positive and significant in HA samples, positive in LA samples. Orange solid edges: interactions positive in HA samples, negative and significant in LA samples. Orange dashed edges: interactions negative in HA samples, negative and significant in LA samples. Dark green solid edges: interactions positive and significant in LA samples, negative in HA samples. Dark green dashed edges: interactions positive and significant in LA samples, positive in HA samples. Light green solid edges: interactions positive in LA samples, negative and significant in HA samples. Light green dashed edges: interactions negative in LA samples, negative and significant in HA samples. Grey edges: Non significant interactions, part of KEGG network data.

4.1.2.3 Sphingolipid metabolism

Sphingolipids are a class of lipids composed of an aliphatic amino alcohols and a sphingosine (long chain base) backbone. These lipids have been established to play a significant role in steroidogenic pathway by acting as secondary messengers, paracrine, autocrine regulators and nuclear receptors (Lucki and Sewer, 2010). Literature evidences (Meroni et al., 2000; Morales et al., 2003) show that ceramides (Cer, N-acylsphingosine), a major class of sphingolipids can suppress testicular StAR gene expression, testosterone biosynthesis and regulate hCG stimulated steroidogenesis in rat Leydig cells. Studies have also shown that sphingosine-1-phosphate (S1P), an intracellular sphingolipid inhibits germ cell apoptosis in human testis (Suomalainen et al., 2003) and modulates lutenizing hormone signaling (Sanz et al., 2013). Sphingomyelin, another sphingolipid is shown to enhance steroid hormone synthesis (Isabella Pörn et al., 1991). It is also suggested that sphingosine (SPH), another sphingolipid class member acts as an antagonist for steroid hormone biosynthesis nuclear receptor SF1 (Urs et al., 2006).

Sphingolipid metabolism was one of the pathways found to be enriched for significant interactions in pathway enrichment results (Table 4.3). A total of 10 interactions in this pathway were positive for HA samples (Figure 4.6). Among these HA positive interactions, gene GALC was involved in 4 interactions (Figure 4.6). The protein encoded by this gene hydrolyzes the galactose ester double bonds of various sphingolipids including galactoceramide and converts into N-acylsphingosine (ceramide) (Schomburg et al., 2003). The first interaction partner of GALC was the gene SMPD1, which encodes a sphingomyelinase enzyme that converts sphingomyelin to ceramide (Schomburg et al., 2003). GBA gene was the second HA positive GALC interaction partner and the product of this gene hydrolyzes D-glucosyl-N-acylsphingosine to D-glucose and N-acylsphingosine. LOC100155321 (ACER2) was the third GALC interactant in HA positive interactions and the product of this gene catalyzes the hydrolysis of N-acylsphingosine to sphingosine (Lennarz and Lane, 2013). In case of gene LOC100525450 (CERS1), the final GALC interaction partner in HA positive interactions, it is speculated that the enzyme encoded by this gene is either a ceramide synthase or a modulator. Although ceramide synthases have been shown to catalyze the de novo synthesis of ceramides (Hannun and Obeid, 2008), the function of the gene LOC100525450 (CERS1) or its product in relation to GALC could not be pinpointed at this time. The results also show that three interactions involving the gene SGMS2 were also positive in HA samples (Figure 4.6). The enzyme encoded by the gene SGMS2 is involved in the synthesis of sphingomyelin from ceramides (Abelson et al., 1999). The interaction partners of SGMS2 in HA positive interactions were the genes LOC100525450 (CERS1), GBA and LOC100511825 (UGT8). As mentioned above, the product of the gene LOC100525450 (CERS1) is speculated to be a ceramide synthase or a modulator and the enzyme encoded by the GBA gene hydrolyzes D-glucosyl-N-acylsphingosine to D-glucose and ceramide. The enzyme encoded by LOC100511825 (UGT8) catalyzes the transfer of galactose to ceramide during the synthesis of galactocerebrosides (Chalfant and Poeta, 2010). An additional HA positive interaction in this pathway was the interaction between the genes LOC100738292 (SPHK2) and SGPL1. LOC100738292 (SPHK2) gene product phosphorylates sphingosine to sphingosine-1-phosphate (McQueen, 2010). The enzyme encoded by the gene SGPL1 cleaves sphingoid bases such as sphingosine-1-phosphate into fatty aldehydes and

phosphoethanolamine (Hirabayashi et al., 2006). From these evidences at the gene level, it could be speculated that in HA samples, ceramides were mainly generated by the conversion/hydrolysis of other sphingolipids such as sphingomyelin or D-glucosyl-N-acylsphingosine and that these ceramides were further converted to galactocerebrosides or to sphingosine and finally into fatty aldehydes and phosphoethanolamine.

In these results, a total of 11 interactions in sphingolipid metabolic pathway were positive for LA samples (Figure 4.6). The gene LOC100152988 (KDSR) was involved in two out of 11 LA positive interactions (Figure 4.6). One of the interaction partners of LOC100152988 (KDSR) was the gene SPTLC3. The enzyme encoded by SPTLC3 converts palmitoyl-CoA and L-serine into 3-ketodihydrosphingosine, initiating de novo synthesis of sphingolipids (Hanada, 2003). The reductase enzyme encoded by LOC100152988 (KDSR) reduces 3-ketodihydrosphingosine into dihydrosphingosine (Chauhan, 2008). The second interaction partner of LOC100152988 (KDSR) was the gene LASS6. LASS6 gene encodes a ceramide synthase enzyme, Ceramide synthase 6 and it is shown that ceramide synthases (CerS) are involved in the acylation of dihydro sphingosine to dihydroceramide, a precursor of ceramide (Hannun and Obeid, 2008). From these interactions it could be speculated that sphingolipid de novo synthesis was active in case of LA samples. Similar to HA samples, an interaction between a gene coding for an enzyme involved in the synthesis of sphingomyelin and a gene coding for ceramide synthase or modulator was found to be positive in LA animals. This interaction was between the genes LASS3 and SGMS1 (Figure 4.6). An interaction between the genes LOC100512419 (PPAP2B) and LOC100622165 (ACER1) was also found to be LA positive. LOC100512419 (PPAP2B) hydrolyzes sphingosine-1-phosphate (Abelson et al., 1999) and LOC100622165 (ACER1) hydrolyzes ceramide to sphingosine.

Literature based evidences (Bartke and Hannun, 2009; Isabella Pörn et al., 1991; Lucki and Sewer, 2010; Meroni et al., 2000; Merrill, 2002; Morales et al., 2003; Sanz et al., 2013; Suomalainen et al., 2003; Urs et al., 2006) indicate that elevated amounts of ceramide negatively affects steroid biosynthesis and the evidences at the genomic level from this analysis suggest active de novo sphingolipid synthesis steps in LA animals. Based on these genomic level evidences, it could be assumed that the elevated concentrations of ceramide in LA animals could be one of the contributing factors to reduced steroid synthesis and possibly reduced androstenone biosynthesis in this phenotype. Although there were several interactions positive in HA animals suggesting the conversion of various sphingolipids to ceramide in these animals, based on the results, it could be speculated that the ceramide levels in these animals were maintained by its conversion either to galactocerebrosides or to fatty aldehydes, mainly by the action of LOC100155321 (ACER2), LOC100738292 (SPHK2) and SGPL1 gene products.

Building around the aforesaid speculations and the literature evidences from model organisms, sphingolipids such as ceramide, sphingosine and sphigosine-1-phosphate and genes involved in sphingolipid metabolic pathway such as GALC, LOC100152988 (KDSR), SGMS1, SGMS2, SMPD1 and SMPD4 could be further investigated as candidate biomarkers for their involvement in porcine steroid hormone biosynthesis and androstenone biosynthesis pathways. From Figure 4.6 it can be seen that several other interactions were positive in either one of the phenotypes, but due to the lack of additional literature or database evidences to support these interactions, these

interactions were dropped from further investigation.

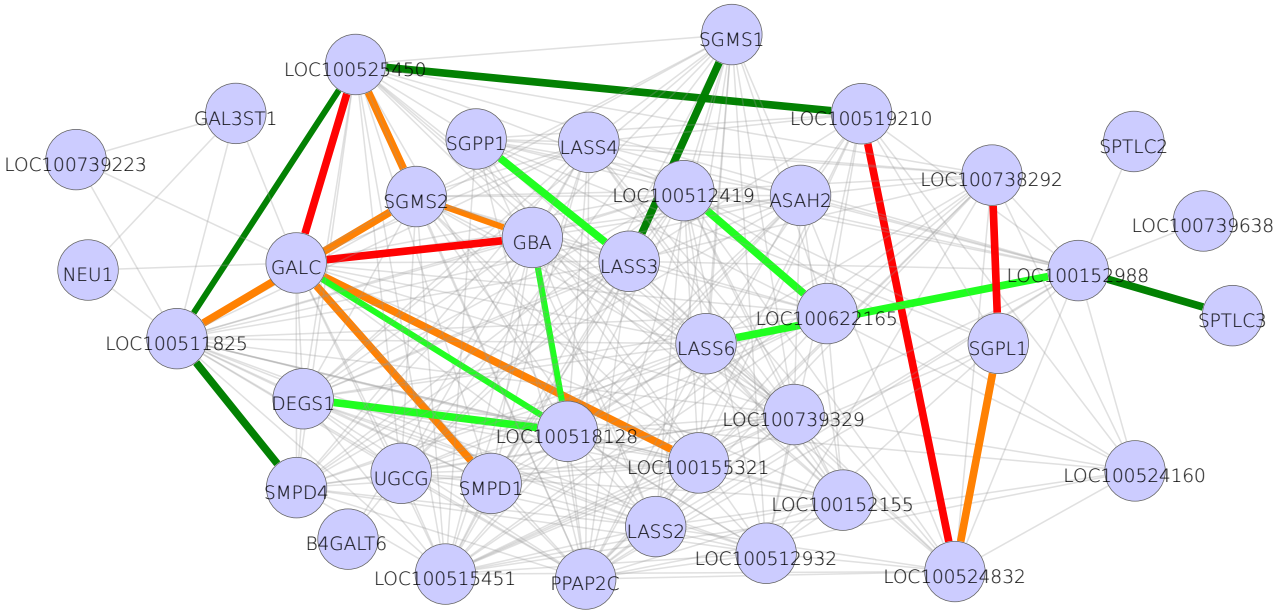

Figure 4.6: Sphingolipid metabolism. *Legend*: Red solid edges: interactions positive and significant in HA samples, negative in LA samples. Red dashed edges: interactions positive and significant in HA samples, positive in LA samples. Orange solid edges: interactions positive in HA samples, negative and significant in LA samples. Orange dashed edges: interactions negative in HA samples, negative and significant in LA samples. Dark green solid edges: interactions positive and significant in LA samples, negative in HA samples. Dark green dashed edges: interactions positive and significant in LA samples, positive in HA samples. Light green solid edges: interactions positive in LA samples, negative and significant in HA samples. Light green dashed edges: interactions negative in LA samples, negative and significant in HA samples. Grey edges: Non significant interactions, part of KEGG network data.

4.1.2.4 Fatty acid metabolism

Fatty acid metabolism was also one of the enriched pathways in the analysis results (Table 4.3). The β oxidation (catabolic) part of fatty acid metabolism breaks down fatty acids to acetyl-CoA which then enters TCA cycle and electron transport chain metabolic pathways for energy generation.

A total of 23 interactions in fatty acid metabolism were significant in pathway enrichment analysis (Figure 4.7). Out of the 23 interactions, 20 interactions were positive in HA samples and 3 interactions were positive in LA samples (Figure 4.7), possibly indicating an active fatty acid metabolic pathway in HA animals. Eight out of the twenty interactions in HA samples had the gene HADHA as one of the interaction partners (Figure 4.7). The gene HADHA codes for mitochondrial trifunctional protein alpha subunit, an enzyme necessary for the final steps mitochondrial beta oxidation of fatty acids (Cheng and Bostwick, 2011). This suggests that the fatty acid oxidation might be highly active in HA samples, oxidizing fatty acids to acetyl-CoA. Acetyl-CoA is also the starting molecule for de novo synthesis of cholesterol. The results also show that the interactions between acetyl-CoA acetyltransferase genes and HADHA were also positive in HA animals. These interactions were: ACAT1 – HADHA interaction and LOC100152303 (ACAT2) – HADHA interaction (Figure 4.7). Enzymes encoded by the genes ACAT1 and LOC100152303 (ACAT2) belongs to the thiolase family of enzymes and the major function of these enzymes is

catalyzing the synthesis of acetoacetyl-CoA from two units of acetyl-CoA (Frey and Hegeman, 2007). Acetoacetyl-CoA generated as a result of this reaction enters mevalonate pathway leading to cholesterol synthesis (Mander and Liu, 2010). It has been shown that cholesterol used in steroidogenesis could be derived from cholesteryl ester mobilization, selective uptake of cholesteryl esters or de novo synthesis of cholesterol in cytosol (Hu et al., 2010). In this regard, based on the results described above, it could be hypothesized that acetoacetyl-CoA derived from an active fatty acid metabolic pathway in HA animals could have enhanced the de novo synthesis of cholesterol in testis tissues of HA animals. Cholesterol synthesized in this manner might be also entering steroidogenic and androstenone biosynthetic pathways finally resulting in higher amounts of androgens in these animals.

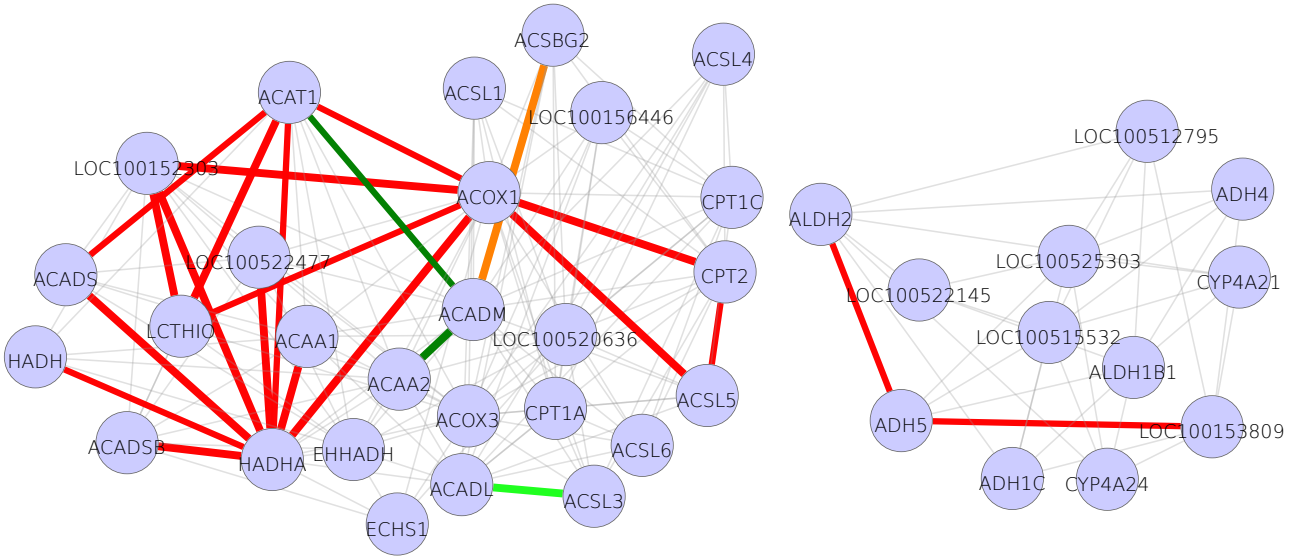

Figure 4.7: Fatty acid metabolism. *Legend*: Red solid edges: interactions positive and significant in HA samples, negative in LA samples. Red dashed edges: interactions positive and significant in HA samples, positive in LA samples. Orange solid edges: interactions positive in HA samples, negative and significant in LA samples. Orange dashed edges: interactions negative in HA samples, negative and significant in LA samples. Dark green solid edges: interactions positive and significant in LA samples, negative in HA samples. Dark green dashed edges: interactions positive and significant in LA samples, positive in HA samples. Light green solid edges: interactions positive in LA samples, negative and significant in HA samples. Light green dashed edges: interactions negative in LA samples, negative and significant in HA samples. Grey edges: Non significant interactions, part of KEGG network data.

4.1.2.5 Cyclic AMP – PKA/PKC signaling

In addition to the interactions in significant pathways, additional interactions which could be relevant in maintaining steroidogenesis in porcine testes tissues were also found in the results. A number of these identified interactions were part of cAMP (cyclic AMP)/PKA signaling, although this pathway was neither represented in the KEGG pathway interaction data used in this analysis nor enriched for significant interactions. Cyclic AMP/PKA signaling pathway is one of the primary signaling cascades maintaining and regulating steroidogenesis (Stocco, 2005). Cyclic-AMP/PKA signaling pathway activation of steroidogenesis is initiated by trophic hormones, which activate G-proteins. G-proteins stimulate adenylate cyclases, thus increasing the levels of intracellular cAMP which further activates protein kinase A (PRKACA). An activated protein kinase A phosphorylates transcription factors such as steroidogenic factor 1 (NR5A1), GATA binding protein 4 (GATA4), cAMP response-element binding protein (CREB) and cAMP

response element modulator (CREM) which activate the genes involved in steroidogenesis (Stocco, 2005).

It was found that the interaction between the genes ADCY9 and PRKCA was significant and positive in HA samples (Figure 4.8). The gene ADCY9 codes for the enzyme adenylate cyclase type 9, which catalyzes the conversion of ATP to cyclic AMP and diphosphate (Hacker et al., 1998). PRKACA, as mentioned above, upon cAMP activation phosphorylates certain transcription factors which activates the genes involved in steroidogenesis. The interaction between the genes PRKCA and CREB3L2 was also found to be significant and positive in HA animals. CREB3L2 is described as cAMP responsive element binding protein (CREB) 3-like 2, but whether the transcription factor encoded by this gene activates the genes involved in steroidogenesis is unknown as of now.

Interestingly, it was also found that two interactions involving adenylate cylases class of genes and guanine nucleotide binding protein class of genes were positive in LA animals. These interactions were: ADCY9 - GNAI2 interaction and LOC100739348 (ADCY8) - GNAI3 interaction (Figure 4.8). Contrary to the interactions observed in HA animals, the interactions found in LA animals were inhibitory. One of the functions of guanine nucleotide binding protein family is the inhibition of adenylate cyclases (Näsman et al., 2002), indicating that GNAI gene products were possibly inhibiting the action of ADCY gene products in LA animals. Another LA positive interaction in these results was the interaction between the genes ADCY2 and PRKCA. ADCY2, similar to other adenylate cyclases, catalyzes the synthesis of cAMP. Gene PRKCA codes for the alpha subunit of the protein protein kinase C (PKC). In a similar manner to PKA, PKC has also been shown to be activated by trophic hormones and stimulates adenylate cyclase activity indicating that in addition to PKA, PKC also influences gonadal steroidogenesis (Manna et al., 2009; Stocco, 2005). But studies done over the years have demonstrated that PRKCA (PKC) is a weak inducer of steroidogenesis and that progesterone synthesis in rat Leydig cells is only moderately elevated by PKC activation (Jo et al., 2005; Manna et al., 2007; Manna and Stocco, 2005). In contrast, Fleury et al. (2004) showed that the mutation of PRKACA (PKA) phosphorylation sites in StAR protein reduced steroidogenesis by 70-80%. These published evidences points out PRKACA (PKA) as a major steroidogenesis activator and PRKCA (PKC) as an auxiliary activator of steroidogenesis. By piecing together the interaction results at the genomic level and information from published articles, it could be speculated that in HA animals an active cAMP/PKA signaling results in higher steroidogenic activity. But in case of LA animals, although cAMP/PKC based signaling of steroidogenesis was active, the inhibition of adenylate cylases by guanine nucleotide binding proteins might be slowing down the steroid hormone synthesis machinery and thus could be affecting androstenone synthesis.

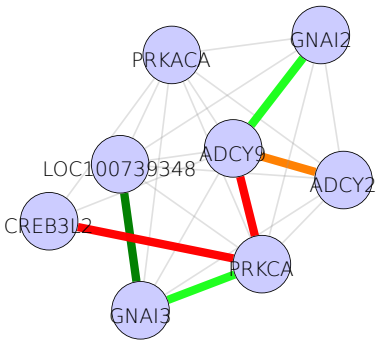

Figure 4.8: Cyclic AMP – PKA/PKC signaling. *Legend*: Red solid edges: interactions positive and significant in HA samples, negative in LA samples. Red dashed edges: interactions positive and significant in HA samples, positive in LA samples. Orange solid edges: interactions positive in HA samples, negative and significant in LA samples. Orange dashed edges: interactions negative in HA samples, negative and significant in LA samples. Dark green solid edges: interactions positive and significant in LA samples, negative in HA samples. Dark green dashed edges: interactions positive and significant in LA samples, positive in HA samples. Light green solid edges: interactions positive in LA samples, negative and significant in HA samples. Light green dashed edges: interactions negative in LA samples, negative and significant in HA samples. Grey edges: Non significant interactions, part of KEGG network data.

Based on the assumptions and speculations discussed above, a hypothetical metabolic pathway was sketched to illustrate the different mechanisms governing the biosynthesis of testosterone, androstenenone and related steroids in the sample HA and LA organisms (Figure 4.9). As represented in Figure 4.9 the proposed hypothesize is as follows: the combined action of cAMP-PKA/PKC signaling, glutathione metabolism, sphingolipid metabolism and fatty acid metabolism was affecting steroid hormone synthesis and therefore androstenone biosynthesis in both HA and LA animals. In HA samples, one of the factors contributing to high androstenone could be that steroidogenesis and hence androstenone synthesis in these animals were activated by trophic hormone signaling through cAMP-PKA (PRKACA) signaling. Additionally, these pathways could have been further boosted by anti lipid peroxidation activity by members of glutathione metabolism pathway and de novo synthesis of cholesterol as a result of an active fatty acid metabolic pathway. In case of LA samples, it could be assumed that a weak cAMP-PKC (PRKCA) based signaling of steroidogenesis activation and synthesis of ceramide by sphingolipid metabolic pathway, which inhibits steroidogenesis could be the reason for a low steroidogenesis and hence low androsteonone synthesis. Since pig and humans share similarities at the genetic, anatomic and physiological level, it can be expected that the hypothetical pathway depicted in Figure 4.9 share a great deal of phylogenetic similarity with the well characterized metabolic interactions in human. In addition to the pathways discussed here, from Table 4.3 it can be seen that a number of other pathways were also enriched. As discussed previously, published literature suggest that these pathways might be activated by steroid hormones and may not have a direct role in maintaining or regulating steroid hormone synthesis. Figure 4.10 depicts the interaction of a number of these over represented pathways (Table 4.3) with steroid hormone biosynthesis pathway. Figure 4.11 gives a detailed illustration of the proposed hypothetical mechanism of androstenone regulation along with significant interactions in each of the metabolic pathways.

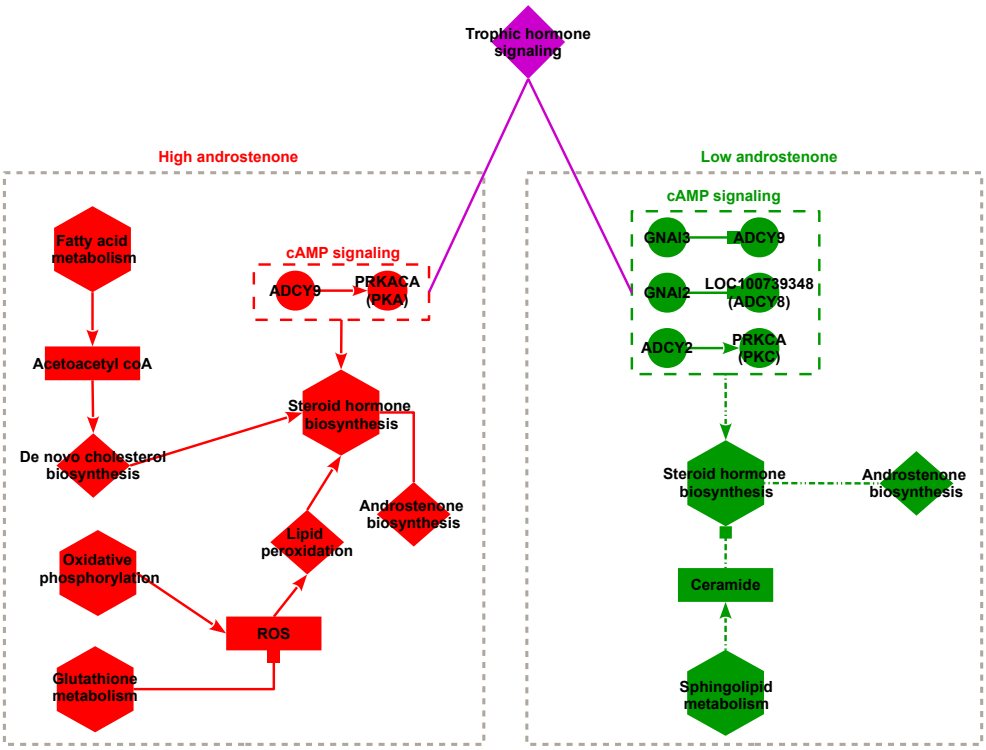

Figure 4.9: Hypothetical pathway showing the different mechanisms governing steroid biosynthesis in HA and LA animals. *Legend*: Circular nodes: genes, hexagonal nodes: enriched pathways, diamond nodes: pathways that might be involved in steroidogenesis, but not found in results, rectangular nodes: metabolites from pathways.

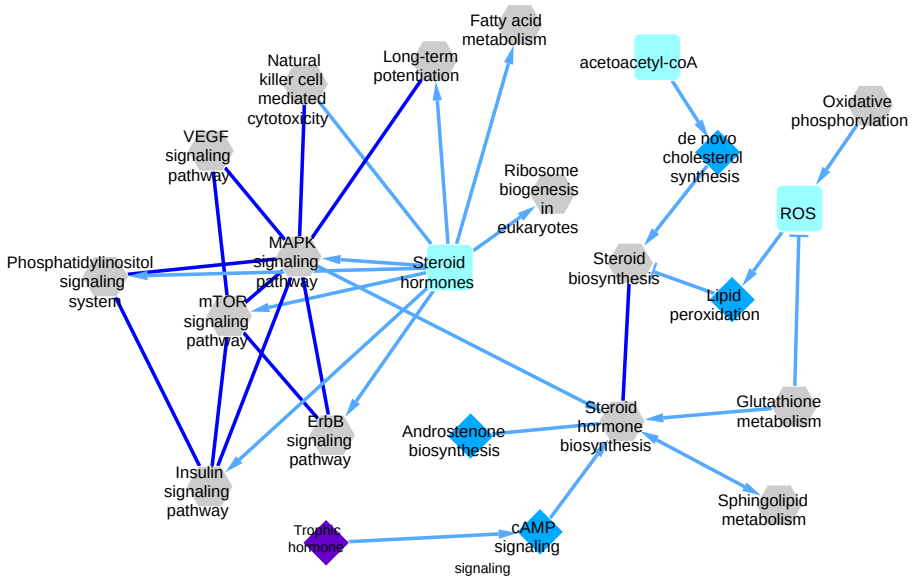

Figure 4.10: Interactions between steroid hormone biosynthesis pathway and enriched pathways. *Legend*: Grey hexagonal nodes: pathways that were enriched for significant interactions. Blue diamond nodes: pathways that might be involved in steroidogenesis, but not found in results. Purple diamond node: external stimulus in the form of hormone signaling. Cyan rectangular nodes: chemical compound or molecules synthesized in pathways. Dark blue solid edges: Interactions between enriched pathways (source: KEGG database). Dark blue solid double line edges: Edge between a compound and a pathway showing a compound synthesized in pathway. Light blue dashed edges: hypothetical interactions based on information from literature.

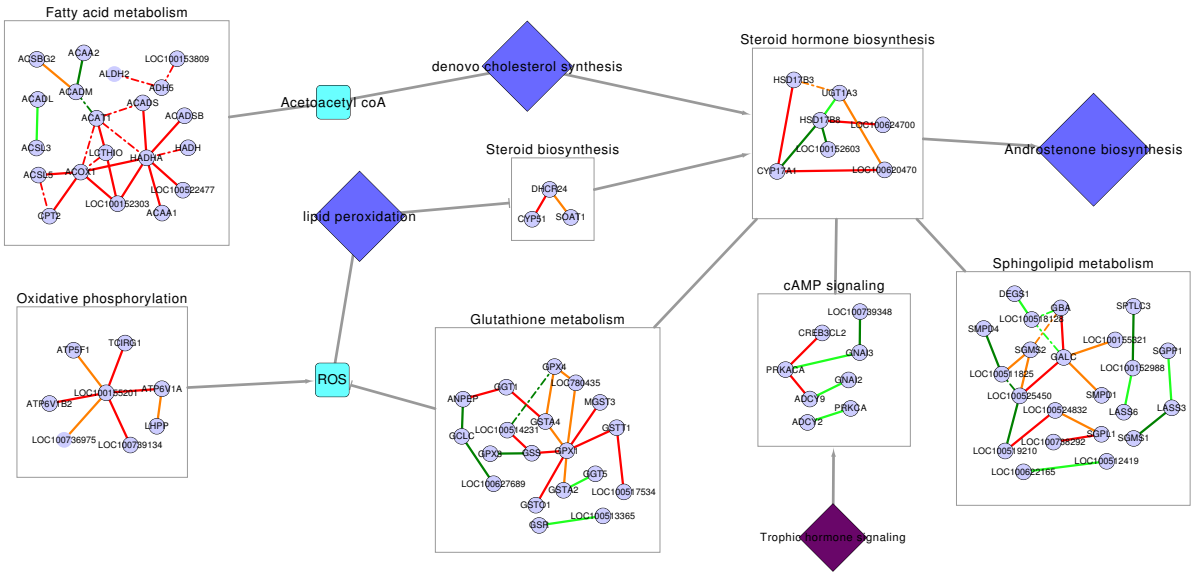

Figure 4.11: Proposed mechanism of androstenone biosynthesis regulation in HA and LA porcine samples. *Legend*: Interactions inside each pathway shows the significant interactions from analysis with genes as nodes and significant KEGG pathway interactions as edges. Blue diamond nodes: pathways that might be involved in steroidogenesis, but not found in results. Cyan colored nodes: chemical compound or molecules synthesized in pathways. Purple node: external stimulus in the form of hormone signaling. Grey solid edges: hypothetical interactions based on information from literature. Very light blue circular nodes: genes involved in significant interactions. Red dashed edges: interactions positive and significant in HA samples, positive in LA samples. Orange solid edges: interactions positive in HA samples, negative and significant in LA samples. Orange dashed edges: interactions negative in HA samples, negative and significant in LA samples. Dark green solid edges: interactions positive and significant in LA samples, negative in HA samples. Dark green dashed edges: interactions positive and significant in LA samples, positive in HA samples. Light green solid edges: interactions positive in LA samples, negative and significant in HA samples. Light green dashed edges: interactions negative in LA samples, negative and significant in HA samples.

4.1.3 Gene polymorphism analysis (Variant calling)

Polymorphism analysis revealed a total of 235,738 polymorphisms in LA samples and a total of 259,991 polymorphisms in HA samples. After filtering for SNP quality score, RMS Phred score and read depth, only the polymorphisms mapped to the exonic positions of 718 interactant genes were retained for further analysis. This filtered list of mutations was further trimmed down to include only those mutations on genes involved in significant interactions in the pathways discussed in previous sections. In this final list of polymorphisms, 11 polymorphisms were specific for LA samples, 8 polymorphisms were specific for HA samples and 50 polymorphisms were present in both LA and HA samples. The final list of polymorphisms along with sample set specific RMS quality score and read depth is given in Table 4.4. The results in Table 4.4 indicate that genes HADHA, ATP5F1, DHCR24, GSTA2, CYP51, DEGS1 and ACAA1 are highly polymorphic. Since the polymorphisms in genes ATP5F1, GSTA2 and DEGS1 were identified in both sample sets, the possible protein changes due to these polymorphisms might not have contributed to the difference in androstenone synthesis in both sample sets. Appendix Table 3 gives read counts for each polymorphism per sample.

Table 4.4: Polymorphisms in genes involved in significant interactions in selected pathways

Gene name	**Chr**	**POS**	**REF**	**ALT**	**LA MQ**	**LA DP**	**HA MQ**	**HA DP**	**Phenotye**	**Effect**
LOC100152303	1	9399735	T	C	50	163	NIL	NIL	LA	UTR 3'
		9399968	G	A	50	221	NIL	NIL	LA	Synonymous
LOC100152988	1	175812592	ATTTTT	ATTTTTT, ATTTTTTT	50	112	50	138	LA,HA	UTR 3'
GPX4	2	77676073	CAAAT	CAAAAAA AAAAAAA AAAAAT	50	155	NIL	NIL	LA	UTR 3'
LOC100736975	3	100117148	A	T	50	326	50	341	LA,HA	new start codon
HADHA	3	119782443	A	G	49	400	49	362	LA,HA	Synonymous
		119782506	T	C	49	307	49	266		
		119782546	A	G	48	349	47	304		
		119782551	T	C	48	345	47	306	LA,HA	UTR 3'
		119782751	A	G	49	468	49	444		
		119782780	T	C	49	458	49	456		
MGST3	4	92725756	T	C	49	537	49	445	LA,HA	Synonymous
ATP5F1	4	119078700	A	T	49	591	48	567		
		119078761	C	G	48	827	48	818		
		119078830	G	T	46	826	46	820		
		119078856	C	A	47	821	46	827	LA,HA	UTR 3'
		119078862	ATTTTT TTTTT	ATTTT TTTT	47	822	47	832		
		119078864	TTTTTTTT	TTTTTTT	47	812	47	824		
		119078865	TTTTTTT	TTTTTT	47	818	47	826		
LOC100514231	4	120827636	A	T	NIL	NIL	49	812	HA	UTR 3'
		120827710	A	G	50	914	48	915	LA,HA	Synonymous
DHCR24	6	145581907	G	A	NIL	NIL	49	111	HA	UTR 3'
		145582020	T	C	50	195	50	223		
		145582255	C	T	48	258	48	250		
		145582258	A	G	48	253	48	249	LA,HA	UTR 3'
		145582458	A	T	49	297	49	308		
		145582665	A	G	48	266	49	310		
		145582785	ATTTTTTT	ATTTTTT	NIL	NIL	50	278	HA	UTR 3'
CPT2	6	146702408	T	C	50	94	NIL	NIL	LA	UTR 3'
LOC100517534	6	147870177	T	G	49	213	50	208	LA,HA	UTR 3'
		147870526	G	A	48	168	49	198		
GALC	7	116349042	A	C	NIL	NIL	50	270	HA	UTR 3'
		116349177	A	G	NIL	NIL	49	238		
		116349201	T	C	49	177	NIL	NIL	LA	UTR 3'
		116349671	A	G	48	143	49	180	LA,HA	Synonymous
GSTA2	7	134289767	C	T	48	556	48	339		
		134289825	A	G	49	556	49	349	LA,HA	UTR 3'
		134289849	C	T	49	539	49	332		
		134289905	T	G	44	521	47	360		
		134289913	G	A	43	548	46	392	LA,HA	Synonymous

Table 4.4: Polymorphisms in genes involved in significant interactions in selected pathways

Gene name	Chr	POS	REF	ALT	LA MQ	LA DP	HA MQ	HA DP	Pheno type	Effect
GSTA4	7	134380269	A	G	47	666	48	507	LA,HA	UTR 3'
		134380285	T	C	47	676	48	512		
		134380456	A	G	48	688	48	537		
HADH	8	122213097	G	A	NIL	NIL	50	147	HA	UTR 3'
		122213121	G	T	49	228	50	157	LA,HA	UTR 3'
HADH	8	130466631	G	A	50	194	NIL	NIL	LA	UTR 3'
		130466820	A	G	50	256	NIL	NIL		
CYP51	9	78792947	C	T	49	211	49	235	LA,HA	UTR 3'
		78792965	G	A	49	272	NIL	NIL	LA	
		78792967	C	A	NIL	NIL	49	314	HA	
		78793035	G	A	49	404	49	490		
		78793339	GTATAT	GTAT	50	456	50	570	LA,HA	
		78793638	A	G	49	711	50	771		
DEGS1	10	15053002	A	G	49	192	48	187	LA,HA	UTR 3'
		15053060	G	A	49	192	49	193		
		15053131	T	C	49	206	48	199		
		15053143	G	A	49	200	48	196		
ACAA1	13	25168976	G	A	49	169	49	144	LA,HA	UTR 3'
		25169066	G	A	48	127	49	131		
		25169119	A	G	48	143	48	125		
		25169195	G	A	49	125	NIL	NIL	LA	Synonymous
		25169225	G	A	49	124	49	118	LA,HA	
ALDH2	14	42379317	T	C	NIL	NIL	49	190	HA	Non synonymous
GSTO1	14	125185652	G	A	49	148	48	173	LA,HA	Synonymous
ACADSB	14	144190025	A	G	50	123	NIL	NIL	LA	UTR 3'
ACSL3	15	138712086	G	A	50	209	50	246	LA,HA	Synonymous
GPX3	16	78290583	C	T	48	207	48	207	LA,HA	UTR 3'
		78290858	A	G	48	164	48	171		
GSS	17	43511491	C	T	50	78	NIL	NIL	LA	UTR 3'

Polymorphism position and function prediction results from SnpEff have shown that a large number (57) of these polymorphisms are on 3' UTR (un-translated region) of the exon and might not have contributed to any change in the protein encoded. The prediction also showed that 10 polymorphisms in the selected genes were synonymous (Table 4.4). The prediction also indicate that the polymorphism g.100117148A>T on gene LOC100736975 (ATP6V1E2) on chromosome 3 resulted in a new start codon and that the only non synonymous SNP in this result was an HA specific SNP: g.42379317T>C on ALDH2 gene on chromosome 14. Deficiency in mitochondrial ALDH2 was shown to be one of the major reasons for oxidative stress in murine cell lines (Ohsawa et al., 2003). These results indicate that these polymorphisms on genes involved in significant interactions might not have contributed to the androstenone phenotype.

4.2 Identification of gene co-expression clusters in liver tissues from multiple porcine populations with high and low backfat androstenone phenotype

This section describes and discusses the results from the data driven analysis performed to identify the common signature gene expression clusters related to androstenone metabolism in three different pig populations with comparable androstenone phenotypes. In this experiment, a total of 17 clusters from LA co-expression network and 12 clusters from HA co-expression network were found be significant with more than 10 nodes per cluster. Table 4.5 shows the number of genes, significance scores and average correlation coefficients of nodes in these clusters across three datasets. The maximum and minimum number of nodes (genes) in LA co-expression clusters were 478 and 20 respectively where as the maximum and minimum number of nodes in HA co-expression clusters were 616 and 11 respectively (Table 4.5).

Table 4.5: Significant clusters in LA and HA co-expression networks.

Cluster Id	**#Genes**	**Significance (p-value)**	**DuF2 cor. coeff.** (mean ± sd)	**Duroc cor. coeff.** (mean ± sd)	**Landrace cor. coeff.** (mean ± sd)
LA 0	478	0.00216	0.758 ± 0.138	0.850 ± 0.115	0.625 ± 0.090
LA 1	316	0.00267	0.742 ± 0.135	0.832 ± 0.122	0.622 ± 0.091
LA 2	134	0.0076	0.776 ± 0.139	0.672 ± 0.100	0.596 ± 0.075
LA 3	116	0.02248	0.741 ± 0.133	0.849 ± 0.111	0.630 ± 0.089
LA 4	96	0.04911	0.773 ± 0.139	0.666 ± 0.101	0.600 ± 0.074
LA 6	86	0.01046	0.793 ± 0.149	0.714 ± 0.108	0.600 ± 0.070
LA 7	87	0.0203	0.736 ± 0.143	0.724 ± 0.115	0.582 ± 0.063
LA 8	72	0.0379	0.765 ± 0.134	0.707 ± 0.132	0.587 ± 0.069
LA 9	68	0.01526	0.765 ± 0.149	0.610 ± 0.081	0.605 ± 0.084
LA 11	61	0.01415	0.729 ± 0.141	0.663 ± 0.126	0.662 ± 0.096
LA 12	40	0.04167	0.739 ± 0.125	0.622 ± 0.085	0.598 ± 0.074
LA 14	39	0.00594	0.736 ± 0.139	0.700 ± 0.116	0.610 ± 0.076
LA 15	30	0.04776	0.768 ± 0.138	0.641 ± 0.104	0.592 ± 0.065
LA 17	21	0.01309	0.748 ± 0.139	0.676 ± 0.131	0.612 ± 0.077
LA 18	28	0.00258	0.749 ± 0.134	0.661 ± 0.117	0.591 ± 0.075
LA 19	20	0.00408	0.726 ± 0.122	0.679 ± 0.100	0.622 ± 0.080
LA 21	21	0.01807	0.758 ± 0.140	0.746 ± 0.107	0.620 ± 0.084
HA 0	616	0.03963	0.780 ± 0.139	0.704 ± 0.115	0.663 ± 0.102
HA 1	75	0.0166	0.812 ± 0.132	0.598 ± 0.077	0.668 ± 0.106
HA 3	23	0.0023	0.815 ± 0.128	0.612 ± 0.081	0.679 ± 0.109
HA 4	18	0.00095	0.826 ± 0.117	0.597 ± 0.065	0.622 ± 0.079
HA 10	207	0.00203	0.770 ± 0.137	0.741 ± 0.116	0.681 ± 0.114
HA 11	22	0.01025	0.773 ± 0.125	0.775 ± 0.098	0.656 ± 0.103
HA 12	13	0.01196	0.776 ± 0.138	0.747 ± 0.105	0.660 ± 0.090
HA 14	75	0.00429	0.750 ± 0.141	0.611 ± 0.086	0.685 ± 0.100
HA 17	40	0.01279	0.821 ± 0.133	0.637 ± 0.088	0.619 ± 0.085
HA 18	25	0.02743	0.770 ± 0.136	0.776 ± 0.094	0.735 ± 0.101
HA 19	25	0.02149	0.767 ± 0.128	0.604 ± 0.080	0.680 ± 0.106
HA 22	11	0.04384	0.744 ± 0.136	0.677 ± 0.121	0.689 ± 0.105

4.2.1 Enrichment analysis and selection of signature co-expression clusters

7 LA co-expression clusters and 5 HA co-expression clusters were enriched for GO biological processes terms, where as 5 LA co-expression clusters and 3 HA co-expression clusters were enriched for KEGG metabolic pathways. Table 4.6 gives the number of GO terms and KEGG pathways enriched per cluster. Appendix Tables 4 and 5 contains GO terms enriched for LA and HA clusters, Appendix Tables 6 and 7 contains KEGG pathways enriched for LA and HA clusters.

Table 4.6: Number of GO terms and KEGG pathways enriched per cluster.

Cluster Id	**#GO enriched terms**	**#KEGG enriched pathways**
LA 0	19	–
LA 1	10	–
LA 2	14	11
LA 3	5	3
LA 4	–	1
LA 6	8	1
LA 7	4	–
LA 8	5	–
LA 9	–	2
HA 0	50	5
HA 1	7	6
HA 3	3	–
HA 10	8	–
HA 17	3	2

Although several LA and HA clusters were enriched for GO processes and KEGG pathways, LA cluster 2 was selected for a detailed analysis based on the enrichment results. This cluster was enriched for GO processes such as oxidation-reduction process, xenobiotic metabolic process, triglyceride metabolic process, lipid metabolic process, cholesterol metabolic process, response to drug, response to hormone stimulus (Table 4.7) as well as KEGG pathways such as PPAR signaling pathway, peroxisome, retinol metabolism, drug metabolism - other enzymes, drug metabolism - cytochrome P450 and metabolism of xenobiotics by cytochrome P450 (Table 4.8). The relationship between various GO biological process terms enriched for LA cluster 2 are depicted in Figure 4.12. It was previously established that steroid metabolism is closely linked to metabolism of drugs/xenobiotics and that the metabolism of steroids, steroid hormones, drugs and other xenobiotics are mediated by phase I and phase II metabolic pathways (Handschin and Meyer, 2003; Schänzer, 1996; Xie et al., 2003; Xu et al., 2005). One of the GO biological processes enriched in LA cluster 2 results is the oxidation reduction process and it was already found that oxidation and reduction metabolic processes constitute phase I metabolism (Gibson and Skett, 2001). Several genes involved in xenobiotic metabolism are also involved in the metabolism of androgens (Xie, 2008) and GO biological process "xenobiotic metabolic processes" was enriched for LA cluster 2 (Table 4.7). In GO and KEGG enrichment results GO term aromatic compound catabolic process and KEGG pathways drug metabolism - cytochrome P450 and metabolism

of xenobiotics by cytochrome P450 were enriched (Table 4.7 and Table 4.8). Cytochrome P450 related enzyme pathways were identified to be involved in metabolism of aromatic compounds, drugs and steroid hormones (de Montellano, 1995; Foye et al., 2008). Since LA cluster 2 GO and KEGG enrichments strongly points to the involvement of the member genes in phase I and II metabolism, LA cluster 2 was chosen for further detailed analysis. A detailed description of the functions of genes in LA cluster 2 is given in the next section.

Table 4.7: GO biological process terms enriched in LA cluster 2

GO.ID	**Term**	**#Enriched genes**	**Enrichment p-value**
GO:0055114	oxidation-reduction process	42	9.6E-011
GO:0051289	protein homotetramerization	6	0.0000016
GO:0006805	xenobiotic metabolic process	8	0.000012
GO:0006641	triglyceride metabolic process	5	0.002
GO:0006629	lipid metabolic process	33	0.00231
GO:0009058	biosynthetic process	40	0.01118
GO:0048869	cellular developmental process	11	0.0115
GO:0006810	transport	34	0.01378
GO:0008203	cholesterol metabolic process	7	0.01502
GO:0042493	response to drug	8	0.01503
GO:0046395	carboxylic acid catabolic process	11	0.02834
GO:0019439	aromatic compound catabolic process	14	0.02987
GO:0006869	lipid transport	5	0.03686
GO:0009725	response to hormone stimulus	7	0.04158

Table 4.8: KEGG pathways enriched in LA cluster 2

KEGG.ID	**Pathway**	**#Enriched genes**	**Enrichment p-value**
ssc00982	Drug metabolism - cytochrome P450	9	0.00000325
ssc00071	Fatty acid degradation	8	0.00001695
ssc00980	Metabolism of xenobiotics by cytochrome P450	7	0.00019518
ssc00830	Retinol metabolism	7	0.00026192
ssc00053	Ascorbate and aldarate metabolism	5	0.00033240
ssc05204	Chemical carcinogenesis	7	0.00082319
ssc00983	Drug metabolism - other enzymes	5	0.00107901
ssc04146	Peroxisome	8	0.00109469
ssc00280	Valine, leucine and isoleucine degradation	6	0.00149421
ssc00380	Tryptophan metabolism	5	0.00343914
ssc03320	PPAR signaling pathway	6	0.00990966

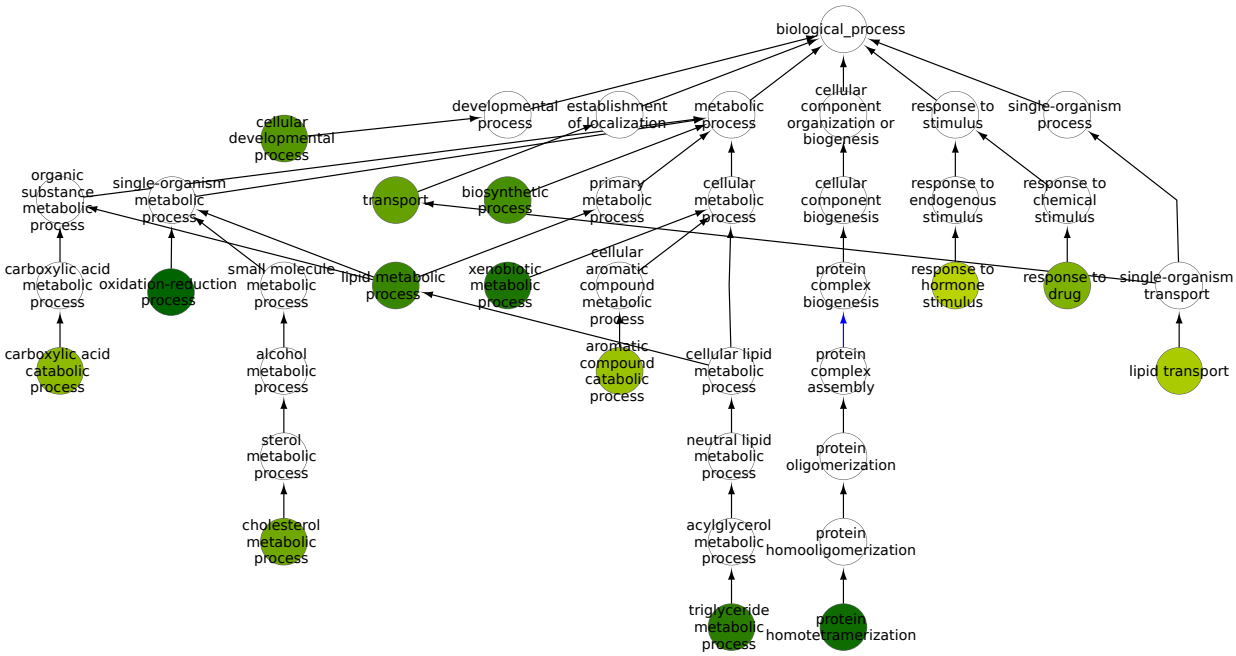

Figure 4.12: Directed acyclic graph showing the relationship between enriched LA cluster 2 enriched GO terms. *Legend*: Green nodes: GO biological process terms enriched in LA cluster 2. The intensity of the node color directly corresponds to the significance of the enriched term. White nodes: part of GO biological process ontology, not enriched in LA cluster 2 results. Black edges: represent *is_a* relationship in GO. Blue edges: represent *part_of* relationship.

4.2.2 Functional roles of LA cluster 2 genes

LA cluster 2 was comprised of 134 nodes (genes) and 1,121 edges (Figure 4.13). Node degree calculations done on the cluster indicated that genes such as PRDX3, LOC100622308 (SCP2), LOC100516628 (UGT2B18-like), PON1 and OTC were the top ranking highly connected nodes in the cluster. Some of the major families of genes in this cluster were: the UGT gene family (UGT2B17, LOC100516628 (UGT2B18-like), LOC100738495 (UGT2B31-like)), HSD/SDR gene family (HSD17B4, HSD17B10, HSD17B13, HSDL2), SLC gene family (LOC100737875 (SLC22A10), SLC25A4), ALDH gene family (ALDH3A2, ALDH5A1) and USP gene family (Usp9x, USP28) (Figure 4.13). Literature references show that UGT, HSD and ALDH gene families are associated with steroids and steroid hormone metabolism (Jin and Penning, 2001; Vasiliou and Nebert, 2005; Yoshida et al., 1998).

Three members of the UGT gene family, UGT2B17, LOC100516628 (UGT2B18-like) and LOC100738495 (UGT2B31-like) were co-expressed in LA cluster 2. Members of the UGT gene family are involved in the metabolism of steroids, biogenic amines, fat soluble vitamins, drugs and xenobiotics (Mackenzie et al., 2005). UGT2B17 was found to be important for hepatic detoxification and involved in androgen metabolism (Jin et al., 2009; Turgeon, 2003). It was shown that UGT2B18 was predominantly active on C19 steroids with a hydroxyl group at the 3α position (Beaulieu et al., 1998). Kojima and Degawa (2013) demonstrated that UGT2B31 expression was higher in male pigs when compared to female pigs and that testosterone treatment of castrated boars increased UGT2B31 expression (Kojima and Degawa, 2013). Canine UGT2B31 catalyzed the glucuronidation of compounds such as steriods, opoids, apliphatic alcohols and phenols (Soars

et al., 2003). Considering that the literatures cited above points to steroid metabolic roles of these genes and that these genes were co-expressed in all the three LA datasets, it could be possible that the UGT family genes mentioned above were involved in androgen/androstenone metabolism in all the three datasets (populations).

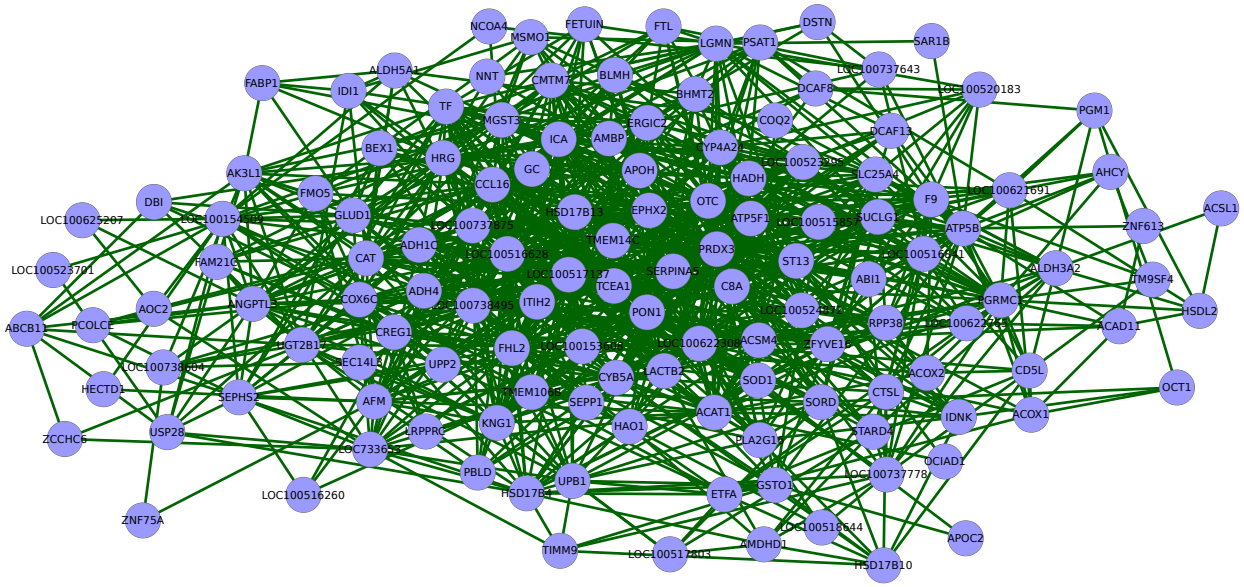

Figure 4.13: LA cluster 2. Figure showing the genes co-expressed in LA cluster 2. *Legend*: light blue nodes indicate genes and green edges indicate node co-expression (cor $\geq +0.50$ in all three populations).

In addition to UGT gene family, 4 members of HSD gene family were also co-expressed in these results. These genes are: HSD17B4, HSD17B10, HSD17B13 and HSDL2. Among these genes, three (HSD17B4, HSD17B10, HSD17B13) are members of 17β-HSD gene family. The reduction reactions catalyzed by 17β-HSDs are necessary for the formation of active androgens where as the oxidative reactions inactivates potent sex steriods (Chen et al., 2002). The enzyme encoded by gene HSD17B4 functions as a steroid inactivating enzyme and is also involved in the beta oxidation of fatty acids (de Launoit and Adamski, 1999). Additionally, it was also demonstrated that the conversion of Δ 5-androstene-3-17-diol to dehydro-epiandrosterone (DHEA) was inactivated by HSD17B4 (Prough et al., 1994). HSD17B10 was shown to be expressed in human liver, gonads, localized to mitochondria and associated with phase I metabolic pathway. The mitochondrial ability to modulate intracellular levels of active sex steroids stem from this localization of HSD17B10 (He et al., 2001). HSD17B13 is expressed in liver across a number of mammalian species. While the functions of HSD17B4 and HSD17B10 could be discussed in detail, it was not possible to find published evidences related to HDS17B13. But, in the light of evidences from SDR (HSD) gene family, it could be hypothesized that HSD17B13 is also involved in the metabolism of sex steroids. Another short chain reductase (SDR/HSD) family member HSDL2 was found to be involved in cholesterol metabolism and homeostasis (Skogsberg et al., 2008). In case of SLC family genes in LA cluster 2, although it is known that LOC100737875 (SLC22A10) gene product transports sulfate conjugates of steroids, estrone sulfate and dehydroepiandrosterone sulfate (DHEAS) with high affinity (Emami Riedmaier et al., 2012) function of SLC25A4 with regard to androgen or sterid metabolism or transport is unknown as of

now. In case of ALDH gene family, although ALDH3A2 is involved in phase I metabolic pathway, known to catalyze the oxidation of long-chain aliphatic aldehydes to fatty acid and ALDH5A1 is involved in γ aminobutyric degradation (Muzio et al., 2012), evidences to link these genes to hepatic androgen/androstenone metabolism were not found.

Another LA cluster 2 member, AKR1C1 is an NADPH dependent ketosteroid reductase. The product of this gene converts progesterone to its inactive form 20-α-dihydroxyprogesterone (Zhang et al., 2000). In androgen metabolism, the conversion of dihydrotestosterone (DHT) to 5α-androstane-3β, 17β-diol is mainly catalyzed by AKR1C1 gene product (Steckelbroeck et al., 2004). It was also shown that AKR1C1 activity can be induced by phase II enzyme inducers (Lou et al., 2006), suggesting a potential role of this gene in phase II metabolic processes. FMO5 was another co-expressed gene in LA cluster 2. The enzyme encoded by this gene is NADPH dependent, upregulated by progesterone and catalyzes the oxidation of drugs, pesticides and xenobiotics (Brooks and Harris, 2006). It was also found that FMO5 is expressed in human liver cells and $\geq 50\%$ of all FMO transcripts in human liver cells are from FMO5 (der Zee and Daly, 2012). STARD4, an LA cluster 2 member is widely expressed in liver and is demonstrated to be an important effector of lipid distribution in body (Riegelhaupt et al., 2010). A functional study postulated that STARD4 might reduce steroid hormone production during murine development (Rodriguez-Agudo et al., 2008) and yet another study (Rodriguez-Agudo et al., 2011) found that STARD4 functions in a rate limiting step in cholesterol ester formation. STARD4 increases intracellular cholesteryl ester formation and is a major component of cholesterol homeostasis regulating mechanism (Mesmin et al., 2011). ADH1C was another gene co-expressed in LA cluster 2. This gene is a member of the alcohol dehyrogenase family which metabolize substrates such as ethanol, retinol, hydroxysteroids and lipid peroxidation products. A study done on human ADH1C allele 2 found that this allele (ADH1C*2) had measurable activity on steroidogenic compounds such as 5β-androstan-17β-ol-3-one, 5β-androstan-3β-ol-17-one, 5β-pregnan-3β-ol-20-one and 5β-pregnan-3, 20-dione (Plapp and Berst, 2003).

PGRMC1, a progesterone steroid receptor is an LA cluster 2 member predominantly expressed in liver and kidney. This gene was found to be involved in sterol metabolism/homeostasis and cell survival (Lösel et al., 2008). DBI, an LA cluster 2 member gene boost steroid synthesis by stimulating delivery of cholesterol to inner mitochondrial membranes (Venturini et al., 1998). The functional roles of DBI include supporting energy metabolism, transcription, membrane production and steroidogenesis (Rasmussen et al., 1993). CRYZ gene, another LA cluster 2 member is associated with lipid, fatty acid and steroid metabolism (Taulan et al., 2004). LOC100622308 (SCP2) gene encodes sterol carrying protein 2 and is also an LA cluster 2 member. This gene is found to be involved in hepatic cholesterol metabolism, biliary lipid secretion and intracellular cholesterol distribution (Stanley et al., 2006) and it is suggested that SCP2 might be involved in regulating steroidogenesis (Fuchs et al., 2001). Yet another LA cluster 2 member gene in this analysis was LOC100523701(aldehyde oxidase like). The richest source of this gene product is liver and is found in a number of mammals. Moreover, aldehyde oxidases are involved in phase I metabolism of a number of compounds and probably functions along with the microsomal cytochrome P450 system (Garattini et al., 2009). FHL2, another LA cluster 2 co-expressed gene is

an androgen responsive gene and a co-activator of androgen receptor (AR) (Heemers et al., 2007; Müller et al., 2000). Further research also found that FHL2 is involved in steroid hormone related pathways and interacts with endoplasmic reticulum (ER) in the presence of 17β-estradiol (Kleiber et al., 2007). An LA cluster 2 member gene, OCT1 interacts with AR and can interact with HNF1 to modulate its capacity to upregulate UGT2B expression in liver (Xie, 2008). Since three UGT2B genes (UGT2B17, LOC100516628 (UGT2B18-like), LOC100738495 (UGT2B31-like)) and OCT1 are found in the same cluster and co-expressed in three different datasets (populations), the potential action of OCT1 on UGT2B genes and their role in androgen/androstenone metabolism could be further investigated. Another LA cluster 2 coexpressed gene was PON1. PON1 is synthesized in liver and is involved in the biotransformation of various xenobiotics as well as protection against lipid peroxidation (Draganov et al., 2005). Table 4.9 gives the summary of functions for a number of other genes in LA cluster 2. The next part of this section describes and discusses the results from cluster similarity assessments.

Table 4.9: Table containing function summaries for genes in LA cluster 2.

Gene name	**Functions**
LOC100517137 (ALDH1A1)	involved in retinol metabolism (Zheng et al., 1993), Androgen receptor might be in involved in regulating levels of ALDH1A1 (Yoshida et al., 1998), involved in non-catalytic interactions with androgen, thyroid hormone, cholesterol and drug compounds including flavopiridol, daunorubicin and quinolone (Marchitti et al., 2008).
IDI1	catalyzes the conversion of isopentenyl diphosphate (IPP) to dimethylallyl diphosphate (DMAPP) an intermediate product in Mevalonate pathway resulting in the formation of cholesterols and sterols (Zheng et al., 2007).
ANGPTL3	expressed specifically in liver (Conklin et al., 1999), in mice overexpression of ANGPTL3 lead to an increased level of circulating plasma lipids and a mutation in the gene is a factor for the low levels of plasma triglycerides in a strain of obese mice (Koishi et al., 2002).
ABCB11	integral part of plasma membrane and product of this gene is a bile acid transporter (Strautnieks et al., 1998), involved in Phase III detoxification system after phase I and phase II metabolism (Sies and Packer, 2005).
ETFA	catalyzes initial step of mitochondrial fatty acid β oxidation (Bartlett and Eaton, 2004).
SEPP1	expressed in liver (Burk and Hill, 2005) androgen responsive gene and shown to have gender specific expression (Takahashi et al., 2006).
GSTO1	proposed to be involved in sterol glucoronidation in liver (Moe et al., 2008), involved in xenobiotic metabolism (Yu et al., 2003).
SOD1	decreases cholesterol biosynthesis processes and HMG-CoA reductase activity in rate hepatocytes and human fibroblast cells (De Felice et al., 2004), hepatic concentrations of total cholesterol, triglyceride and non-esterified fatty acids were were significantly increased in $SOD1^{-/-}$ mice (De Felice et al., 2004).
MGST3	down regulated by testosterone (Bagchi et al., 2011), involved in drug metabolism and disposition in liver and kidney (Choudhuri et al., 2003; Lu et al., 2010).
ACAT1	overexpression of ACAT1 stimulates assembly and release of VLDL from liver cells (Liang et al., 2004), involved in esterification process in cholesterol metabolism (An et al., 2006).
GLYAT	catalyzes the conjugation of glycine with acetyl-coA substrates and is involved in the detoxification of endogenous and xenobiotic acyl CoA's (Mawal and Qureshi, 1994), differentially expressed in Duroc liver tissues with divergent androstenone content (Moe et al., 2008).

Table 4.9: Table containing function summaries for genes in LA cluster 2 (continued...)

Gene name	Functions
SEC14L3 (hTAP2)	expressed in human liver, involved in lipid transportation, regulation of lipid dependent events and cholesterol biosynthesis (Zingg et al., 2008).
FABP1	promote cellular uptake, transport and metabolism of fatty acids and peroxisomal oxidation of long chain fatty acids (Antonenkov et al., 2006), can bind to a variety of ligands including fatty-acyl CoAs, lysophospholipids, bile acids and drugs (Hagan et al., 2002; Myszka and Swenson, 1991).
ACOX1	catalyzes the rate limit limiting step in peroxisomal β oxidation pathway (Fan et al., 1996), human ACOX1 isoforms are involved in cholesterol homeostasis (Vluggens et al., 2010).
ACOX2	involved in the oxidation of long chain straight fatty acids and bile acid intermediates (Baumgart et al., 1996).
HADH	overexpression of HADH resulted in higher rates of mitochondrial β oxidation and HADH is also associated with down regulating insulin release (Martens et al., 2007), shows increased activity in response to drugs and hormones (Furuhashi et al., 2002; Seiva et al., 2008).
PRDX3	majority of PRDX3 is localized to mitochondria, but PRDX3 located in cell membranes is regulated by androgens (Whitaker et al., 2013), PRDX3 is a c-Myc target gene essential for maintaining mitochondrial functions (Wonsey et al., 2002), possibly involved in mitochondrial response to hydrogen peroxide and reactive oxygen species (ROS) and protects mitochondria from ROS generated through electron leakage in cytochrome P-450 system (Chae et al., 1999; Lee et al., 2007; Simoni et al., 2008).
MSMO1	associated with metabolism of sterols and small molecules (Li and Kaplan, 1996).

4.2.3 Cluster similarity analysis

Hypergeometric test for cluster node overlap assessment showed that 15 LA clusters and 13 HA clusters had significant node overlap between them (Figure 4.14). The highest node overlap was between clusters LA 0 and HA 0 with 280 common nodes followed by the overlap between clusters LA 1 and HA 10 with 152 common nodes (Figure 4.14). LA cluster 2 showed significant node overlap between 6 HA clusters: HA 0, HA 1, HA 3 , HA 14, HA 17 and HA 22. Among these clusters, the highest overlap was with cluster HA 0, with 35 nodes in common where as HA cluster 1 with 33 common nodes showed the next highest overlap with LA cluster 2 (Figure 4.15). It can also be seen from Figure 4.15 that LA cluster 2 showed the least physical overlap with HA cluster 22 with only 4 nodes in common. The results from functional similarity assessment showed that 12 LA and HA clusters had significant functional similarity overlap (Figure 4.16). Out of these 12 clusters, 7 clusters were from LA network and 5 clusters were from HA network. The highest functional similarity (0.626) was between clusters LA 1 and HA 10 (Figure 4.16). These clusters also showed the second highest physical similarity (node overlap) (Figure 4.14). The second highest functional similarity (0.603) was between clusters HA 3 and HA 17, indicating that irrespective of having no physical overlap, the clusters showed significant functional similarity. The third highest functional similarity (0.586) was between clusters LA 0 and HA 0, the clusters with highest physical overlap (Figure 4.16, Figure 4.14). LA cluster 2 showed significant functional similarity with one LA cluster, LA 0 and 4 HA clusters: HA 0, HA 1, HA 3 and HA 17. Interestingly, the four HA clusters with significant functional similarity also

showed significant physical similarity (node overlap) with LA cluster 2 (Figure 4.15).

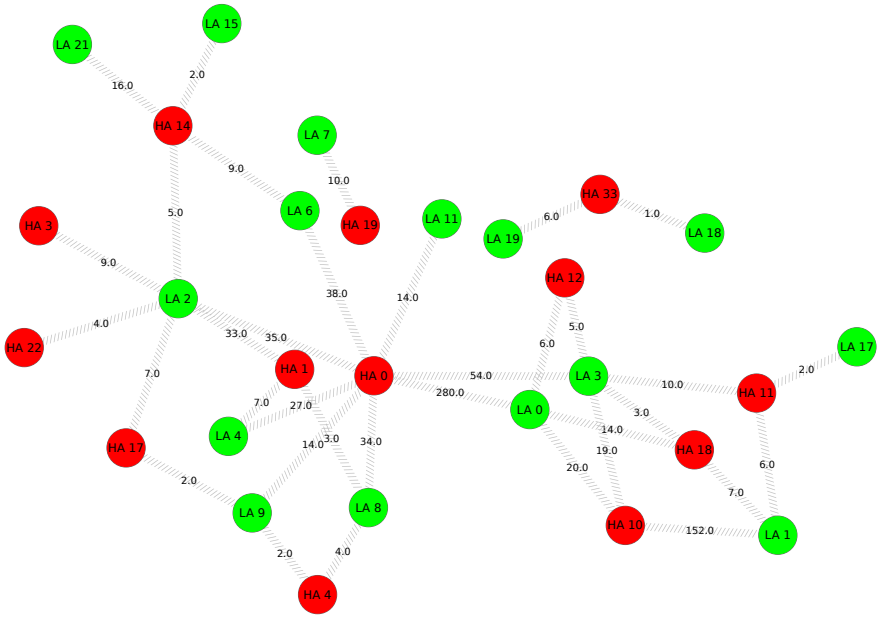

Figure 4.14: Cluster physical overlap. Figure showing significant node overlap between LA and HA clusters. *Legend*: Green nodes indicate LA clusters and red nodes indicates HA clusters. Grey forward slashed edges indicate significant physical overlap and edge labels indicate the number of common nodes between two clusters.

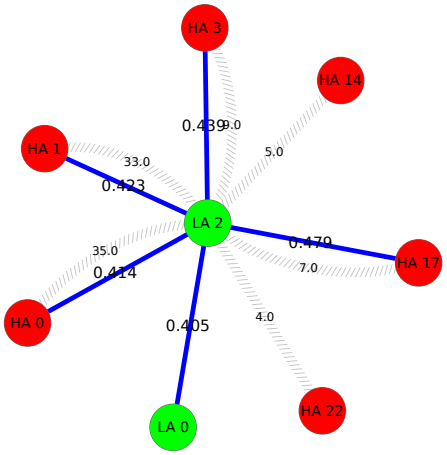

Figure 4.15: LA 2 cluster similarity. Figure showing physical and functional similarity of LA cluster 2 with other clusters. Legend: Green nodes indicate LA clusters and red nodes indicates HA clusters. Grey forward slashed edges indicate significant physical overlap and edge labels indicate the number of common nodes between two clusters.

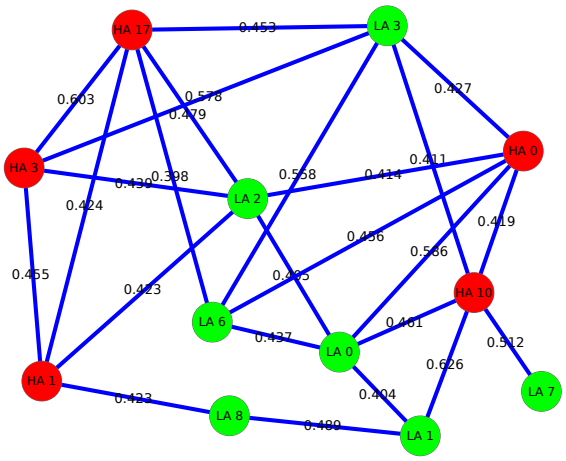

Figure 4.16: Cluster functional overlap. Figure showing functional overlap between LA and HA clusters with significant GO enrichmentr. Legend: Green nodes indicate LA clusters and red nodes indicates HA clusters. Grey forward slashed edges indicate significant physical overlap and edge labels indicate the number of common nodes between two clusters.

Based on the evidences from GO and KEGG enrichment results (Table 4.7, Table 4.8), it could be postulated that LA cluster 2 could be one of the signature co-expression clusters in boars with low androstenone measurement since enrichment results indicate the role of the cluster member genes in phase I and phase II metabolism. Additionally, literature references on functions of cluster members (as mentioned above) also hint the potential role of these genes in hepatic metabolism. Since this analysis incorporate data/metadata from three different populations (datasets) it could further be postulated that this cluster commonly occurs in all the three populations used in this analysis and functions in a similar manner across the three populations. Physical similarity analysis reveals that LA cluster 2 shows significant physical similarity (node overlap) with 6 HA clusters. Although these similarities were significant, even the HA cluster with highest physical similarity (HA cluster 0) shares only 26% of LA cluster 2 nodes (Table 4.5, Figure 4.15). From these results it could be hypothesized that LA cluster 2 members exhibit strong co-expression, cluster behavior and probably similar functions in LA dataset where as in HA datasets, these genes show a weak co-expression and weak cluster behavior. The dispersion of genes occurring in LA cluster 2 across multiple clusters in HA network could be an indication of weak co-expression of these genes in HA datasets. The functional similarity assessment also supports this hypothesis. As can be seen from Figure 4.15, although the functional similarity between LA cluster 2 and HA clusters are statistically significant, based on GO semantic similarity range (0-1) it can be seen that the functional similarities are only moderate. The strong co-expression of LA cluster 2 member genes in low androstenone animals and enrichment analysis results from this cluster is an indication that these genes are highly involved in hepatic phase I and phase II metabolism of these animals and since these genes do not exhibit a strong co-expression in high androstenone animals, it could be assumed that this co-expression cluster is a signature cluster for hepatic androgen and hence androstenone metabolism in low androstenone animals. Consolidating these analysis results, LA cluster 2 can be proposed as one of the signature co-expression cluster for low androstenone boars and that the combined action of LA cluster 2 member genes might be contributing to hepatic androstenone and androgen metabolism in the populations used in this

experiment. Since these results are based on gene expression data from three pig populations (datasets), it could be further postulated that this co-expression cluster might be functioning in a similar manner in all the three pig populations (Duroc $\times$ F_2, Duroc and Norwegian Landrace) used in the experiment.

In general, experiment 1 was performed to identify the interaction patterns that are significantly different between LA and HA phenotype. This experiment utilized the existing knowledge by integrating gene expression data from a previous RNA-seq experiment metabolic interactions retrieved from KEGG database. Based on the result from this experiment, this thesis extends the current knowledge in androstenone biosynthesis by proposing a hypothetical model of androstenone biosynthesis. This model postulates that there are differences in the interaction patterns involved in signaling and regulating testicular androstenone biosynthesis in low and high androstenone phenotypes. In the light of the results from experiment 1, it was postulated that pathways such as glutathione metabolism, sphingolipid metabolism, fatty acid metabolism and cAMP-PKA/PKC signaling were fundamental in maintaining and regulating steroidogenesis and hence androstenone biosynthesis in both high and low androstenone animals. The proposed model theorized that in high androstenone animals, steroidogenesis was activated by cAMP-PKA signaling and that the anti lipid peroxidation activity of glutathione metabolism and de novo synthesis of cholesterol as a result of an active fatty acid metabolism activity might have boosted steroidogenesis and androstenone metabolism. In low androstenone animals, a weak cAMP-PKC activation of steroidogenesis and regulatory action of ceramides on steroidogenesis might have contributed to a weak steroid hormone synthesis and hence, low levels of androstenone synthesis. The combined effect of these key differences in the metabolic and signaling pathways could have a "cascading effect" in determining the levels of androstenone synthesis in the sample population.

In experiment 2, due to the short comings in current knowledge about androstenone metabolism in porcine liver tissues, a data driven approach was used to analyze the common patterns of gene expression in multiple porcine populations. Meta data from three different porcine populations were used to generate low and high androstenone co-expression clusters. The results from this experiment broaden the current understanding of hepatic androstenone metabolism by identifying the common genes involved in hepatic androstenone metabolism of the selected porcine populations. The co-expression cluster selected in this experiment (LA cluster 2) could be used as a signature co-expression cluster for low androstenone hepatic metabolism. Cluster similarity assessments done in this experiment indicate that the strong clustering behavior exhibited by the (LA 2) cluster genes in low androstenone expression network is absent in high androstenone expression network, hinting a weak co-expression of these genes in high androstenone animals. Additionally, since the analysis combined meta data from three different pig populations, it could be theorized that this signature cluster functions in a similar manner across all the three porcine populations.

The hypothetical model pathway from experiment 1 and common gene expression cluster in experiment 2 are in silico results based on gene expression data and external database information. These results demonstrate that by using integrative analysis approaches following the concepts of knowledge discovery and data mining, new knowledge can be gained from existing datasets in livestock genomics. These results presented here also shows how integrating multiple data types

or data resources can add an additional dimension to the results. The metabolic interactions from KEGG database add this additional dimension in first experiment where as datasets from multiple experiments contribute an improved granularity to the results in the second experiment.

5. Conclusion

Integrative analysis methods have long been used in human and model organism genomics to extract new knowledge from existing high-throughput datasets and other data sources. Although livestock genomics have been using modern high-throughput techniques for data generation, very few researches in this field have made use of integrative analysis techniques utilizing data mining and knowledge discovery concepts to extract new knowledge. This thesis was aimed at demonstrating the capabilities of integrative analysis methods to extract new knowledge and to generate novel hypothesis from existing datasets to answer major research questions in livestock genomics. For this purpose, a porcine meat quality related phenotype, androstenone content in backfat was selected as a target analysis trait. Despite being an active research topic in porcine genomics for the last few years, the current understanding of testicular synthesis and hepatic degradation of androstenone is limited to a handful of biomarkers and a selected number of metabolic pathways.

Based on the two experiments performed, this thesis extends the current understanding of testicular androstenone synthesis and hepatic androstenone degradation by proposing a novel hypothetical model for the difference in androstenone biosynthesis with divergent androstenone measurements and introducing a novel hepatic gene co-expression cluster as a signature cluster in animals with low androstenone content. The analysis methods and results presented in this thesis is one of the forerunning attempts in livestock genomics and androstenone genomics to make use of integrative analysis approaches for knowledge extraction and hypothesis generation. A major challenge in the application of integrative analysis methods in livestock genomics is the reduced availability of high-throughput datasets and other data resources in livestock genomics in comparison to human or other model organism genomics. Hence, the extend of metabolic interactions and diversity of gene expression variations covered in this thesis is limited. Another limitation is that is that the second experiment in this thesis integrates multiple datasets, but not multiple data types and hence lacks an additional dimension in results, which would be theoretically possible. But, in spite of these limitations, this thesis showed that the available datasets in livestock genomics can be used for the extraction of new knowledge and for hypothesis generation.

In the future, these analysis methods can be extended in a number of directions. The results presented in thesis are based on in-silico models and therefore are hypothetical. By using data from multiple porcine populations and multiple experiments at genomic, proteomic and metabolomic level these results can either be validated/challenged or extended. Proceeding in this direction

would enable the androstenone research community to develop standard models to explain testicular androstenone synthesis or hepatic metabolism. These standard models can be then established as "ground truth" models for androstenone metabolism in porcine. The second direction to take would be the extension of current models using multiple data types such as experimentally determined protein - protein interaction networks, exon expression, gene polymorphisms and epigenetic data such as transcription factor binding and histone modifications. Further, these analysis strategies can be generalized for a global livestock genomics perspective. Experimental genomic or proteomic data in livestock can be used in integrative analysis strategies and can be enriched using metabolic, protein interaction and gene variation data from publicly available biological databases and information mined from published literature. Such an integration of multiple flavors of data would enable livestock genomics researchers to visualize the phenotype of interest based on various layers of cellular mechanisms with an increased level of granularity. In the long run, integrative analysis methods can be used in two major research scenarios in livestock genomics. In the first scenario, integrative analysis methods can be utilized as a platform for the extraction of new knowledge and the generation of hypothesis. This generated knowledge and hypothesis can be used as pointers for further laboratory experiments thus facilitating a detailed understanding of the molecular mechanisms involved in the manifestation of various economically important traits in livestock. In another scenario, integrative analysis strategies could also promote the generation of computable models in livestock genomics to enable the replication of results from computational biology experiments in livestock genomics. In a yet another future direction, researches in livestock genomics could greatly benefit from open source and online analysis platforms specialized for livestock genomics species. Currently only a limited number of public analysis platforms hosts livestock genomic data and very few livestock genomics analysis platforms are available for data mining, knowledge discovery and integrative analysis.

The analysis results in this thesis show that although data availability is a major confounding factor, data mining and knowledge discovery methods can be successfully used for gaining new knowledge in livestock genomics. The knowledge gained through these methods can boost the current understanding about cellular processes involved in the development of various economically important traits in livestock species and can ultimately aid in improving these traits and future proofing the livestock animals against the challenges ahead. The limited availability of publicly available data in livestock genomics in comparison to model organism species is the major factor impeding the wide spread use of integrative analysis methods, data mining and knowledge discovery concepts in livestock genomics. But, the increasing amounts of data generated by high throughput technologies and the steadily decreasing cost of such data generation technologies will aid in the availability of large volumes of publicly available data in livestock genomics enabling the widespread use of the integrative analysis methods in livestock genomics in the coming years.

6. References

Abelson JN, Simon MI, Merrill AH and Hannun YA (1999): Sphingolipid metabolism and cell signaling. pt. 1, Elsevier Science, 748 .

Alexa A and Rahnenfuhrer J (2010): topGO: Enrichment analysis for Gene Ontology. Tech. rep.

Alexa A, Rahnenführer J and Lengauer T (2006): Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22(13), 1600–1607.

Almeida-de Macedo MM, Ransom N, Feng Y, Hurst J and Wurtele ES (2013): Comprehensive analysis of correlation coefficients estimated from pooling heterogeneous microarray data. BMC Bioinformatics 14(1), 214.

Alon U (2003): Biological Networks: The Tinkerer as an Engineer. Science 301(5641), 1866–1867.

Altamirano F, Oyarce C, Silva P, Toyos M, Wilson C, Lavandero S, Uhlén P and Estrada M (2009): Testosterone induces cardiomyocyte hypertrophy through mammalian target of rapamycin complex 1 pathway. Journal of Endocrinology 202(2), 299–307.

Amar D and Shamir R (2014): Constructing module maps for integrated analysis of heterogeneous biological networks. Nucleic Acids Research 42(2), 4208–4219.

An S, Cho KH, Lee WS, Lee JO, Paik YK and Jeong TS (2006): A critical role for the histidine residues in the catalytic function of acyl-CoA:cholesterol acyltransferase catalysis: Evidence for catalytic difference between ACAT1 and ACAT2. FEBS Letters 580(11), 2741–2749.

Andresen O (1976): Concentrations of fat and plasma 5α-androstenone and plasma testosterone in boars selected for rate of body weight gain and thickness of back fat during growth, sexual maturation and after mating. Journal of Reproduction and Fertility 48(1), 51–59.

Antonenkov VD, Sormunen RT, Ohlmeier S, Amery L, Fransen M, Mannaerts GP and Hiltunen JK (2006): Localization of a portion of the liver isoform of fatty-acid-binding protein (L-FABP) to peroxisomes. Biochemical Journal 394(2), 475–484.

Archibald AL, Bolund L, Churcher C, Fredholm M, Groenen MAM, Harlizius B, Lee KT, Milan D, Rogers J, Rothschild MF, Uenishi H, Wang J and Schook LB (2010): Pig genome sequence–analysis and publication strategy. BMC Genomics 11(1), 438.

Bagchi G, Zhang Y, Stanley K and Waxman D (2011): Complex modulation of androgen responsive gene expression by methoxyacetic acid. Reproductive Biology and Endocrinology 9(1), 42.

Bartke N and Hannun YA (2009): Bioactive sphingolipids: metabolism and function. Journal of Lipid Research 50 Suppl, S91–S96.

Bartlett K and Eaton S (2004): Mitochondrial beta-oxidation. European Journal of Biochemistry 271(3), 462–469.

Basarab JA, Beauchemin KA, Baron VS, Ominski KH, Guan LL, Miller SP and Crowley JJ (2013): Reducing GHG emissions through genetic improvement for feed efficiency: effects on economically important traits and enteric methane production. Animal 7(Supplement s2), 303–315.

Bauer BK, Isom SC, Spate LD, Whitworth KM, Spollen WG, Blake SM, Springer GK, Murphy CN and Prather RS (2010): Transcriptional Profiling by Deep Sequencing Identifies Differences in mRNA Transcript Abundance in In Vivo-Derived Versus In Vitro-Cultured Porcine Blastocyst Stage Embryos. Biology of Reproduction 83(5), 791–798.

Baumgart E, Vanhooren JCT, Fransen M, Marynen P, Puype M, Vandekerckhove J, Leunissen JAM, Fahimi HD, Mannaerts GP and Van Veldhoven PP (1996): Molecular characterization of the human peroxisomal branchedchain acyl-CoA oxidase: cDNA cloning, chromosomal assignment, tissue distribution, and evidence for the absence of the protein in Zellweger syndrome. Proceedings of the National Academy of Sciences 93(24), 13748–13753.

Beaulieu M, Lévesque E, Barbier O, Turgeon D, Bélanger G, Hum DW and Bélanger A (1998): Isolation and characterization of a simian UDP-glucuronosyltransferase UGT2B18 active on 3-hydroxyandrogens. Journal of Molecular Biology 275(5), 785–794.

Billen MJ and Squires EJ (2009): The role of porcine cytochrome b5A and cytochrome b5B in the regulation of cytochrome P45017A1 activities. Journal of Steroid Biochemistry 113(1-2), 98–104.

Blondel VD, Guillaume JL, Lambiotte R and Lefebvre E (2008): Fast unfolding of communities in large networks. Journal of Statistical Mechanics 2008(10), P10008.

Boer HMT, Apri M, Molenaar J, Stötzel C, Veerkamp R and Woelders H (2012): Candidate mechanisms underlying atypical progesterone profiles as deduced from parameter perturbations in a mathematical model of the bovine estrous cycle. Journal of Dairy Science 95(7), 3837–3851.

Bonneau M (1982): Compounds responsible for boar taint, with special emphasis on androstenone: A review. Livestock Production Science 9(6), 687–705.

Bonneau M, Le Denmat M, Vaudelet J, Veloso Nunes J, Mortensen A and Mortensen H (1992): Contributions of fat androstenone and skatole to boar taint: I. Sensory attributes of fat and pork meat. Livestock Production Science 32(1), 63–80.

Boron WF and Boulpaep EL (2005): Medical physiology: a cellular and molecular approach. Elsevier Science Health Science Division, 1319 .

Boulliou-Robic A, Feve K, Larzul C, Billon Y, Van Son M, Liaubet L, Sarry J, Milan D, Grindflek E, Bidanel JP and Riquet J (2011): Expression levels of 25 genes in liver and testis located in a QTL region for androstenone on SSC7q1.2. Animal Genetics 42(6), 662–665.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS and Crawford GE (2008): High-resolution mapping and characterization of open chromatin across the genome. Cell 132(2), 311–322.

Brooks SA and Harris A (2006): Breast Cancer Research Protocols. Biomed Protocols, Humana Press, 517 .

Brunner RM, Srikanchai T, Murani E, Wimmers K and Ponsuksili S (2012): Genes with expression levels correlating to drip loss prove association of their polymorphism with water holding capacity of pork. Molecular Biology Reports 39(1), 97–107.

Brusic V and Zeleznikow J (1999): Knowledge discovery and data mining in biological databases. The Knowledge Engineering Review 14(03), 257–277.

Burk RF and Hill KE (2005): Selenoprotein P: an extracellular protein with unique physical characteristics and a role in selenium homeostasis. Annual Review of Nutrition 25, 215–235.

Burrows M and Wheeler DJ (1994): A block-sorting lossless data compression algorithm. Systems Research Research R(124), 24, 0908.0239.

Calenge F, Legarra A and Beaumont C (2011): Genomic selection for carrier-state resistance in chicken commercial lines. BMC Proceedings 5(Suppl 4), S24.

Chae HZ, Kim HJ, Kang SW and Rhee SG (1999): Characterization of three isoforms of mammalian peroxiredoxin that reduce peroxides in the presence of thioredoxin. Diabetes Research and Clinical Practice 45(2-3), 101–112.

Chalfant C and Poeta MD (2010): Sphingolipids as signaling and regulatory molecules. Advances in experimental medicine and biology, Springer, 7–8 .

Chandra R, Aneja R, Rewal C, Konduri R, Dass SK and Agarwal S (2000): An opium alkaloid-papaverine ameliorates ethanol-induced hepatotoxicity: Diminution of oxidative stress. Indian Journal of Clinical Biochemistry 15(2), 155–160.

Chang LC, Lin HM, Sibille E and Tseng GC (2013): Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. BMC Bioinformatics 14(1), 368.

Chang R, Brauer W and Stetter M (2008): Modeling semantics of inconsistent qualitative knowledge for quantitative Bayesian network inference. Neural Networks 21(2-3), 182–192.

Chauhan BS (2008): Principles of Biochemistry and Biophysics. Laxmi Publications Pvt Limited, 580 .

Chen C, Ai H, Ren J, Li W, Li P, Qiao R, Ouyang J, Yang M, Ma J and Huang L (2011): A global view of porcine transcriptome in three tissues from a full-sib pair with extreme phenotypes in growth and fat deposition by paired-end RNA sequencing. BMC Genomics 12(1), 448.

Chen H, Pechenino AS, Liu J, Beattie MC, Brown TR and Zirkin BR (2008a): Effect of glutathione depletion on Leydig cell steroidogenesis in young and old brown Norway rats. Endocrinology 149(5), 2612–2619.

Chen K and Rajewsky N (2006): Natural selection on human microRNA binding sites inferred from SNP data. Nature Genetics 38(12), 1452–1456.

Chen M and Hofestädt R (2006): A medical bioinformatics approach for metabolic disorders: Biomedical data prediction, modeling, and systematic analysis. Journal of Biomedical Informatics 39(2), 147–159.

Chen W, Thiboutot D and Zouboulis CC (2002): Cutaneous androgen metabolism: basic research and clinical perspectives. Journal of Investigative Dermatology 119(5), 992–1007.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL and Ng HH (2008b): Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133(6), 1106–17.

Cheng L and Bostwick DG (2011): Essentials of anatomic pathology. Springer New York, 573 .

Choudhuri S, Cherrington NJ, Li N and Klaassen CD (2003): Constitutive expression of various xenobiotic and endobiotic transporter mRNAs in the choroid plexus of rats. Drug Metabolism and Disposition 31(11), 1337–1345.

Chuang HY, Hofree M and Ideker T (2010): A decade of systems biology. Annual Review of Cell and Developmental Biology 26, 721–744.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X and Ruden DM (2012): A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6(2), 80–92.

Cios KJ, Pedrycz W, Swiniarski RW and Kurgan LA (2007): Data Mining: A Knowledge Discovery Approach. Springer, 621 .

Clauset A, Newman M and Moore C (2004): Finding community structure in very large networks. Physical Review E 70(6), 066111.

Cleveland MA, Hickey JM and Forni S (2012): A common dataset for genomic analysis of livestock populations. G3 2(4), 429–435.

Conklin D, Gilbertson D, Taft DW, Maurer MF, Whitmore TE, Smith DL, Walker KM, Chen LH, Wattler S, Nehls M and Lewis KB (1999): Identification of a mammalian angiopoietin-related protein expressed specifically in liver. Genomics 62(3), 477–482.

Couture O, Callenberg K, Koul N, Pandit S, Younes R, Hu ZL, Dekkers J, Reecy J, Honavar V and Tuggle C (2009): ANEXdb: an integrated animal ANnotation and microarray EXpression database. Mammalian Genome 20(11-12), 768–777.

Creighton C and Hanash S (2003): Mining gene expression databases for association rules. Bioinformatics 19(1), 79–86.

Csardi G and Nepusz T (2006): The igraph software package for complex network research. InterJournal Complex Sy, 1695.

Daetwyler HD, Hickey JM, Henshall JM, Dominik S, Gredler B, van der Werf JHJ and Hayes BJ (2010): Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. Animal Production Science 50(12), 1004–1010.

Daetwyler HD, Kemper KE, van der Werf JHJ and Hayes BJ (2012): Components of the accuracy of genomic prediction in a multi-breed sheep population. Journal of Animal Science 90(10), 3375–3384.

D'Antonio M and Ciccarelli FD (2011): Modification of gene duplicability during the evolution of protein interaction network. PLoS Computational Biology 7(4), e1002029.

Davis S and Meltzer PS (2007): GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 23(14), 1846–1847.

Davis SM and Squires EJ (1999): Association of cytochrome b5 with 16-androstene steroid synthesis in the testis and accumulation in the fat of male pigs. Journal of Animal Science 77(5), 1230–1235.

De Felice B, Santillo M, Serù R, Damiano S, Matrone G, Wilson RR and Mondola P (2004): Modulation of 3-hydroxy-3-methylglutaryl-CoA reductase gene expression by CuZn superoxide dismutase in human fibroblasts and HepG2 cells. Gene Expression 12(1), 29–38.

de Haas Y, Calus MPL, Veerkamp RF, Wall E, Coffey MP, Daetwyler HD, Hayes BJ and Pryce JE (2012): Improved accuracy of genomic prediction for dry matter intake of dairy cattle from combined European and Australian data sets. Journal of Dairy Science 95(10), 6103–12.

de Haas Y, Windig JJ, Calus MPL, Dijkstra J, de Haan M, Bannink A and Veerkamp RF (2011): Genetic parameters for predicted methane production and potential for reducing enteric emissions through genomic selection. Journal of Dairy Science 94(12), 6122–6134.

de Launoit Y and Adamski J (1999): Unique multifunctional HSD17B4 gene product: 17beta-hydroxysteroid dehydrogenase 4 and D-3-hydroxyacyl-coenzyme A dehydrogenase/hydratase involved in Zellweger syndrome. Journal of Molecular Endrocrinology 22(3), 227–240.

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K and Cotes JM (2009): Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182(1), 375–385.

de Montellano PRO (1995): Cytochrome P450: Structure, Mechanism, and Biochemistry. Springer, 652 .

Dekkers JCM (2012): Application of genomics tools to animal breeding. Current Genomics 13(3), 207–212.

der Zee AHM and Daly AK (2012): Pharmacogenetics and Individualized Therapy. Wiley, 432 .

Doeschl-Wilson AB (2011): The role of mathematical models of host-pathogen interactions for livestock health and production - a review. Animal 5(6), 895–910.

Donetti L and Muñoz MA (2005): Improved spectral algorithm for the detection of network communities. AIP Conference Proceedings 779(1).

Doran E, Whittington FM, Wood JD and McGivan JD (2004): Characterisation of androstenone metabolism in pig liver microsomes. Chemico-biological interactions 147(2), 14114–14119.

Draganov DI, Teiber JF, Speelman A, Osawa Y, Sunahara R and La Du BN (2005): Human paraoxonases (PON1, PON2, and PON3) are lactonases with overlapping and distinct substrate specificities. Journal of Lipid Research 46(6), 1239–1247.

Duchemin SI, Colombani C, Legarra A, Baloche G, Larroque H, Astruc JM, Barillet F, Robert-Granié C and Manfredi E (2012): Genomic selection in the French Lacaune dairy sheep breed. Journal of Dairy Science 95(5), 2723–2733.

Dudley JT and Butte AJ (2009): Identification of discriminating biomarkers for human disease using integrative network biology. In: Pacific Symposium on Biocomputing, Hawaii, 27–38.

Duijvesteijn N, Knol EF, Merks JWM, Crooijmans RPMA, Groenen MAM, Bovenhuis H and Harlizius B (2010): A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. BMC Genetics 11, 42.

Dunshea FR, Colantoni C, Howard K, McCauley I, Jackson P, Long KA, Lopaticki S, Nugent EA, Simons JA, Walker J and Hennessy DP (2001): Vaccination of boars with a GnRH vaccine (Improvac) eliminates boar taint and increases growth performance. Journal of Animal Science 79(10), 2524–2535.

Efroni S, Schaefer CF and Buetow KH (2007): Identification of key processes underlying cancer phenotypes using biologic pathway analysis. PLoS ONE 2(5), e425.

Emami Riedmaier A, Nies AT, Schaeffeler E and Schwab M (2012): Organic anion transporters and their implications in pharmacotherapy. Pharmacological Reviews 64(3), 421–449.

Esteve-Codina A, Kofler R, Palmieri N, Bussotti G, Notredame C and Perez-Enciso M (2011): Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. BMC Genomics 12(1), 552.

Fadiel A, Anidi I and Eichenbaum KD (2005): Farm animal genomics and informatics: an update. Nucleic Acids Research 33(19), 6308–6318.

Fan CY, Pan J, Chu R, Lee D, Kluckman KD, Usuda N, Singh I, Yeldandi AV, Rao MS, Maeda N and Reddy JK (1996): Hepatocellular and hepatic peroxisomal alterations in mice with a disrupted peroxisomal fatty acyl-coenzyme A oxidase gene. Journal of Biological Chemistry 271(40), 24698–24710.

Fayyad U, Piatetsky-Shapiro G and Padhraic S (1996a): Knowledge Discovery and data mining : Towards a unifying framework. AAAI Press, Portland, 82–88.

Fayyad U, Piatetsky-shapiro G and Smyth P (1996b): From Data Mining to Knowledge Discovery in Databases. AI Magazine 17, 37–54.

Fayyad UM, Piatetsky-Shapiro G and Smyth P (1996c): Advances in Knowledge Discovery and Data Mining. chap. From Data, American Association for Artificial Intelligence, Menlo Park, CA, USA, 1–34.

Fellows I (2013): wordcloud: Word Clouds. Tech. rep., R package version 2.4.

Ferragina P and Manzini G (2000): Opportunistic data structures with applications. In: Proceedings of the 41st Annual Symposium on Foundations of Computer Science, IEEE Computer Society, Redondo Beach, CA, USA, 390–398.

Ferragina P and Manzini G (2001): An experimental study of an opportunistic index. In: Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 269–278.

Fishbein JC (2011): Advances in molecular toxicology. No. v. 5 in Advances in molecular toxicology, Elsevier Science Serials, 124–125 .

Fix C, Jordan C, Cano P and Walker WH (2004): Testosterone activates mitogen-activated protein kinase and the cAMP response element binding protein transcription factor in Sertoli cells. Proceedings of the National Academy of Sciences 101(30), 10919–10924.

Fleury A, Mathieu AP, Ducharme L, Hales DB and LeHoux JG (2004): Phosphorylation and function of the hamster adrenal steroidogenic acute regulatory protein (StAR). Journal of Steroid Biochemistry 91(4-5), 259–271.

Flori L, Gao Y, Laloë D, Lemonnier G, Leplat JJ, Teillaud A, Cossalter AM, Laffitte J, Pinton P, de Vaureix C, Bouffaud M, Mercat MJ, Lefèvre F, Oswald IP, Bidanel JP and Rogel-Gaillard C (2011): Immunity traits in pigs: substantial genetic variation and limited covariation. PLoS ONE 6(7), e22717.

Foye WO, Lemke TL and Williams DA (2008): Foye's Principles of Medicinal Chemistry. Lippincott Williams & Wilkins, 1377 .

Freeman T, Ivens A, Baillie JK, Beraldi D, Barnett M, Dorward D, Downing A, Fairbairn L, Kapetanovic R, Raza S, Tomoiu A, Alberio R, Wu C, Su A, Summers K, Tuggle C, Archibald A and Hume D (2012): A gene expression atlas of the domestic pig. BMC Biology 10(1), 90.

Frey PA and Hegeman AD (2007): Enzymatic reaction mechanisms. Oxford University Press, USA, 627–628 .

Frieden L, Looft C and Tholen E (2011): Breeding for reduced boar taint. Lohmann Information 46(1), 21–27.

Fuchs M, Hafer A, Münch C, Kannenberg F, Teichmann S, Scheibner J, Stange EF and Seedorf U (2001): Disruption of the sterol carrier protein 2 gene in mice impairs biliary lipid and hepatic cholesterol metabolism. Journal of Biological Chemistry 276(51), 48058–48065.

Furuhashi M, Ura N, Murakami H, Hyakukoku M, Yamaguchi K, Higashiura K and Shimamoto K (2002): Fenofibrate improves insulin sensitivity in connection with intramuscular lipid content, muscle fatty acid-binding protein, and beta-oxidation in skeletal muscle. Journal of Endocrinology 174(2), 321–329.

Gad A, Schellander K, Hoelker M and Tesfaye D (2012): Transcriptome profile of early mammalian embryos in response to culture environment. Animal Reproduction Science 134(1-2), 76–83.

Garattini E, Fratelli M and Terao M (2009): The mammalian aldehyde oxidase gene family. Human Genomics 4(2), 119–130.

Gatti DM, Sypa M, Rusyn I, Wright FA and Barry WT (2009): SAFEGUI: resampling-based tests of categorical significance in gene expression data made easy. Bioinformatics 25(4), 541–542.

GEO Datasets B taurus (2014): http://www.ncbi.nlm.nih.gov/sra/?term=txid9913[Organism:noexp].

GEO Datasets S scrofa (2014): http://www.ncbi.nlm.nih.gov/gds/?term=txid9823[Organism:noexp].

Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, Gill CA, Green RD, Hamernik DL, Kappes SM, Lien Sr, Matukumalli LK, McEwan JC, Nazareth LV, Schnabel RD, Weinstock GM, Wheeler DA, Ajmone-Marsan P, Boettcher PJ, Caetano AR, Garcia JF, Hanotte O, Mariani P, Skow LC, Sonstegard TS, Williams JL, Diallo B, Hailemariam L, Martinez ML, Morris CA, Silva LOC, Spelman RJ, Mulatu W, Zhao K, Abbey CA, Agaba M, Araujo FR, Bunch RJ, Burton J, Gorni C, Olivier H, Harrison BE, Luff B, Machado MA, Mwakaya J, Plastow G, Sim W, Smith T, Thomas MB, Valentini A, Williams P, Womack J, Woolliams JA, Liu Y, Qin X, Worley KC, Gao C, Jiang H, Moore SS, Ren Y, Song XZ, Bustamante CD, Hernandez RD, Muzny DM, Patil S, San Lucas A, Fu Q, Kent MP, Vega R, Matukumalli A, McWilliam S, Sclep G, Bryc K, Choi J, Gao H, Grefenstette JJ, Murdoch B, Stella A, Villa-Angulo R, Wright M, Aerts J, Jann O, Negrini R, Goddard ME, Hayes BJ, Bradley DG, Barbosa da Silva M, Lau LPL, Liu GE, Lynn DJ, Panzitta F and Dodds KG (2009): Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324(5926), 528–532.

Gibson GG and Skett P (2001): Introduction to drug metabolism. illustrate edn., Nelson Thornes Publishers, Cheltenham, Great Britain, 256 .

Girvan M and Newman MEJ (2002): Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99(12), 7821–7826.

Glass EJ, Crutchley S and Jensen K (2012): Living with the enemy or uninvited guests: Functional genomics approaches to investigating host resistance or tolerance traits to a protozoan parasite, Theileria annulata, in cattle. Veterinary Immunology and Immunopathology 148(1–2), 178–189.

Goddard ME and Hayes BJ (2007): Genomic selection. Journal of Animal Breeding and Genetics 124(6), 323–330.

Goder A and Filkov V (2008): Consensus Clustering Algorithms: Comparison and Refinement. In: Proceedings of the 9th workshop on Algorithm Engineering and Experiments, SIAM, San Francisco, California, USA, 109–117.

Golik W, Dameron O, Bugeon J, Fatet A, Hue I, Hurtaud C, Reichstadt M, Salauen MC, Vernet J, Joret L, Papazian F, Nédellec C and Le Bail PY (2012): ATOL: the multi-species livestock trait ontology. In: 6th International Conference on Metadata and Semantic Research, Cadiz, Spain.

González-Recio O, Gianola D, Rosa GJ, Weigel KA and Kranis A (2009): Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. Genetics Selection Evolution 41(1), 3.

Gregersen VR, Conley LN, Sø rensen KK, Guldbrandtsen B, Velander IH and Bendixen C (2012): Genome-wide association scan and phased haplotype construction for quantitative trait loci affecting boar taint in three pig breeds. BMC Genomics 13(1), 22.

Grindflek E, Berget I, Moe M, Oeth P and Lien Sr (2010): Transcript profiling of candidate genes in testis of pigs exhibiting large differences in androstenone levels. BMC Genetics 11(1), 4.

Grindflek E, Lien Sr, Hamland H, Hansen MH, Kent M, Van Son M and Meuwissen TH (2011): Large scale genome-wide association and LDLA mapping study identifies QTLs for boar taint and related sex steroids. BMC Genomics 12(1), 362.

Gu J, Orr N, Park SD, Katz LM, Sulimova G, MacHugh DE and Hill EW (2009): A genome scan for positive selection in thoroughbred horses. PLoS ONE 4(6), e5767.

Guan Y, Ackert-Bicknell CL, Kell B, Troyanskaya OG and Hibbs MA (2010): Functional genomics complements quantitative genetics in identifying disease-gene associations. PLoS Computational Biology 6(11), e1000991.

Guimerà R, Sales-Pardo M and Amaral LAN (2004): Modularity from fluctuations in random graphs and complex networks. Physical Review E 70(2), 25101.

Gunawan A, Sahadevan S, Neuhoff C, Groß e Brinkhaus C, Gad A, Frieden L, Tesfaye D, Tholen E, Looft C, Uddin MJ, Schellander K and Cinar MU (2013): RNA deep sequencing reveals novel candidate genes and polymorphisms in boar testis and liver tissues with divergent androstenone levels. PLoS ONE 8(5), e63259.

Guo J, Jorjani H and Carlborg O (2012): A genome-wide association study using international breeding-evaluation data identifies major loci affecting production traits and stature in the Brown Swiss cattle breed. BMC Genetics 13(1), 82.

Hacker BM, Tomlinson JE, Wayman GA, Sultana R, Chan G, Villacres E, Disteche C and Storm DR (1998): Cloning, chromosomal mapping, and regulatory properties of the human type 9 adenylyl cyclase (ADCY9). Genomics 50(1), 97–104.

Hagan R, Davies J and Wilton D (2002): The effect of charge reversal mutations in the α-helical region of liver fatty acid binding protein on the binding of fatty-acyl CoAs, lysophospholipids and bile acids. Molecular and Cellular Biochemistry 239(1-2), 55–60.

Han J and Kamber M (2011): Data Mining: Concepts and Techniques: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems, Elsevier Science, 744 .

Hanada K (2003): Serine palmitoyltransferase, a key enzyme of sphingolipid metabolism. Biochimica et Biophysica Acta 1632(1-3), 16–30.

Handschin C and Meyer UA (2003): Induction of drug metabolism: the role of nuclear receptors. Pharmacological Reviews 55(4), 649–673.

Hannun YA and Obeid LM (2008): Principles of bioactive lipid signalling: lessons from sphingolipids. Molecular Cell Biology 9(2), 139–150.

Hartley JL, Temple GF and Brasch MA (2000): DNA cloning using in vitro site-specific recombination. Genome Research 10(11), 1788–1795.

Hartwell LH, Hopfield JJ, Leibler S and Murray AW (1999): From molecular to modular cell biology. Nature , C47–C52.

Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D and Brown P (2000): 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biology 1(2), research0003.1–research0003.21.

Hatem A, Bozdag D, Toland A and Catalyurek U (2013): Benchmarking short sequence mapping tools. BMC Bioinformatics 14(1), 184.

Haugen JE, Brunius C and Zamaratskaia G (2012): Review of analytical methods to measure boar taint compounds in porcine adipose tissue: the need for harmonised methods. Meat Science 90(1), 9–19.

Hawkins RD, Hon GC and Ren B (2010): Next-generation genomics: an integrative approach. Nature Reviews Genetics 11(7), 476–486.

Hayashi T and Iwata H (2010): EM algorithm for Bayesian estimation of genomic breeding values. BMC Genetics 11, 3.

Hayes BJ, Bowman PJ, Chamberlain AJ, Savin K, van Tassell CP, Sonstegard TS and Goddard ME (2009): A validated genome wide association study to breed cattle adapted to an environment altered by climate change. PLoS ONE 4(8), e6676.

Hayes BJ, Lewin HA and Goddard ME (2013): The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. Trends in Genetics 29(4), 206–214.

He XY, Merz G, Yang YZ, Mehta P, Schulz H and Yang SY (2001): Characterization and localization of human type10 17beta-hydroxysteroid dehydrogenase. European Journal of Biochemistry 268(18), 4899–4907.

Heemers HV, Regan KM, Dehm SM and Tindall DJ (2007): Androgen induction of the androgen receptor coactivator four and a half LIM domain protein-2: evidence for a role for serum response factor in prostate cancer. Cancer Research 67(21), 10592–10599.

Herrmann C, de Sande B, Potier D and Aerts S (2012): i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. Nucleic Acids Research 40(15), e114.

Hill DP, Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Sherlock G, Issel-Tarver L, Lewis S and Rubin GM (2000): Gene Ontology: tool for the unification of biology. Nature Genetics 25(1), 25–29.

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson RK and Mardis ER (2008): Whole-genome sequencing and variant discovery in C. elegans. Nature Methods 5(2), 183–188.

Hirabayashi Y, Igarashi Y and Merrill AHJ (2006): Sphingolipid Biology. Springer, 219–220 .

Holloway L and Morris C (2008): Boosted bodies: Genetic techniques, domestic livestock bodies and complex representations of life. Geoforum 39(5), 1709–1720.

Hu J, Zhang Z, Shen WJ and Azhar S (2010): Cellular cholesterol delivery, intracellular processing and utilization for biosynthesis of steroid hormones. Nutrition & Metabolism 7(1), 47.

Hu Z, Kumar D and Reecy JM (2013a): CorrDB: A Livestock Animal Genetic/Phenotypic Trait Correlation Database. In: Plant & Animal Genomes XX Conference, San Diego, CA, P0960.

Hu ZL, Koltes JE, Park CA, Fritz ER and Reecy JM (2011): Bioinformatics approaches to livestock animal genomics research. CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources 6, 1–15.

Hu ZL, Park CA, Wu XL and Reecy JM (2013b): Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Research 41(D1), D871–D879.

Humblot P, Le Bourhis D, Fritz S, Colleau JJ, Gonzalez C, Guyader Joly C, Malafosse A, Heyman Y, Amigues Y, Tissier M and Ponsart C (2010): Reproductive technologies and genomic selection in cattle. Veterinary Medicine International 2010, 1–8.

Isabella Pörn M, Tenhunen J and Peter Slotte J (1991): Increased steroid hormone secretion in mouse Leydig tumor cells after induction of cholesterol translocation by sphingomyelin degradation. Biochimica et Biophysica Acta 1093(1), 7–12.

Islam MA, Uddin MJ, Tholen E, Tesfaye D, Looft C, Schellander K and Cinar MU (2013): Age-associated differential production of IFN-γ, IL-10 and GM-CSF by porcine alveolar macrophages in response to lipopolysaccharide. Veterinary Journal 198(1), 245–251.

James Squires E (2010): Metabolism of androstenone and skatole. In: Applied Animal Endocrinology, 2nd edn., chap. 1.2, Cambridge University Press, Cambridge, 103–106.

Jiang JJ and Conrath DW (1997): Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the international conference on research in computational linguistics, 19–33.

Jiang L, Sø rensen P, Thomsen B, Edwards SM, Skarman A, Rø ntved CM, Lund MS and Workman CT (2012): Gene prioritization for livestock diseases by data integration. Physiological Genomics 44(5), 305–317.

Jin Y, Duan L, Lee SH, Kloosterboer HJ, Blair IA and Penning TM (2009): Human cytosolic hydroxysteroid dehydrogenases of the aldo-ketoreductase superfamily catalyze reduction of conjugated steroids: implications for phase I and phase II steroid hormone metabolism. Journal of Biological Chemistry 284(15), 10013–10022.

Jin Y and Penning TM (2001): Steroid 5alpha-reductases and 3alpha-hydroxysteroid dehydrogenases: key enzymes in androgen metabolism. Best Practice & Research Clinical Endocrinology & Metabolism 15(1), 79–94.

Jo Y, King SR, Khan SA and Stocco DM (2005): Involvement of protein kinase C and cyclic adenosine 3',5'-monophosphate-dependent kinase in steroidogenic acute regulatory protein expression and steroid biosynthesis in Leydig cells. Biology of Reproduction 73(2), 244–55.

Johnson DS, Mortazavi A, Myers RM and Wold B (2007): Genome-wide mapping of in vivo protein-DNA interactions. Science 316(5830), 1497–1502.

Johnson KA and Johnson DE (1995): Methane emissions from cattle. Journal of Animal Science 73(8), 2483–2492.

Jung WY, Kwon SG, Son M, Cho ES, Lee Y, Kim JH, Kim BW, Park DH, Hwang JH, Kim TW, Park HC, Park BY, Choi JS, Cho KK, Chung KH, Song YM, Kim IS, Jin SK, Kim DH, Lee SW, Lee KW, Bang WY and Kim CW (2012): RNA-Seq Approach for Genetic Improvement of Meat Quality in Pig and Evolutionary Insight into the Substrate Specificity of Animal Carbonyl Reductases. PLoS ONE 7(9), e42198.

Kanehisa M and Goto S (2000): KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28(1), 27–30.

Kappes SM, Keele JW, Stone RT, McGraw RA, Sonstegard TS, Smith TP, Lopez-Corrales NL and Beattie CW (1997): A second-generation linkage map of the bovine genome. Genome Research 7(3), 235–249.

Khatkar MS, Thomson PC, Tammen I and Raadsma HW (2004): Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genetics Selection Evolution 36(2), 163–190.

Kleiber K, Strebhardt K and Martin BT (2007): The biological relevance of FHL2 in tumour cells and its role as a putative cancer target. Anticancer Research 27(1A), 55–61.

Koishi R, Ando Y, Ono M, Shimamura M, Yasumo H, Fujiwara T, Horikoshi H and Furukawa H (2002): Angptl3 regulates lipid metabolism in mice. Nature Genetics 30(2), 151–157.

Kojima M and Degawa M (2013): Sex Differences in the Constitutive Gene Expression of Sulfotransferases and UDP-glucuronosyltransferases in the Pig Liver: Androgen-Mediated Regulation. Drug Metabolism and Pharmacokinetics [Epub].

Krallinger M and Valencia A (2005): Text-mining and information-retrieval services for molecular biology. Genome Biology 6(7), 224.

Kugler KG, Mueller LAJ, Graber A and Dehmer M (2011): Integrative network biology: graph prototyping for co-expression cancer networks. PLoS ONE 6(7), e22843.

Kumar AS (2011): Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains. Premier reference source, 390 .

Kumar TR, Wiseman AL, Kala G, Kala SV, Matzuk MM and Lieberman MW (2000): Reproductive defects in gamma-glutamyl transpeptidase-deficient mice. Endocrinology 141(11), 4270–4277.

Kuuranne T, Kurkela M, Thevis M, Schänzer W, Finel M and Kostiainen R (2003): Glucuronidation of anabolic androgenic steroids by recombinant human UDP-glucuronosyltransferases. Drug Metabolism and Disposition 31(9), 1117–1124.

Lancichinetti A and Fortunato S (2009): Community detection algorithms: A comparative analysis. Physical Review E 80(5), 12, 0908.1062.

Lancichinetti A and Fortunato S (2012): Consensus clustering in complex networks. Scientific Reports 2, 336.

Lancichinetti A, Fortunato S and Radicchi F (2008): Benchmark graphs for testing community detection algorithms. Physical Review E 78(4), 46110.

Lancichinetti A, Radicchi F and Ramasco JJ (2010): Statistical significance of communities in networks. Physical Review E 81(4), 046110, 0907.3708.

Lancichinetti A, Radicchi F, Ramasco JJ and Fortunato S (2011): Finding statistically significant communities in networks. PLoS ONE 6(4), e18961.

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

Langmead B, Trapnell C, Pop M and Salzberg SL (2009): Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10(3), R25.

Lassmann T, Hayashizaki Y and Daub CO (2011): SAMStat: monitoring biases in next generation sequencing data. Bioinformatics 27(1), 130–131.

Law CW, Chen Y, Shi W and Smyth GK (2013): Voom! Precision weights unlock linear model analysis tools for RNA-seq read counts. Tech. rep., Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia.

Lecerf F, Bretaudeau A, Sallou O, Desert C, Blum Y, Lagarrigue S and Demeure O (2011): AnnotQTL: a new tool to gather functional and comparative information on a genomic region. Nucleic Acids Research 39(suppl 2), W328–W333.

Lee GJ, Archibald AL, Law AS, Lloyd S, Wood J and Haley CS (2005): Detection of quantitative trait loci for androstenone, skatole and boar taint in a cross between Large White and Meishan pigs. Animal Genetics 36(1), 14–22.

Lee W, Wells T and Kantorow M (2007): Localization and H2O2-specific induction of PRDX3 in the eye lens. Molecular Vision 13, 1469–1474.

Lennarz WJ and Lane D (2013): Encyclopedia of biological chemistry. 2nd edn., Elsevier Science, 289 .

Lervik S, Oskam I, Krogenæ s A, Andresen Oy, Dahl E, Haga HA, Tajet Hv, Olsaker I and Ropstad E (2013): Androstenone and testosterone levels and testicular morphology of Duroc boars related to estimated breeding value for androstenone. Theriogenology 79(6), 986–94.

Leung MCK, Bowley KL and Squires EJ (2010): Examination of testicular gene expression patterns in Yorkshire pigs with high and low levels of boar taint. Animal Biotechnology 21(2), 77–87.

Li H (2011): A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21), 2987–2993.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R (2009): The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16), 2078–2079.

Li H, Ruan J and Durbin R (2008a): Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18(11), 1851–1858.

Li L and Kaplan J (1996): Characterization of Yeast Methyl Sterol Oxidase (ERG25) and Identification of a Human Homologue. Journal of Biological Chemistry 271(28), 16927–16933.

Li R, Li Y, Kristiansen K and Wang J (2008b): SOAP: short oligonucleotide alignment program. Bioinformatics 24(5), 713–714.

Liang JJ, Oelkers P, Guo C, Chu PC, Dixon JL, Ginsberg HN and Sturley SL (2004): Overexpression of Human Diacylglycerol Acyltransferase 1, Acyl-CoA:Cholesterol Acyltransferase 1, or Acyl-CoA:Cholesterol Acyltransferase 2 Stimulates Secretion of Apolipoprotein B-containing Lipoproteins in McA-RH7777 Cells. Journal of Biological Chemistry 279(43), 44938–44944.

Lillehammer M, Meuwissen THE and Sonesson AK (2013): Genomic selection for two traits in a maternal pig breeding scheme. Journal of Animal Science 91(7), 3079–3087.

Lim D, Kim NK, Lee SH, Park HS, Cho YM, Chai HH and Kim H (2014): Characterization of genes for beef marbling based on applying gene coexpression network. International Journal of Genomics 2014, 708562.

Lin D (1998): An Information-Theoretic Definition of Similarity. In: ICML, Morgan Kaufmann, Madison, 296–304.

Lin M, Shen X and Chen X (2011): PAIR: the predicted Arabidopsis interactome resource. Nucleic Acids Research 39(1), D1134–D1140.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B and Ecker JR (2009): Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462(7271), 315–322.

Lord PW, Stevens RD, Brass A and Goble CA (2003): Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19(10), 1275–1283.

Loscalzo J and Barabasi AL (2011): Systems biology and the future of medicine. Wiley Interdisciplinary Reviews: Systems Biology and Medicine 3(6), 619–627.

Lösel RM, Besong D, Peluso JJ and Wehling M (2008): Progesterone receptor membrane component 1–many tasks for a versatile protein. Steroids 73(9-10), 929–934.

Losel RM, Falkenstein E, Feuring M, Schultz A, Tillmann HC, Rossol-Haseroth K and Wehling M (2003): Nongenomic steroid action: controversies, questions, and answers. Physiological Reviews 83(3), 965–1016.

Lou H, Du S, Ji Q and Stolz A (2006): Induction of AKR1C2 by phase II inducers: identification of a distal consensus antioxidant response element regulated by NRF2. Molecular Pharmacology 69(5), 1662–1672.

Lu H, Gonzalez FJ and Klaassen C (2010): Alterations in Hepatic mRNA Expression of Phase II Enzymes and Xenobiotic Transporters after Targeted Disruption of Hepatocyte Nuclear Factor 4 Alpha. Toxicological Sciences 118(2), 380–390.

Lucki NC and Sewer MB (2010): The interplay between bioactive sphingolipids and steroid hormones. Steroids 75(6), 390–399.

Ma J, Yang J, Zhou L, Zhang Z, Ma H, Xie X, Zhang F, Xiong X, Cui L, Yang H, Liu X, Duan Y, Xiao S, Ai H, Ren J and Huang L (2013): Genome-Wide Association Study of meat quality traits in a white Duroc×Erhualian F2 intercross and chinese sutai pigs. PLoS ONE 8(5), e64047.

Mackenzie PI, Bock KW, Burchell B, Guillemette C, Ikushiro Si, Iyanagi T, Miners JO, Owens IS and Nebert DW (2005): Nomenclature update for the mammalian UDP glycosyltransferase (UGT) gene superfamily. Pharmacogenetics and Genomics 15(10), 677–685.

Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N and Chinnaiyan AM (2009): Transcriptome sequencing to detect gene fusions in cancer. Nature 458(7234), 97–101.

Mander L and Liu HW (2010): Comprehensive natural products II: Chemistry and Biology. Comprehensive natural products II : chemistry and biology, Elsevier Science, 494–495 .

Manna PR, Huhtaniemi IT and Stocco DM (2009): Mechanisms of protein kinase C signaling in the modulation of 3',5'-cyclic adenosine monophosphate-mediated steroidogenesis in mouse gonadal cells. Endocrinology 150(7), 3308–3317.

Manna PR, Jo Y and Stocco DM (2007): Regulation of Leydig cell steroidogenesis by extracellular signal-regulated kinase 1/2: role of protein kinase A and protein kinase C signaling. Journal of Endocrinology 193(1), 53–63.

Manna PR and Stocco DM (2005): Regulation of the steroidogenic acute regulatory protein expression: functional and physiological consequences. Current Drug Targets - Immune Endocrine & Metabolic Disorders 5(1), 93–108.

Marchitti SA, Brocker C, Stagos D and Vasiliou V (2008): Non-P450 aldehyde oxidizing enzymes: the aldehyde dehydrogenase superfamily. Expert Opinion on Drug Metabolism & Toxicology 4(6), 697–720.

Marot G, Foulley JL, Mayer CD and Jaffrézic F (2009): Moderated effect size and P-value combinations for microarray meta-analyses. Bioinformatics 25(20), 2692–2699.

Martens GA, Vervoort A, de Casteele M, Stangé G, Hellemans K, Van Thi HV, Schuit F and Pipeleers D (2007): Specificity in beta cell expression of l-3-hydroxyacyl-CoA dehydrogenase, short chain, and potential role in down-regulating insulin release. Journal of Biological Chemistry 282(29), 21134–21144.

Martin M (2011): Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17(1), pp. 10–12.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE and Wingender E (2006): TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. Nucleic Acids Research 34(1), D108–D110.

Mawal YR and Qureshi IA (1994): Purification to Homogeneity of Mitochondrial Acyl CoA: Glycine N-Acyltransferase from Human Liver. Biochemical and Biophysical Research Communications 205(2), 1373–1379.

McCarthy FM, Wang N, Magee GB, Nanduri B, Lawrence ML, Camon EB, Barrell DG, Hill DP, Dolan ME, Williams WP, Luthe DS, Bridges SM and Burgess SC (2006): AgBase: a functional genomics resource for agriculture. BMC Genomics 7, 229.

McIntyre L, Lopiano K, Morse A, Amin V, Oberg A, Young L and Nuzhdin S (2011): RNA-seq: technical variability and sampling. BMC Genomics 12(1), 293.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and DePristo MA (2010): The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20(9), 1297–1303.

McQueen CA (2010): Comprehensive toxicology. Elsevier Science, 199 .

Megens HJ, Crooijmans R, Larson G, Scandura M, Iacolina L, Apollonio M, Bertorelle G, Triantafyllidis A, Alexandri P, Muir W, Semiadi G, Perez-Enciso M, Archibald A, Groenen M and Schook L (2010): The Porcine HapMap projects: genome-wide analysis of pig, wild boar and suiforme diversity. In: European Wild Boar 8th International Symposium on Wild Boars and Other Suids, York, UK.

Meroni SB, Pellizzari EH, Cánepa DF and Cigorraga SB (2000): Possible involvement of ceramide in the regulation of rat Leydig cell function. Journal of Steroid Biochemistry 75(4-5), 307–313.

Merrill AH (2002): De novo sphingolipid biosynthesis: a necessary, but dangerous, pathway. Journal of Biological Chemistry 277(29), 25843–25846.

Mesmin B, Pipalia NH, Lund FW, Ramlall TF, Sokolov A, Eliezer D and Maxfield FR (2011): STARD4 abundance regulates sterol transport and sensing. Molecular Biology of the Cell 22(21), 4004–4015.

Meuwissen THE, Hayes BJ and Goddard ME (2001): Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4), 1819–1829.

Meuwissen THE, Solberg TR, Shepherd R and Woolliams JA (2009): A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genetics Selection Evolution 41, 2.

Minoche AE, Dohm JC and Himmelbauer H (2011): Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biology 12(11), R112.

Mitra K, Carvunis AR, Ramesh SK and Ideker T (2013): Integrative approaches for finding modular structure in biological networks. Nature Reviews Genetics 14(10), 719–732.

Moe M, Grindflek E and Doran O (2007a): Expression of 3beta-hydroxysteroid dehydrogenase, cytochrome P450-c17, and sulfotransferase 2B1 proteins in liver and testis of pigs of two breeds: relationship with adipose tissue androstenone concentration. Journal of Animal Science 85(11), 2924–2931.

Moe M, Lien Sr, Bendixen C, Hedegaard J, Hornshø j H, Berget I, Meuwissen THE and Grindflek E (2008): Gene expression profiles in liver of pigs with extreme high and low levels of androstenone. BMC Veterinary Research 4, 29.

Moe M, Meuwissen T, Lien Sr, Bendixen C, Wang X, Conley LN, Berget I, Tajet Hv and Grindflek E (2007b): Gene expression profiles in testis of pigs with extreme high and low levels of androstenone. BMC Genomics 8(1), 405.

Molloy M and Reed B (1995): A critical point for random graphs with a given degree sequence. Random Structures & Algorithms 6(2-3), 161–180.

Morales V, Santana P, Díaz R, Tabraue C, Gallardo G, López Blanco F, Hernández I, Fanjul LF and Ruiz de Galarreta CM (2003): Intratesticular delivery of tumor necrosis factor-alpha and ceramide directly abrogates steroidogenic acute regulatory protein expression and Leydig cell steroidogenesis in adult rats. Endocrinology 144(11), 4763–4772.

Mori K, Kaido M, Fujishiro K and Inoue N (1989): Testicular toxicity and alterations of glutathione metabolism resulting from chronic inhalation of ethylene oxide in rats. Toxicology and Applied Pharmacology 101(2), 299–309.

Mörlein D, Grave A, Sharifi AR, Bücking M and Wicke M (2012): Different scalding techniques do not affect boar taint. Meat Science 91(4), 435–440.

Mount DW (2007): Using the Basic Local Alignment Search Tool (BLAST). CSH protocols 2007(7), pdb.top17.

Mujibi FDN, Nkrumah JD, Durunna ON, Grant JR, Mah J, Wang Z, Basarab J, Plastow G, Crews DH and Moore SS (2011): Associations of marker panel scores with feed intake and efficiency traits in beef cattle using preselected single nucleotide polymorphisms. Journal of Animal Science 89(11), 3362–3371.

Müller JM, Isele U, Metzger E, Rempel A, Moser M, Pscherer A, Breyer T, Holubarsch C, Buettner R and Schüle R (2000): FHL2, a novel tissue-specific coactivator of the androgen receptor. EMBO Journal 19(3), 359–369.

Muzio G, Maggiora M, Paiuzzi E, Oraldi M and Canuto RA (2012): Aldehyde dehydrogenases and cell proliferation. Free Radical Biology & Medicine 52(4), 735–746.

Myszka DG and Swenson RP (1991): Identification by photoaffinity labeling of fatty acid-binding protein as a potential warfarin receptor in rat liver. Journal of Biological Chemistry 266(31), 20725–20731.

Näsman J, Kukkonen JP, Holmqvist T and Akerman KEO (2002): Different roles for Gi and Go proteins in modulation of adenylyl cyclase type-2 activity. Journal of Neurochemistry 83(6), 1252–1261.

Newman M and Girvan M (2004): Finding and evaluating community structure in networks. Physical Review E 69(2), 026113.

Newman MEJ and Leicht EA (2007): Mixture models and exploratory analysis in networks. Proceedings of the National Academy of Sciences 104(23), 9564–9569.

Nguyen H, Thompson JD, Schutz P and Poch O (2013): Intelligent integrative knowledge bases: Bridging genomics, systems biology and personalized medicine. In: 2nd International Conference on Translational & Personalized Medicine, Chicago, USA.

Nguyen N and Caruana R (2007): Consensus Clusterings. In: Seventh IEEE International Conference on Data Mining, IEEE Xplore, Omaha, 607–612.

Nicolau-Solano SI, McGivan JD, Whittington FM, Nieuwhof GJ, Wood JD and Doran O (2006): Relationship between the expression of hepatic but not testicular 3beta-hydroxysteroid dehydrogenase with androstenone deposition in pig adipose tissue. Journal of Animal Science 84(10), 2809–2817.

Nicolazzi E, Picciolini M, Strozzi F, Schnabel R, Lawley C, Pirani A, Brew F and Stella A (2014): SNPchiMp: a database to disentangle the SNPchip jungle in bovine livestock. BMC Genomics 15(1), 123.

Ohsawa I, Nishimaki K, Yasuda C, Kamino K and Ohta S (2003): Deficiency in a mitochondrial aldehyde dehydrogenase increases vulnerability to oxidative stress in PC12 cells. Journal of Neurochemistry 84(5), 1110–7.

Ostersen T, Christensen OF, Henryon M, Nielsen B, Su G and Madsen P (2011): Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. Genetics Selection Evolution 43(1), 38.

Palace B (1996): Data Mining. Tech. rep., Anderson Graduate School of Management (UCLA), Los Angeles.

Palla G, Derenyi I, Farkas I and Vicsek T (2005): Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043), 814–818.

Park PJ (2009): ChIP-seq: advantages and challenges of a maturing technology. Nature Reviews Genetics 10(10), 669–680.

Payne AH and Hardy MP (2007): The leydig cell in health and disease. Contemporary endocrinology, 2nd edn., Humana Press Incorporated, 355 .

Petersen JL, Mickelson JR, Rendahl AK, Valberg SJ, Andersson LS, Axelsson J, Bailey E, Bannasch D, Binns MM, Borges AS, Brama P, da Câmara Machado A, Capomaccio S, Cappelli K, Cothran EG, Distl O, Fox-Clipsham L, Graves KT, Guérin G, Haase B, Hasegawa T, Hemmann K, Hill EW, Leeb T, Lindgren G, Lohi H, Lopes MS, McGivney BA, Mikko S, Orr N, Penedo MCT, Piercy RJ, Raekallio M, Rieder S, Rø ed KH, Swinburne J, Tozaki T, Vaudin M, Wade CM and McCue ME (2013): Genome-Wide Analysis Reveals Selection for Important Traits in Domestic Horse Breeds. PLoS Genetics 9(1), e1003211.

Plapp BV and Berst KB (2003): Specificity of human alcohol dehydrogenase 1C*2 (gamma2gamma2) for steroids and simulation of the uncompetitive inhibition of ethanol metabolism. Chemico-biological interactions 143-144, 183–193.

Polineni P, Aragonda P, Xavier S, Furuta R and Adelson D (2006): The bovine QTL viewer: a web accessible database of bovine Quantitative Trait Loci. BMC Bioinformatics 7(1), 283.

Prough RA, Webb SJ, Wu HQ, Lapenson DP and Waxman DJ (1994): Induction of microsomal and peroxisomal enzymes by dehydroepiandrosterone and its reduced metabolite in rats. Cancer Research 54(11), 2878–2886.

Pryce JE, Arias J, Bowman PJ, Davis SR, Macdonald KA, Waghorn GC, Wales WJ, Williams YJ, Spelman RJ and Hayes BJ (2012): Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. Journal of Dairy Science 95(4), 2108–2119.

Pryce JE and Daetwyler HD (2012): Designing dairy cattle breeding schemes under genomic selection: a review of international research. Animal Production Science 52(3), 107–114.

Quinlan AR and Hall IM (2010): BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6), 841–842.

Quintanilla R, Demeure O, Bidanel JP, Milan D, Iannuccelli N, Amigues Y, Gruand J, Renard C, Chevalet C and Bonneau M (2003): Detection of quantitative trait loci for fat androstenone levels in pigs. Journal of Animal Science 81(2), 385–94.

R Development Core Team RDC (2013): R: A Language and Environment for Statistical Computing. Tech. rep., R Foundation for Statistical Computing, Vienna, Austria.

Radicchi F, Castellano C, Cecconi F, Loreto V and Parisi D (2004): Defining and identifying communities in networks. Proceedings of the National Academy of Sciences 101(9), 2658–2663.

Raghavan U, Albert R and Kumara S (2007): Near linear time algorithm to detect community structures in large-scale networks. Physical Review E 76(3), 036106.

Ramayo-Caldas Y, Mach N, Esteve-Codina A, Corominas J, Castello A, Ballester M, Estelle J, Ibanez-Escriche N, Fernandez A, Perez-Enciso M and Folch J (2012): Liver transcriptome profile in pigs with extreme phenotypes of intramuscular fatty acid composition. BMC Genomics 13(1), 547.

Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND and Betel D (2013): Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biology 14(9), R95.

Rasmussen JT, Rosendal J and Knudsen J (1993): Interaction of acyl-CoA binding protein (ACBP) on processes for which acyl-CoA is a substrate, product or inhibitor. Biochemical Journal 292 (Pt.3), 907–913.

Ray S, Johnston R, Campbell DC, Nugent S, McDade SS, Waugh D and Panov KI (2013): Androgens and estrogens stimulate ribosome biogenesis in prostate and breast cancer cells in receptor dependent manner. Gene 526(1), 53–46.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP and Young RA (2000): Genome-wide location and function of DNA binding proteins. Science 290(5500), 2306–2309.

Resnik P (1999): Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11, 95–130.

Reverter A, Hudson NJ, Wang Y, Tan SH, Barris W, Byrne KA, McWilliam SM, Bottema CDK, Kister A, Greenwood PL, Harper GS, Lehnert SA and Dalrymple BP (2006): A gene coexpression network for bovine skeletal muscle inferred from microarray data. Physiological Genomics 28(1), 76–83.

Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A and Chinnaiyan AM (2005): Probabilistic model of the human protein-protein interaction network. Nature Biotechnology 23(8), 951–959.

Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A and Chinnaiyan AM (2004): ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 6(1), 1–6.

Riegelhaupt JJ, Waase MP, Garbarino J, Cruz DE and Breslow JL (2010): Targeted disruption of steroidogenic acute regulatory protein D4 leads to modest weight reduction and minor alterations in lipid metabolism. Journal of Lipid Research 51(5), 1134–1143.

Ripley BD (1987): Stochastic simulation. John Wiley and Sons, Inc., New York, NY, USA, 237 .

Robic A, Larzul C and Bonneau M (2008): Genetic and metabolic aspects of androstenone and skatole deposition in pig adipose tissue: A review. Genetics Selection Evolution 40(1), 129.

Robinson MD, McCarthy DJ and Smyth GK (2010): edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1), 139–140.

Rodriguez-Agudo D, Calderon-Dominguez M, Ren S, Marques D, Redford K, Medina-Torres MA, Hylemon P, Gil G and Pandak WM (2011): Subcellular localization and regulation of StarD4 protein in macrophages and fibroblasts. Biochimica et Biophysica Acta 1811(10), 597–606.

Rodriguez-Agudo D, Ren S, Wong E, Marques D, Redford K, Gil G, Hylemon P and Pandak WM (2008): Intracellular cholesterol transporter StarD4 binds free cholesterol and increases cholesteryl ester formation. Journal of Lipid Research 49(7), 1409–1419.

Rogers FB (1963): Medical subject headings. Bulletin of the Medical Library Association 51, 114–116.

Ronhovde P and Nussinov Z (2009): Multiresolution community detection for megascale networks by information-based replica correlations. Physical Review E 80(1), 16109.

Rosvall M, Axelsson D and Bergstrom CT (2010): The map equation. European Physical Journal Special topics 178(1), 13–23.

Rosvall M and Bergstrom CT (2007): An information-theoretic framework for resolving community structure in complex networks. Proceedings of the National Academy of Sciences 104(18), 7327–7331.

Rosvall M and Bergstrom CT (2008): Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 105(4), 1118–1123.

Saatchi M, McClure MC, McKay SD, Rolf MM, Kim J, Decker JE, Taxis TM, Chapple RH, Ramey HR, Northcutt SL, Bauck S, Woodward B, Dekkers JCM, Fernando RL, Schnabel RD, Garrick DJ and Taylor JF (2011): Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genetics Selection Evolution 43(1), 40.

Sakata K, Tokue A and Kawai N (2000): Altered synaptic transmission in the hippocampus of the castrated male mouse is reversed by testosterone replacement. Journal of Urology 163(4), 1333–1338.

Samur MK, Yan Z, Wang X, Cao Q, Munshi NC, Li C and Shah PK (2013): canEvolve: A web portal for integrative oncogenomics. PLoS ONE 8(2), e56228.

Sanz E, Evanoff R, Quintana A, Evans E, Miller JA, Ko C, Amieux PS, Griswold MD and McKnight GS (2013): RiboTag analysis of actively translated mRNAs in Sertoli and Leydig cells in vivo. PLoS ONE 8(6), e66179.

Sarkar IN, Schenk R, Miller H and Norton CN (2009): LigerCat: using "MeSH Clouds" from journal, article, or gene citations to facilitate the identification of relevant biomedical literature. In: Proceedings of AMIA annual Symposium 2009, vol. 2009, San Francisco, California, USA, 563–567.

Sauer U, Heinemann M and Zamboni N (2007): Genetics. Getting closer to the whole picture. Science 316(5824), 550–551.

Schaeffer LR (2006): Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics 123(4), 218–223.

Schänzer W (1996): Metabolism of anabolic androgenic steroids. Clinical Chemistry 42(7), 1001–1020.

Schlicker A, Domingues FS, Rahnenführer J and Lengauer T (2006): A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics 7, 302.

Schomburg D, Schomburg I and Chang A (2003): Springer handbook of enzymes. No. v. 12 in Springer Handbook of Enzymes, 2nd edn., Springer, 625 .

Schomburg I, Chang A and Schomburg D (2002): BRENDA, enzyme data and metabolic information. Nucleic Acids Research 30(1), 47–49.

Seiva FRF, Ebaid GMX, Castro AVB, Okoshi K, Nascimento A, Rocha KKH, Padovani CR, Cicogna AC and Novelli ELB (2008): Growth hormone and heart failure: Oxidative stress and energetic metabolism in rats. Growth Hormone & IGF Research 18(4), 275–283.

Shabalin AA, Tjelmeland Hk, Fan C, Perou CM and Nobel AB (2008): Merging two gene-expression studies via cross-platform normalization. Bioinformatics 24(9), 1154–1160.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B and Idekker T (2003): Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research 13, 2498–2504.

Sharifi AM and Mottaghi S (2012): Finasteride as a potential tool to improve mesenchymal stem cell transplantation for myocardial infarction. Medical Hypotheses 78(4), 465–467.

Sharma R, Ellis B and Sharma A (2011): Role of alpha class glutathione transferases (GSTs) in chemoprevention: GSTA1 and A4 overexpressing human leukemia (HL60) cells resist sulforaphane and curcumin induced toxicity. Phytotherapy Research 25(4), 563–568.

Sheng J, Deng HW, Calhoun VD and Wang YP (2011): Integrated Analysis of Gene Expression and Copy Number Data on Gene Shaving Using Independent Component Analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8(6), 1568–1579.

Shepherd RK, Meuwissen THE and Woolliams JA (2010): Genomic selection and complex trait prediction using a fast EM algorithm applied to genome-wide markers. BMC Bioinformatics 11, 529.

Sies H and Packer L (2005): Phase II Conjugation Enzymes and Transport Systems. No. v. 400 in Methods in enzymology, illustrate edn., Elsevier Academic Press, 695 .

Simoni SD, Goemaere J and Knoops B (2008): Silencing of peroxiredoxin 3 and peroxiredoxin 5 reveals the role of mitochondrial peroxiredoxins in the protection of human neuroblastoma SH-SY5Y cells toward MPP+. Neuroscience Letters 433(3), 219–224.

Sinclair PA, Hancock S, Gilmore WJ and Squires EJ (2005): Metabolism of the 16-androstene steroids in primary cultured porcine hepatocytes. Journal of Steroid Biochemistry 96(1), 79–87.

Skogsberg J, Lundström J, Kovacs A, Nilsson R, Noori P, Maleki S, Köhler M, Hamsten A, Tegnér J and Björkegren J (2008): Transcriptional profiling uncovers a network of cholesterol-responsive atherosclerosis target genes. PLoS Genetics 4(3), e1000036.

Smith GW and Rosa GJM (2007): Interpretation of microarray data: trudging out of the abyss towards elucidation of biological significance. Journal of Animal Science 85(13 Suppl), E20–23.

Smyth GK (2005): Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions Using {R} and Bioconductor, Springer, New York, 397–420.

Soars MG, Fettes M, O'Sullivan AC, Riley RJ, Ethell BT and Burchell B (2003): Cloning and characterisation of the first drug-metabolising canine UDP-glucuronosyltransferase of the 2B subfamily. Biochemical Pharmacology 65(8), 1251–1259.

Soneson C and Delorenzi M (2013): A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14(1), 91.

Spoolder H, Bracke M, Mueller-Graf C and Edwards S (2011): Report 1: Preparatory work for the future development of animal based measures for assessing the welfare of sow, boar and piglet including aspects related to pig castration. Tech. rep., European Food safety authority, Parma.

SRA Datasets B taurus (2014): http://www.ncbi.nlm.nih.gov/sra/?term=txid9913[Organism:noexp].

SRA Datasets S scrofa (2014): http://www.ncbi.nlm.nih.gov/sra/?term=txid9823[Organism:exp].

Stanley WA, Filipp FV, Kursula P, Schüller N, Erdmann R, Schliebs W, Sattler M and Wilmanns M (2006): Recognition of a functional peroxisome type 1 target by the dynamic import receptor Pex5p. Molecular Cell 24(5), 653–663.

Steckelbroeck S, Jin Y, Gopishetty S, Oyesanmi B and Penning TM (2004): Human cytosolic 3alpha-hydroxysteroid dehydrogenases of the aldo-keto reductase superfamily display significant 3beta-hydroxysteroid dehydrogenase activity: implications for steroid hormone metabolism and action. Journal of Biological Chemistry 279(11), 10784–10795.

Stocco DM (2005): Multiple signaling pathways regulating steroidogenesis and steroidogenic acute regulatory protein expression: more complicated than we thought. Molecular Endocrinology 19(11), 2647–2659.

Strautnieks SS, Bull LN, Knisely AS, Kocoshis SA, Dahl N, Arnell H, Sokal E, Dahan K, Childs S, Ling V, Tanner MS, Kagalwalla AF, Németh A, Pawlowska J, Baker A, Mieli-Vergani G,

Freimer NB, Gardiner RM and Thompson RJ (1998): A gene encoding a liver-specific ABC transporter is mutated in progressive familial intrahepatic cholestasis. Nature Genetics 20(3), 233–238.

Strehl A, Ghosh J and Cardie C (2002): Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. Journal of Machine Learning Research 3, 583–617.

Stuart JM, Segal E, Koller D and Kim SK (2003): A gene-coexpression network for global discovery of conserved genetic modules. Science 302(5643), 249–255.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and Mesirov JP (2005): Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102(43), 15545–15550.

Sun X, Qu L, Garrick DJ, Dekkers JCM and Fernando RL (2012): A Fast EM Algorithm for BayesA-Like Prediction of Genomic Breeding Values. PLoS ONE 7(11), e49157.

Suomalainen L, Hakala JK, Pentikäinen V, Otala M, Erkkilä K, Pentikäinen MO and Dunkel L (2003): Sphingosine-1-phosphate in inhibition of male germ cell apoptosis in the human testis. Journal of Clinical Endocrinology and Metabolism 88(11), 5572–5579.

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ and von Mering C (2011): The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Research 39(suppl 1), D561–D568.

Takahashi Y, Hursting SD, Perkins SN, Wang TC and Wang TTY (2006): Genistein affects androgen-responsive genes through both androgen- and estrogen-induced signaling pathways. Molecular Carcinogenesis 45(1), 18–25.

Taulan M, Paquet F, Maubert C, Delissen O, Demaille J and Romey MC (2004): Renal toxicogenomic response to chronic uranyl nitrate insult in mice. Environmental Health Perspectives 112(16), 1628–1635.

te Pas M, Hulsegge I, Schokker D, Smits M, Fife M, Zoorob R, Endale ML and Rebel J (2012): Meta-analysis of Chicken - Salmonella infection experiments. BMC Genomics 13(1), 146.

The ENCODE Project Consortium (2004): The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306(5696), 636–640.

Tindall D and Mohler J (2009): Androgen action in prostate cancer. Viii edn., Springer, 128–129 .

Topchy A, Jain AK and Punch W (2005): Clustering ensembles: models of consensus and weak partitions. IEEE transactions on pattern analysis and machine intelligence 27(12), 1866–1881.

Trapnell C, Pachter L and Salzberg SL (2009): TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9), 1105–1111.

Tseng CW, Lin CC, Chen CN, Huang HC and Juan HF (2011): Integrative network analysis reveals active microRNAs and their functions in gastric cancer. BMC systems biology 5(1), 99.

Turgeon D (2003): Glucuronidation activity of the UGT2B17 enzyme toward xenobiotics. Drug Metabolism and Disposition 31(5), 670–676.

Turnbull AK, Kitchen RR, Larionov AA, Renshaw L, Dixon J and Sims AH (2012): Direct integration of intensity-level data from Affymetrix and Illumina microarrays improves statistical power for robust reanalysis. BMC Medical Genomics 5(1), 35.

Tzfadia O, Amar D, Bradbury LMT, Wurtzel ET and Shamir R (2012): The MORPH algorithm: ranking candidate genes for membership in Arabidopsis and tomato pathways. Plant Cell 24(11), 4389–406.

Uddin MJ, Cinar MU, Groß e Brinkhaus C, Tesfaye D, Tholen E, Juengst H, Looft C, Wimmers K, Phatsara C and Schellander K (2011): Mapping quantitative trait loci for innate immune response in the pig. International Journal of Immunogenetics 38(2), 121–131.

Urs AN, Dammer E and Sewer MB (2006): Sphingosine regulates the transcription of CYP17 by binding to steroidogenic factor-1. Endocrinology 147(11), 5249–5258.

Utsunomiya Y, do Carmo A, Carvalheiro R, Neves H, Matos M, Zavarez L, Perez O'Brien A, Solkner J, McEwan J, Cole J, Van Tassell C, Schenkel F, da Silva M, Porto Neto L, Sonstegard T and Garcia J (2013): Genome-wide association study for birth weight in Nellore cattle points to previously described orthologous genes affecting human and bovine height. BMC Genetics 14(1), 52.

van Dongen S (2000): Graph Clustering by Flow Simulation. Ph.D. thesis, University of Utrecht.

Van Eenennaam AL, van der Werf JHJ and Goddard ME (2011): The value of using DNA markers for beef bull selection in the seedstock sector. Journal of Animal Science 89(2), 307–320.

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF and Schenkel FS (2009): Invited review: reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92(1), 16–24.

Vasiliou V and Nebert DW (2005): Analysis and update of the human aldehyde dehydrogenase (ALDH) gene family. Human Genomics 2(2), 138–143.

Venturini I, Zeneroli ML, Corsi L, Baraldi C, Ferrarese C, Pecora N, Frigo M, Alho H, Farina F and Baraldi M (1998): Diazepam binding inhibitor and total cholesterol plasma levels in cirrhosis and hepatocellular carcinoma. Regulatory peptides 74(1), 31–34.

Vluggens A, Andreoletti P, Viswakarma N, Jia Y, Matsumoto K, Kulik W, Khan M, Huang J, Guo D, Yu S, Sarkar J, Singh I, Rao MS, Wanders RJ, Reddy JK and Cherkaoui-Malki M (2010): Functional significance of the two ACOX1 isoforms and their crosstalks with PPAR[alpha] and RXR[alpha]. Laboratory Investigation 90(5), 696–708.

Wang J, He X, Ruan J, Dai M, Chen J, Zhang Y, Hu Y, Ye C, Li S, Cong L, Fang L, Liu B, Li S, Wang J, Burt DW, Wong GKS, Yu J, Yang H and Wang J (2005): ChickVD: a sequence variation database for the chicken genome. Nucleic Acids Research 33(Database issue), D438–441.

Wang JZ, Du Z, Payattakool R, Yu PS and Chen CF (2007): A new method to measure the semantic similarity of GO terms. Bioinformatics 23(10), 1274–1281.

Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD and Morris Q (2010): The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Research 38(suppl 2), W214–W220.

Weber KL, Thallman RM, Keele JW, Snelling WM, Bennett GL, Smith TPL, McDaneld TG, Allan MF, Van Eenennaam AL and Kuehn LA (2012): Accuracy of genomic breeding values in multibreed beef cattle populations derived from deregressed breeding values and phenotypes. Journal of Animal Science 90(12), 4177–4190.

West AP, Shadel GS and Ghosh S (2011): Mitochondria in innate immune responses. Nature Reviews Immunology 11(6), 389–402.

Whitaker HC, Patel D, Howat WJ, Warren AY, Kay JD, Sangan T, Marioni JC, Mitchell J, Aldridge S, Luxton HJ, Massie C, Lynch AG and Neal DE (2013): Peroxiredoxin-3 is overexpressed in prostate cancer and promotes cancer cell survival by protecting cells from oxidative stress. British Journal of Cancer 109(4), 983–993.

Wickham B (2012): An information infrastructure for facilitating the delivery of improved profits on Irish cattle farms and improving the commercial viability of the Irish breeding industry. In: Proceedings of the 38th International Committee for Animal Recording conference, Cork (Ireland), 1–6.

Wiggans GR, Vanraden PM and Cooper TA (2011): The genomic evaluation system in the United States: past, present, future. Journal of Dairy Science 94(6), 3202–3211.

Womack JE (2005): Advances in livestock genomics: opening the barn door. Genome Research 15(12), 1699–705.

Wong AK, Park CY, Greene CS, Bongo LA, Guan Y and Troyanskaya OG (2012): IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. Nucleic Acids Research 40(W1), W484–W490.

Wonsey DR, Zeller KI and Dang CV (2002): The c-Myc target gene PRDX3 is required for mitochondrial homeostasis and neoplastic transformation. Proceedings of the National Academy of Sciences 99(10), 6649–6654.

Xie W (2008): Nuclear Receptors in Drug Metabolism. Wiley, New Jersey, U.S.A, 336 .

Xie W, Yeuh MF, Radominska-Pandya A, Saini SPS, Negishi Y, Bottroff BS, Cabrera GY, Tukey RH and Evans RM (2003): Control of steroid, heme, and carcinogen metabolism by nuclear pregnane X receptor and constitutive androstane receptor. Proceedings of the National Academy of Sciences 100(7), 4150–4155.

Xu C, Li CYT and Kong ANT (2005): Induction of phase I, II and III drug metabolism/transport by xenobiotics. Archives of Pharmacal Research 28(3), 249–268.

Xu L, Tan AC, Winslow RL and Geman D (2008): Merging microarray data from separate breast cancer studies provides a robust prognostic test. BMC Bioinformatics 9(1), 125.

Yang L, Zhang X, Chen J, Wang Q, Wang L, Jiang Y and Pan Y (2010): ReCGiP, a database of reproduction candidate genes in pigs based on bibliomics. Reproductive Biology and Endocrinology 8(1), 96.

Yi N and Banerjee S (2009): Hierarchical generalized linear models for multiple quantitative trait locus mapping. Genetics 181(3), 1101–1113.

Yoshida A, Rzhetsky A, Hsu LC and Chang C (1998): Human aldehyde dehydrogenase gene family. European Journal of Biochemistry 251(3), 549–557.

Yu G, Li F, Qin Y, Bo X, Wu Y and Wang S (2010): GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics 26(7), 976–978.

Yu L, Kalla K, Guthrie E, Vidrine A and Klimecki WT (2003): Genetic variation in genes associated with arsenic metabolism: glutathione S-transferase omega 1-1 and purine nucleoside phosphorylase polymorphisms in European and indigenous Americans. Environmental Health Perspectives 111(11), 1421–1427.

Zaki MJ, Morishita S and Rigoutsos I (2003): Data mining in Bioinformatics: Report on BIOKDD'03. Tech. rep., Seattle, USA.

Zhang X, Huang S, Sun W and Wang W (2012): Rapid and robust resampling-based multiple-testing correction with application in a genome-wide expression quantitative trait loci study. Genetics 190(4), 1511–1520.

Zhang Y, Dufort I, Rheault P and Luu-The V (2000): Characterization of a human 20alpha-hydroxysteroid dehydrogenase. Journal of Molecular Endrocrinology 25(2), 221–228.

Zheng CF, Wang TTY and Weiner H (1993): Cloning and Expression of the Full-Length cDNAS Encoding Human Liver Class 1 and Class 2 Aldehyde Dehydrogenase. Alcoholism: Clinical and Experimental Research 17(4), 828–831.

Zheng W, Sun F, Bartlam M, Li X, Li R and Rao Z (2007): The crystal structure of human isopentenyl diphosphate isomerase at 1.7 A resolution reveals its catalytic mechanism in isoprenoid biosynthesis. Journal of Molecular Biology 366(5), 1447–1458.

Zingg JM, Kempna P, Paris M, Reiter E, Villacorta L, Cipollone R, Munteanu A, Pascale CD, Menini S, Cueff A, Arock M, Azzi A and Ricciarelli R (2008): Characterization of three human sec14p-like proteins: α-Tocopherol transport activity and expression pattern in tissues. Biochimie 90(11–12), 1703–1715.

Appendices

.1 Publications

Thesis publications

Methodology and analysis results from Experiment 1 except the variant calling pipeline and results are published as:
Sahadevan S, Gunawan A, Tholen E, Große-Brinkhaus C, Tesfaye D, Schellander K, Hofmann-Apitius M, Cinar MU, Uddin MJ (2014): Pathway based analysis of genes and interactions influencing porcine testis samples from boars with divergent androstenone content in back fat. PLoS ONE 9(3). e91077.

Methodology and analysis results from Experiment 2 were submitted as:
Sahadevan S, Tholen E, Große-Brinkhaus C, Tesfaye D, Schellander K, Hofmann-Apitius M, Cinar MU, Gunawan A, Hölker M, Neuhoff C. Identification of gene co-expression clusters in liver tissues from multiple porcine populations with high and low backfat androstenone phenotype. [BMC Genetics].

Other publications

Salilew-Wondim D, Ahmed I, Gebremedhn S, **Sahadevan S**, Hossain MD, Rings F, Hölker M, Tholen E, Neuhoff C, Looft C, Schellander K, Tesfaye D. The expression pattern of microRNAs in granulosa cells of subordinate and dominant follicles during the early luteal phase of the bovine estrous cycle. [Under review: PLoS ONE]

Gunawan A, **Sahadevan S**, Cinar MU, Neuhoff C, Große-Brinkhaus C, Frieden L, Tesfaye D, Tholen E, Looft D, Salilew Wondim D, Hölker M, Schellander K, Uddin MJ (2013): Identification of the novel candidate genes and variants in boar liver tissues with divergent skatole levels using RNA deep sequencing. PLoS ONE 8(5): e72298.

Gunawan A, **Sahadevan S**, Neuhoff C, Große-Brinkhaus C, Gad A, Frieden L, Tesfaye D, Tholen E, Looft C, Uddin MJ, Schellander K, Cinar MU (2013): RNA deep sequencing reveals novel candidate genes and polymorphisms in boar testis and liver tissues with divergent androstenone levels. PLoS ONE 8(5): e63259.

Sahadevan S, Hofmann-Apitius M, Schellander K, Tesfaye D, Fluck J, Friedrich CM. (2012): Text mining in livestock animal science: introducing the potential of text mining to animal sciences. Journal of Animal Science 90(10): 3666–3676.

.2 Literature review: analysis approaches in livestock genomics

Table 1: Appendix Table Analysis approaches in livestock genomics literature.

Pmid	Year	Organism	High throughput platform	Analysis approaches
24631266	2014	*G. gallus*	Agilent 4 × 44K chicken microarray	differential expression analysis, GeneSpring GX
24548287	2014	*B. taurus*	Agilent 8 × 15 K miRNA arrays	correlation network, GeneSpring GX, Multi Experiment Viewer
24467805	2014	*B. taurus*	Affymetrix Bovine GeneChip	differential expression analysis, ANOVA
24496830	2014	*S. scrofa*	Agilent 4 × 44K procine microarray	differential expression analysis, network analysis
24341289	2013	*S. scrofa*	Custom microarray (GEO GPL7151)	Principal Component Analysis, hierarchical clustering, differential expression analysis, limma R package
24104205	2013	*B. taurus*	Agilent 44K bovine microarray	differential expression analysis, mixed model analysis, REML
23893995	2013	*B. taurus*	CombiMatrix bovine microarray	differential expression analysis, local pooled error analysis
23786935	2013	*S. scrofa*	μ Paraflo Microfluidics chip	differential expression analysis, Student's t test
23758853	2013	*O. aries*	Illumina HiSeq 2000	differential expression analysis, Fisher's Exact Test
23550144	2013	*G. gallus*	Illumina 60 K chicken SNP BeadChip	GenABEL, Mann-Whitney U-test
23451171	2013	*S. scrofa*	miRCURY LNA Array	differential expression analysis
24024930	2013	*B. taurus*	Illumina 50 K bovine SNP BeadChip	association analysis, univariate model analysis, PLINK
23803555	2013	*B. taurus*	Affymetrix GeneChip miRNA microarray	differential expression analysis, ANOVA, Principal Component Analysis, hierarchical clustering
23437186	2013	*S. scrofa*	Illumina HiSeq 2000	differential expression analysis, ANOVA, Mann-Whitney U test
23642483	2013	*B. taurus*	Agilent Bovine-Four-Plex G2519F	differential expression anaysis, Student's t test, Principal Component Analysis
23530236	2013	*B. taurus*	Affymetrix bovine GeneChip	differential expression analysis, GeneSpring
23363372	2013	*G. gallus*	Illumina GA II	differential expression analysis, DESeq, SNP calling, mixed model analysis
23355796	2013	*S. scrofa*	Affymetrix porcine GeneChip	differential expression analysis
23284895	2012	*S. scrofa*	Solexa sequencing	reference mapping, prediction
23226446	2012	*G. gallus*	Solexa G1 sequencer, μ Paraflo Microfluidics chip	reference mapping, prediction, differential expression analysis, Audic and Claverie test, Fisher's exact test, and Chi-squared test

Table 1: Appendix table: Analysis approaches in livestock genomics literature (continued...)

Pmid	**Year**	**Organism**	**High throughput platform**	**Analysis approaches**
22844420	2012	*G. gallus*	Illumina 60 K chicken SNP BeadChip	association analysis, PLINK
22567158	2012	*S. scrofa*	Roche NimbleGen Porcine Genome Expression Array	differential expression analysis, linear models, empirical Bayes method, interaction network analysis
22530940	2012	*G. gallus*	Agilent 4 × 44K chicken microarray	differential expression analysis, ANOVA
22848698	2012	*S. scrofa*	Roche 454 GS-FLX pyrosequencing	de novo assembly, prediction
22607119	2012	*B. taurus*	Illumina GAII	differential expression analysis, DESeq
22308471	2012	*G. gallus*	Agilent 4 × 44K chicken microarray	differential expression analysis, ANOVA, SAM
23097340	2012	*G. gallus*	Agilent chicken 44K oligo microarray	differential expression analysis, linear models, empirical Bayes method
22701814	2012	*B. taurus*	BOTL-5 cDNA microarray	differential expression analysis, empirical Bayes model
22531008	2012	*G. gallus*	multiple platforms	differential expression analysis, meta analysis, metaMA
22337866	2012	*S. scrofa*	DJF Pig oligo 27K1 (GPL5972)	differential expression analysis, linear models, Principal component analysis, hierarchical clustering
22270015	2012	*S. scrofa*	Affymetrix porcine GeneChip	differential expression analysis, GeneChip, heirarchical clustering
22234994	2012	*B. taurus*	–	network analysis, gene prioritization, interaction networks, text mining, relevancy scores
22190712	2012	*G. gallus*	Nimblegen chicken genome array	survival analysis, Cox's proportional hazards model, correlation networks, hierarchical clustering
21994447	2011	*E. f. caballus*	Illumina equine SNP50 BeadChip	association analysis, Golden Helix SNP and Variation Suite 7
22099820	2011	*S. scrofa*	Affymetrix porcine GeneChip	differential expression analysis, limma, GenMapp, MAPPFinder
22140460	2011	*G. gallus*	avian IEL array	differential expression analysis, ANOVA, Student's t-test, GeneSpring
20732839	2010	*B. taurus*	Bovine oligonucleotide 24 K chip	differential expression analysis, GeneSifter
20302897	2010	*S. scrofa*	Agilent 244 K porcine microarray	differential expression analysis, ANOVA, Acuity 4.0 Enterprise Microarray Informatics software
20214824	2010	*G. gallus*	Arizona *G. gallus* 20.7K Oligo Array	differential expression analysis, ANOVA

Table 1: Appendix table: Analysis approaches in livestock genomics literature (continued...)

Pmid	**Year**	**Organism**	**High throughput platform**	**Analysis approaches**
20138717	2010	*S. scrofa*	Agilent chicken 44K oligo microarray	differential expression analysis, GeneSpring
19644847	2009	*B. taurus*	Custom miRNA microarray	differential expression analysis
19421343	2009	*B. taurus*	Affymetrix porcine GeneChip	differential expression analysis, paired t-test, Wilcoxon rank sum test, Student's t-test
19366786	2009	*S. scrofa*	Affymetrix porcine GeneChip	linear model analysis
19056128	2009	*O. aries*	Ruminant Immunoinflammatory Gene Universal Array	differential expression analysis
20494844	2008	*B. taurus*	Custom microarray	Student's T-test, differential expression analysis, ANOVA, GeneSifter
18818466	2008	*B. taurus*	NCode Multi-Species miRNA Microarray	differential expression analysis, Significance Analysis of Microarray
17594506	2007	*G. gallus*	Affymetrix chicken GeneChip	differential expression analysis, Significance Analysis of Microarray, hierarchical clustering, Multi Experiment Viewer
17974019	2007	*B. taurus*	Bovine Total Leukocyte cDNA microarray (GPL 363)	differential expression analysis, mixed model analysis
16091418	2005	*B. taurus*	Cattle 7,872-element cDNA (GPL2108)	differential expression analysis, k-means clustering, correlation analysis

Table 2: Appendix Table Number of times each analysis method is mentioned in 50 random full text articles.

Method	**Count**
differential expression analysis	39
ANOVA	9
hierarchical clustering	6
Student's t-test	6
linear models	4
Principal Component Analysis	4
association analysis	3
empirical Bayes method	3
GeneSpring	3
mixed model analysis	3
prediction	3
correlation network	2
DESeq	2
Fisher's exact test	2
GeneSifter	2
GeneSpring GX	2
limma R package	2
Mann-Whitney U test	2

Table 2: Number of times each analysis method is mentioned in 50 random full text articles (continued...)

Method	**Count**
Multi Experiment Viewer	2
network analysis	2
PLINK	2
reference mapping	2
Significance Analysis of Microarray	2
Acuity 4.0 Enterprise Microarray Informatics software	1
Audic and Claverie test	1
Chi-squared test	1
correlation analysis	1
Cox's proportional hazards model	1
de novo assembly	1
GenABEL	1
GeneChip	1
gene prioritization	1
GenMapp	1
Golden Helix SNP and Variation Suite 7	1
interaction network analysis	2
k-means clustering	1
local pooled error analysis	1
MAPPFinder	1
meta analysis	1
metaMA	1
relevancy scores	1
REML	1
SAM	1
SNP calling	1
survival analysis	1
text mining	1
univariate model analysis	1
Wilcoxon rank sum test	1

.3 Results and discussion: Experiment 1 Variant calling

Table 3: Appendix Table Variant calling. *Legend*: NIL indicates polymorphism was absent in the sample. LA. 1 read depth for the polymorphism in sample 1 in LA phenotype, HA. 1 read depth of the polymorphisms in sample 1 in HA phenotype.

Gene name	**Chr**	**POS**	**LA 1**	**LA 2**	**LA 3**	**LA 4**	**LA 5**	**HA 1**	**HA 2**	**HA 3**	**HA 4**	**HA 5**
LOC100152303	1	9399735	23	24	37	46	33	NIL	NIL	NIL	NIL	NIL
LOC100152303	1	9399968	27	27	59	61	46	NIL	NIL	NIL	NIL	NIL
LOC100152988	1	175812592	24	15	34	20	13	26	37	22	22	18
GPX4	2	77676073	15	11	22	18	16	NIL	NIL	NIL	NIL	NIL
LOC100736975	3	100117148	53	22	97	80	72	93	107	45	36	58
HADHA	3	119782443	49	60	103	109	77	86	119	64	45	42
HADHA	3	119782506	45	50	52	85	71	48	104	48	40	22
HADHA	3	119782546	52	62	59	91	80	59	105	63	41	32
HADHA	3	119782551	48	61	62	86	81	57	105	67	45	31
HADHA	3	119782751	69	82	113	117	84	111	126	81	61	58
HADHA	3	119782780	67	73	109	113	85	102	131	82	71	59
MGST3	4	92725756	80	105	110	131	108	99	131	76	64	74
ATP5F1	4	119078700	112	107	127	123	115	113	115	109	104	116
ATP5F1	4	119078761	154	149	176	177	165	169	173	158	149	162
ATP5F1	4	119078830	157	152	170	170	160	173	175	145	145	164
ATP5F1	4	119078856	155	146	170	168	155	172	173	150	145	160
ATP5F1	4	119078862	136	132	160	152	141	159	158	137	130	144
ATP5F1	4	119078864	137	131	155	153	141	162	155	139	133	146
ATP5F1	4	119078865	137	131	158	153	141	162	155	138	133	146
LOC100514231	4	120827636	NIL	NIL	NIL	NIL	NIL	163	164	160	159	162
LOC100514231	4	120827710	179	183	182	182	184	182	180	183	178	184
DHCR24	6	145581907	NIL	NIL	NIL	NIL	NIL	19	37	22	13	20
DHCR24	6	145582020	31	33	38	44	48	45	61	42	54	20
DHCR24	6	145582255	63	53	37	52	44	23	57	50	85	27
DHCR24	6	145582258	63	53	38	53	44	26	59	50	85	26
DHCR24	6	145582458	70	38	43	68	73	43	78	49	96	38
DHCR24	6	145582665	68	49	44	57	42	61	74	48	89	36
DHCR24	6	145582785	NIL	NIL	NIL	NIL	NIL	61	64	33	65	39
CPT2	6	146702408	11	17	32	13	21	NIL	NIL	NIL	NIL	NIL
LOC100517534	6	147870177	34	22	71	46	38	41	75	34	29	29
LOC100517534	6	147870526	32	14	53	32	32	44	76	30	21	26
GALC	7	116349042	NIL	NIL	NIL	NIL	NIL	62	64	41	42	51
GALC	7	116349177	NIL	NIL	NIL	NIL	NIL	56	70	36	33	39
GALC	7	116349201	23	21	55	38	38	NIL	NIL	NIL	NIL	NIL
GALC	7	116349671	23	18	45	33	24	41	48	28	18	41
GSTA2	7	134289767	155	160	86	98	49	68	65	145	27	31
GSTA2	7	134289825	166	167	77	97	49	72	64	153	34	26
GSTA2	7	134289849	166	164	78	90	41	71	59	151	31	19
GSTA2	7	134289905	154	152	74	91	46	58	96	134	32	36
GSTA2	7	134289913	153	153	75	96	59	66	110	134	37	41
GSTA4	7	134380269	112	129	145	144	119	104	150	99	77	74
GSTA4	7	134380285	111	136	151	147	122	109	151	99	76	72
GSTA4	7	134380456	116	137	150	140	132	115	146	116	75	77
HADH	8	122213097	NIL	NIL	NIL	NIL	NIL	19	63	28	13	24

Table 3: Appendix Table Variant calling. *Legend*: NIL indicates polymorphism was absent in the sample. LA. 1 read depth for the polymorphism in sample 1 in LA phenotype, HA. 1 read depth of the polymorphisms in sample 1 in HA phenotype (continued...)

Gene name	**Chr**	**POS**	**LA 1**	**LA 2**	**LA 3**	**LA 4**	**LA 5**	**HA 1**	**HA 2**	**HA 3**	**HA 4**	**HA 5**
HADH	8	122213121	29	48	50	49	51	23	65	30	12	26
ADH5	8	130466631	27	36	53	44	34	NIL	NIL	NIL	NIL	NIL
ADH5	8	130466820	42	48	69	66	31	NIL	NIL	NIL	NIL	NIL
CYP51	9	78792947	35	39	55	50	32	58	61	46	32	35
CYP51	9	78792965	46	52	71	66	37	NIL	NIL	NIL	NIL	NIL
CYP51	9	78792967	NIL	NIL	NIL	NIL	NIL	79	79	60	42	52
CYP51	9	78793035	75	62	114	99	48	120	120	93	69	85
CYP51	9	78793339	74	67	116	106	70	121	136	95	75	110
CYP51	9	78793638	133	105	174	152	143	168	173	139	147	135
DEGS1	10	15053002	21	23	57	48	42	35	59	37	17	38
DEGS1	10	15053060	20	25	57	46	44	35	69	28	18	43
DEGS1	10	15053131	26	23	75	46	35	41	63	36	21	35
DEGS1	10	15053143	26	23	78	42	31	44	59	38	22	32
ACAA1	13	25168976	17	31	46	42	32	26	54	30	13	20
ACAA1	13	25169066	16	21	36	28	24	31	52	18	10	17
ACAA1	13	25169119	14	30	39	30	28	29	50	18	12	14
ACAA1	13	25169195	16	29	30	21	29	NIL	NIL	NIL	NIL	NIL
ACAA1	13	25169225	17	29	28	22	28	21	43	23	10	19
ALDH2	14	42379317	NIL	NIL	NIL	NIL	NIL	38	63	40	25	24
GSTO1	14	125185652	23	22	32	41	27	31	54	38	17	31
ACADSB	14	144190025	14	19	43	27	20	NIL	NIL	NIL	NIL	NIL
ACSL3	15	138712086	33	30	65	50	31	65	65	40	34	41
GPX3	16	78290583	20	51	42	62	28	70	69	32	11	10
GPX3	16	78290858	22	53	40	21	27	60	49	29	18	15
GSS	17	43511491	10	17	20	15	13	NIL	NIL	NIL	NIL	NIL

.4 Results and discussion: Experiment 2 Enrichment Tables

Table 4: Appendix Table LA cluster GO enrichment.

GO.ID	Term	# Annotated	# Significant	Expected	Enrichment p.value
LA cluster 0					
GO:0032259	methylation	195	11	6.48	0.00014
GO:0040011	locomotion	934	23	31.04	0.00076
GO:0022008	neurogenesis	877	19	29.14	0.00303
GO:0002119	nematode larval development	26	5	0.86	0.00378
GO:0006396	RNA processing	559	31	18.57	0.00589
GO:0043069	negative regulation of programmed cell death	406	15	13.49	0.00633
GO:0006839	mitochondrial transport	99	11	3.29	0.01591
GO:0051225	spindle assembly	45	5	1.5	0.01653
GO:0016485	protein processing	84	8	2.79	0.01892
GO:0045454	cell redox homeostasis	51	5	1.69	0.02652
GO:0051436	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	69	6	2.29	0.02675
GO:0006366	transcription from RNA polymerase II promoter	1128	25	37.48	0.03064
GO:0000216	M/G1 transition of mitotic cell cycle	72	6	2.39	0.03213
GO:0007584	response to nutrient	92	6	3.06	0.0354
GO:0007067	mitosis	243	13	8.07	0.03709
GO:0031647	regulation of protein stability	104	7	3.46	0.03904
GO:0018279	protein N-linked glycosylation via asparagine	96	7	3.19	0.04046
GO:0007017	microtubule-based process	382	17	12.69	0.04098
GO:0034660	ncRNA metabolic process	236	17	7.84	0.04187
LA cluster 1					
GO:0021915	neural tube development	90	8	1.9	0.00016
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	381	17	8.05	0.003
GO:0010923	negative regulation of phosphatase activity	49	5	1.04	0.00362
GO:0035239	tube morphogenesis	220	9	4.65	0.00853
GO:0031929	TOR signaling cascade	47	6	0.99	0.01116
GO:0007155	cell adhesion	615	19	13	0.02183
GO:0030308	negative regulation of cell growth	115	7	2.43	0.0286
GO:0006897	endocytosis	293	8	6.19	0.02905
GO:0043065	positive regulation of apoptotic process	333	12	7.04	0.04146
GO:0035023	regulation of Rho protein signal transduction	146	5	3.09	0.04441

Table 4: Appendix Table LA cluster GO enrichment (continued...)

GO.ID	Term	# Annotated	# Significant	Expected	Enrichment p.value
LA cluster 2					
GO:0055114	oxidation-reduction process	807	42	8.31	9.6E-011
GO:0051289	protein homotetramerization	36	6	0.37	0.0000016
GO:0006805	xenobiotic metabolic process	105	8	1.08	0.000012
GO:0006641	triglyceride metabolic process	82	5	0.84	0.002
GO:0006629	lipid metabolic process	911	33	9.38	0.00231
GO:0009058	biosynthetic process	3614	40	37.22	0.01118
GO:0048869	cellular developmental process	2038	11	20.99	0.0115
GO:0006810	transport	2743	34	28.25	0.01378
GO:0008203	cholesterol metabolic process	100	7	1.03	0.01502
GO:0042493	response to drug	269	8	2.77	0.01503
GO:0046395	carboxylic acid catabolic process	157	11	1.62	0.02834
GO:0019439	aromatic compound catabolic process	953	14	9.82	0.02987
GO:0006869	lipid transport	150	5	1.55	0.03686
GO:0009725	response to hormone stimulus	555	7	5.72	0.04158
LA cluster 3					
GO:0006415	translational termination	103	5	0.89	0.0019
GO:0022900	electron transport chain	101	7	0.87	0.0023
GO:0044281	small molecule metabolic process	2035	19	17.5	0.0038
GO:0006401	RNA catabolic process	215	7	1.85	0.0076
GO:0007267	cell-cell signaling	693	7	5.96	0.0416
LA cluster 6					
GO:0006415	translational termination	103	29	0.67	$<$ 1e-30
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	120	29	0.78	$<$ 1e-30
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	128	29	0.84	$<$ 1e-30
GO:0006414	translational elongation	130	29	0.85	$<$ 1e-30
GO:0019083	viral transcription	165	29	1.08	$<$ 1e-30
GO:0006413	translational initiation	177	29	1.16	$<$ 1e-30
GO:0006364	rRNA processing	88	7	0.58	0.000012
GO:0042592	homeostatic process	903	8	5.9	0.038
LA cluster 7					
GO:0048585	negative regulation of response to stimulus	566	5	3.14	0.02706
GO:0006195	purine nucleotide catabolic process	629	5	3.49	0.02722
GO:0006355	regulation of transcription, DNA-templated	1848	13	10.26	0.03237
GO:0048699	generation of neurons	794	5	4.41	0.04386
LA cluster 8					
GO:0010951	negative regulation of endopeptidase activity	109	5	0.59	0.0024
GO:0007243	intracellular protein kinase cascade	723	5	3.89	0.01
GO:0007599	hemostasis	396	5	2.13	0.0155
GO:0005975	carbohydrate metabolic process	651	6	3.5	0.0196

Table 4: Appendix Table LA cluster GO enrichment (continued...)

GO.ID	Term	# Annotated	# Significant	Expected	Enrichment p.value
GO:0043065	positive regulation of apoptotic process	333	6	1.79	0.0493

Table 5: Appendix Table HA cluster GO enrichment.

GO.ID	Term	# Annotated	# Significant	Expected	Enrichment p.value
HA cluster 0					
GO:0006415	translational termination	103	35	4.55	2.8E-022
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	120	37	5.3	8.1E-022
GO:0006414	translational elongation	130	39	5.74	1.4E-021
GO:0019083	viral transcription	165	35	7.29	4.5E-021
GO:0006413	translational initiation	177	41	7.82	1.3E-020
GO:000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	128	34	5.65	7.1E-018
GO:0022904	respiratory electron transport chain	92	21	4.06	5.1E-009
GO:0042273	ribosomal large subunit biogenesis	17	6	0.75	0.000059
GO:0006364	rRNA processing	88	14	3.89	0.000061
GO:0040010	positive regulation of growth rate	20	6	0.88	0.00016
GO:0006099	tricarboxylic acid cycle	28	7	1.24	0.00017
GO:0019430	removal of superoxide radicals	15	5	0.66	0.00063
GO:0007067	mitosis	243	18	10.73	0.0007
GO:0002119	nematode larval development	26	6	1.15	0.00146
GO:0055114	oxidation-reduction process	807	70	35.63	0.00188
GO:0042274	ribosomal small subunit biogenesis	25	7	1.1	0.00257
GO:0072593	reactive oxygen species metabolic process	92	14	4.06	0.00426
GO:0032259	methylation	195	11	8.61	0.00438
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	37	6	1.63	0.00521
GO:0045454	cell redox homeostasis	51	7	2.25	0.00673
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	81	9	3.58	0.00919
GO:0043524	negative regulation of neuron apoptotic process	82	9	3.62	0.00994
GO:0051436	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	69	8	3.05	0.01072
GO:0051437	positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	71	8	3.14	0.01265
GO:009792	embryo development ending in birth or egg hatching	452	14	19.96	0.01336
GO:0042542	response to hydrogen peroxide	62	8	2.74	0.01384

Table 5: Appendix Table HA cluster GO enrichment (continued...)

GO.ID	Term	# Annotated	# Significant	Expected	Enrichment p.value
GO:0042127	regulation of cell proliferation	866	33	38.24	0.01664
GO:0006184	GTP catabolic process	448	26	19.78	0.01677
GO:0042255	ribosome assembly	20	6	0.88	0.01757
GO:0045839	negative regulation of mitosis	35	5	1.55	0.01769
GO:0040017	positive regulation of locomotion	204	8	9.01	0.018
GO:0006412	translation	453	65	20	0.02002
GO:0034660	ncRNA metabolic process	236	24	10.42	0.02413
GO:0009615	response to virus	177	10	7.82	0.02509
GO:0006396	RNA processing	559	42	24.68	0.02512
GO:0051402	neuron apoptotic process	129	12	5.7	0.02564
GO:0090068	positive regulation of cell cycle process	140	12	6.18	0.02886
GO:0045471	response to ethanol	68	7	3	0.02983
GO:0034341	response to interferon-gamma	91	6	4.02	0.03125
GO:0006521	regulation of cellular amino acid metabolic process	55	6	2.43	0.03348
GO:0006749	glutathione metabolic process	42	5	1.85	0.03709
GO:0051591	response to cAMP	57	6	2.52	0.03896
GO:0043154	negative regulation of cysteine-type endopeptidase activity involved in apoptotic process	57	6	2.52	0.03896
GO:0000216	M/G1 transition of mitotic cell cycle	72	7	3.18	0.03908
GO:0022008	neurogenesis	877	27	38.72	0.04097
GO:0044281	small molecule metabolic process	2035	121	89.86	0.04356
GO:0006119	oxidative phosphorylation	54	9	2.38	0.04481
GO:0009790	embryo development	730	29	32.23	0.04586
GO:0048869	cellular developmental process	2038	60	89.99	0.04755
GO:0006200	ATP catabolic process	142	11	6.27	0.04959
HA cluster 1					
GO:0010951	negative regulation of endopeptidase activity	109	7	0.68	0.00000067
GO:0006879	cellular iron ion homeostasis	58	5	0.36	0.00003
GO:0055114	oxidation-reduction process	807	18	5.06	0.000036
GO:0006956	complement activation	47	5	0.29	0.0002
GO:0046395	carboxylic acid catabolic process	157	5	0.98	0.01788
GO:0006875	cellular metal ion homeostasis	238	8	1.49	0.03965
GO:0006508	proteolysis	799	10	5.01	0.04925
HA cluster 3					
GO:0055114	oxidation-reduction process	807	9	1.45	0.00039
GO:0044281	small molecule metabolic process	2035	13	3.65	0.00281
GO:0006082	organic acid metabolic process	757	9	1.36	0.01032
HA cluster 10					
GO:0007156	homophilic cell adhesion	63	5	0.84	0.0014
GO:0009166	nucleotide catabolic process	634	8	8.4	0.0018
GO:0035239	tube morphogenesis	220	6	2.92	0.0035
GO:0021915	neural tube development	90	5	1.19	0.0057

Table 5: Appendix Table HA cluster GO enrichment (continued...)

GO.ID	Term	# Annotated	# Significant	Expected	Enrichment p.value
GO:0009987	cellular process	10095	140	133.82	0.0146
GO:0008152	metabolic process	7747	99	102.69	0.0212
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	381	10	5.05	0.0302
GO:0051726	regulation of cell cycle	657	17	8.71	0.0355
HA cluster 10					
GO:0006457	protein folding	204	7	0.66	0.000013
GO:0006869	lipid transport	150	6	0.48	0.00336
GO:0055114	oxidation-reduction process	807	10	2.6	0.03544

Table 6: Appendix Table LA cluster KEGG enrichment.

KEGG.ID	Pathway	# Enriched genes	Enrichment p.value
LA cluster 2			
ssc03320	PPAR signaling pathway	6	0.00990966
ssc04146	Peroxisome	8	0.00109469
ssc00280	Valine, leucine and isoleucine degradation	6	0.00149421
ssc00071	Fatty acid degradation	8	0.00001695
ssc00830	Retinol metabolism	7	0.00026192
ssc05204	Chemical carcinogenesis	7	0.00082319
ssc00983	Drug metabolism - other enzymes	5	0.00107901
ssc00982	Drug metabolism - cytochrome P450	9	0.00000325
ssc00380	Tryptophan metabolism	5	0.00343914
ssc00980	Metabolism of xenobiotics by cytochrome P450	7	0.00019518
ssc00053	Ascorbate and aldarate metabolism	5	0.00033240
LA cluster 3			
ssc00190	Oxidative phosphorylation	7	0.01474052
ssc04932	Non-alcoholic fatty liver disease (NAFLD)	7	0.04177778
ssc05012	Parkinsons disease	8	0.00499836
LA cluster 4			
ssc01200	Carbon metabolism	5	0.041088727
LA cluster 6			
ssc03010	Ribosome	29	5.09E-025
LA cluster 9			
ssc03013	RNA transport	6	0.011340915
ssc03015	mRNA surveillance pathway	5	0.002345787

Table 7: Appendix Table HA cluster KEGG enrichment.

KEGG.ID	**Pathway**	**# Enriched genes**	**Enrichment p.value**
HA cluster 0			
ssc05016	Huntingtons disease	29	0.0134605056
ssc00190	Oxidative phosphorylation	24	0.0018550649
ssc05010	Alzheimers disease	27	0.0122442626
ssc05012	Parkinsons disease	24	0.0027797706
ssc03010	Ribosome	36	1.3182E-006
HA cluster 1			
ssc04610	Complement and coagulation cascades	12	3.4967E-010
ssc00830	Retinol metabolism	6	7.5598E-005
ssc05204	Chemical carcinogenesis	6	0.0002134292
ssc00860	Porphyrin and chlorophyll metabolism	5	0.0002234595
ssc00982	Drug metabolism - cytochrome P450	6	6.6350E-005
ssc00980	Metabolism of xenobiotics by cytochrome P450	6	0.000058064
HA cluster 17			
ssc03320	PPAR signaling pathway	5	0.0002034416
ssc04141	Protein processing in endoplasmic reticulum	6	0.0014178496

Acknowledgement

First of all, I wish to express my sincere gratitude to Prof. Dr. Karl Schellander for providing me with the opportunity to pursue my doctoral thesis at the Institute of Animal Sciences and supporting me under all conditions during these years. I would also like to show my indebtedness to Prof. Dr. Martin Hofmann-Apitius for allowing me to work at Fraunhofer SCAI Bioinformatics and his support and ideas throughout my thesis. I am obliged to both of them equally for freedom of work I enjoyed during these four years.

I would also like to thank Dr. Ernst Tholen, Dr. Christine Große-Brinkhaus and Dr. Christiane Neuhoff for their constructive criticisms and help during my thesis. I am also thankful to former colleagues Dr. Mehmet Ulas Cinar, Dr. Jasim Uddin and Dr. Asep Gunawan for their scientific support and Ms. Maren Julia Pröll for her help with German abstract translation and final thesis submission procedures. I would also like to take this opportunity to thank Ms. Bianca Peters, Ms. Ulrike Schröter and Ms. Meike Knieps for supporting me with the official matters. I would also like to thank my friends and colleagues at the Institute of Animal Sciences and Fraunhofer SCAI for their co-operation and friendly working environments I enjoyed in both institutes.

Last but not least, I owe my deepest gratitude to my family and friends for help and support throughout the past four years.