HANIEL CEDRAZ DE OLIVEIRA

APPLYING NEXT GENERATION SEQUENCE DATA IN ANIMAL BREEDING

Thesis presented to the Animal Science Graduate Program of Universidade Federal de Viçosa, in partial fulfillment of the requirements for the degree of *Doctor Scientiae.*

Advisor: Simone Eliza Facioni Guimarães

VIÇOSA – MINAS GERAIS
2020

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

O48a
2020

Oliveira, Haniel Cedraz de, 1985-
Applying next generation sequence data in animal breeding / Haniel Cedraz de Oliveira. - Viçosa, MG, 2020.
85 f. : il. (algumas color.) ; 29 cm.

Orientador: Simone Eliza Facioni Guimarães.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Melhoramento genético. 2. Fenótipo. 3. Genômica. 4. Sequenciamento de nucleotídeo. 5. Transcriptoma. I. Universidade Federal de Viçosa. Departamento de Zootecnia. Programa de Pós-Graduação em Zootecnia. II. Título.

CDD 22. ed. 636.0821

HANIEL CEDRAZ DE OLIVEIRA

APPLYING NEXT GENERATION SEQUENCE DATA IN ANIMAL BREEDING

Thesis presented to the Animal Science Graduate Program of the Universidade Federal de Viçosa, in partial fulfillment of the requirements for the degree of *Doctor Scientiae.*

APPROVED: September 30th, 2020.

Assent:

Haniel Cedraz de Oliveira
Author

Simone Eliza Facioni Guimarães
Adviser

Aos meus pais, Levi Cedraz de Oliveira e Telma Patrício de Oliveira, meus exemplos de vida e meus heróis e à minha irmã Marêssa Cedraz de Oliveira.

Dedico.

ACKNOWLEDGMENTS

First of all, I would like to thank God for protection and strength in all these years during my studies.

Especially my mother Telma Patricio de Oliveira and my father Levi Cedraz de Oliveira, who always gave me a good education, encouraging, being always present and doing everything with much love;

All my family for always cheer with my achievements;

To my adviser, Professor Simone Eliza Facioni Guimarães, for patience and orientation along all these years;

The Federal University of Viçosa and the Animal Science Department for support and the opportunity of carrying out the course;

To CNPq for financial support with scholarships;

Embrapa Suinos and Aves Research Center for the partnership and for providing the data needed to carry out one of the chapters in this thesis;

My co-advisors Jane de Oliveira Peixoto and Marcos Soares Lopes for the incentive and confidence in my work, being patient, advising and helping me with the analyses;

Wageningen University & Research, for accept me for doing my doctorate sandwich at the Animal Breeding and Genomics (ABG) Group, providing all support that I needed;

My supervisor and daily advisor during the sandwich, Martien Groenen and Ole Madson, and all the people from ABG for good discussions and support during my training period, also during the coffee break;

My friends from LABTEC (Animal Biotechnology Laboratory) for continuous knowledge, learning exchange and pleasant workplace.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

“Em um lugar escuro nos encontramos, e um pouco mais de conhecimento ilumina nosso caminho"

Mestre Yoda

ABSTRACT

OLIVEIRA, Haniel Cedraz, D.Sc., Universidade Federal de Viçosa, September, 2020. **Applying next generation sequence data in animal breeding.** Advisor: Simone Eliza Facioni Guimarães. Co-advisors: Jane de Oliveira Peixoto and Marcos Soares Lopes.

In Animal Breeding there is a history of using phenotypic data of the best animals from one generation to produce the next one. However, it is known that the heritability of interest traits depends on a genetic basis and can be explained by the DNA. In quantitative genetics, the basic model says that the phenotype is the sum of the genotype (genetic composition of the individual) and the environment, however, different genotypes can perform superior or inferior results depending on the environment in which they are inserted, with that, there was the need to include an interaction factor at the genotype and the environment. The development of Next Generation Sequencing (NGS) has brought advances in the understanding of complex traits in animal breeding, with the Whole Genome Sequencing (WGS) of the domestic animals it was possible to better understand how causal and structural variants have influenced the phenotype. With the use of NGS it is possible to quantify the mRNA that is being expressed of a certain gene that influences a certain trait (transcriptome - RNA-Seq), discover epigenetic changes (ChIP-Seq) and, then, better understand the molecular mechanisms that modulate the trait and tissue in question; increase the density of Simple Nucleotide Polymorphism Markers (SNP) with WGS and then discover more significant and more accurate genomic associations. With the advent of NGS came the need for the development of computational power and new tools capable of analyzing and storing the amount of data generated. As a result, the development of software capable of performing quality control, mapping sequences against the reference genome and quantifying the transcripts (in the case of RNA-Seq) has become necessary and, dealing with this type of analysis can be difficult for those unfamiliar with computer systems. For this reason, a user-friendly pipeline was developed to analyze RNA-Seq data, carrying out quality control, mapping and sequence counting. BAQCOM (Bioinformatics Analysis of Quality Control and Mapping) has been shown to be efficient and fast. Using RNA-Seq data, it was aimed to know the global gene expression profile involved in Bacterial Chondronecrosis with Osteomyelitis (BCO) in chicken, which is developed in the bone growth plates of the femur and tibia followed by opportunistic bacteria. Using tibia samples from six commercial broilers (three affected by BCO and three healthy), 192 differentially expressed genes (FDR

<0.05) were found (63 upregulated and 129 downregulated). 26 genes and seven transcription factors were found downregulated, explaining BCO in tibia, concluding that BCO in chickens may be caused by the low expression of genes related to bone growth and that bacterial proliferation seems to be a secondary process. Another scenario for the use of NGS is the fine mapping of QTL (quantitative Traits Loci) regions. A QTL region on chromosome 5 associated with backfat thickness was found in four commercial pig lines (synthetic based on large-white, Pietrain, Landrace and Large-White). Backfat is important for feed efficiency and meat quality, as well as being a storage of energy. Although the Genome Wide Association Study (GWAS) has been widely used in association with phenotypes, there is still a need to reduce the associated region if the goal is to find causal mutations. For this reason, it was aimed to reduce this QTL region using data from WGS, RNA-Seq and ChIP-Seq. From the most significant SNP of GWAS (leadSNP - SSC5:66103958), haplotypes were phased and haplotypes of size 41 SNPs (20 <leadSNP> 20) were selected. The most frequent haplotype was selected among the four breeds, so it was possible to identify a region of 5 SNPs (2 <leadSNP> 2) cross breed. Fine mapping was carried out using this region, first WGS was used and three candidate variants were identified (SSC5:66097445, SSC5:66099282 and SSC5:66103958). In this region between SSC5:66097445-66103958 epigenetic marks H3K27me3 and H3K4me3 were found using ChIP-Seq data from pig alveolar macrophages, indicating regulation of gene expression during the prenatal development period. From this result, RNA-Seq from embryos and fetuses were used, where it was possible to find high expression of the CCND2 gene, which is related to the development and differentiation of adipose tissue, being a strong candidate for a causal gene for backfat. It is also possible to conclude that in this region there may be regulatory elements involved in the expression of the CCND2 gene in embryonic development and that the epigenetic impact during embryonic life can impact the production of traits in adult life. In this thesis, sequencing, genotyping, phenotype, transcriptome and ChIP-Seq data were used, integrating them all to narrow down regions of interest in the genome, thus resulting and proposing tools to improve animal breeding in the future.

Keywords: Genomics. Whole Genome Sequencing. Transcriptome. ChIP-Sequencing. Phenotype.

RESUMO

OLIVEIRA, Haniel Cedraz, D.Sc., Universidade Federal de Viçosa, setembro de 2020. **Aplicação de sequenciamento de próxima geração no melhoramento animal.** Orientador: Simone Eliza Facioni Guimarães. Co-orientadores: Jane de Oliveira Peixoto e Marcos Soares Lopes.

No Melhoramento Animal existe um histórico de utilizar dados fenotípicos dos melhores animais de uma geração para produzir a próxima geração. Entretanto, é sabido que a herdabilidade das características de interesse depende de uma base genética e podem ser explicadas através do DNA. Na genética quantitativa é usado o modelo básico que diz que o fenótipo é a soma do genótipo (composição genética do indivíduo) e do ambiente, entretanto, genótipos diferentes podem desempenhar resultados superiores ou inferiores dependendo do ambiente em que estão inseridos, com isso, houve a necessidade de incluir o fator interação entre o genótipo e o ambiente. Com o desenvolvimento tecnológico, o sequenciamento de nova geração (NGS) tem trazido avanços no entendimento de características complexas no melhoramento animal, com o sequenciamento completo do genoma dos animais domésticos foi possível melhor entender como variantes causais e estruturais tem influenciado o fenótipo. Com o uso do NGS é possível quantificar o mRNA que está sendo expresso de um determinado gene que influencia determinada característica (transcriptoma – RNA-Seq), descobrir alterações epigenéticas (ChIP-Seq) e, então, conhecer melhor os mecanismos moleculares que modulam a característica e tecido em questão; aumentar a densidade de marcadores de nucleotídeos simples (SNP) com o sequenciamento do genoma completo (WGS) e então, descobrir associações genômicas mais significativas e mais precisas. Com o advento do NGS veio a necessidade de o desenvolvimento poder computacional e novas ferramentas capazes de analisar e armazenar a quantidade de dados gerada. Em função disso, o desenvolvimento de softwares capazes de realizar o controle de qualidade, mapeamento das sequencias contra o genoma de referencia e, quantificar os transcritos (no caso de RNA-Seq) tem se tornado necessário e, lidar com esse tipo de análise pode ser difícil para aqueles que não tem familiaridade com sistemas computacionais. Visto isso foi desenvolvido um pipeline de uso amigável para analisar dados de RNA-Seq, realizando controle de qualidade, mapeamento e contagem de sequencias. O BAQCOM (*Bioinformatics Analysis of Quality Control and Mapping*) tem se mostrado eficiente e rápido. Utilizando dados de RNA-Seq objetivou-se conhecer o perfil global de expressão gênica envolvido em Condronecrose

Bacteriana com Osteomielite em frangos (BCO), que é desenvolvida nas placas de crescimento ósseo do fêmur e tíbia e colonizada por bactérias oportunistas. Utilizando amostras de tíbia de seis frangos de linhagem comercial (três afetados por BCO e três sadios), foram encontrados 192 genes diferencialmente expressos (FDR < 0.05) divididos em 63 upregulados e 129 downregulados. 26 genes e sete fatores de transcrição foram encontrados downregulados explicando BCO em tíbia, podendo concluir que a BCO em frangos pode ser causada pela baixa expressão de genes relacionados ao crescimento ósseo e, que, a proliferação bacteriana parece ser um processo secundário. Outro cenário para o uso de NGS é o mapeamento fino de região de QTL (*quantitative Traits Loci*). Uma região de QTL no cromossomo 5 associada com espessura de toucinho foi encontrada em quatro linhagens comercial de suínos (sintético baseado em large-white, Pietrain, Landrace e Large-White). Espessura de toucinho é importante para eficiência alimentar e qualidade de carne, além de ser reservatório energético. Embora *Genome Wide Association Study* (GWAS) tem sido amplamente usada associada à fenótipos, ainda há a necessidade de reduzir a região de associação se o objetivo é achar mutações causais. Por essa razão objetivou-se reduzir esta região de QTL usando dados de WGS, RNA-Seq e ChIP-Seq. A partir do SNP mais significativo do GWAS (leadSNP – SSC5:66103958), os haplótipos foram postos em fase e selecionados haplótipos de tamanho 41 SNPs (20 < leadSNP > 20). Foi selecionado o haplótipo mais frequente entre as linhagens utilizadas, então foi possível identificar uma região de 5 SNPs (2 < leadSNP > 2) presente em todas as linhagens. A partir desta região foi realizado o mapeamento fino, primeiro foi utilizado WGS e foi identificado três variantes candidatas (SSC5:66097445, SSC5:66099282 e SSC5c66103958). Nesta região entre SSC5:66097445-66103958 foram encontradas marcas epigenéticas H3K27me3 e H3K4me3 utilizado dados de ChIP-Seq de macrófagos alveolar de suínos, indicando regulação de expressão gênica durante período de desenvolvimento pre-natal. Em função disso foram utilizados RNA-Seq de embriões e fetos, onde foi possível encontrar alta expressão do gene CCND2, que está relacionado ao desenvolvimento e diferenciação de tecido adiposo, sendo um forte candidato a gene causal para espessura de toucinho. É possível, também, concluir que nesta região pode ter elementos regulatórios envolvidos na expressão do gene CCND2 no desenvolvimento embrionário e que, o impacto epigenético durante a vida embrionária pode impactar a produção de características na vida adulta. Nessa tese foi usado dados de sequenciamento, genotipagem, fenótipo, transcriptoma e ChIP-Seq, integrando todos para

estreitar regiões de interesse no genoma, assim, resultando e propondo ferramentas para melhorar o melhoramento animal no futuro.

Palavras-chave: Genômica. Sequenciamento de Genoma Completo. Transcriptoma. ChIP-Seq. Fenótipo.

SUMMARY

GENERAL INTRODUCTION

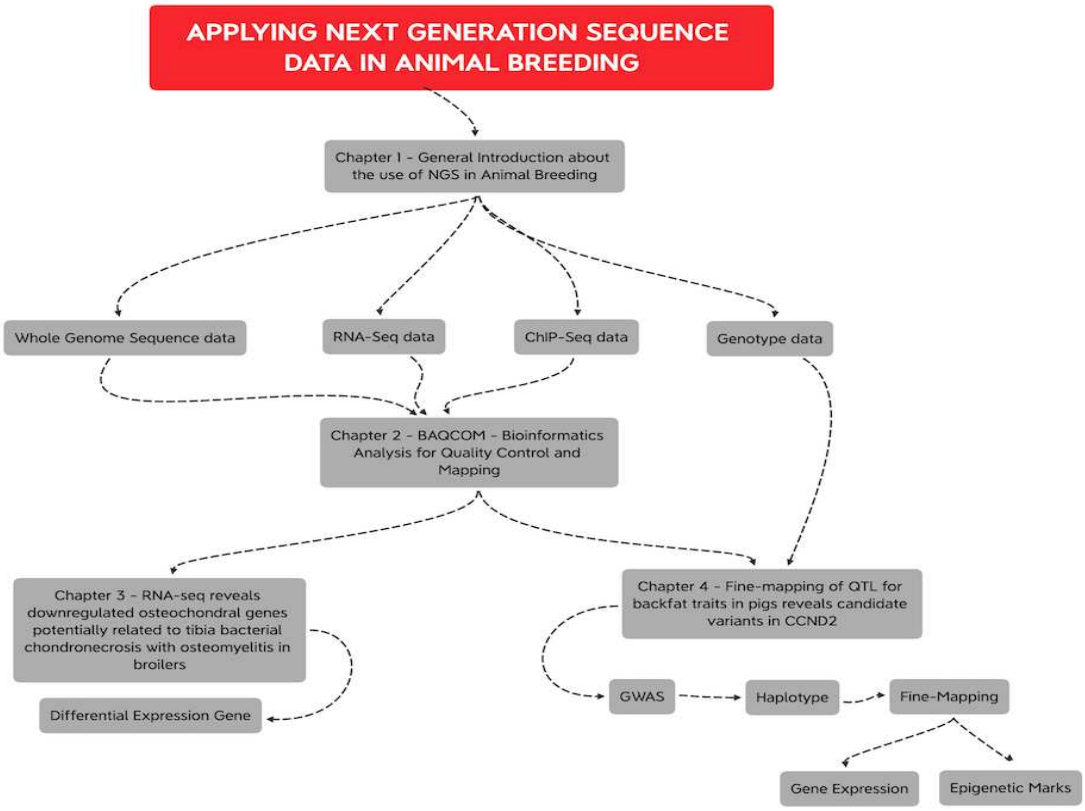

Figure 1: Scheme showing the step by step of BAQCOM pipeline

Introduction to Animal Breeding

Animal Breeding aims to select domestic animals for improving desirable and heritable traits for the next generation, using the best males and females that have passed for a quality control (FALCONER; MACKAY, 1996). The heritable trait depends on the genetic background and differences performance between animals can be explained by the DNA. Hill, (2008) has demonstrated that selection in dairy cattle has resulted in an increase of average milk yield of approximately 50% over a period of 50 years.

Although traits can be selected one by one, most of the animal breeding programs select more than one traits simultaneously, for that, the animal must be superior in the combination of two or more traits related to health, reproduction and production. This process has been repeated for many generations improving the average of the population to the desired direction. However, the major improvement in animal breeding is the use of modern statistical theory and based on the principle of heredity with the quantitative genetics (OLTENACU; BROOM, 2010).

In quantitative genetics, the basic model says that the phenotype (P) is the sum of genotype (G) and environment (E), $P = G + E$ (FALCONER; MACKAY, 1996). It means that the differences in the phenotype comes from the portion of the individual genetic plus the environment where the animal lives. However, the same genotype could have different phenotype according to the environment, then it is necessary to include the interaction between the genotype and the environment (GE) in the basic model: $P = G + E + GE$.

Selecting animals for quantitative traits is traditionally based on the phenotype of the individual and its relatives (MEUWISSEN; HAYES; GODDARD, 2001), estimating breeding values (EBV) by Best Linear Unbiased Prediction (BLUP) (HENDERSON, 1984). However, the inclusion of DNA information have led for more accurate EBV and detection of Quantitative Trait Locus (QTL) when genetic markers has been included into BLUP (FERNANDO; GROSSMAN, 1989; GEORGES et al., 1995; MEUWISSEN; GODDARD, 1996).

In the past decades animal breeders has focused on high-intensified selection to improve genetically commercial traits in pigs such as backfat thickness, lean meat content, growth, and feed conversion efficiency (BOSI; RUSSO, 2016). Among the most important traits, backfat thickness affects indirectly carcass lean percentage, fat deposition and meat quality (FONTANESI et al., 2012). Fatness tissue is the main storage of energy in the organism and it is important for hormones production, which acts in the physiological and physiopathology processes (FONSECA-ALANIZ et al., 2006). Fatness deposition traits are important in animal breeding and, knowing the molecular mechanism for this trait is essential to lead breeding programs in order to improve it.

Due to the high selection based on production traits, few attention is given to secondary ones such as health, metabolism and locomotion traits. Cows with high milk yield production has negative correlation with reproductive performance and production diseases (INGVARTSEN; DEWHURST; FRIGGENS, 2003; NEBEL; MCGLLLIARD, 1993), chicken with selection for muscular growth has low tolerance to high temperatures (CEDRAZ et al., 2017; CLASSEN, 2000; LAGANÁ, 2005; WANG et al., 2015) and poor development of the skeletal structure (GONZALES; MACARI; PAZ, 2009; JULIAN, 1998) such as Bacterial Chondronecrosis with Osteomyelitis (BCO).

BCO has been associated with the fast growing of the growth plates on femur and proximal tibia (WIDEMAN, 2016) where the osteochondrosis is more susceptible and it has been considered as the main cause of lameness in broilers, causing great economic loss (COOK, 2000; MCNAMEE; SMYTH, 2000; WIDEMAN et al., 2012). The chicken growth plates are thicker than those from mammals, for example and chondrocytes' columns are aligned in a irregular way due to the higher growth associated to the turnover of the chickens' growth plate (WIDEMAN; PRISBY, 2013).

Some studies have shown that the BCO pathogenesis has been initiated by the poor mineralization of chondrocytes followed by proliferation of oportunistic bacteria such as *Staphylococcus aureus*, *Escherichia coli*, *Staphylococcus* negative-coagulase and *Enterococcus spp* (MCNAMEE; SMYTH, 2000; WIDEMAN; PRISBY, 2013; WIJESURENDRA et al., 2017). Our group has shown that the main cause of BCO seems to be the downregulation of some osteochondral genes which was supposed to be regulated in normal bones (DE OLIVEIRA et al., 2020 – chapter 2).

For these reasons, breeding companies are facing a pressure for selecting animals disease resistance and better welfare conditions (BISHOP, 2014). Furthermore, the development of genomic tools allowed to use genetic markers such as Single Nucleotide Polymorphism (SNP) to search for QTL regions and haplotype blocks which could explain other effects in the interest traits and lead the selection to get more healthy animals.

Linkage disequilibrium and haplotype blocks

Linkage Disequilibrium (LD) is the non-random allele association between different loci. It occurs by the proximity among alleles and can be influenced by the evolutive process and the history of the population (PÉREZ O'BRIENet al., 2014). Its extension depends on the recombination rates, when low recombination rates regions such as chromosome Y and part of the chromosome X have higher LD than regions where the recombination rates are high such as hotspots (JEFFREYS; KAUPPI; NEUMANN, 2001). Moreover, the LD

extension has great importance on the prediction of the breeding values in Genomic Selection (GS), Genome-Wide Association Study (GWAS), Marker-Assisted Selection (MAS) and Fine-Mapping (PÉREZ O'BRIEN et al., 2014).

Thus, a previous analysis of the LD is important for determining the number of markers which will be necessary to have good results in GWAS (KHATKAR et al., 2008; MCKAY et al., 2007). LD estimation tends to decrease due to low density of markers, in addition, differences between genomic regions can have influence in the LD extension (WALL; PRITCHARD, 2003). However, studies have shown high LD in haplotype blocks separated by recombination hotspots (ARDLIE; KRUGLYAK; SEIELSTAD, 2002; DALY et al., 2001).

Haplotypes Block is defined as the recombination of inherited regions, without recombination in the previous generation with a minimum number of genetic markers (GABRIEL et al., 2002; PATIL et al., 2001). SNP are the most used markers in Animal Breeding. SNPs with high LD segregate together in the haplotype blocks but can be separated along the generations by recombination (DALY et al., 2001). These markers have great importance on the GWAS allowing select a set of SNPs in a block of SNPs with high LD in a Haplotype Block (JOHNSON et al., 2001), carrying a better comprehension about the genome (WANG et al., 2002).

Next Generation Sequence (NGS)

Next Generation Sequence (NGS) has brought progress in the comprehension of complex traits in Animal Breeding. Since the whole genome has been sequenced it is an excellent method to understand how causal/structure variants have influence on the phenotype (CHARLIER et al., 2016).

NGS has become more popular and has been used in several research field (ARUMUGAM; ULI; ANNAVI, 2019). Applying the high throughput technology, it allows to generate millions of sequences which can be used in studies such as *de novo* genome sequence, transcriptome analysis, and epigenomic sequence such as ChIP-Sequence.

Since the domestic animals genome such as pig and chicken were sequenced, many studies have been used NGS for a better understanding of the quantitative traits in the Animal Breeding (CRESPO-PIAZUELO et al., 2019; LI et al., 2020), moreover, it has brought new methods such as marker imputation which has been implemented in Animal Breeding to improve the accuracy in GWAS and GS (HABIER; FERNANDO; DEKKERS, 2009; MARCHINI et al., 2007; MARCHINI; HOWIE, 2010), using reference panel genotyped at high density or the whole genome sequence to impute markers not present in the low-density marker panel without loss in accuracy and increasing the density of markers (DASSONNEVILLE et al., 2011;

LARMER et al., 2017; MARCHINI; HOWIE, 2010; MULDER et al., 2012; WEIGEL et al., 2010).

With a higher density of maker, the detection of QTL has become more accurate allowing to identify SNPs located in expressed regions. Furthermore, it allows the comprehension of molecular mechanism and annotation of candidate genes related with quantitative traits (MINOZZI et al., 2013). However, the comprehension of a trait is not limited to the phenotype, whereas a small difference on the mRNA amount could change the gene expression and alter the traits formation (SCHADT et al., 2003). For quantifying the amount of mRNA is necessary to use transcriptome data, which is the collection of all mRNA and noncoding RNA (ncRNA) transcripts produced by a cell in a specific tissue, development stage or physiologic condition. The RNA sequence (RNA-Seq) allows to identify and quantify transcripts (CLOONAN et al., 2008; MORTAZAVI et al., 2008), allowing the description of a set of genes (OZSOLAK; MILOS, 2011). In this way it is possible to find out which gene has expressed due an experimental condition as well as the tissue studied.

Transcriptome (RNA-Seq) analysis starts from a messenger RNA (mRNA) sample converted into Complementary DNA (cDNA) generating a library which is sequenced into millions of DNA sequences (VAN VERK et al., 2013). Following, it is necessary a set of bioinformatics software to analyze the RNA-Seq data in four steps: Quality Control (QC), Alignment, Quantification and, Differential Expression (DE) (VAN VERK et al., 2013).

The first step, quality control, is performed to remove short reads, barcodes, adapters, and bases with low quality (LOWE et al., 2017). There are several tools for performing quality control, however, the most used is Trimmomatic (BOLGER; LOHSE; USADEL, 2014) which corrects abnormalities occurred during the sequencing. The second step, alignment, aims to combine the gene sequences with the reference genome or *de novo* assembly when the reference genome is not available (LOWE et al., 2017).

After the alignment it is necessary to quantify the gene expression (VAN VERK et al., 2013). The quantification algorithms seeks for genes which have known splices and those potentially novels (WANG; GERSTEIN; SNYDER, 2009). There is a complexity in estimating the expression of genes due to the set of exon and, therefore, just a fraction of reads aligned on the specific regions (VAN VERK et al., 2013) demanding probabilistic methods to estimate the transcript abundancies from the information of the short reads (TRAPNELL et al., 2011). Once the gene expression is quantified, the differential expression analysis aim to identify the differential gene expressed between the experimental conditions (VAN VERK et al., 2013; WANG; GERSTEIN; SNYDER, 2009).

ChIP-Sequencing (ChIP-Seq) has also been used in animal research improving genome annotation and searching regulatory elements into the genome regions (BEIKI et al., 2019; HAN

et al., 2019). Combining chromatin immunoprecipitation with high throughput sequence it allows the identification of transcription factors structure proteins, chromatin associated proteins and specific DNA sequences which bind into the proteins which may cause influence on the phenotype of the interest traits (DE SÁ et al., 2018).

Identifying these proteins and Transcription Factor (TF) allow a better understanding of the biological process and the regulation of the gene expression (METZKER, 2010; WORSLEY HUNT et al., 2014). Moreover, it also allows the understanding of the interaction of TF's binding sites, promoters of target genes and, epigenetics mechanism which guide the cellular and physiological differentiation (KANTERAKIS; POTAMIAS; PATRINOS, 2018; SHIN et al., 2013).

GWAS and Fine-Mapping

The used SNP markers and genotyping has allowed to detect QTL and candidate genes with more accuracy. Moreover, NGS has become cheaper, which makes it more popular in the last decade increasing the number of studies using GWAS to find associations between phenotypes and genotypes of the interest traits in animal breeding (ALILOO et al., 2015; PRYCE et al., 2010).

For a better understanding of the quantitative traits in a molecular level it is necessary to increase the number of SNPs to increase the marker density in the Quantitative Trait Loci (QTL). After the release of Illumina PorcineSNP60 Chip (RAMOS et al., 2009), important traits such as fatness related have been studied using GWAS (DO et al., 2014; FONTANESI et al., 2012; FOWLER et al., 2013; OKUMURA et al., 2013), allowing to map the genome regions that affect these traits.

Although GWAS can identify associated markers, the current Illumina PorcineSNP60 BeadChip can lead to clusters of markers which may cover regions that are still too large, making difficult the identification of genes and variants. Therefore, there are still necessity of reduce the regions to find causal relation between genes or variants which affect economic traits.

Moreover, there is a tendency to consider the gene nearest of the leadSNP (the smallest P-value) to be the causal, however, the physical distance between a variant and a gene is not enough evidence for causality (SPAIN; BARRETT, 2015). Additionally, the majority of associated variants do not change protein sequence. In this reason, several researches have focused on the improvement of functional annotation to identify variants which affect the expression level of a gene and change the phenotype.

Fine-mapping uses NGS data such as whole genome sequence, transcriptome, ChIP-Seq, etc.,

for seeking genetic variants which are liable for the expressed phenotype starting from at least one high significant SNP from previous GWAS results to more specific results finding one or more variants for identifying the potential causal variant (SCHAID; CHEN; LARSON, 2018; SPAIN; BARRETT, 2015).

Different approaches can be used to fine-map genome regions to identify important variants and genes associated to diseases and traits (SCHAID; CHEN; LARSON, 2018; SPAIN; BARRETT, 2015; TAK; FARNHAM, 2015; WEISSENKAMPEN et al., 2020). Fine-mapping from a variant approach is the process of narrow down a genomic region, assigning one or more variants involved in a trait. For that, an understanding of the mechanism by which variant contribute for a phenotype is required for fine-mapping variants from a GWAS loci (BROEKEMA; BAKKER; JONKERS, 2020). Moreover, it is important to identify the most significant variants and calculate the LD to understand and identify which genes and pathways are affecting the traits.

GWAS variants which overlap in high LD with regulatory elements is the most straightforward approach. CEDAR; BERGMAN, (2009) have shown that interactions between DNA methylation and histone modifications can have importance regulating chromatin and biological process. The use of ChIP-seq can measure the histone marks modification to determine the location of the functional elements (BERNSTEIN et al., 2010; KUNDAJE et al., 2015). Histone markers such as H3K27me3 and H3K4me3 have been widely used to fine-map genome regions and are associated with gene repression and activation, respectively (RINGROSE; PARO, 2004; SCHUETTENGRUBER et al., 2007). Therefore, these marks are associated in the regulation of developmental genes during development stage in mammals (BOGLIOTTI; ROSS, 2012; LI, 2002), indicating that there are histone modifications during the embryo development (ERHARDT et al., 2003; VAN DER HEIJDEN et al., 2005; WANG; DEY, 2006; ZHANG et al., 2009).

Studies have used fine-mapping in different approaches in Animal Breeding (DERKS et al., 2017, 2019; DRAG et al., 2019; JIANG et al., 2019; LIU et al., 2019; VAN SON et al., 2019). Although ChIP-Seq has been used for fine-mapping the human genome (ÇALIŞKAN et al., 2019; KUMASAKA; KNIGHTS; GAFFNEY, 2016), there is still a lack of studies using this approach in animal breeding.

References

ALILOO, H. et al. Validation of markers with non-additive effects on milk yield and fertility in Holstein and Jersey cows. **BMC Genetics**, v. 16, n. 1, p. 89, 22 dez. 2015.

ARDLIE, K. G.; KRUGLYAK, L.; SEIELSTAD, M. Patterns of linkage disequilibrium in the

human genome. **Nature Reviews Genetics**, v. 3, n. 4, p. 299–309, 2002.

ARUMUGAM, R.; ULI, J. E.; ANNAVI, G. A Review of the Application of Next Generation Sequencing (NGS) in Wild Terrestrial Vertebrate Research. **Annual Research & Review in Biology**, v. 31, n. 5, p. 1–9, 2019.

BEIKI, H. et al. Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. **BMC Genomics**, v. 20, n. 1, p. 344, 7 dez. 2019.

BERNSTEIN, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium complex. **Nature Publishing Group**, v. 28, n. 10, p. 1045–1048, 2010.

BISHOP, S. **Disease Genetics: From Concepts to Utilisation**. Short Course at Iowa State University. **Anais**…Ames, IA, USA: 2014

BOGLIOTTI, Y. S.; ROSS, P. J. Mechanisms of histone H3 lysine 27 trimethylation remodeling during early mammalian development. **Epigenetics**, v. 7, n. 9, p. 976–981, 7 set. 2012.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 2014.

BOSI, P.; RUSSO, V. The production of the heavy pig for high quality processed products The production of the heavy pig for high quality processed products. 2016.

BROEKEMA, R. V; BAKKER, O. B.; JONKERS, I. H. A practical view of fine-mapping and gene prioritization in the post-genome- wide association era. 2020.

ÇALIŞKAN, M. et al. Genetic and Epigenetic Fine Mapping of Complex Trait Associated Loci in the Human Liver. **The American Journal of Human Genetics**, v. 105, n. 1, p. 89–107, jul. 2019.

CEDAR, H.; BERGMAN, Y. Linking DNA methylation and histone modification: patterns and paradigms. v. 10, n. mAy, p. 295–304, 2009.

CEDRAZ, H. et al. Heat stress induces expression of HSP genes in genetically divergent chickens. **PLoS ONE**, v. 12, n. 10, 2017.

CHARLIER, C. et al. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. **Genome Research**, v. 26, n. 10, p. 1333–1341, out. 2016.

CLASSEN, H. L. Managing metabolic disease in rapidly growing strains of poultry. In: HILL, W. G. et al. (Eds.). **The challenge of genetic change in animal production**. Publicatio ed. Edinburgh: Journal of the British Society of Animal Science, 2000. p. 63–64.

CLOONAN, N. et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. **Nature Methods**, v. 5, n. 7, p. 613–619, 2008.

COOK, M. E. Skeletal deformities and their causes: introduction. **Poultry science**, v. 79, n. 7, p. 982–4, 2000.

CRESPO-PIAZUELO, D. et al. Indel detection from Whole Genome Sequencing data and association with lipid metabolism in pigs. **PLOS ONE**, v. 14, n. 6, p. e0218862, 27 jun. 2019.

DALY, M. J. et al. High-resolution haplotype structure in the human genome. **Nature Genetics**, v. 29, n. 2, p. 229–232, 2001.

DASSONNEVILLE, R. et al. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. **Journal of Dairy Science**, v. 94, n. 7, p. 3679–3686, jul. 2011.

DE OLIVEIRA, H. C. et al. RNA-seq reveals downregulated osteochondral genes potentially related to tibia bacterial chondronecrosis with osteomyelitis in broilers. **BMC Genetics**, v. 21, n. 1, p. 58, 2020.

DE SÁ, P. H. C. G. et al. Next-generation sequencing and data analysis: Strategies, tools, pipelines and protocols. **Omics Technologies and Bio-engineering: Towards Improving Quality of Life**, v. 1, p. 191–207, 2018.

DERKS, M. F. L. et al. A systematic survey to identify lethal recessive variation in highly managed pig populations. **BMC Genomics**, v. 18, n. 1, 2017.

DERKS, M. F. L. et al. Loss of function mutations in essential genes cause embryonic lethality in pigs. **PLoS Genetics**, v. 15, n. 3, p. 1–22, 2019.

DO, D. N. et al. Genome-wide association and systems genetic analyses of residual feed intake, daily feed consumption, backfat and weight gain in pigs. **BMC Genetics**, v. 15, n. 1, p. 1–15, 2014.

DRAG, M. H. et al. Characterization of eqtls associated with androstenone by rna sequencing in porcine testis. **Physiological Genomics**, v. 51, n. 10, p. 488–499, 2019.

ERHARDT, S. et al. Consequences of the depletion of zygotic and embryonic enhancer of zeste 2 during preimplantation mouse development. **Development**, v. 130, n. 18, p. 4235–4248, 2003.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to quantitative genetics**. Harlow: Longman group Ltda, 1996. FERNANDO, R.; GROSSMAN, M. Marker assisted selection using best linear unbiased prediction. **Genetics Selection Evolution**, v. 21, n. 4, p. 467, 1989.

FONSECA-ALANIZ, M. H. et al. O tecido adiposo como centro regulador do metabolismo. **Arquivos Brasileiros de Endocrinologia & Metabologia**, v. 50, n. 2, p. 216–229, abr. 2006.

FONTANESI, L. et al. A genome wide association study for backfat thickness in Italian Large White pigs highlights new regions affecting fat deposition including neuronal genes. **BMC Genomics**, v. 13, n. 1, 2012.

FOWLER, K. E. et al. Genome wide analysis reveals single nucleotide polymorphisms associated with fatness and putative novel copy number variants in three pig breeds. **BMC Genomics**, v. 14, n. 1, p. 1, 2013.

GABRIEL, S. B. et al. The structure of haplotype blocks in the human genome. **Science**, v. 296, n. 5576, p. 2225–2229, 2002.

GEORGES, M. et al. Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. **Genetics**, v. 139, n. 2, p. 907–920, 1995.

GONZALES, E.; MACARI, M.; PAZ, I. C. DE L. A. Enfermidades metabólicas em frangos de corte. In: SILVA, E. N. et al. (Eds.). . **Doenças das aves**. 2. ed. Campinas: FACTA, 2009. p. 972–996.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. Genomic Selection Using Low-Density Marker Panels. **Genetics**, v. 182, n. 1, p. 343–353, maio 2009.

HAN, K. et al. Genome-Wide Identification of Histone Modifications Involved in Placental Development in Pigs. **Frontiers in Genetics**, v. 10, 29 mar. 2019.

HENDERSON, C. R. **Applications of Linear Models in Animal Breeding**. Guelph, Ontario, Canada: University of Guelph, 1984.

INGVARTSEN, K. L.; DEWHURST, R. J.; FRIGGENS, N. C. O n the relationship between lactational performance and health : is it yield or metabolic imbalance that cause production diseases in dairy cattle ? A position paper q. v. 83, n. August 2000, p. 277–308, 2003.

JEFFREYS, A. J.; KAUPPI, L.; NEUMANN, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. **Nature Genetics**, v. 29, n. 2, p. 217–222, 2001.

JIANG, J. et al. Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. **Communications Biology**, v. 2, n. 1, p. 212, 18 dez. 2019.

JOHNSON, G. C. L. et al. Haplotype tagging for the identification of common disease genes. **Nature Genetics**, v. 29, n. 2, p. 233–237, 2001.

JULIAN, R. J. Rapid Growth Problems: Ascites and Skeletal Deformities in Broilers. **Poultry Science**, v. 77, n. 12, p. 1773–1780, 1998.

KANTERAKIS, A.; POTAMIAS, G.; PATRINOS, G. P. An Introduction to Tools, Databases, and Practical Guidelines for NGS Data Analysis. In: **Human Genome Informatics**. [s.l.] Elsevier, 2018. p. 61–89.

KHATKAR, M. S. et al. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. **BMC Genomics**, v. 9, n. 1, p. 187, 2008.

KUMASAKA, N.; KNIGHTS, A. J.; GAFFNEY, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. **Nature Genetics**, v. 48, n. 2, p. 206–213, 14 fev. 2016.

KUNDAJE, A. et al. Integrative analysis of 111 reference human epigenomes. **Nature**, v. 518, p. 317–330, 2015.

LAGANÁ, C. Otimização da produção de frango de corte em condições de estresse por calor. 2005.

LARMER, S. G. et al. Novel methods for genotype imputation to whole-genome sequence and a simple linear model to predict imputation accuracy. **BMC Genetics**, v. 18, n. 1, p. 120, 27 dez. 2017.

LI, E. Chromatin modification and epigenetic reprogramming in mammalian development. **Nature Reviews Genetics**, v. 3, n. 9, p. 662–673, set. 2002.

LI, X. et al. Whole-genome sequencing identifies potential candidate genes for reproductive traits in pigs. **Genomics**, v. 112, n. 1, p. 199–206, jan. 2020.

LIU, Z. et al. Fine-mapping sequence mutations with a major effect on oligosaccharide content in bovine milk. **Scientific Reports**, v. 9, n. 1, p. 2137, 14 dez. 2019.

LOWE, R. et al. Transcriptomics technologies. **PLoS Computational Biology**, v. 13, n. 5, p. 1–23, 2017.

MARCHINI, J. et al. A new multipoint method for genome-wide association studies by imputation of genotypes. **Nature Genetics**, v. 39, n. 7, p. 906–913, 17 jul. 2007.

MARCHINI, J.; HOWIE, B. Genotype imputation for genome-wide association studies. **Nature Reviews Genetics**, v. 11, n. 7, p. 499–511, 2 jul. 2010.

MCKAY, S. D. et al. Whole genome linkage disequilibrium maps in cattle. **BMC Genetics**, v. 8, p. 1–12, 2007.

MCNAMEE, P. T.; SMYTH, J. A. Bacterial chondronecrosis with osteomyelitis (‘femoral head necrosis’) of broiler chickens: A review. **Avian Pathology**, v. 29, n. 4, p. 253–270,

2000.

METZKER, M. L. Sequencing technologies — the next generation. **Nature Reviews Genetics**, v. 11, n. 1, p. 31–46, 8 jan. 2010.

MEUWISSEN, T. H. E.; GODDARD, M. E. The use of marker haplotypes in animal breeding schemes. **Genetics Selection Evolution**, v. 28, n. 2, p. 161–176, 1996.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819–1829, 2001.

MINOZZI, G. et al. Genome Wide Analysis of Fertility and Production Traits in Italian Holstein Cattle. **PLoS ONE**, v. 8, n. 11, p. e80219, 12 nov. 2013.

MORTAZAVI, A. et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. **Nature Methods**, v. 5, n. 7, p. 621–628, 2008.

MULDER, H. A. et al. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. **Journal of Dairy Science**, v. 95, n. 2, p. 876–889, fev. 2012.

NEBEL, R. L.; MCGLLLIARD, M. L. Interactions of High Milk Yield and Reproductive Performance In Dairy Cows. **Journal of Dairy Science**, v. 76, n. 10, p. 3257–3268, 1993.

OKUMURA, N. et al. Genomic regions affecting backfat thickness and cannon bone circumference identified by genome-wide association study in a Duroc pig population. **Animal Genetics**, v. 44, n. 4, p. 454–457, ago. 2013.

OLTENACU, P. A.; BROOM, D. M. The impact of genetic selection for increased milk yield on the welfare of dairy cows * Department of Animal Science , Oklahoma State University , Stillwater , OK + Centre for Animal Welfare and Anthrozoology , Department of Veterinar. **Animal Welfare**, v. 19, n. S, p. 39–49, 2010.

OZSOLAK, F.; MILOS, P. M. RNA sequencing: advances, challenges and opportunities. **Nature Reviews Genetics**, v. 12, n. 2, p. 87–98, fev. 2011.

PATIL, N. et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. **Science**, v. 294, n. 5547, p. 1719–1723, 2001.

PÉREZ O'BRIEN, A. M. et al. Linkage disequilibrium levels in Bos indicus and Bos taurus cattle using medium and high density SNP chip data and different minor allele frequency distributions. **Livestock Science**, v. 166, n. 1, p. 121–132, ago. 2014.

PRYCE, J. E. et al. A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. **Journal of Dairy Science**, v.

93, n. 7, p. 3331–3345, jul. 2010.

RAMOS, A. M. et al. Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. **PLoS ONE**, v. 4, n. 8, p. e6524, 5 ago. 2009.

RINGROSE, L.; PARO, R. Epigenetic Regulation of Cellular Memory by the Polycomb and Trithorax Group Proteins. **Annual Review of Genetics**, v. 38, n. 1, p. 413–443, dez. 2004.

SCHADT, E. E. et al. Genetics of gene expression surveyed in maize, mouse and man. **Nature**, v. 422, n. 6929, p. 297–302, 2003.

SCHAID, D. J.; CHEN, W.; LARSON, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. **Nature Reviews Genetics**, v. 19, n. 8, p. 491–504, 29 ago. 2018.

SCHUETTENGRUBER, B. et al. Genome Regulation by Polycomb and Trithorax Proteins. **Cell**, v. 128, n. 4, p. 735–745, fev. 2007.

SHIN, H. et al. Computational methodology for ChIP-seq analysis. **Quantitative Biology**, v. 1, n. 1, p. 54–70, 2 mar. 2013.

SPAIN, S. L.; BARRETT, J. C. Strategies for fine-mapping complex traits. **Human Molecular Genetics**, v. 24, n. R1, p. R111–R119, 15 out. 2015.

TAK, Y. G.; FARNHAM, P. J. Making sense of GWAS : using epigenomics and genome engineering to understand the functional relevance of SNPs in non □ coding regions of the human genome. **Epigenetics & Chromatin**, p. 1–18, 2015.

TRAPNELL, C. et al. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. **Nature Biotechnology**, v. 28, n. 5, p. 511–515, 2011.

VAN DER HEIJDEN, G. W. et al. Asymmetry in Histone H3 variants and lysine methylation between paternal and maternal chromatin of the early mouse zygote. **Mechanisms of Development**, v. 122, n. 9, p. 1008–1022, 2005.

VAN SON, M. et al. A QTL for number of teats shows breed specific effects on number of vertebrae in pigs: Bridging the gap between molecular and quantitative genetics. **Frontiers in Genetics**, v. 10, n. MAR, 2019.

VAN VERK, M. C. et al. RNA-Seq: Revelation of the messengers. **Trends in Plant Science**, v. 18, n. 4, p. 175–179, 2013.

WALL, J. D.; PRITCHARD, J. K. Haplotype blocks and linkage disequilibrium in the human

genome. **Nature Reviews Genetics**, v. 4, n. 8, p. 587–597, 2003.

WANG, H.; DEY, S. K. Roadmap to embryo implantation: Clues from mouse models. **Nature Reviews Genetics**, v. 7, n. 3, p. 185–199, 2006.

WANG, N. et al. Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. **American Journal of Human Genetics**, v. 71, n. 5, p. 1227–1234, 2002.

WANG, S. H. et al. Acute heat stress induces differential gene expressions in the testes of a broiler-type strain of Taiwan country chickens. **PLoS ONE**, v. 10, n. 5, p. 1–14, 2015.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, v. 10, n. 1, p. 57–63, 16 jan. 2009.

WEIGEL, K. A. et al. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. **Journal of Dairy Science**, v. 93, n. 11, p. 5423–5435, nov. 2010.

WEISSENKAMPEN, J. D. et al. Methods for the analysis and interpretation for rare variants associated with complex traits. **Curr Protoc Hum Genet.**, v. 101, n. 1, p. 1–20, 2020.

WIDEMAN, R. F. et al. A wire-flooring model for inducing lameness in broilers: Evaluation of probiotics as a prophylactic treatment. **Poultry Science**, v. 91, n. 4, p. 870–883, 1 abr. 2012.

WIDEMAN, R. F. Bacterial chondronecrosis with osteomyelitis and lameness in broilers: A review. **Poultry Science**, v. 95, n. 2, p. 325–344, 2016.

WIDEMAN, R. F.; PRISBY, R. D. Bone circulatory disturbances in the development of spontaneous bacterial chondronecrosis with osteomyelitis: a translational model for the pathogenesis of femoral head necrosis. **Frontiers in Endocrinology**, v. 3, n. JAN, p. 1–15, 2013.

WIJESURENDRA, D. S. et al. Pathological and microbiological investigations into cases of bacterial chondronecrosis and osteomyelitis in broiler poultry. **Avian Pathology**, v. 46, n. 6, p. 683–694, 2017.

WORSLEY HUNT, R. et al. Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. **BMC Genomics**, v. 15, n. 1, p. 472, 2014.

ZHANG, M. et al. Defective chromatin structure in somatic cell cloned mouse embryos. **Journal of Biological Chemistry**, v. 284, n. 37, p. 24981–24987, 2009.

CHAPTER 1

BAQCOM - Bioinformatics Analysis for Quality Control and Mapping

Haniel Cedraz de Oliveira[1]; Mauricio Egidio Cantão[2];
Simone Eliza Facioni Guimarães[1]

[1]Federal University of Viçosa, Viçosa, Minas Gerais;
[2]Embrapa Suínos e Aves, Concórdia, Santa Catarina, Brazil

Abstract

Bioinformatic analysis has become more demanded and new software has been developed for quality control, mapping and quantify transcripts, however, dealing with these software could be hard for people who do not have experience with computational environment. For this reason, we aimed to develop a user-friendly pipeline for RNA-Seq analysis for quality control, mapping and counting reads. Our pipeline uses a set of software such as Trimmomatic, STAR, HISAT2, HTSeq-count and featureCounts. All pipelines run automatically with a few basic options and the output of each script is prepared for the input to the next step. BAQCOM pipelines have shown to be efficient and fast. Although the pipeline has being designed for RNA-Seq analysis, there are the flexibility to integrate other genomic tools in the future, such as fine-mapping analysis.

keywords: Pipeline; RNA-Seq; Sequencing; STAR; Trimmomatic

Introduction

With the advanceof the Next Generation Sequencing (NGS), bioinformatics analysis has become possible and more demanded. Moreover, the development of new software for high-throughput sequencing analysis has provided different methods for quality control, mapping and quantifying transcripts. RNA-Seq has become popular and has revolutionized the manner in which the transcriptome is analyzed (WANG; GERSTEIN; SNYDER, 2009). Transcriptome is the set of transcripts in a specific cell or tissue, development stage or physiological condition and it is essential for understanding the molecular level of a trait or a disease.

Using transcriptome, it is possible to list all transcripts, including non-coding RNA and small RNAs, determining the structure of genes, splicing patterns and quantify the expression level of each gene under different conditions and discover novel transcripts. It is also important for improving the genome annotation. Many studies have used RNA-Seq to improve genome annotation and better understand the molecular level of quantitative traits in animal breeding (BEIKI et al., 2019; KIM et al., 2020; MARTÍNEZ-MONTES et al., 2017).

However, applying NGS analysis requires a certain level of knowledge of programing and a large computer structure for handing large datasets and specific algorithms. Running an NGS analysis it is necessary to find the best tools, install and configure them, manage libraries, write scripts to convert outputs from one tool into input to the next tool. It could be a challenge for those who attempt NGS analysis, specially without any computational experience (KESH; RAGHUPATHI, 2004). For these reasons it is important the use of automated scripts which can guide the analysis.

The aim of bioinformatic pipelines is to provide a user-friendly interface that allows anyone to run NGS analysis. Bioinformatics Analysis for Quality Control and Mapping (BAQCOM) is an user-friendly pipeline for RNA-Seq analysis from quality control to counting reads steps, running tools such as Trimmomatic (BOLGER; LOHSE; USADEL, 2014), STAR (DOBIN et al., 2013), HISAT2 (KIM et al., 2019), HTSeq-count (ANDERS; PYL; HUBER, 2015) and featureCounts (LIAO; SMYTH; SHI, 2014).

Material and Methods

Starting from raw fastq reads, the pipelines were developed for analyzing RNA-Seq Illumina data. The pipeline uses open sourced available tools to perform quality control, mapping and read counts. Two paired-end RNA-Seq samples of hernia from pig's were analyzed using the methodologies in BAQCOM. We edited the samples decreasing the number of reads to 100000 to facilitate the analysis. For the reference genome and gene annotation we have used

the chromosome 1 of *Sus scrofa* reference genome version 11.1 and ensembl release 98.

BAQCOM is a set of command line pipelines for Unix-based systems calling functions and software to perform a specific data process or analysis. Each pipeline is a R script which calls one or more installed software, takes a variety of parameters through command line flags and generates specific outputs which will be used as input into the next analyses. BAQCOM is hosted on GitHub https://github.com/hanielcedraz/BAQCOM) and can be easily installed locally or/and on remote clusters by a shell script which download the necessary software and install them.

Quality Control

For quality Control (QC) BAQCOM uses baqcomTrimmomatic pipeline. This Rscript receive raw fastq (RNA-Seq or DNA-Seq) files and calls trimmomatic (BOLGER; LOHSE; USADEL, 2014) software using default parameters, however, it can be changed according to user's necessity. For QC statistics the pipeline runs FastQC (ANDREWS, 2010) and/or MultiQC (EWELS et al., 2016) software. BaqcomTrimmomatic pipeline reates cleaned fastq files as output and the statistics from all samples are joined into one single file.

Mapping

For the mapping step, two pipelines are available in BAQCOM. Both Rscripts receive cleaned fastq files (outputs from baqcomTrimmomatic) as input. BaqcomHisat2 uses Hisat2 (KIM; LANGMEAD; SALZBERG, 2015) software for mapping, while baqcomSTAR uses STAR (DOBIN et al., 2013) software for mapping and/or counting reads. Moreover, baqcomHisat2 uses samtools (LI; DURBIN, 2009) to convert sam into bam files as output. Both pipelines use default parameters of the software, however, according to the user's necessity it can be changed including more options. Moreover, either baqcomSTAR and baqcomHisat2 get all mapping statistics and generate it from all samples into one single file.

Counting Reads

For counting reads step BAQCOM also provides two pipelines, both receives files from baqcomSTAR and/or baqcomHisat2. In baqcomHtseq is implemented HTSeq-count (ANDERS; PYL; HUBER, 2015) software while baqcomFeatureCounts uses featureCounts (LIAO;

SMYTH; SHI, 2014) software. The outputs from these pipelines are already done for input into a differential expression gene script, such as edgeR or DEseq.

All pipelines, except baqcomHtseq run multithreads, parallelizing process and running as many samples as possible at the same time. Moreover all pipelines run MultiQC (EWELS et al., 2016) software for statistical parameters.

Results and Discussion

BAQCOM pipeline is an automated approach for analyzing RNA-Seq data. We do not expect to have different results running BAQCOM or the programs themselves, however, it is much more user-friendly and faster, since it is possible to analyze several samples at same time and short command line (Additional file 1).

We have run analysis using edited datasets with 100 thousand of reads against chromosome 1 of the pig's reference genome and genome annotation. After the QC, the mapping and reads counting can be running in one or two steps (Mapping and Counting reads using baqcom-STAR or Mapping using baqcomHisat2 and Counting read using baqcomFeatureCounts or baqcomHtseq) (**Fig. 1**).

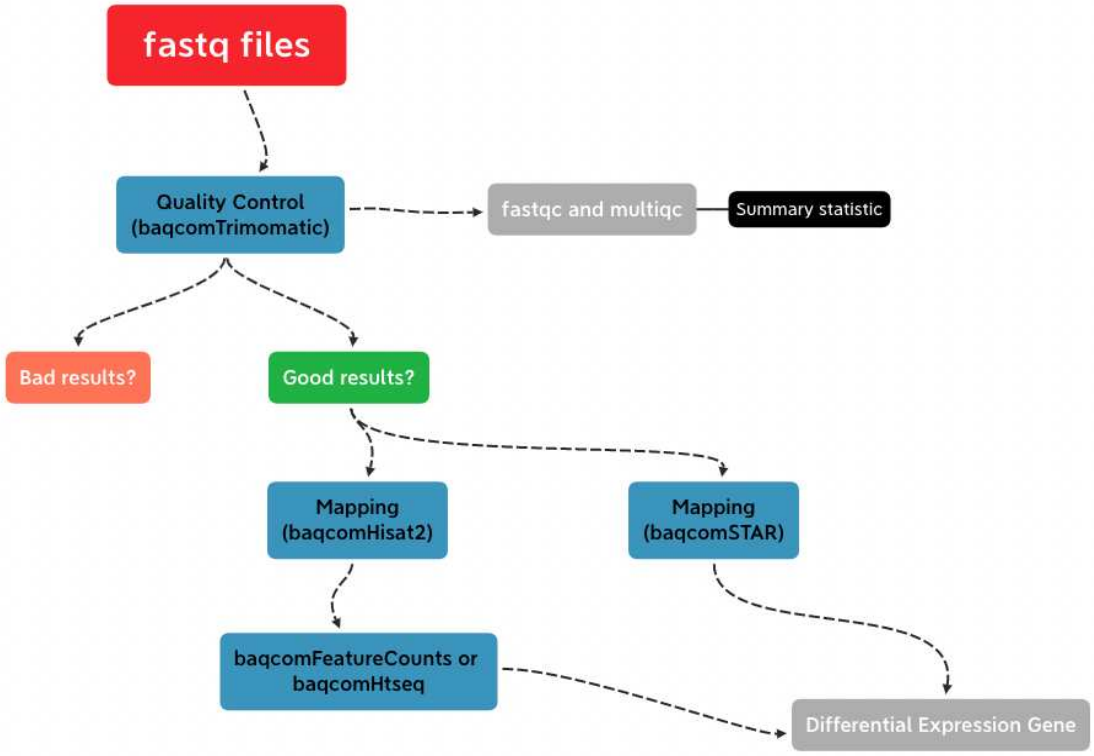

Figure 1: Scheme showing the step by step of BAQCOM pipeline

Quality Control

BaqcomTrimmomatic is used for quality control. It has generated a folder named 01-CleanedReads/ contained four files (two Paired-End (PE) and two Single-End (SE) for each sample (*PE1.fastq, *PE2.fastq, *SE1.fastq and *SE2.fastq) and a folder named 02-Reports which is stored results from Multiqc, fastqc and the statistical summary from the quality control (**Table 1**).

Table 1: Summary statistics from baqcomTrimmmomatic showing number of input and output reads, percentage of output reads and other parameters

Sample	Input Read Pairs	Pairs Reads	Pairs Read Percent	Forward Only Surviving Reads	Forward Only Surviving Read Percent	Reverse Only Surviving Reads	Reverse Only Surviving Read Percent	Dropped Reads	Dropped Read Percent
HE20-100K	1e+05	92194	92.19	7283	7.28	357	0.36	166	0.17
HE21-100K	1e+05	92147	92.15	7348	7.35	365	0.36	140	0.14

On the fastqc analysis it is created two folders containing statistical before (FastQCBefore/) and statistical after (FastQCAfter/) quality control. In these folders it is possible to find a *.html and *.zip files for each sample in 00-Fastq/ and 01-CleanedReads/, respectively. On the Multiqc analysis it is created a folder named multiqc_data/ containing a statistical summary and a html file (multiqc_report.html) showing a graphical statistical summary for the previous analysis.

Running quality control using Trimmomatic software it is not possible to run more than one sample at same time, using the basic command line, which can increase the time for analyzing the data. Moreover, Trimmomatic does not create a statistical summary joining all samples, it creates one statistical file for each sample, generating more files.

Mapping

For aligning the reads against the reference genome (mapping) it could be done using baqcomSTAR or baqcomHisat2. By default, baqcomSTAR does mapping and counting reads. Using default parameters, it is created three folders containing the mapped reads (02-MappedReadsSTAR/), unmapped reads (03-UnmappedReadsSTAR/) and the counted reads (04-GeneCountsSTAR/).

Report folder is renamed from 02-Reports to 05-Reports for a better organization of the folders. Running baqcomHisat2, the mapping will be done by HISAT2 software, which does not perform counting reads, for this reason it is generate just one folder named 02-MappedReadsHISAT2/. In this folder contains one *.bam file and a *summary.log for each sample. Both pipelines perform the genome indexation by default when it is the first time it is ran, moreover they join all statistical summary into a single file (**Table 2 and 3**). The differences among the results is due to the different algorithms used in each software.

Table 2: Tables showing statistical summary from baqcomSTAR

Sample	Number Input reads	Number Uniquely Mapped	Percent Uniquely Mapped	Number Mapped multiLoci	Percent Mapped multiLoci	Number Mapped manyLoci	Percent Mapped manyLoci	Number Reads unmapped short	Percent reads unmapped short	Number reads unmapped other	Percent reads unmapped other
HE20-100K	92194	78228	84.85%	13789	14.96%	30	0.03%	146	0.16%	1	0.00%
HE21-100K	92147	78007	84.65%	13966	15.16%	27	0.03%	147	0.16%	0	0.00%

Table 3: Table showing statistical summary from baqcomHisat2

Sample	Input Read Pairs	Mapped reads	Percent Mapped reads	Reads unmapped	Percent reads unmapped	Reads multi mapped	Percent reads uniquely mapped
HE20-100K	92194	58081	(63.00%)	31690	(34.37%)	2423	(2.63%)
HE21-100K	92147	57752	(62.67%)	32020	(34.75%)	2375	(2.58%)

In 05-Reports folder, besides the previous files in 01-Reports folder it will contain the statistical summary for each mapping pipeline as well as the graphical summary from Multiqc if it was enabled.

Counting reads

Counting reads is performed using baqcomFeatureCounts or baqcomHtseq. BaqcomFeatureCounts uses FeatureCounts software and accepts output both from baqcomSTAR, if it was used just for mapping the reads, and baqcomHisat2.

BaqcomFeatureCounts creates a folder named 04-GeneCountsFeatCounts/ which contains a file *.summary with statistical parameters and *.outputs with seven columns showing the geneID, chromosome number, gene start, gene end, strand, gene length and number of reads for each sample. The pipeline edits the *.outputs file selecting the first and the last columns creating a new file named *.counts for each sample to make it suitable for the next analysis.

BaqcomHtseq uses htseq-count software and accepts outputs just from baqcomHisat2. Does not make sense to use htseq-count to count reads mapped with STAR, since they use the same algorithm for counting reads according to the STAR's manual. This pipeline creates a folder named 04-GeneCountsHTSeq/ containing *.out file for log and *.counts file for number of reads. The statistical summary for both pipelines is stored in the 05-Reports folder created in previous analysis (**Table 4 and 5)**.

Table 4: Table showing the statistical summary from baqcomFeatureCounts

Sample	Reads in feature	Reads NOT in.feature	Reads ambiguous	Reads too low qual	Percent Assigned To Feature	Number of Features	Number of 0 count.features
HE20-100K	42745	16093	954	6329	0.46	209	55
HE21-100K	42583	16024	984	6525	0.45	209	54

Table 5: Table showing the statistical summary from baqcomHtseq

Sample	Reads in feature	Reads NOT in.feature	Reads ambiguous	Reads too low qual	Percent Assigned To Feature	Number of Features	Number of 0 count.features
HE20-100K	42745	16093	954	6329	0.46	209	55
HE21-100K	42583	16024	984	6525	0.45	209	54

Conclusions

We have introduced a set of pipelines suitable for RNA-Seq analysis which could facilitate the command line for those who are not comfortable with the UNIX environment. Using multithread processing and multi samples, BAQCOM pipelines have shown to be efficient and fast. Although the pipeline has being designed for RNA-Seq analysis, there are the flexibility to integrate other genomic tools in the future, such as haplotype phasing and fine-mapping analysis.

References

ANDERS, S.; PYL, P. T.; HUBER, W. HTSeq-A Python framework to work with high-throughput sequencing data. **Bioinformatics**, v. 31, n. 2, p. 166–169, 2015.

ANDREWS, SI. **A Quality Control tool for High Throughput Sequence Data**. Disponível em: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/%3E. Acesso em: 29 maio.

2020.

BEIKI, H. et al. Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. **BMC Genomics**, v. 20, n. 1, p. 344, 7 dez. 2019.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 2014.

DOBIN, A. et al. STAR: ultrafast universal RNA-seq aligner. **Bioinformatics**, v. 29, n. 1, p. 15–21, jan. 2013.

EWELS, P. et al. MultiQC: summarize analysis results for multiple tools and samples in a single report. **Bioinformatics**, v. 32, n. 19, p. 3047–3048, 1 out. 2016.

KESH, S.; RAGHUPATHI, W. Critical issues in bioinformatics and computing. **Perspectives in health information management**, v. 1, n. 1, p. 9, 2004.

KIM, D. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. **Nature Biotechnology**, v. 37, n. 8, p. 907–915, 2 ago. 2019.

KIM, D.; LANGMEAD, B.; SALZBERG, S. L. HISAT: a fast spliced aligner with low memory requirements. **Nature Methods**, v. 12, n. 4, p. 357–360, 9 abr. 2015.

KIM, S. et al. Molecular and genetic regulation of pig pancreatic islet cell development. **Development**, v. 147, n. 6, p. dev186213, 15 mar. 2020.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754–1760, 2009. LIAO, Y.; SMYTH, G. K.; SHI, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. **Bioinformatics**, v. 30, n. 7, p. 923–930, 1 abr. 2014.

MARTÍNEZ-MONTES, A. M. et al. Using RNA-Seq SNP data to reveal potential causal mutations related to pig production traits and RNA editing. **Animal Genetics**, v. 48, n. 2, p. 151–165, abr. 2017.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, v. 10, n. 1, p. 57–63, 16 jan. 2009.

CHAPTER 2

[Published in: BMC Genetcs 21, 58 (2020)]
Original Research Paper
Doi: https:10.1186/s12863-020-00862-2

RNA-seq reveals downregulated osteochondral genes potentially related to tibia bacterial chondronecrosis with osteomyelitis in broilers

Haniel Cedraz de Oliveira[1]; Adriana Mércia Guaratini Ibelli[2,3]; Simone Eliza Facioni Guimaraes[1]; Mauricio Egidio Cantão[2]; Jane de Oliveira Peixoto[2,3]; Luiz Lehmann Coutinho[4]; Mônica Corrêa Ledur*[2,5]

[1]Federal University of Viçosa, Viçosa, Minas Gerais;
[2]Embrapa Suínos e Aves, Concórdia, Santa Catarina, Brazil;
[3]Programa de Pós-Graduação em Ciências Veterinárias, Universidade Estadual do Centro-Oeste, Guarapuava, Paraná, Brazil;
[4]Universidade de São Paulo, Piracicaba, SP, Brazil;
[5]Programa de Pós-Graduação em Zootecnia, UDESC-Oeste, Chapecó, SC, Brazil

*Correspondence: monica.ledur@embrapa.br

Abstract

Background:

Bacterial chondronecrosis with osteomyelitis (BCO) develops in the growth plate (GP) of the proximal femur and tibia and is initiated by damage to the less mineralized chondrocytes followed by colonization of opportunistic bacteria. This condition affects approximately 1% of all birds housed, being considered one of the major causes of lameness in fast growing broilers. Although several studies have been previously performed aiming to understand its pathogenesis, the molecular mechanisms involved with BCO remains to be elucidated. Therefore, this study aimed to generate a profile of global differential gene expression involved with BCO in the tibia of commercial broilers, through RNA sequencing analysis to identity genes and molecular pathways involved with BCO in chickens.

Results:

Our data showed 192 differentially expressed (DE) genes: 63 upregulated and 129 downregulated in the GP of the tibia proximal epiphysis of BCO-affected broilers. Using all DE genes, six Biological Processes (BP) were associated with bone development (connective tissue development, cartilage development, skeletal system development, organ morphogenesis, system development and skeletal system morphogenesis). The analyses of the upregulated genes did not indicate any significant BP (FDR < 0.05). However, with the downregulated genes, the same BP were identified when using all DE genes in the analysis, with a total of 26 coding genes explaining BCO in the tibia: *ACAN, ALDH1A2, CDH7, CHAD, CHADL, COL11A1, COMP, CSGALNACT1, CYR61, FRZB, GAL3ST1, HAPLN1, IHH, KIF26B, LECT1, LPPR1, PDE6B, RBP4A, SERINC5, SFRP1, SOX8, SOX9, TENM2, THBS1, UCHL1* and *WFIKKN2*. In addition, seven transcription factors were also associated to BCO: *NFATC2, MAFB, HIF1A-ARNT, EWSR1-FLI1, NFIC, TCF3* and *NF-KAPPAB*.

Conclusions:

Our data show that osteochondral downregulated genes are potential molecular causes of BCO in broilers, and the bacterial process seems to be, in fact, a secondary condition. Sixteen genes responsible for bone and cartilage formation were downregulated in BCO-affected broilers being strong candidate genes to trigger this disorder.

Keywords: RNA-Seq, Gallus gallus, BCO, Lameness, Cartilage

Background

The broiler chickens have undergone intense genetic and nutritional improvement, exponentially increasing their weight gain and decreasing the slaughter age. However, the improvement focused on the high performance has overlooked some physiological characteristics, such as the skeletal structure [1]. The bacterial chondronecrosis with osteomyelitis (BCO) affects up to 1% of all birds housed, being considered a worldwide major cause of lameness in commercial broilers, generating economic losses and impacting negatively the animal welfare [2,3,4]. The prevalence of lameness due the BCO can extent up to 50%, and 5% of mortality [5], however the incidence of BCO is unknown in most countries [1,3,4,6].

BCO affect the chicken GP, where the chondrocyte columns are irregularly aligned due to the high longitudinal growth rates, being associated with the growth plate turnover in this specie [7]. The BCO pathogenesis is supposed to be initiated by the damage of the poorly mineralized chondrocytes followed by colonization of opportunistic bacteria, such as Staphylococcus aureus, Escherichia coli, Coagulase-negative Staphylococcus and Enterococcus spp., in the osteochondrotic clefts [2,6,7] of both femur and tibia [8, 9].

Despite the fact that many opportunistic microorganisms have been identified in studies focused on the involvement of bacteria in the process of BCO [3, 6,7,8,9,10,11,12], there are no studies exploring the genetics and molecular mechanisms involved with BCO in tibia, although, previous studies have shown the importance of some genes involved with BCO in the femur [13,14,15]. Therefore, the aim of this study was to identify the global differential gene expression profile in the tibia GP involved with BCO in chickens, through the transcriptome analysis using RNA-seq of normal and BCO-affected broilers.

Results

RNA-Seq data

An average of 17,5 million reads/sample (2 × 100 base paired-end reads) was generated and about an average of 13 million reads/sample were kept after the data quality control. The transcriptome had in average more than 99% of the reads mapped against the chicken reference genome Galgal5, ranging from 98.83 to 99.26% to each individual sample, with 73% of the reads present in genes.

Differential gene expression

A total of 12,576 transcripts were obtained. From those, 1018 transcripts were selected with

FDR < 0.05, and grouped in three categories: protein-coding (973), lncRNA (44) and snoRNA (1). The transcripts with logFC < − 2.0 and > 2.0 were selected, obtaining 192 DE genes (12 lncRNA and 180 protein-coding), being 63 upregulated (logFC> 2.0) (8 lncRNA and 55 protein-coding) and 129 downregulated (logFC<− 2.0) (4 lncRNA and 125 protein-coding (112 annotated genes) in the affected compared to healthy broilers (**Additional File 1**).

Gene ontology

In the first step, 192 DE genes were used in the GO analysis and a total of 304 Biological Processes (BP) were identified. From those, only nine BP were significant with FDR ≤ 0.05 (**Table 1**). Among these nine BP, only six were specifically related to bone development: connective tissue development, cartilage development, skeletal system development, organ morphogenesis, system development, and skeletal system morphogenesis.

Table 1: Biological Process (PB) found using DAVID 6.8.Table 4a shows nine PB using all 192 Differentially Expressed (DE genes) with six PB associated with bone formation. Table 4b shows six PB using only 129 downregulated genes in the affected broilers. In bold are the BP that were the same in both analyses

Category	Term	Pathways	Count	FDR
Table 1a.				
GOTERM_BP_ALL	**GO:0061448**	**connective tissue development**	**12**	**1.89E-05**
GOTERM_BP_ALL	**GO:0051216**	**cartilage development**	**11**	**2.46E-05**
GOTERM_BP_ALL	**GO:0001501**	**skeletal system development**	**15**	**3.81E-05**
GOTERM_BP_ALL	**GO:0009887**	**organ morphogenesis**	**17**	**6.34E-04**
GOTERM_BP_ALL	**GO:0048731**	**system development**	**35**	**0.00825838**
GOTERM_BP_ALL	**GO:0048705**	**skeletal system morphogenesis**	**9**	**0.008659783**
GOTERM_BP_ALL	GO:0007275	multicellular organism development	36	0.019958649
GOTERM_BP_ALL	GO:0048856	anatomical structure development	39	0.026663753
GOTERM_BP_ALL	GO:0044707	single-multicellular organism process	40	0.031747676
Table 1b.				
GOTERM_BP_ALL	**GO:0001501**	**skeletal system development**	**15**	**2,32E-02**
GOTERM_BP_ALL	**GO:0061448**	**connective tissue development**	**12**	**3,26E-02**
GOTERM_BP_ALL	**GO:0051216**	**cartilage development**	**11**	**6,14E-02**
GOTERM_BP_ALL	**GO:0048705**	**skeletal system morphogenesis**	**9**	**0.000509513**
GOTERM_BP_ALL	**GO:0009887**	**organ morphogenesis**	**14**	**0.001261476**
GOTERM_BP_ALL	**GO:0048731**	**system development**	**26**	**0.049749525**

In a second step, the GO analyses were performed separating the up and downregulated genes. Using the 63 upregulated genes, the GO did not show any significant BP. However, when the 129 downregulated genes were used, the GO showed the same six BP that were identified when all 192 DE genes were used, with a total of 26 coding genes among the BPs (**Table 2**). The most enriched BP were system development (26 genes), skeletal system development (15 genes), organ morphogenesis (14 genes), connective tissue development (12 genes), cartilage development (11 genes) and skeletal system morphogenesis (9 genes) (**Table 1**).

Table 2: Twenty-six downregulated genes associated with bone formation enriched in the Biological Processes found using DAVID 6.8 tool

Gene symbol	Gene ID	logFC	FDR	Gene Name
TENM2	ENSGALG00000001768	−4.61	4.17e-05	teneurin transmembrane protein 2
WFIKKN2	ENSGALG00000007367	−4.32	1.70e-06	WAP, follistatin/kazal, immunoglobulin, kunitz and netrin domain containing 2
RBP4A	ENSGALG00000006629	−3.83	0.00e+00	retinol binding protein 4
LECT1	ENSGALG00000016945	−3.19	3.99e-03	chondromodulin
COMP	ENSGALG00000003283	−3.18	4.80e-02	cartilage oligomeric matrix protein
UCHL1	ENSGALG00000014261	−3.12	0.00e+00	Ubiquitin C-Terminal Hydrolase L1
ACAN	ENSGALG00000006725	−3.09	3.72e-03	aggrecan
CDH7	ENSGALG00000013782	−3.05	3.04e-05	cadherin 7
IHH	ENSGALG00000011347	−3.00	3.10e-04	Indian hedgehog protein
HAPLN1	ENSGALG00000015627	− 2.97	1.51e-03	Hyaluronan and proteoglycan link protein 1
KIF26B	ENSGALG00000010664	− 2.55	2.62e-02	kinesin 32 Family member 26B
PLPPR1	ENSGALG00000015542	−2.52	2.02e-05	phospholipid phosphatase related 1
COL11A1	ENSGALG00000005180	−2.47	1.51e-02	Collagen Type XI Alpha 1 Chain
GAL3ST1	ENSGALG00000007781	−2.45	1.19e-02	galactose-3-O-sulfotransferase 1
SOX9	ENSGALG00000004386	−2.40	4.58e-05	SRY-box 9
SFRP1	ENSGALG00000003473	−2.37	1.47e-03	secreted frizzled related protein 1
CHADL	ENSGALG00000011970	−2.34	2.00e-07	chondroadherin like
CSGALNACT1	ENSGALG00000010125	−2.33	4.60e-06	chondroitin sulfate N-acetylgalactosaminyltransferase 1
FRZB	ENSGALG00000002763	−2.33	1.35e-04	frizzled-related protein
CYR61	ENSGALG00000008661	−2.28	4.00e-07	Protein CYR61
SOX8	ENSGALG00000005263	−2.27	1.29e-04	SRY-box 8
THBS1	ENSGALG00000009626	−2.22	2.35e-02	thrombospondin-1 precursor
CHAD	ENSGALG00000019761	−2.12	7.28e-03	chondroadherin
PDE6B	ENSGALG00000015373	−2.11	1.67e-03	phosphodiesterase 6B
ALDH1A2	ENSGALG00000004270	−2.11	0.00e+00	Retinal dehydrogenase 2
SERINC5	ENSGALG00000014798	−2.09	6.48e-05	serine incorporator 5

Gene and Transcription factor (TF) network

A gene network using Cytoscape v.3.6 [16]. was performed to identify connections among genes enriched in bone bioprocesses (**Fig. 1**). Three genes (*COMP, COL11A1* and *SOX9*) were shared in all six BP, four genes (*CDH7, CSGALNACT1*, *ACAN* and *IHH*) were shared by five BP, three genes (*CHADL, CYRG1* and *LECT1*) were shared by four BPs, four genes (*SFRP1, CHAB, FRZB* and *SOX8*) were shared by three BPs and three genes (*TENM2, WFIKKN2* and *ALDH1A2*) were shared by two BPs (**Additional file 2**). The other genes were in only one BP (**Fig. 1**). In addition, regulatory sequence analyses for all detected genes were performed for the 26 genes used as input at TFM-explorer [17]. A total of 21 transcription factors (TF) were identified from the downregulated genes. Based on the BP from DAVID and the literature review, we selected the most putative BCO associated TF. The main TF, from the most to the least connected were: *NFATC2, MAFB, HIF1A::ARNT, EWSR1-FLI1, NFIC, TCF3 and NF-KAPPAB*, which were used to construct a TF network highlighting the most connected genes (**Fig. 2**).

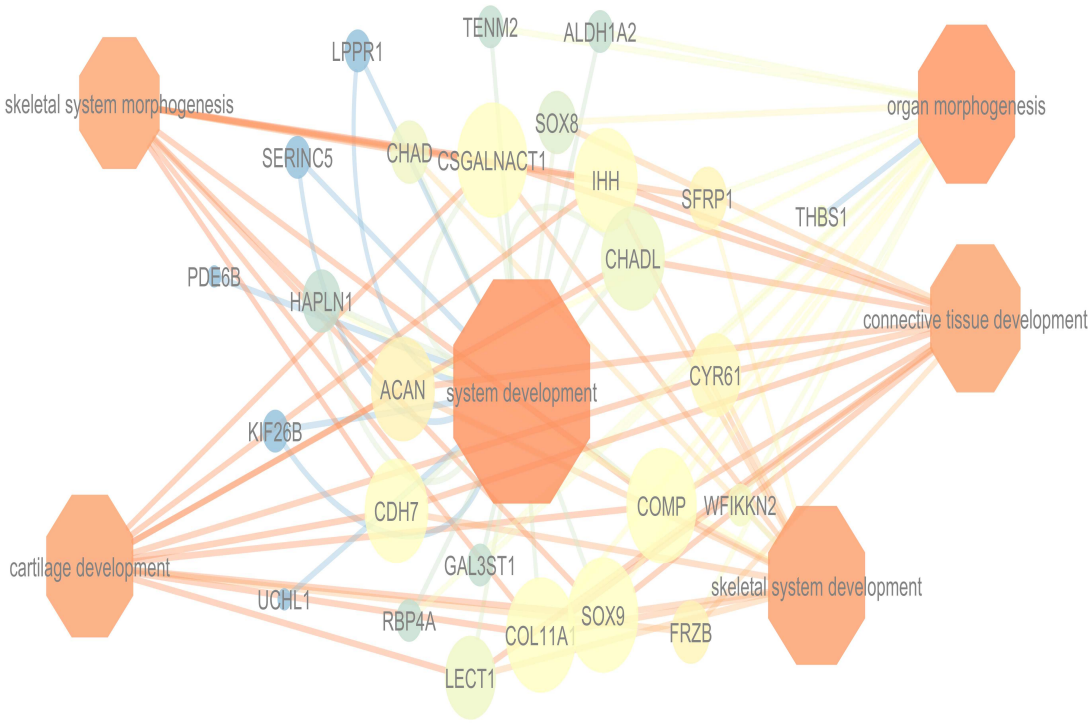

Figure 1: Gene network showing connections between downregulated genes (circles) and Biological Process (diamonds). The size (low values are small size) and the colors (low values are in bright colors) of the nodes indicate the number of directed edges and the neighborhood connectivity, respectively. The Edge colors indicate the betweenness of the edges (low values are in bright colors).

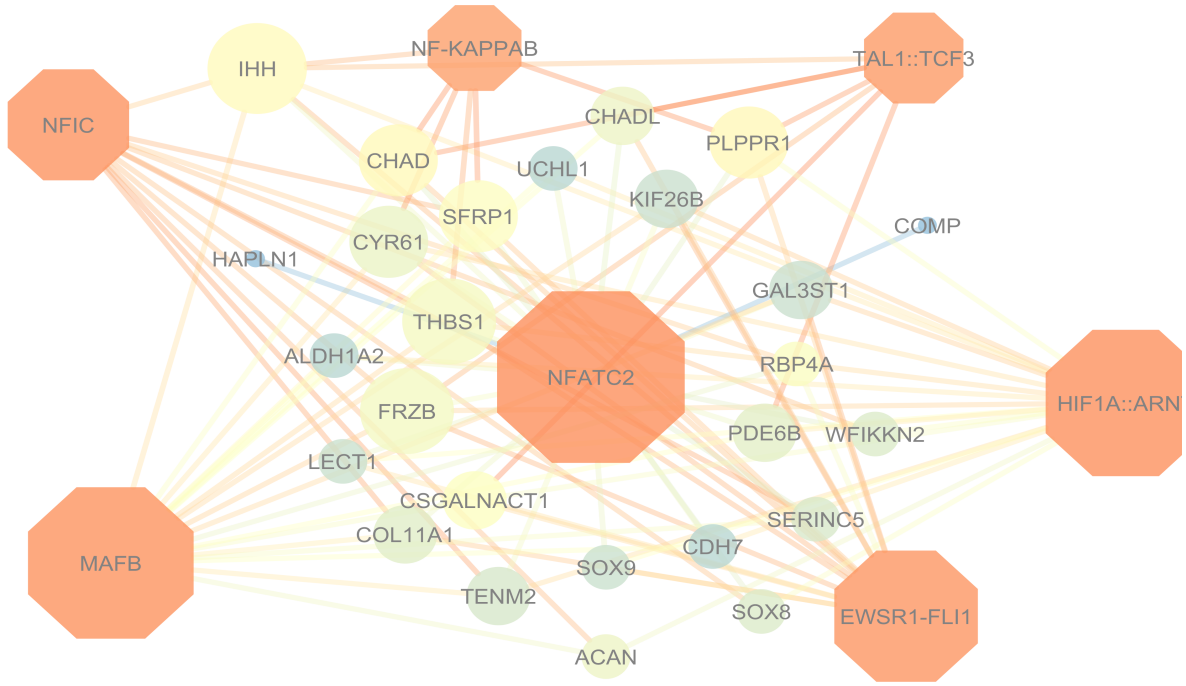

Figure 2: Transcription Factor network showing connections between downregulated genes (circles) and Transcription Factors (diamonds). The size (low values are small size) and the colors (low values are in bright colors) of the nodes indicate the number of directed edges and the neighborhood connectivity, respectively. The Edge colors indicate the betweenness of the edges (low values are in bright colors).

Discussion

Tibia is a high-mineralized bone in the body, being considered a good indicator of mineralization in ossification studies [18, 19]. The bacterial chondronecrosis with osteomyelitis (BCO) has been diagnosed worldwide and its pathogenesis seems to be related to the poor mineralization of chondrocytes and cartilage of the femur and tibia bones in fast-growing chicken, with proliferation of opportunistic bacteria [7]. Although several studies are available evaluating femur, it is known that the lesion status differ between bones [3], and that high incidences of tibia BCO has been previously reported in broilers [20]. Nowadays, there are just few studies addressing the molecular mechanisms involved with femur BCO [13,14,15] and no information regarding tibia BCO is available which highlights the importance of our study, since it brings novel and relevant knowledge to the field. Furthermore, this condition is observed in males and females, and an influence of sire line on susceptibility for this trait has already been found [20]. However, it is consensus that high performance broilers are widely affected by

BCO [3, 20,21,22], which reinforces the importance of new studies considering tibia BCO.

Although only one commercial line was used in this study, it is expected common molecular pathways involved in the BCO development in different lines, since this is a health condition. However, the line or sex effect could influence the gene expression profile related to this condition, leading to a variation in the BCO manifestation, and should be further confirmed. According to Wideman et al. (2012) [3], no difference was found on the incidence of lameness between males and females and among lines in the same type of floor. Regarding tibia BCO, no differences between males and females were found until 40 days, but an effect of sire was observed, especially when broilers were older than 6 weeks of age [20]. In our study, we used males from a high performance commercial broiler line since they are fast growing, heavy weight and therefore more probable to get affected with BCO. Therefore, if we have used other high performance commercial line, we expect that similar results would be obtained.

In the tibia GP transcriptome, 1018 genes were DE between normal and BCO-affects broilers. A total of 44 lncRNA were identified and these transcripts may be related to the regulation of adjacent genes [23]. According to the position in the genome of the NCBI database, the lncRNAs ENSGALG00000040226, ENSGALG00000033872 and ENSGALG00000032326 are close to the genes SOX9, COL11A1 and SOX8, respectively, and may be involved in the regulation of these important genes for bone development. The SOX9 and SOX8, two transcriptions factors acting on endochondral ossification, have recently been associated to femur epiphysiolysis [15]. In the current study, it was possible to identify several DE genes involved with BCO development, even using a small number of samples per group, as can be seen in several RNA-Seq studies, when groups are well characterized [24,25,26,27,28].

Initially, using 112 downregulated genes, a total of 26 genes enriched six BP directly related to bone and cartilage development (Table 2): system development (GO: 0048731), skeletal system development (GO: 0001501), organ morphogenesis (GO: 0009887), connective tissue development (GO: 0061448), cartilage development (GO: 0051216) and skeletal system morphogenesis (GO: 0048705). These are important BP for bone development and their malfunction can cause poor ossification, leading to cases of lameness, commonly observed in BCO-affected chickens [12]. The genes related to these biological pathways are ACAN, ALDH1A2, CDH7, CHAD, CHADL, COL11A1, COMP, CSGALNACT1, CYR61, FRZB, GAL3ST1, HAPLN1, IHH, KIF26B, LECT1, LPPR1, PDE6B, RBP4A, SERINC5, SFRP1, SOX8, SOX9, TENM2, THBS1, UCHL1 and WFIKKN2 (**Fig. 1**). These are functional candidate genes associated to BCO, as discussed below.

The aggrecan gene (ACAN) is one of the most important cartilage aggregating chondroitin [29]. This gene is essential to the regulation of growth factors and cartilage development [30, 31], and it is associated with diseases such as skeletal dysplasia [32], osteoarthritis [33],

osteochondritis [34, 35] and chondrodystrophy in chickens [36]. In addition, SOX8 and SOX9 play an important role as transcription factors in the development of cartilage, controlling the differentiation of chondrocytes and osteoblasts in the development of progenitor cells [30, 31, 37,38,39,40,41,42].

The CHAD gene (Chondroadherin) is highly expressed in cartilage and was first isolated in cattle [43], direct binding with calcium phosphate [44]. Hessle et al. (2013) [38] showed that CHAD has an influence on the enlargement of the epiphyseal plate and its inactivation compromises the hypertrophic differentiation of the chondrocytes, as well as the cartilage composition in rats. An important CHAD paralog is CHADL, which can negatively regulate chondrocyte differentiation, inhibiting cartilage fibrillogenesis and can be used as a marker for cartilage diseases [45].

Collagen type XI (COL11) constitutes a triple helix formed by COL11a1, COL11a2 and COL12a1, being an essential component of the extracellular matrix-regulating the diameter of the fibrils [46,47,48,49]. In the absence of COL11a1, an alternate triple helix is formed with COL11a2 and COL5a1, but it is unable to compensate for the functional deficiency of Collagen 11a1 [50]. COL11 is essential for the cartilage collagen fibrils formation and for differentiation and organization of growth plate chondrocytes [51]. The COL11a1 plays a key role in endochondral ossification, and the low expression of this gene results in alterations in the mineralization of newly formed bone [40].

The Cartilage Oligomeric Matrix Protein (COMP) is a non-collagenous protein present in the extracellular matrix of bone and cartilage, and mutation in this gene can cause death of the chondrocytes [52]. Chondrocyte death can be caused by the retention of COMP and other extracellular matrix proteins within an enlarged rugged endoplasmic reticulum, and this can lead to an abnormal matrix that is easily eroded over time [53]. In addition, COMP may be a potential molecular marker for bone diseases [54,55,56].

The CSGALNACT1 gene encodes the N-acetilgalactosaminiltransferase enzyme, which has activity in beginning and the elongation of the chondroitin sulfate synthesis [57]. In rats, reduced levels of CSGALNACT1 may cause post-natal lethality due to respiratory failure, mild dwarfism and the cartilage has an abnormality of endochondral ossification [58, 59].

Cysteine-Rich Angiogenic Inducer 61 (CYR61 or CCN1) is part of the group of matricellular proteins [60], being associated with extracellular matrix secretion [61]. Furthermore, CYR61 has important role in regulating inflammation and possible cell repair [62]. This gene potentiates DNA synthesis for cell proliferation induced by other mitogens [63] in response to bacterial or viral infection [64, 65].

Frizzled Related Protein (FRZB/SFRP-3) is part of the set of Secreted Frizzled-Related Pro-

tein (SFRP1), which comprises five members that modulate negatively and positively the Wingless-type (Wnt) signaling cascade [66]. Wnt signaling cascade is important for regulating the development, maintenance and homeostasis of bone and cartilage [67]. Therefore, dysregulation of Wnt signaling may lead to the development of osteoarthritis in rodents [67,68,69,70,71].

Hyaluronan and Proteoglycan Link Protein 1 (HAPLN1/CRTL1) is an abundant polysaccharide in the extracellular matrix of cartilage under normal conditions and is involved in cell differentiation and morphogenesis [72,73,74]. Recent studies have shown that in bone diseases in humans, such as osteoarthritis and allograft transplantation, there is always a low HAPLN1 expression [75, 76].

The Indian hedgehog (IHH) gene is expressed in cartilage, and it is recognized as regulator of bone development and morphogenesis [77]. IHH induces the differentiation of the osteogenic cell of the periosteum, being essential in the differentiation of the osteoblasts and in the maintenance of the growth plate, articular cartilage and adjacent endochondral bone formation [78,79,80]. Jin et al. [81] demonstrated that a deletion in the IHH gene was responsible for the Creeper phenotype in broilers, which is characterized by short and stunted legs. In addition, a mutation in this gene causes brachydactyly in humans [82].

Chondromodulin (CNMD) regulates the rapid growth of cartilage and vascular invasion prior to the process of cartilage replacement by the endochondral bone [83]. New evidence on the mechanism of differentiation of mesenchymal stem cells (CTMs) into chondrocytes induced by ChM-I shows that the main pathways involved in the process are focal adhesion, glycolysis, regulation of the actin cytoskeleton and the ribosome [84]. In osteoarthritis condition, the expression of ChM-I is decreased. Therefore, the regulation of this gene in cartilage may be a potential treatment, because it protects the chondrocytes from hypertrophy and delays the progression of osteoarthritis [85].

Retinol-binding protein (RBP4A) has the function of transporting vitamin A in the blood, from the liver to the peripheral tissues [86, 87]. Vitamin A plays an important role in the development of various organs, including bone growth [88], contributing to bone health [89, 90], and abnormal levels of Vitamin A may have a negative impact on bone growth [91,92,93,94]. Moreover, RBP4 is expressed during limb growth and this gene is also involved in chondrogenesis, collagen X transcription and bone mineral density [95,96,97].

Thrombospondin (THBS1) is a glycoprotein that regulates the structure of the extracellular matrix [98] being involved in the protection of chondrocytes, since it exerts pro-chondrogenic and anti-inflammatory function [99]. Furthermore, THBS1 seems to be important in homeostasis and maintenance of bone matrix integrity, and in the regulation of osteoclast formation [100].

From the 26 DE genes enriching the bone and cartilage development BP, 10 (ALDH1A2, CDH7, GAL3ST1, KIF26B, PLPPR1, PDE6B, SERINC5, TENM2, UCHL1 and WFIKKN2) appear to have no direct relationship to bone or cartilage formation. However, due to the complexity of these processes, the function of these genes in the regulation of bone cells could still be unknown, especially in chicken.

Transcription Factors

Nine TF associated to downregulated DE genes were found (**Fig. 2**), being three of those highly connected (NFATC2, MAFB and HIF1A:TNA). Several studies have shown that these TF were related to osteoblast and osteoclast differentiation, and osteolysis [101,102,103,104,105,106,107,108] The NFATC2 plays important role in the immune response, and in the differentiation and regulation of osteoclast growth [109]. Zanoti and Canalis [102] concluded that the activation of NFATC2 may inhibit the function of osteoblasts and decrease the volume of spongy bone. Furthermore, NFATC2 was associated with 24 out of 26 DE genes detected as downregulated in the affected chickens (Fig. 2). Moreover, the MAFB TF, although expressed selectively in monocytes [110], can be found in several tissues, and has been related to osteolysis in humans and rats [111] by the negative regulation of important cytokines for osteoclast differentiation [105]. HIF1A regulates oxygen homeostasis, glucose transport and generation of anaerobic energy in joints and chondrocytes, and may play an important role in the osteoarthritis metabolism [107, 112]. Four other TF were also found with a high number of connections and appear to be associated with osteosarcoma (EWSR1-FLI1 and NFIC), cranium formation (TCF3), soft tissue calcification and chondrocyte differentiation (NF-KAPPAB) [113,114,115,116,117].

Clinical implications related to BCO include, but are not limited to claudication. There is no efficient treatment and the diagnosis is not possible in early stages, being necessary to perform a necropsy. The DE genes and transcription factors found in this study play a fundamental role in bone and cartilaginous development in broilers. Their low expression in the affected chickens seems to be related to incomplete bone formation, initiating the BCO process in the tibia growth plate of fast-growing chickens. The current findings confirm our hypothesis and indicate that improvement in ossification and cartilage formation have to be addressed in poultry breeding programs.

Conclusions

We found 16 differentially expressed genes in the tibia GP transcriptome that are directly responsible for bone and cartilage formation, which were downregulated in BCO-affected broilers. According to our data, the lack of ossification might be the main cause of BCO in broilers, and the bacterial process seems to be a secondary condition. Moreover, our results highlight the pathogenesis of BCO, and show that to pursue prevention and control of such condition, breeding strategies have to focus on the improvement of ossification and cartilage formation.

Methods

Animals and sample collection

This study was performed at the Embrapa Swine and Poultry National Research Center. Approximately 50 male chicks from Cobb500 commercial broiler line were raised from 1 to 42 days of age, with a density of 12 broilers/m^2 with infrared lamps. The drinkers, feeders, curtains, light, and chickens were managed following the recommendations for the commercial line. Diets containing 3150 kcal/kg AME and 21% CP (1–21 days), 3200 kcal/kg AME and 20% CP (22–35 days), and 3200 kcal/kg AME and 18.5% CP (36–42 days) were provided. The animals had free access to feed and water. Before sample collection, the broilers were weighted at 42 days of age and euthanized by cervical dislocation followed by bleeding, according to the approval of the Embrapa Swine and Poultry Ethical Committee of Animal Use (CEUA), under protocol number 012/2012.

A classification of the tibia was performed according to the presence or absence of different levels of BCO, by visual observation of compatible necrosis lesions, according to Wideman et al. (2012) [3]. Tibia samples with adhesion between the growth plate (GP) and cartilage (CA) were considered in the normal group and those presenting separation between GP and CA were classified as the affected group. Only those with the initial level of BCO and present in both tibias were used in this study. For sample collection of the normal group, all CA was removed to access the GP. The entire GP of samples were collected, stored in liquid nitrogen and transferred to the − 80 °C freezer for further RNA analysis. Six samples (three normal and three affected) were collected and prepared for RNA-Seq analysis.

RNA extraction and library preparation

About 100 mg of the tibia GP tissue was used for RNA extraction using Trizol Reagent (Invitrogen, Carlsbad, CA) following the manufacturer's instructions. The RNA cleanup was performed using Rneasy mini kit (Qiagen, Germany) following manufacturer's instructions. After RNA extraction, the RNA was quantified in Nanodrop spectrophotometer (Thermo Scientific; Waltham, MA, USA) and the Agilent 2100 BioAnalyzer (Agilent Technologies; Santa Clara, CA, USA) was used for integrity measurement, where samples with RNA integrity number (RIN) higher than 8 were used for library preparation. A total of six tibia samples, three normal and three BCO-affected were prepared for RNA-sequencing using the TruSeq™ RNA Sample Prep Kit v2 (Illumina, Inc.; San Diego CA, USA), according to the manufacturer's recommendations, with 2 μg of total RNA.

Sequencing, quality control, assembly and differential expression analysis

The libraries were sent to the Functional Genomics Center, ESALQ, University of São Paulo, Piracicaba, SP, Brazil for sequencing in Illumina HiSeq2500 equipment (Illumina, Inc.; San Diego CA, EUA), all in the same lane, following the 2x100bp paired-end protocol.

The quality control was performed using SeqyClean tool (https://github.com/ibest/seqyclean)) with the raw FASTQ data for removing short reads (<70pb), low quality reads (Qphred < 24), PCR artifacts and adapter sequences. The sequence reads were mapped against the chicken reference genome (Gallus gallus, assembly 5.0) available in (www.ensembl.org), using BWA-MEM software [118]. The read counting was performed with Htseq software [119] using Ensembl annotation release 89. The edgeR package [120] in R environment [121] was used to identify differentially expressed (DE) genes from BCO-affected and unaffected groups. Significance threshold for DE genes was set at a False Discovery Rate (FDR) ≤ 0.05 after multiple correction tests to reduce type I error. Smear plots of DE genes were generated using the expression data for each gene within each sample in the edgeR package from the R environment [121].

Gene Ontology (GO) and Network Analyses

In a first step, a total of 192 DE genes with log Fold Change (logFC) < − 2.0 and > 2.0 was used as input on DAVID Bioinformatics Resources 6.8 tool (https://david.ncifcrf.gov/summary), and all the identified expressed genes were used as background. Genes with positive or negative logFC values were considered, respectively, upregulated or downregulated when the BCO-affected group was compared to the unaffected group. In a second step, a GO analysis was performed using only upregulated DE genes (63 genes) and only downregulated DE genes (129 genes).

The gene and transcription factor (TF) network analysis was built using Cytoscape 3.6.1 [16]. The output of GO analysis was used as input on Cytoscape with each gene associated with the correspondent biological process. Twenty-six genes were submitted to the TFM-Explorer program (http://bioinfo.lifl.fr/tfm-explorer/form.php) to identify the TF related to them. From sequence of a set of gene promoters, TFM-Explorer searches for locally overrepresented transcription factor binding sites (TFBS) using weight matrices from JASPAR database [122] to detect all potential TFBS, and extracts significant clusters by calculating a score function.

List of abbreviations

BCO: Bacterial Chondronecrosis with Osteomyelitis

DE: Differentially Expressed

BP: Biological Processes

lincRNA: Long Non-Coding RNAs

snoRNA: Small Nucleolar RNAs

GO: Gene Ontology

FDR: False Discovery Rate

TF: Transcription Factors

Declarations

Ethics approval

This study was performed with the approval of the Embrapa Swine and Poultry Ethical Committee of Animal Use (CEUA) under protocol number 012/2012.

Consent for publication

Not applicable

Availability of data and materials

All data generated or analyzed during this study are included in this published article (and its additional files). The transcriptome sequences are available in the SRA database with BioProject number PRJNA352716 and biosample numbers: SAMN05992341, SAMN05992340, SAMN05992339, SAMN05992338, SAMN05992337 and SAMN05992336.

Competing Interests

The authors declare that they have no competing interests

Funding

This study was funded by the Brazilian Agricultural Research Corporation (Embrapa), project number 01.11.07.002.04.03, from the Brazilian Government.

Author's Contribution

Conceived and designed the experiment: MCL, JOP, AMGI

Performed the experiment: HCO, AMGI, JOP, LLC, MEC, MCL

Data analysis and curation: HCO, AMGI, JOP, LLC, MEC, MCL, SEFG

Writing Original Draft preparation: HCO, AMGI, JOP, LLC, MEC, MCL, SEFG

Funding Acquisition and supervision of the research: JOP, MCL

Acknowledgments

The authors thank to Alexandre L. Tessmann for technical assistance. HCO received a scholarship from CNPq at Universidade Federal de Viçosa. MCL, SEFG and LLC are CNPq fellows. This study was financed in part by the Cordenação de Aperfeiçoamento de Pessoal de Nivel Superior-Brasil (CAPES)-Finance Code 001.

Additional files

Additional file 1. Detailed expressed values of all Differentially Expressed Genes, showing the name and description of each annotated gene, LogFoldChange, logCPM, FDR and gene Ensembl ID.

https://static-content.springer.com/esm/art%3A10.1186%2Fs12863-020-00862-2/MediaObjects/12863_2020_862_MOESM1_ESM.xlsx

Additional file 2. Detailed table showing genes among all Biological Process.

https://static-content.springer.com/esm/art%3A10.1186%2Fs12863-020-00862-2/MediaObjects/12863_2020_862_MOESM2_ESM.xlsx

References

1. Julian RJ. Rapid growth problems: ascites and skeletal deformities in broilers. Poult Sci. 1998;77:1773–80.

2. McNamee PT, Smyth JA. Bacterial chondronecrosis with osteomyelitis ('femoral head necrosis') of broiler chickens: a review. Avian Pathol. 2000;29:253–70.

3. Wideman RF, Hamal KR, Stark JM, Blankenship J, Lester H, Mitchell KN, et al. A wire-flooring model for inducing lameness in broilers: evaluation of probiotics as a prophylactic treatment. Poult Sci. 2012;91:870–83 https://doi.org/10.3382/ps.2011-01907.

4. Cook ME. Skeletal deformities and their causes: introduction. Poult Sci. 2000;79:982–4 http://www.ncbi.nlm.nih.gov//10901198.

5. Nairn ME, Watson ARA. Leg weakness of poultry—a clinical and PATHOLOGICAL characterisation. Aust Vet J. 1972;48:645–56 https://doi.org/10.1111/j.1751-0813.1972.tb09237.x.

6. Wijesurendra DS, Chamings AN, Bushell RN, Rourke DO, Stevenson M, Marenda MS, et al. Pathological and microbiological investigations into cases of bacterial chondronecrosis and osteomyelitis in broiler poultry. Avian Pathol. 2017;46:683–94.

7. Wideman RF, Prisby RD. Bone circulatory disturbances in the development of spontaneous bacterial chondronecrosis with osteomyelitis: a translational model for the pathogenesis of femoral head necrosis. Front Endocrinol (Lausanne). 2013;3:1–15 https://doi.org/10.3389/fendo.2012.00183.

8. Jiang T, Mandal RK, Wideman RF, Khatiwara A, Pevzner I, Kwon YM. Molecular survey of bacterial communities associated with bacterial chondronecrosis with osteomyelitis (BCO) in broilers. PLoS One. 2015;10:1–20.

9. Mandal RK, Jiang T, Al-Rubaye AA, Rhoads DD, Wideman RF, Zhao J, et al. An investigation into blood microbiota and its potential association with Bacterial Chondronecrosis with Osteomyelitis (BCO) in Broilers. Sci Rep. 2016;6:1–11 https://doi.org/10.1038/srep25882.

10. Al-Rubaye AAK, Ekesi NS, Zaki S, Emami NK, Wideman RF, Rhoads DD. Chondronecrosis with osteomyelitis in broilers: further defining a bacterial challenge model using the wire flooring model. Poult Sci. 2017;96:332–40.

11. Gaußmann B, Hess C, Grafl B, Kovacs M, Troxler S, Stessl B, et al. Escherichia coli isolates from femoral bone marrow of broilers exhibit diverse pheno- and genotypic characteristics that do not correlate with macroscopic lesions of bacterial chondronecrosis with osteomyelitis. Avian Pathol. 2018;9457:1–10.

12. Dinev I. Clinical and morphological investigations on the prevalence of lameness associated with femoral head necrosis in broilers. Br Poult Sci. 2009;50:284–90.

13. Petry B, Savoldi IR, Ibelli AMG, Paludo E, de Oliveira PJ, Jaenisch FRF, et al. New genes involved in the bacterial Chondronecrosis with osteomyelitis in commercial broilers. Livest Sci. 2018;208(December 2017):33–9 https://doi.org/10.1016/j.livsci.2017.12.003.

14. Paludo E, Ibelli AMG, Peixoto JO, Tavernari FC, Lima-Rosa CAV, Pandolfi JRC, et al. The involvement of RUNX2 and SPARC genes in the bacterial chondronecrosis with osteomyelitis in broilers. Animal. 2017;11:1063–70.

15. de Oliveira PJ, Savoldi IR, Ibelli AMG, Cantão ME, Jaenisch FRF, Giachetto PF, et al. Proximal femoral head transcriptome reveals novel candidate genes related to epiphysiolysis in broiler chickens. BMC Genomics. 2019;20:1031 https://doi.org/10.1186/s12864-019-6411-9.

16. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504 https://doi.org/10.1101/gr.1239303.

17. Tonon L, Touzet H, Varre J-S. TFM-Explorer: mining cis-regulatory regions in genomes. Nucleic Acids Res. 2010;38(Web Server):W286–92 https://doi.org/10.1093/nar/gkq473.

18. Angel R. Metabolic disorders: limitations to growth of and mineral deposition into the broiler skeleton after hatch and potential implications for leg problems. J Appl Poult Res. 2007;16:138–49 https://doi.org/10.1093/japr/16.1.138.

19. Sanchez-Rodriguez E, Benavides-Reyes C, Torres C, Dominguez-Gasca N, Garcia-Ruiz AI, Gonzalez-Lopez S, et al. Changes with age (from 0 to 37 D) in tibiae bone mineralization, chemical composition and structural organization in broiler chickens. Poult Sci. 2019; https://doi.org/10.3382/ps/pez363.

20. Wideman RF, Al-Rubaye A, Reynolds D, Yoho D, Lester H, Spencer C, et al. Bacterial chondronecrosis with osteomyelitis in broilers: influence of sires and straight-run versus sex-separate rearing. Poult Sci. 2014;93:1675–87 https://doi.org/10.3382/ps.2014-03912.

21. Gilley AD, Lester H, Pevzner IY, Anthony NB, Wideman RF. Evaluating portable wire-flooring models for inducing bacterial chondronecrosis with osteomyelitis in broilers1. Poult Sci. 2014;93:1354–67 https://doi.org/10.3382/ps.2013-03781.

22. Wideman RF. Bacterial chondronecrosis with osteomyelitis and lameness in broilers: a review. Poult Sci. 2016;95:325–44.

23. Milligan MJ, Lipovich L. Pseudogene-derived lncRNAs: emerging regulators of gene expression. Front Genet. 2015;5 https://doi.org/10.3389/fgene.2014.00476.

24. Yu S, Wang G, Liao J, Tang M. Transcriptome profile analysis identifies candidate genes for the melanin pigmentation of breast muscle in Muchuan black-boned chicken. Poult Sci. 2018;97:3446–55 https://doi.org/10.3382/ps/pey238.

25. Ding X, Zhang S, Li X, Feng C, Huang Q, Wang S, et al. Profiling expression of coding genes, long noncoding RNA, and circular RNA in lung adenocarcinoma by ribosomal RNA-depleted RNA sequencing. FEBS Open Bio. 2018;8:544–55 https://doi.org/10.1002/2211-5463.12397.

26. Jaing C, Rowland RRR, Allen JE, Certoma A, Thissen JB, Bingham J, et al. Gene expression analysis of whole blood RNA from pigs infected with low and high pathogenic African swine fever viruses. Sci Rep. 2017;7:10115 https://doi.org/10.1038/s41598-017-10186-4.

27. Kwon SG, Hwang JH, Park DH, Kim TW, Kang DG, Kang KH, et al. Identification of differentially expressed genes associated with litter size in Berkshire pig placenta. PLoS One. 2016;11:e0153311 https://doi.org/10.1371/journal.pone.0153311.

28. Li HL, Lin HR, Xia JH. Differential gene expression profiles and alternative isoform regulations in gill of Nile Tilapia in response to acute hypoxia. Mar Biotechnol. 2017;19:551–62 https://doi.org/10.1007/s10126-017-9774-4.

29. Ovadia BM, Parker CH, Lash JW. Changing patterns of proteoglycan synthesis during chondrogenic differentiation. J Embryol Exp Morpholog. 1980;56:59–70.

30. Hu G, Codina M, Fisher S. Multiple enhancers associated with ACAN suggest highly

redundant transcriptional regulation in cartilage. Matrix Biol. 2012;31:328–37 https://doi.org/10.1016/j.matbio.2012.06.001.

31. Lauing KL, Cortes M, Domowicz MS, Henry JG, Baria AT, Schwartz NB. Aggrecan is required for growth plate cytoarchitecture and differentiation. Dev Biol. 2014;396:224–36 https://doi.org/10.1016/j.ydbio.2014.10.005.

32. Tompson SW, Merriman B, Funari VA, Fresquet M, Lachman RS, Rimoin DL, et al. A recessive skeletal dysplasia, SEMD Aggrecan type, results from a missense mutation affecting the C-type Lectin domain of Aggrecan. Am J Hum Genet. 2009;84:72–9 https://doi.org/10.1016/j.ajhg.2008.12.001.

33. Stattin E-L, Tegner Y, Domellöf M, Dahl N. Familial osteochondritis dissecans associated with early osteoarthritis and disproportionate short stature. Osteoarthr Cartil. 2008;16:890–6 https://doi.org/10.1016/j.joca.2007.11.009.

34. Stattin E-L, Wiklund F, Lindblom K, Önnerfjord P, Jonsson B-A, Tegner Y, et al. A missense mutation in the Aggrecan C-type Lectin domain disrupts extracellular matrix interactions and causes dominant familial Osteochondritis Dissecans. Am J Hum Genet. 2010;86:126–37 https://doi.org/10.1016/j.ajhg.2009.12.018.

35. Xu M, Stattin E-L, Shaw G, Heinegård D, Sullivan G, Wilmut I, et al. Chondrocytes derived from Mesenchymal stromal cells and induced pluripotent cells of patients with familial Osteochondritis Dissecans exhibit an endoplasmic reticulum stress response and defective matrix assembly. Stem Cells Transl Med. 2016;5:1171–81 https://doi.org/10.5966/sctm.2015-0384.

36. Primorac D, Stover ML, Clark SH, Rowe DW. Molecular basis of nanomelia, a heritable chondrodystrophy of chicken. Matrix Biol. 1994;14:297–305 https://doi.org/10.1016/0945-053X(94)90195-3.

37. Bi W, Deng JM, Zhang Z, Behringer RR, de Crombrugghe B. Sox9 is required for cartilage formation. Nat Genet. 1999;22:85–9 https://doi.org/10.1038/8792.

38. Dy P, Wang W, Bhattaram P, Wang Q, Wang L, Ballock RT, et al. Sox9 directs hypertrophic maturation and blocks osteoblast differentiation of growth plate chondrocytes. Dev Cell. 2012;22:597–609 https://doi.org/10.1016/j.devcel.2011.12.024.

39. Serrano RL, Chen L-Y, Lotz MK, Liu-Bryan R, Terkeltaub R. Proteasomal function is impaired in human osteoarthritic chondrocytes and this can contribute to decreased SOX9 and aggrecan. Arthritis Rheum. 2018; https://doi.org/10.1002/art.40456.

40. Schmidt K, Schinke T, Haberland M, Priemel M, Schilling AF, Mueldner C, et al. The high mobility group transcription factor Sox8 is a negative regulator of osteoblast differentiation.

J Cell Biol. 2005;168:899–910 https://doi.org/10.1083/jcb.200408013.

41. Finzsch M, Stolt CC, Lommes P, Wegner M. Sox9 and Sox10 influence survival and migration of oligodendrocyte precursors in the spinal cord by regulating PDGF receptor expression. Development. 2008;135:637–46 https://doi.org/10.1242/dev.010454.

42. Umeda K, Oda H, Yan Q, Matthias N, Zhao J, Davis BR, et al. Long-term expandable SOX9+ chondrogenic ectomesenchymal cells from human pluripotent stem cells. Stem Cell Reports. 2015;4:712–26 https://doi.org/10.1016/j.stemcr.2015.02.012.

43. Larsson T, Sommarin Y, Paulsson M, Antonsson P, Hedbom E, Wendel M, et al. Cartilage matrix proteins. A basic 36-kDa protein with a restricted distribution to cartilage and bone. J Biol Chem. 1991;266:20428–33.

44. H-Y ZHOU. Proteomic analysis of hydroxyapatite interaction proteins in bone. Ann N Y Acad Sci. 2007;1116:323–6 https://doi.org/10.1196/annals.1402.023.

45. Tillgren V, Ho JCS, Önnerfjord P, Kalamajski S. The novel small leucine-rich protein chondroadherin-like (CHADL) is expressed in cartilage and modulates chondrocyte differentiation. J Biol Chem. 2015;290:918–25.

46. Van Der Rest M, Garrone R. Collagen family of proteins. FASEB J. 1991;5:2814–23.

47. Holmes DF, Kadler KE. The 10+4 microfibril structure of thin cartilage fibrils. Proc Natl Acad Sci. 2006;103:17249–54 https://doi.org/10.1073/pnas.0608417103.

48. Blaschke UK, Eikenberry EF, Hulmes DJS, Galla HJ, Bruckner P. Collagen XI nucleates self-assembly and limits lateral growth of cartilage fibrils. J Biol Chem. 2000;275:10370–8.

49. Gregory KE, Oxford JT, Chen Y, Gambee JE, Gygi SP, Aebersold R, et al. Structural organization of distinct domains within the non-collagenous N-terminal region of collagen type XI. J Biol Chem. 2000;275:11498–506.

50. Fernandes RJ, Weis MA, Scott MA, Seegmiller RE, Eyre DR. Collagen XI chain misassembly in cartilage of the chondrodysplasia (cho) mouse. Matrix Biol. 2007;26:597–603.

51. Li Y, Lacerda DA, Warman ML, Beier DR, Yoshioka H, Ninomiya Y, et al. A Fibrillar Collagen Gene , Collla 1 , Is Essential for Skeletal Morphogenesis. Cell. 1995;80:423–30.

52. Valdes AM. Osteoarthritis – Genetic Studies of Monogenic and Complex Forms. In: Eisman J, Thakker RV, Igarashi T, Whyte MP, editors. Genetics of Bone Biology and Skeletal Disease: Academic Press; 2013. p. 275–93. https://www.sciencedirect.com/science//pii/B9780123878298000378?via%3Dihub.

53. Posey K, Hecht J. The role of cartilage Oligomeric matrix protein (COMP) in skeletal

disease. Curr Drug Targets. 2008;9:869–77 https://doi.org/10.2174/138945008785909293.

54. Roman-Blas J, Dion AS, Seghatoleslami MR, Giunta K, Oca P, Jimenez SA, et al. MED and PSACH COMP mutations affect chondrogenesis in chicken limb bud micromass cultures. J Cell Physiol. 2010;224:817–26.

55. Saghafi M, Khodashahi M, Saadati N, Azarian A, Rezaieyazdi Z, Salehi M, et al. Relationship between cartilage oligomeric matrix protein (COMP) and rheumatoid arthritis severity. Electron Physician. 2017;9:5940–7.

56. Firner S, Willwacher S, de Marées M, Bleuel J, Zaucke F, Brüggemann GP, et al. Effect of increased mechanical knee joint loading during running on the serum concentration of cartilage oligomeric matrix protein (COMP). J Orthop Res. 2018;36(7):1937–46.

57. Gotoh M, Sato T, Akashima T, Iwasaki H, Kameyama A, Mochizuki H, et al. Enzymatic synthesis of chondroitin with a novel chondroitin sulfate N-acetylgalactosaminyltransferase that transfers N-acetylgalactosamine to glucuronic acid in initiation and elongation of chondroitin sulfate synthesis. J Biol Chem. 2002;277:38189–96 https://doi.org/10.1074/jbc.M203619200.

58. Vodopiutz J, Mizumoto S, Lausch E, Rossi A, Unger S, Janocha N, et al. Chondroitin sulfate N -acetylgalactosaminyltransferase-1 (CSGalNAcT-1) deficiency results in a mild skeletal dysplasia and joint laxity. Hum Mutat. 2017;38:34–8 https://doi.org/10.1002/humu.23070.

59. Shimbo M, Suzuki R, Fuseya S, Sato T, Kiyohara K, Hagiwara K, et al. Postnatal lethality and chondrodysplasia in mice lacking both chondroitin sulfate N-acetylgalactosaminyltransferase-1 and -2. PLoS One. 2017;12:e0190333 https://doi.org/10.1371/journal.pone.0190333.

60. Bornstein P. Diversity of function is inherent in matricellular proteins: an appraisal of thrombospondin 1. J Cell Biol. 1995;130:503–6.

61. Yang GP, Lau LF. Cyr61, product of a growth factor-inducible immediate early gene, is associated with the extracellular matrix and the cell surface. Cell Growth Differ. 1991;2:351–7.

62. Jun J, Lau LF. The matricellular protein CCN1 induces fibroblast senescence and restricts fibrosis in cutaneous wound healing. Nat Cell Biol. 2010;12:676–85.

63. Kireeva ML, Mo F-E, Yang GP, Lau LF. Cyr61, a product of a growth factor-inducible immediate-early gene, promotes cell proliferation, migration, and adhesion. Mol Cell Biol. 1996;16:1326–34.

64. Kim S, Park J, Chung S, Hwang H, Chung K, Jo I, et al. Coxsackievirus B3 Infection Induces cyr61 Activation via JNK To Mediate Cell Death Coxsackievirus B3 Infection Induces cyr61 Activation via JNK To Mediate Cell Death. J Virol. 2004;78:13479–88.

65. Wiedmaier N, Müller S, Köberle M, Manncke B, Krejci J, Autenrieth IB, et al. Bacteria induce CTGF and CYR61 expression in epithelial cells in a lysophosphatidic acid receptor-dependent manner. Int J Med Microbiol. 2008;298:231–43.

66. Bovolenta P, Esteve P, Ruiz JM, Cisneros E, Lopez-Rios J. Beyond Wnt inhibition: new functions of secreted frizzled-related proteins in development and disease. J Cell Sci. 2008;121:737–46 https://doi.org/10.1242/jcs.026096.

67. Lories RJ, Corr M, Lane NE. To Wnt or not to Wnt: the bone and joint health dilemma. Nat Rev Rheumatol. 2013;9:328–39 https://doi.org/10.1038/nrrheum.2013.25.

68. Lories RJU, Peeters J, Bakker A, Tylzanowski P, Derese I, Schrooten J, et al. Articular cartilage and biomechanical properties of the long bones in Frzb -knockout mice. Arthritis Rheum. 2007;56:4095–103.

69. Thysen S, Luyten FP, Lories RJ. Loss of Frzb and Sfrp1 differentially affects joint homeostasis in instability-induced osteoarthritis. Osteoarthr Cartil. 2015;23:275–9 https://doi.org/10.1016/j.joca.2014.10.010.

70. Carrillo-López N, Panizo S, Alonso-Montes C, Román-García P, Rodríguez I, Martínez-Salgado C, et al. Direct inhibition of osteoblastic Wnt pathway by fibroblast growth factor 23 contributes to bone loss in chronic kidney disease. Kidney Int. 2016;90:77–89 https://doi.org/10.1016/j.kint.2016.01.024.

71. Shaw AT, Maeda Y, Gravallese EM. IL-17A deficiency promotes periosteal bone formation in a model of inflammatory arthritis. Arthritis Res Ther. 2016;18:104 https://doi.org/10.1186/s13075-016-0998-x.

72. Fraser JRE, Laurent TC, Laurent UBG. Hyaluronan: its nature, distribution, functions and turnover. J Intern Med. 1997;242:27–33.

73. Oohashi T, Edamatsu M, Bekku Y, Carulli D. The hyaluronan and proteoglycan link proteins: organizers of the brain extracellular matrix and key molecules for neuronal function and plasticity. Exp Neurol. 2015;274:134–44 https://doi.org/10.1016/j.expneurol.2015.09.010.

74. Dicker KT, Gurski LA, Pradhan-Bhatt S, Witt RL, Farach-Carson MC, Jia X. Hyaluronan: a simple polysaccharide with diverse biological functions. Acta Biomater. 2014;10:1558–70 https://doi.org/10.1016/j.actbio.2013.12.019.

75. Wang W, Yu Y, Hao J, Wen Y, Han J, Hou W, et al. Genome-wide DNA methylation profiling of articular cartilage reveals significant epigenetic alterations in Kashin-Beck disease and osteoarthritis. Osteoarthr Cartil. 2017;25:2127–33 https://doi.org/10.1016/j.joca.2017.08.002.

76. Lin Y, Lewallen EA, Camilleri ET, Bonin CA, Jones DL, Dudakovic A, et al. RNA-seq analysis of clinical-grade osteochondral allografts reveals activation of early response genes. J Orthop Res. 2016;34:1950–9 https://doi.org/10.1002/jor.23209.

77. Bitgood MJ, McMahon AP. Hedgehog and bmp genes are Coexpressed at many diverse sites of cell–cell interaction in the mouse embryo. Dev Biol. 1995;172:126–38 https://doi.org/10.1006/dbio.1995.0010.

78. de Crombrugghe B, Lefebvre V, Nakashima K. Regulatory mechanisms in the pathways of cartilage and bone formation. Curr Opin Cell Biol. 2001;13:721–8 https://doi.org/10.1016/S0955-0674(00)00276-3.

79. Maeda Y, Nakamura E, Nguyen M-T, Suva LJ, Swain FL, Razzaque MS, et al. Indian hedgehog produced by postnatal chondrocytes is essential for maintaining a growth plate and trabecular bone. Proc Natl Acad Sci U S A. 2007;104:6382–7.

80. Kurio N, Saunders C, Bechtold TE, Salhab I, Nah HD, Sinha S, et al. Roles of Ihh signaling in chondroprogenitor function in postnatal condylar cartilage. Matrix Biol. 2018;67:15–31 https://doi.org/10.1016/j.matbio.2018.02.011.

81. Jin S, Zhu F, Wang Y, Yi G, Li J, Lian L, et al. Deletion of Indian hedgehog gene causes dominant semi-lethal Creeper trait in chicken. Sci Rep. 2016;6:1–10 https://doi.org/10.1038/srep30172.

82. Ma G, Yu J, Xiao Y, Chan D, Gao B, Hu J, et al. Indian hedgehog mutations causing brachydactyly type A1 impair hedgehog signal transduction at multiple levels. Cell Res. 2011;21:1343–57 https://doi.org/10.1038/cr.2011.76.

83. Hiraki Y, Mitsui K, Endo N, Takahashi K, Hayami T, Inoue H, et al. Molecular cloning of human chondromodulin-I, a cartilage-derived growth modulating factor, and its expression in Chinese hamster ovary cells. Eur J Biochem. 1999;260:869–78 https://doi.org/10.1046/j.1432-1327.1999.00227.x.

84. Chun XS, Xin DL, Zhou W, Qiang HY, Feng Y, Feng LH, et al. Proteomic analysis of chondromodulin-I-induced differentiation of mesenchymal stem cells into chondrocytes. J Proteome. 2017;159:1–18 https://doi.org/10.1016/j.jprot.2017.02.017.

85. Zhang X, Prasadam I, Fang W, Crawford R, Xiao Y. Chondromodulin-1 ameliorates osteoarthritis progression by inhibiting HIF-2α activity. Osteoarthr Cartil. 2016;24:1970–80

https://doi.org/10.1016/j.joca.2016.06.005.

86. Peterson PA. Studies on the interaction between Prealbumin, retinol-binding protein, and vitamin a. J Biol Chem. 1971;246:44–9.

87. Berry DC, Byrne SMO, Vreeland AC, Blaner WS, Noy N. Cross Talk between Signaling and Vitamin A Transport by the the Retinol-Binding Protein Receptor STRA6. Mol Cell Biol. 2012;32:3164–75.

88. Leitch VD, Dwivedi PP, Anderson PJ, Powell BC. Retinol-binding protein 4 downregulation during osteogenesis and its localization to non-endocytic vesicles in human cranial suture mesenchymal cells suggest a novel tissue function. Histochem Cell Biol. 2013;139:75–87.

89. Hatfield JT, Anderson PJ, Powell BC. Retinol-binding protein 4 is expressed in chondrocytes of developing mouse long bones: implications for a local role in formation of the secondary ossification center. Histochem Cell Biol. 2013;139:727–34.

90. NIH NI of H. Vitamin A and Bone Health National. 2015.

91. Johansson S, Lind P, Hakansson H, Oxlund H, Rberg J, Melhus H. Subclinical hypervitaminosis a causes fragile bones in rats. Bone. 2002;31:685–9 https://doi.org/10.1016/S8756-3282(02)00910-9.

92. Joo N-S, Yang S-W, Song B, Yeum K-J. Vitamin a intake, serum vitamin D and bone mineral density: analysis of the Korea National Health and nutrition examination survey (KNHANES, 2008–2011). Nutrients. 2015;7:1716–27 https://doi.org/10.3390/nu7031716.

93. WOLBACH SB. Vitamin-a deficiency and excess in relation to skeletal growth. J Bone Joint Surg Am. 1947;29:171–92 http://www.ncbi.nlm.nih.gov//20284696.

94. Binkley N, Krueger D. Hypervitaminosis a and bone. Nutr Rev. 2000;58:138–44 http://www.ncbi.nlm.nih.gov//10860393.

95. De Luca F, Uyeda JA, Mericq V, Mancilla EE, Yanovski JA, Barnes KM, Zile MH, Baron J. Retinoic acid is a potent regulator of growth plate chondrogenesis. Endocrinology. 2000;141:346–53.

96.Cohen AJ, Lassová L, Golden EB, Niu Z, Adams SL. Retinoids directly activate the collagen X promoter in prehypertrophic chondrocytes through a distal retinoic acid response element. J Cell Biochem. 2006;99:269–78 https://doi.org/10.1002/jcb.20937.

97. Huang N, Zhou J, Wang W, Wang Q, Tang Y, Sun Y, et al. Retinol-binding protein 4 is positively associated with bone mineral density in patients with type 2 diabetes and osteopenia or osteoporosis. Clin Endocrinol. 2018;2017:659–64.

98. Kvansakul M, Adams JC, Hohenester E. Structure of a thrombospondin C-terminal fragment reveals a novel calcium core in the type 3 repeats. EMBO J. 2004;23:1223–33.

99. Maumus M, Manferdini C, Toupet K, Chuchana P, Casteilla L, Gachet M, et al. Thrombospondin-1 partly mediates the cartilage protective effect of adipose-derived mesenchymal stem cells in osteoarthritis. Front Immunol. 2017;8.

100. Amend SR, Uluckan O, Hurchla M, Leib D, Veis Novack D, Silva M, et al. Thrombospondin-1 regulates bone homeostasis through effects on bone matrix integrity and nitric oxide signaling in osteoclasts. J Bone Miner Res. 2015;30:106–15 https://doi.org/10.1002/jbmr.2308.

101. Kim RY, Seong Y, Cho TH, Lee B, Kim IS, Hwang SJ. Local administration of nuclear factor of activated T cells (NFAT) c1 inhibitor to suppress early resorption and inflammation induced by bone morphogenetic protein-2. J Biomed Mater Res Part A. 2018;106:1299–310 https://doi.org/10.1002/jbm.a.36332.

102. Zanotti S, Canalis E. Activation of Nfatc2 in osteoblasts causes osteopenia. J Cell Physiol. 2015;230:1689–95 https://doi.org/10.1002/jcp.24928.

103. Lei Z, Sabine A, Daniel A, M VP, Beat T, et al. Swiss Med Wkly. 2017;147 https://doi.org/10.4414/smw.2017.14529.

104. Nishikomori R, Kawai T, Toshiyuki K, Oda H, Yasumi T, Izawa K, et al. Remarkable improvement of articular pain by biologics in a Multicentric carpotarsal osteolysis patient with a mutation of MAFB gene. Pediatr Rheumatol. 2015;13(Suppl 1):152 https://doi.org/10.1186/1546-0096-13-S1-P152.

105. Kim K, Kim JH, Lee J, Jin HM, Kook H, Kim KK, et al. MafB negatively regulates RANKL-mediated osteoclast differentiation. Blood. 2007;109:3253–9 https://doi.org/10.1182/blood-2006-09-048249.

106. Chen X, Gu S, Chen B-F, Shen W-L, Yin Z, Xu G-W, et al. Nanoparticle delivery of stable miR-199a-5p agomir improves the osteogenesis of human mesenchymal stem cells via the HIF1a pathway. Biomaterials. 2015;53:239–50 https://doi.org/10.1016/J.BIOMATERIALS.2015.02.071.

107. Fernández-Torres J, Hernández-Díaz C, Espinosa-Morales R, Camacho-Galindo J, del Carmen G-SN, López-Macay Á, et al. Olymorphic variation of hypoxia inducible factor-1 a (HIF1A) gene might contribute to the development of knee osteoarthritis: a pilot study. BMC Musculoskelet Disord. 2015;16:218 https://doi.org/10.1186/s12891-015-0678-z.

108. Sato Y, Miyauchi Y, Yoshida S, Morita M, Kobayashi T, Kanagawa H, et al. The vitamin D analogue ED71 but not 1,25(OH) 2 D 3 targets HIF1a protein in osteoclasts; 2014. https:

//doi.org/10.1371/journal.pone.0111845.

109. Zanotti S, Smerdel-Ramoya A, Canalis E. Nuclear factor of activated T-cells (NFAT)C2 inhibits notch receptor signaling in osteoblasts. J Biol Chem. 2013;288:624–32 https://doi.org/10.1074/jbc.M112.340455.

110. Sieweke MH, Tekotte H, Frampton J, Graf T. MafB is an interaction partner and repressor of Ets-1 that inhibits Erythroid differentiation. Cell. 1996;85:49–60 https://doi.org/10.1016/S0092-8674(00)81081-8.

111. Park JG, Tischfield MA, Nugent AA, Cheng L, Di Gioia SA, Chan W-M, et al. Loss of MAFB function in humans and mice causes Duane syndrome, aberrant Extraocular muscle innervation, and inner-ear defects. Am J Hum Genet. 2016;98:1220–7 https://doi.org/10.1016/j.ajhg.2016.03.023.

112. Grimmer C, Pfander D, Swoboda B, Aigner T, Mueller L, Hennig FF, et al. Hypoxia-inducible factor 1α is involved in the prostaglandin metabolism of osteoarthritic cartilage through up-regulation of microsomal prostaglandin E synthase 1 in articular chondrocytes. Arthritis Rheum. 2007;56:4084–94 https://doi.org/10.1002/art.23136.

113. Grohar PJ, Janeway KA, Mase LD, Schiffman JD. Advances in the Treatment of Pediatric Bone Sarcomas. Am Soc Clin Oncol Educ B. 2017;37:725–35 https://doi.org/10.14694/EDBK_175378.

114. Zhang H, Mai Q, Chen J. MicroRNA-210 is increased and it is required for dedifferentiation of osteosarcoma cell line. Cell Biol Int. 2017;41:267–75 https://doi.org/10.1002/cbin.10721.

115. Kim CH, Oda T, Itoh M, Jiang D, Artinger KB, Chandrasekharappa SC, et al. Repressor activity of headless/Tcf3 is essential for vertebrate head formation. Nature. 2000;407:913–6.

116. Al-Huseini I, Ashida N, Nakao T, Ono K, Kimura T. P672Inhibition of NF-kappaB in smooth muscle cells promotes calcification. Eur Heart J. 2017;38(suppl_1) https://doi.org/10.1093/eurheartj/ehx501.P672.

117. Olivotto E, Otero M, Marcu KB, Goldring MB. Pathophysiology of osteoarthritis: canonical NF-κB/IKKβ-dependent and kinase-independent effects of IKKα in cartilage degradation and chondrocyte differentiation. RMD Open. 2015;1(Suppl 1):e000061 https://doi.org/10.1136/rmdopen-2015-000061.

118. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25:1754–60.

119. Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput

sequencing data. Bioinformatics. 2015;31:166–9.

120. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26:139–40.

121. R Core Team. A language and environment for statistical computing: R Foundation for Statistical Computing; 2018. http://www.r-project.org/.

122. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004;32(Database issue):D91–4 https://doi.org/10.1093/nar/gkh012.

CHAPTER 3

[To be submitted in: Frontiers in Genetics]
Original Research Paper

Fine-mapping of QTL for backfat in pigs reveals candidate variants in the CCND2 gene using large-scale genotype and phenotype data

Haniel C. Oliveira[1]; Marcos S. Lopes[2,3]; Martijn F. L. Derks[2,4]; Ole Madsen[4]; Barbara Harlizius[2]; Maren van Son[5]; Eli H. Grindflek[5]; Arne B. Gjuvsland[5]; Martien A. M. Groenen[4]; Simone Eliza Facioni Guimaraes[1]

[1]Federal University of Viçosa, Viçosa, Minas Gerais;
[2]Topigs Norsvin Research Center, Beuningen, Netherlands;
[3]Topigs Norsvin, Curitiba, Brazil;
[4]nimal Breeding and Genomics, Wageningen University & Research, Wageningen, the Netherlands;
[5]Norsvin SA, Hamar, Norway

ABSTRACT

Backfat is an important trait and has been chosen as a trait to improve feed efficiency and high lean meat yield. Although adipose tissue is a good energy storage, excessive fat leads to obesity, diseases in humans and economic losses either in human medicine, as well as animal health and production. Genome Wide Association Studies (GWAS) have been widely used for finding variants, however, there is still the needed to reduce the map regions to find causative mutations. For this reason, we aimed to fine map a QTL region on chromosome 5 related to backfat in 4 commercial pig breeds. For GWAS we used 122084 genotyped animals distributed among the breeds Synthetic = 31,183; Landrace = 38,298; Pietrain = 16,189; Large White = 36,414 and, for fine-mapping it was used 659 whole genome sequenced animals (Synthetic = 187; Pietrain = 40; Landrace = 227; Large White = 205). From the leadSNP (most significant SNP) on GWAS we phased the haplotype and found a core haplotype with 7 SNPs across breeds. In this region we have identified three variants in different intronic regions in the *CCND2* gene. Using RNA-Seq and ChIP-Seq dataset we have found peaks of H3K27me3 and H3K4me3 markers, which indicates regulation of gene expression during the embryonic development and RNA-Seq dataset from embryos and fetuses have confirmed that. In this study a strong QTL region is associated with backfat and we have showed that there are three candidate variants associated with this trait located in intronic regions of *CCND2* gene indicating that this region seems to have promoters involved on the *CCND2* expression during pre-natal development.

Keywords: sequencing; ChIP-Seq; GWAS; haplotype; RNA-Seq

INTRODUCTION

Backfat (BFE) is an important trait in pork production and therefore it has been included in the breeding objectives of genetic companies for decades (MERKS, 2000). Due to the (negative) genetic correlation, by decreasing BFE it is also possible to increase traits such as feed efficiency and meat yield, which is very valuable for the pork chain. However a strong selection for decreasing BFE also decrease other fat content such as intra-muscular fat which is important for meat quality (CHEN et al., 2014; FONTANESI et al., 2012; LONERGAN et al., 2001).

Adipose tissue has an important role as storage of energy (WANG; PENG; YI, 2016), however, excessive fat may cause obesity leading to diseases in humans such as hypertension, diabetes and cancer (HANSON et al., 2014; RAJENDER RAO; LAL; GIRIDHARAN, 2014) and economic implications in pig breeding affecting the feed efficiency and meat quality (LONERGAN et al., 2001). Furthermore, fat and fatty acids contribute importantly to the quality and nutritional content of the meat (WOOD et al., 2008). The amount of saturated fatty acids (SFA), monounsaturated fatty acids (MUFA) and polyunsaturated fatty acids (PUFA) is important for the processing industry. Thickness and fatty acid components of subcutaneous fat play an important role in dry-cured hams and BFE prevent excessive loss during seasoning period (BOSI; RUSSO, 2016).

For a better understanding of quantitative traits suchas BFE, a further look on the genetics basis of these traits is required. The development of commercial SNP chips has enabled GWAS to efficiently map regions throughout the genome affecting fat-related traits (DO et al., 2014; FONTANESI et al., 2012; FOWLER et al., 2013; OKUMURA et al., 2013). While GWAS identify significant marker associations, the current commercial SNP chip density often leads to clusters of markers (due to extended linkage disequilibrium) covering a region that is still too large to allow accurate identification of the responsible genes or variants. Hence, there is still the need of fine mapping these clusters if the aim is to find causative relations between gene(s) or variants that affect economically important traits such as BFE.

In this study, we identified a QTL region on chromosome 5 with strong association with BFE. Further, we aimed to fine map this QTL region in order to identify the causative mutation(s) related to BFE and better understand the biological process of this trait.

MATERIALAND METHODS

Data

Data from five pig populations, Synthetic line , Pietrain, Landrace, Large-White and Duroc were available for this study. The trait evaluated was backfat (BFE), measured with at the end of the test period (on average 120 Kg of live weight). Two datasets from each population were used: ALL and GENOTYPED. The dataset ALL consisted of all genotyped animals and their contemporaries (Synthetic: 125,319, Landrace: 223,723, 86,361, Pietrain: 97,358, and Large-White: 175,757) animals **(Table 1).** Using ALL, the phenotypes were pre-corrected for all non-genetic effects. The pre-corrected phenotype was used as the response variable in further analysis. The non-genetic effects were estimated within line using the following animal linear model in ASReml v3.0 (GILMOUR et al., 2009):

$$BFE_{ijkl} = \mu + sex_i + hym_j + a_k + litter_l + e_{ijkl} \quad (1)$$

where BFE_{ijkl} is the phenotypic record of the $k$animal; μ is the overall mean, sex_i is the fixed effect of sex i, hym_j is the fixed effect of the herd-year-month j of birth, a_k is the random additive genetic effect of animal k, $litter_l$ is the random effect of litter l and e_{ijkl} is the random residual effect. The vector of additive genetic effects was assumed to be distributed as $\sim N(0, A\sigma_a^2)$, which accounted for the (co)variances between animals due to relationships by formation of an **A** matrix (pedigree-based numerator relationship matrix), σ_a^2 being the additive genetic variance. The vector of litter effects was assumed to be distributed as $\sim N(0, I\sigma_I^2)$, with **I** being an identity matrix and σ_I^2 the litter variance. The vector of residual effects was assumed to be distributed as $\sim N(0, I\sigma_e^2)$, σ_e^2 being the residual variance.

Table 1: Summary statistics of the datasets used in this study.

Population	Dataset	N	Mean(mm)	SD(mm)	Min(mm)	Max(mm)
Synthetic	All[1]	125,319	9.98	2.28	3.5	29.4
	Genotyped[2]	31,183	9.48	1.97	3.5	23.5
Pietrain	All	97,358	7.72	1.58	3.5	28.8
	Genotyped	16,189	7.32	1.38	3.5	19.9
Landrace	All	223,723	7.01	1.66	3.5	27.0
	Genotyped	38,298	7.59	1.68	3.9	20.8
Large-White	All	175,757	12.62	2.74	3.5	29.9
	Genotyped	36,414	12.52	2.60	4.0	29.5
Duroc	All	37,991	8.70	2.04	4.0	20.0
	Genotyped	11,274	8.13	2.19	4.0	20.0

[1] ALL: consisted of all genotyped animals and their contemporaries, which was used for pre-correction of the phenotypes.
[2] Genotyped: a subset of ALL that includes only genotyped animals which was used for the genome-wise association analyses.

Genotypes

The dataset GENOTYPED was a subset of ALL consisting of all animals that had both phenotypes and genotypes (Synthetic: 31,183, Landrace: 38,298, Duroc: 11,832, Pietrain: 16,189 and Large-White: 36,414) (**Table 1**). All animals from the GENOTYPED dataset were genotyped using a medium density SNP chip, which was mainly the (Illumina) GeneSeek custom 50K SNP chip (Lincoln, NE, United States). However, a part of the animals from the Synthetic, Duroc and Pietrain populations were genotyped using either the (Illumina) GeneSeek custom 80K SNP chip (Lincoln, NE, United States) or the Illumina Porcine SNP60 Beadchip (Illumina, San Diego, CA, United States). In addition, 319 Synthetic, 228 Pietrain, 441 Landrace, 140 Duroc, and 401 Large-White genotyped animals were also genotyped using the Axiom porcine 660K array from Affymetrix (Affymetrix Inc., Santa Clara, CA, United States) (**Table 2**).

Table 2: Number of animals genotyped with each chip density and DNA sequenced

Population	50K	60K	80K	660K	All	Sequenced
Synthetic	25,193	739	5,251	319	31,183	187
Pietrain	10,396	1,414	4,379	228	16,189	40
Landrace	38,298	-	-	441	38,298	227
Large-White	36,414	-	-	401	36,414	205
Duroc	6,139	1,174	3,961	140	11,274	-
Total	116440	3327	13591	1529	133358	659

Quality control was performed within population and SNP chip in order to exclude SNPs with

GenCall < 0.15 (Illumina Inc., 2005), call rate < 0.95, minor allele frequency < 0.01, strong deviation from Hardy-Weinberg equilibrium (χ^2> 600), SNPs located on sex chromosomes and unmapped SNPs. The positions of the SNPs were based on the Sscrofa11.1 assembly of the reference genome. Animals with frequency of missing genotypes ≥ 0.05 would be removed from the dataset. However, all genotyped animals had a frequency of missing genotypes < 0.05 and were therefore kept for further analyses.

After the quality control, the medium density genotypes where imputed towards the high density genotypes within population using Fimpute v2.2 (SARGOLZAEI; CHESNAIS; SCHENKEL, 2014). After imputation, a total of 31,183; 16,189; 38,298; 36,414 and 11,274 animals and 488,585, 504,525, 443,619, 518,659 and 417,532 SNPs were available for the Synthetic, Pietrain, Landrace, Large-White and Duroc populations, respectively, were available for the genome-wide association study (GWAS).

GWAS

A single-SNP GWAS was performed with the GENOTYPED dataset and the imputed high density genotypes within each population using the following linear animal model in GCTA (YANG et al., 2011, 2014):

$$y_k^* = \mu + X\hat{\beta} + u_k + e_k \tag{2}$$

where y_k^* is the pre-corrected phenotype of the k animal (pre-corrected for all non-genetic effects); μ is the average of the pre-corrected phenotype; X is the genotype (0, 1, 2) of the k animal for the evaluated SNP; $\hat{\beta}$ is the unknown allele substitution effect of the evaluated SNP; u_k is the random additive genetic effect, being that the vector of additive genetic effects was assumed to be distributed as $\sim N(0, G\sigma_a^2)$, which accounted for the (co)variances between animals due to relationships by formation of an **G** matrix (genomic numerator relationship matrix build using the imputed 660K genotypes), σ_a^2being the additive genetic variance; and e_k is the random residual effect which was assumed to be distributed as $\sim N(0, I\sigma_e^2)$.

The proportion of phenotypic variance explained by a SNP was defined as $\sigma_{SNP}^2/\sigma_p^2$, where σ_p^2 is total phenotypic variance (sum of the additive and residual variances) which was estimated based on model (2) without a SNP effect, and $\sigma_{SNP}^2 = 2pq\alpha^2$, being p and q the allele frequencies and α the estimated allele substitution effect of the evaluated SNP. The association between a SNP and the phenotype was declared significant when $-log_{10}(P) > 8$.

Fine-mapping

In order to search for possible causal mutations, we chose the most significant analysis that overlap across the largest number of populations for more in-depth analysis. For this specific region, which turned up to be the around 66 mega bases on *Sus Scrofa* chromosome (SSC) 5, we performed further haplotype analysis using whole-genome sequence, ChIP-Seq and RNA-Seq data.

Haplotype Analysis

Beagle 5.1 (BROWNING; BROWNING, 2007) was used to phase the imputed high density genotypes taking 0.5 Mb downstream and 0.5 Mb upstream from the lead SNP (SSC5:66103958) in the QTL region SSC5:66 ± 0.5 Mb, and we focus on the 40 SNPs surround the lead SNP (20<leadSNP>20) using HaploPi pipeline (https://github.com/hanielcedraz/HaploPi) which has built and calculated the haplotype frequencies.

Sequence Data

Whole-genome sequence (WGS) was used to access a sequence level SNP for the haplotype region (SSC5:66094630-66140348). For this purpose, 659 animals (Synthetic = 187; Pietrain = 40; Landrace = 227; Large White = 205) were available (**Table 2**). Genomic DNA from ear biopsies was extracted using BioSprint DNA Kit (Qiagen, Hilden, Germany) and its concentration and quality was measured using NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, DE, USA). The data was sequenced on Illumina Hiseq 2000, 2×100 paired-end reads. For alignment of the sequences to the Sscrofa11.1 genome build (WARR et al., 2019) we used BWA-MEM version 0.7.15 (LI; DURBIN, 2009) with an average mappability of 96.11% and a sample coverage ranging from 6.6–22.7X (10X average). Samtools dedup function was used to remove PCR duplicates (LI et al., 2009). Variant calling was performed with Freebayes v1.1.0 with following settings: -min-base-quality 10 min-alternate-fraction 0.2 haplotype-length 0 min-alternate-count 2 (GARRISON; MARTH, 2012). Variants with phred quality score < 20, and within 3 bp of an indel were discarded (LI et al., 2009). Variants were annotated using the Ensembl variant effect predictor (VEP, release 90) (MCLAREN et al., 2016). The impact of missense variants was predicted using SIFT (KUMAR; HENIKOFF; NG, 2009).

ChIP-Seq and RNA-Seq Analysis

ChIP-Seq and RNA-seq data were downloaded from European Nucleotide Archive (ENA) data base. ChIP-Seq project PRJEB31483 (pig alveolar macrophages), PRJNA436619 (pig muscle) and PRJEB31482 (pig liver). ChIP-Seq data was aligned against the pig reference genome (S. *Scrofa* 11.1v) using BWA-mem algorithm (LI, 2013) and the Peak calling was performed using MACS2 (ZHANG et al., 2008).

For RNA-Seq analysis was used datasets from ENA project PRJNA428757 (pig backfat) and project PRJNA240057 (pig muscle) and dataset provided by Prof. Dr. Simone Eliza Facioni Guimarães (from pig fetuses and embryos). BAQCOM pipeline (https://github.com/hanielcedraz/BAQCOM)) was used to perform the Quality Control using Trimmomatic v. 0.39 and Mapping against the pig reference genome (S. Scrofa 11.1v) using STAR v. 2.7.2b. Aligned files were visualized using Jbrowse 1.16.8 (WESTESSON; SKINNER; HOLMES, 2013). Also, cufflinks v2.2.1 (TRAPNELL et al., 2010) was used to obtain the Fragments Per Kilobase Million (FPKM).

Haplotype-based association

Using the core haplotype, a haplotype-based association analysis was performed by the number of copies, using the mixed-model described previously. Animals with two copies of the core haplotype were codified as 2, animals with one copy were codified as 1 and, animals which do not have the haplotype as 0. This analysis was performed using all individuals which have at least one haplotype GC**A**AG or GC**G**AG, being 31,085; 16,158; 38,220 and 36,352 animals from synthetic, Pietrain, Landrace and Large-White populations.

RESULTS

In this study, using large-scale genotype and phenotype data, we identified a QTL region on SSC5 with a strong association with backfat in four pig breeds (Synthetic, Pietrain, Landrace and Large White) (**Figure 1**).

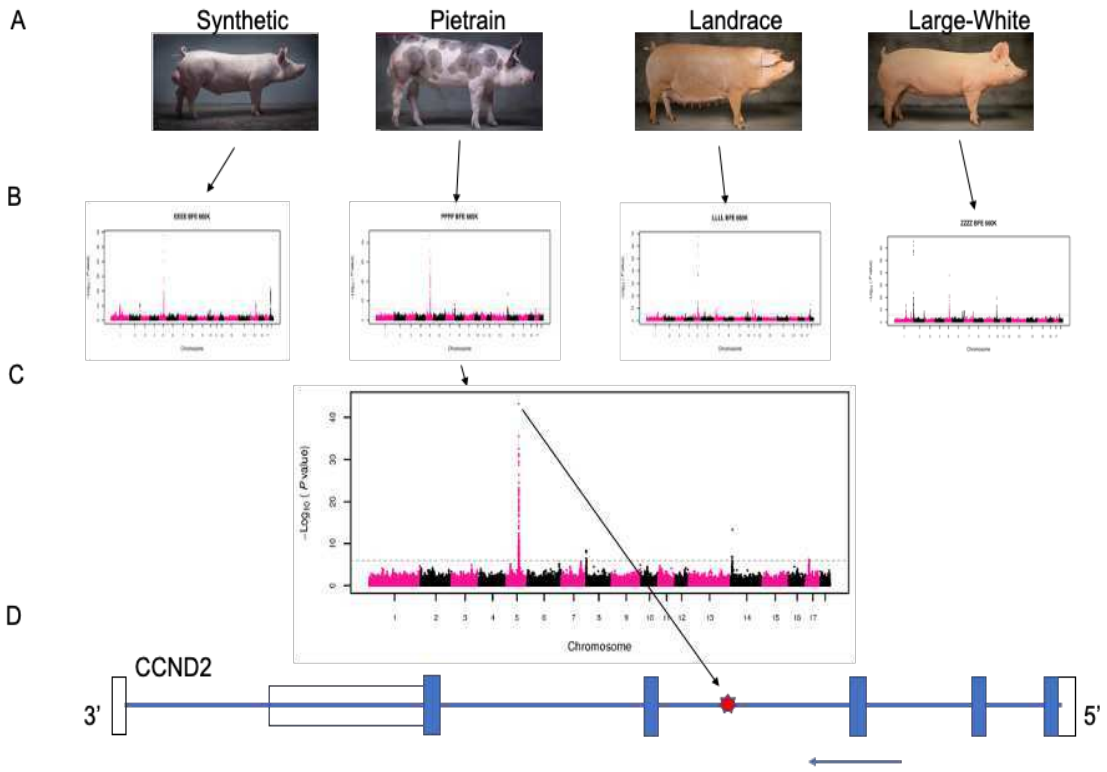

Figure 1: A schemmatic representation of our study. A) Four pig breeds were used in this study: Synthetic (Large White based), Pietrain, Landrace and Large-White. B) GWAS peaks on chromosome 5 associated with backfat for each evaluated breed. C) Manhatan plot showing the most significant SNP (leadSNP - AX-116265245) in Pietrain breed. D) CCND2 gene model. The location of the mutation on the intron 3 is indicated with a red star. Blue arrow indicates the reading frame

The QTL region identified on SSC5 explains up to 6% of the genetic variance of backfat in four out of the five breeds evaluated in this study (**Table 3** and **Figure 2**) in the same region. This region was not identified only in the Duroc population in which the most significant SNP identified for the other breeds is not segregating anymore (**Figure 3**). The most significant SNP AX-116265245 (SSC5:66103958) for breeds Synthetic (logP = 57.67), Pietrain (logP = 43.29), Landrace (logP = 67.61) and Large-White (logP = 37.91) was used as leadSNP to further fine mapping analyses of this region. The frequency of the allele associated with increased BFE (G) is the highest in the in Large-White population (0.79). In the other population where this SNP is still segregating the frequency of the G allele is below 0.40 (0.27 in the Synthetic, 0.32 in the Landrace and 0.35 in the Pietrain population) (**Table 3**).

Table 3: Most significant SNP identified in the within-breed association analysis.

Line	SNP	Chr	Position	A1	Frq	LogP	b	SE	VgExp	VpExp
Synthetic	AX-116265245	5	66103958	G	0.27	57.67	0.36	0.02	0.04	0.02
Pietrain	AX-116265245	5	66103958	G	0.35	43.29	0.27	0.02	0.06	0.03
Landrace	AX-116265245	5	66103958	G	0.32	67.61	0.24	0.01	0.05	0.02
Large-White	AX-116265245	5	66103958	G	0.79	37.91	0.33	0.03	0.02	0.01

SNP: Single Nucleotide Polymorphism; Chr: Chromosome; A1: Reference allele; Frq: Frequency; LogP: Log P-value; b: Beta coefficient; SE: Standart Error; VgExp: Genetic Variance explained; VpExp: Phenotypic Variance explained

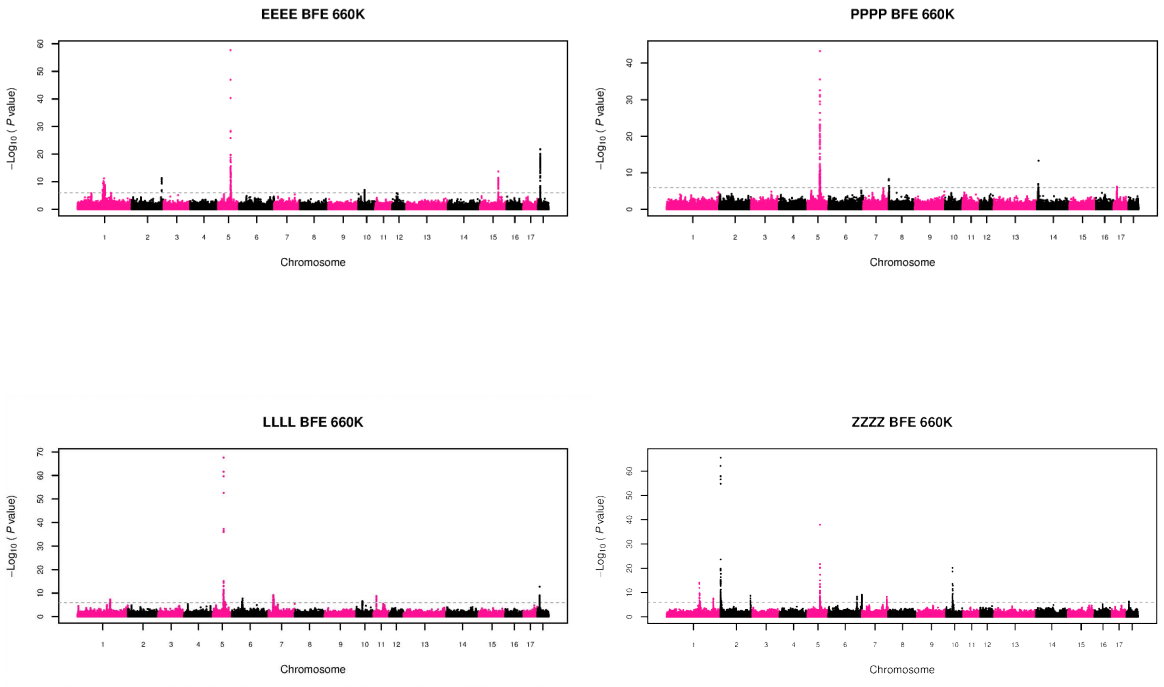

Figure 2: GWAS pead on chromosome 5 showing association with backfat for each breed. EEEE = Synthetic; LLLL = Landrace; PPPP = Pietrain; ZZZZ = Large-White

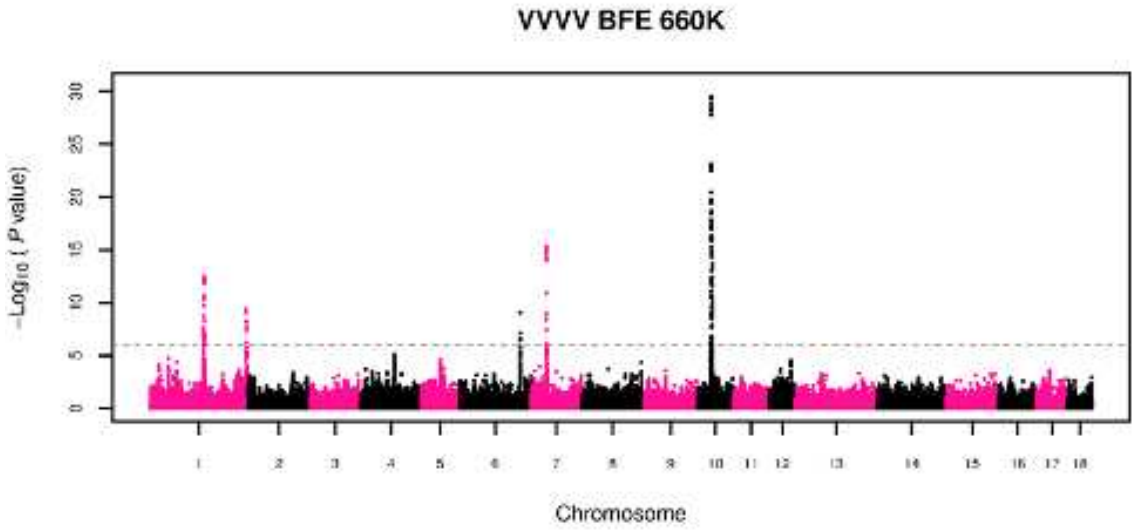

Figure 3: Manhattan plot for Duroc breed showing that there is not an association on chromosome 5

The heritability (h^2) for BFE ranged from 0.42 to 0.54 among the four breeds. Synthetic (0.54 ± 0.01), Pietrain (0.50 ± 0.01), Landrace (0.42 ± 0.01), Large White (0.52 ± 0.01) and Duroc (0.43 ± 0.01) (**Table 4**).

Table 4: Summary of the Genetic parameters for backfat trait of each pig breed.

Population	Va	Ve	Vp	h2
Synthetic	1.19 ± 0.04	1.01 ± 0.01	2.21 ± 0.04	0.54 ± 0.01
Pietrain	0.56 ± 0.02	0.57 ± 0.01	1.13 ± 0.02	0.50 ± 0.01
Landrace	0.55 ± 0.02	0.77 ± 0.01	1.32 ± 0.02	0.42 ± 0.01
Large-White	1.68 ± 0.05	1.56 ± 0.01	3.24 ± 0.05	0.52 ± 0.01
Duroc	0.78 ± 0.04	1.03 ± 0.02	1.82 ± 0.04	0.43 ± 0.01

Va Variance Additive;
Ve Variance environment;
Vp Variance Phenotipic;
h2 Heritability

Haplotype Analysis

We identified a haplotype of four SNPS (GC**A**AG) surrounding the leadSNP present in all breeds (size ~28kb). Using beagle 5.1 (BROWNING; BROWNING, 2007) we phased the haplotypes and selected the most frequent haplotype associated with less backfat in the four breeds, which has the A allele as leadSNP. We included two more SNPS (AGC**A**AGC =

Synthetic; AGC**A**AGG = Pietrain; GGC**A**AGA = Landrace; AGC**A**AGG = Large-White) to increase the region to 45kb (minimum size to search regulatory elements). The frequencies of these haplotypes are 57.70%, 35.87%, 22.67% and 16.85% for each population (Synthetic, Pietrain, Landrace and Large-White), respectively (**Table 5**). LD between the variants in the core haplotype and the leadSNP range from 0.0217944 and 0.956603 (**Table 6**)

Table 5: Most frequent haplotypes. In red, lead SNPs (Alternative allele (A) which decreases backfat). In orange, the core region shared by all breeds. In green, the neighbor nucleotides included to increase the region to 45kb.

haplotypes	N.of Haplotypes	Frequency	Breed
CCATTAGTTACAGAGTGAGCAAGCTATCGGGAGTCGTGTGT	35869	57,70	Sinthetic
CCATTAGTTACAGAGTGAGCAAGGCTATCGGGAGTCGTGTG	11591	35,87	Pietrain
TCAAGCTTGACTCAGAAGGCAAGACCTATCGGTCGTGTACC	17326	22,67	Landrace
GCCATTATTACAGAGTGAGCAAGGACCTATCGGGAGTCGTG	12252	16,85	Large-White

Table 6: Linkage Disequilibrium between the leadSNP and the variants presented in the core haplotype

	SSC5:66103958			
Variants	Synthetic	Pietrain	Landrace	Large-White
SSC5:66094630	0.353006	0.814286	-	0.0300595
SSC5:66097126	0.339592	0.814286	-	0.0283620
SSC5:66100317	0.278956	0.172024	0.143658	0.4840810
SSC5:66107719	0.517851	0.914765	0.956603	0.2363540
SSC5:66126122	0.290814	0.814286	-	0.0217944
SSC5:66140348	0.380046	0.814286	0.143658	0.0487122

Animals homozygous for the A allele have corrected backfat (as explained in the material and methods section) levels ranging from 3.5 to 27.1mm (Synthetic = 3.8 to 20.5 mm; Pietrain = 3.5 to 19.6mm; Landrace = 3.9 to 18.0mm; Large-White = 4.2 to 27.1mm), heterozygous animals have levels ranging from 3.5 to 29.5mm (Synthetic = 3.5 to 23.5mm; Pietrain = 3.8 to 19.9mm; Landrace = 4.0 to 20.8mm; Large-White = 4.0 to 29.5mm) and animals homozygous for the G allele have levels ranging from 4.0 to 27.3mm (Synthetic = 4.8 to 19.6mm; Pietrain = 4.2 to 13.5mm; Landrace = 4.0 to 18.0mm; Large-White = 4.6 to 127.3mm) (**Figure 4**).

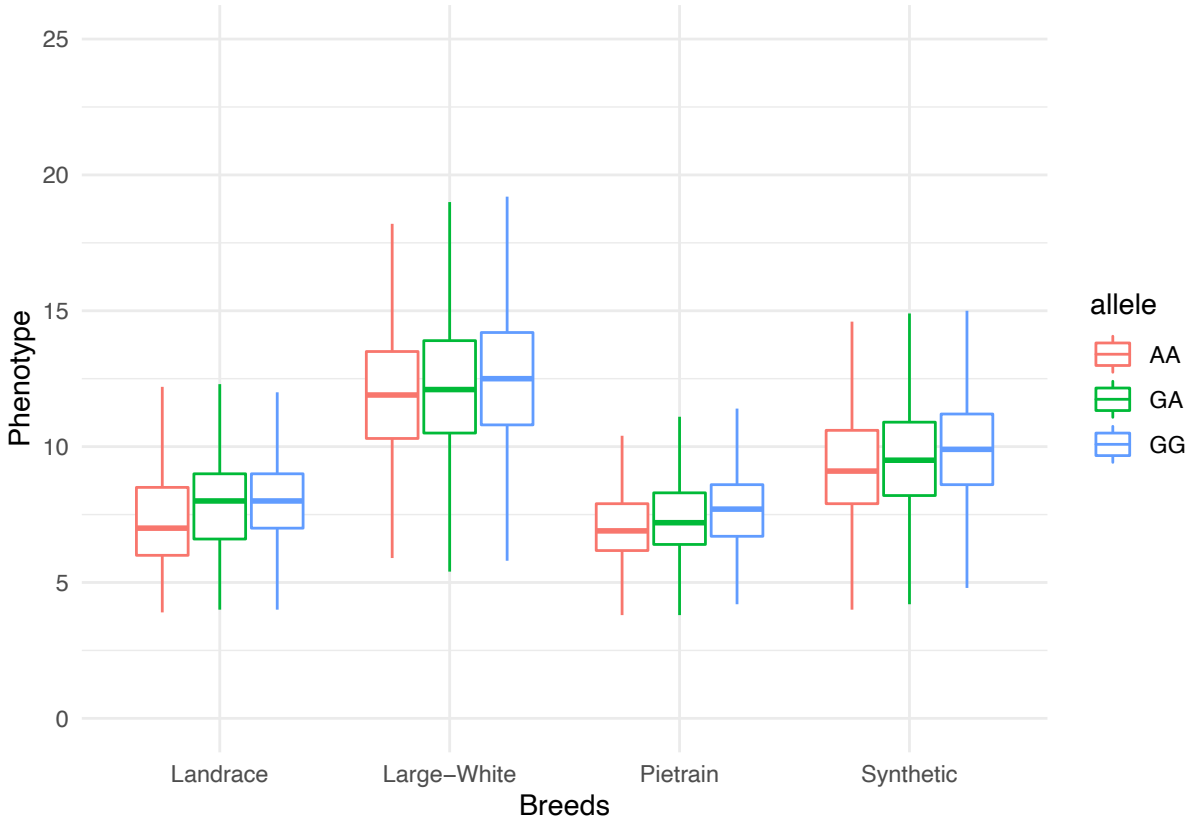

Figure 4: Boxplot showing the phenotypic variation of backfat between the genotypes among the breeds

Whole-Genome Sequence Data Analysis

For fine mapping, we evaluated a total of 659, being 101 homozygous for the A allele and 38 homozygous for the G allele, all with whole-genome sequence data available. Analyzing the data of these subset of animals, we found two variants (besides the leadSNP) with high frequency across breeds (SSC5:66103958 (leadSNP) = 1, SSC5:66097445 = 0.924426, and SSC5:66099282 = 0.7114135) (**Table 7**).

Table 7: Difference of frequency between the reference and alternative allele for all breeds. Subtraction of allele frequency (AF) from homozygous A and homozygous G for the lead SNP of each breed (DeltaAF), then the average from those DeltaAF (avgDeltaAF) across all breeds.

Positions	REF	ALT	$\Delta AF_{Synthetic}$	$\Delta AF_{Landrace}$	$\Delta AF_{Pietrain}$	$\Delta AF_{Large\text{-}White}$	AvgΔAF
66103958	G	A	1.000000	1.000000	1.0000	1.00	1.0000000
66097445	C	G	0.867347	0.892857	0.9375	1.00	0.9244260
66099282	A	AG	0.677083	0.428571	1.0000	0.74	0.7114135

$^{\Delta AFBreed}$ Frequency of A allele minus frequency of G allele;
$^{Avg\Delta AF}$ Average of each ΔAF from each line;
^{REF} Reference allele;
^{ALT} Alternative allele;

The LD between these variants and the leadSNP varies among the breeds. Variant SSC5:66099282 ($r^2_{Synthetic}$ = 0.466352; $r^2_{Pietrain}$ = 0.533249; $r^2_{Large\text{-}White}$ = 0.386646; this variant is not in LD in Landrace) and the variant SSC5:66097445 ($r^2_{Synthetic}$ = 0.632067; $r^2_{Pietrain}$ = 0.676933; $r^2_{Landrace}$=0.472127; $r^2_{Large\text{-}White}$ = 0.910731) (**Table 8**). The three variants are annotated within the intronic region of the Cyclin D2 (*CCND2*) gene (ENSSSCG00000038694).

Table 8: Linkage Disequilibrium between the leadSNP and the variants presented in the core haplotype

	SSC5:66103958			
Variants	Synthetic	Pietrain	Landrace	Large-White
SSC5:66099282	0.466352	0.533249	-	0.386646
SSC5:66097445	0.632067	0.676933	0.472127	0.910731

CCND2 Regulatory region

We have identified (three) variants in different intronic regions in the *CCND2* gene. Le et al., (2017) describe the *CCND2* as a possible candidate gene for back conformation. Furthermore, D-type Cyclin regulate postnatal islet growth pathway (KUSHNER et al., 2005). The *CCND2* gene in human is located at the chromosome 12. Although the human genome is better annotated than pig genome, we could not identify regulatory elements just looking at the reference genome and annotation.

All identified variants are in intronic regions. Using ChIp-Seq data from pig alveolar macrophages cultured for 2 and 6 hours we identified active and inactive promotors in the variant regions. On Chr5:66103958 (leadSNP) we found peaks for histone H3 lysine 27 trimethylation (H3K27me3) and histone H3 lysine 4 trimethylation (H3K4me3) marks in macrophages cultured for 2 and 6 hours, respectively while on Chr5:66099282 we just found peaks for H3K4me3 in macrophages cultured for 2 hours (Supplementary figure S1).

Using RNA-Seq the average expression level (measured as FPKM) of *CCND2* in backfat, muscle, embryo and fetus was 0.8937968, 0.6865253, 8.2841239 and 8.2732024, respectively (**Figure 5**).

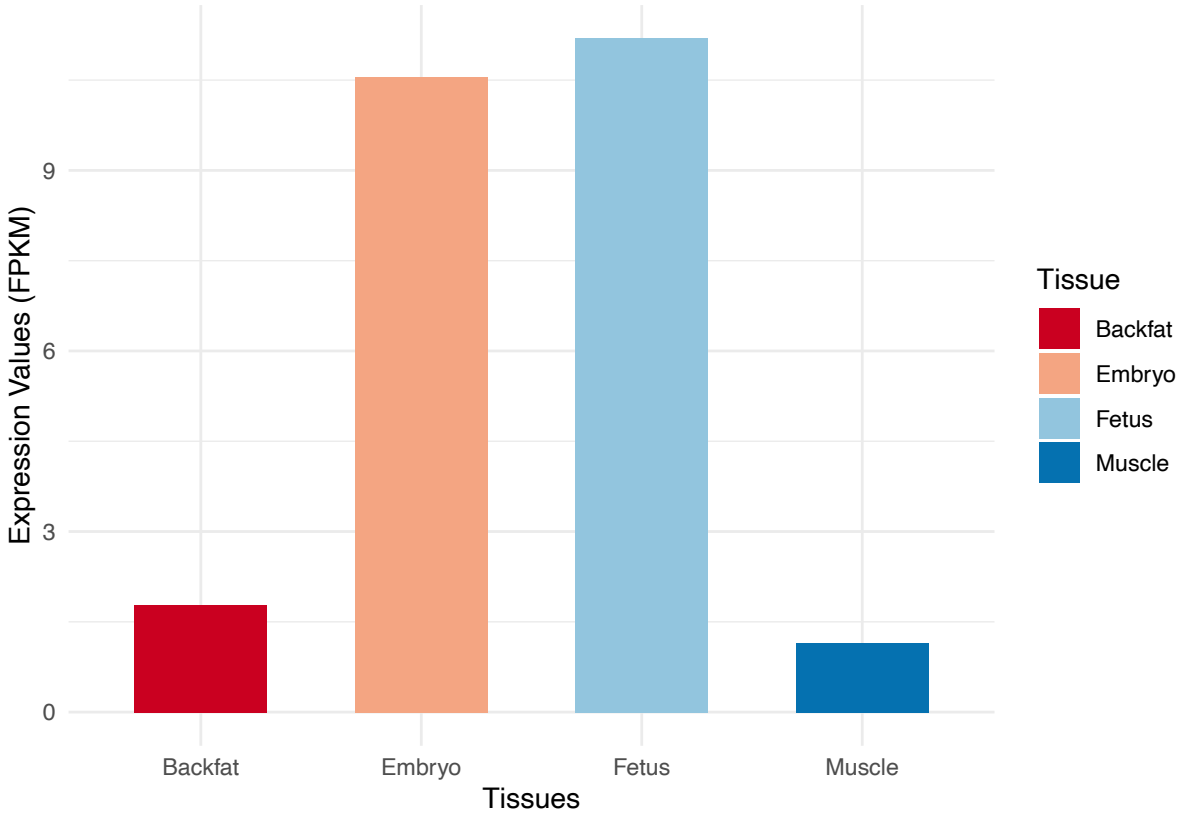

Figure 5: Expression values as FPKM of CCND2 gene in different tissues.

Haplotype-based association

We have performed a haplotype-based association analysis which we compared the association with the leadSNP and the core haplotype (GC**A**AG). The association with the core haplotype is lower than the leadSNP (Synthetic: $LogP_{LeadSNP}$ = 57.67; $LogP_{coreHaplotype}$ = 45.92, Pietrain: $LogP_{LeadSNP}$ = 43.29; $LogP_{coreHaplotype}$ = 40.08, Landrace: $LogP_{LeadSNP}$ = 67.61; $LogP_{coreHaplotype}$ = 40.68 and Large-White: $LogP_{LeadSNP}$ = 37.91; $LogP_{coreHaplotype}$ = 32.27). it means that the leadSNP may be the causal variant itself (**Table 9**).

Table 9: Haplotype-based association using number of copies of the core haplotype shows association lower than when used the leadSNP.

Line	SNP/haplotype	LogP	b	SE	VgExp	VpExp
Synthetic	AX-116265245[1]	57.67	0.36	0.02	0.04	0.02
	GCAAG[2]	45.92	0.29	0.02	0.03	0.02
Pietrain	AX-116265245	43.29	0.27	0.02	0.06	0.03
	GCAAG	40.08	0.26	0.02	0.05	0.03
Landrace	AX-116265245	67.61	0.24	0.01	0.05	0.02
	GCAAG	40.68	0.17	0.01	0.02	0.01
Large-White	AX-116265245	37.91	0.33	0.03	0.02	0.01
	GCAAG	32.27	0.30	0.03	0.02	0.01

LogP: Log P-value; b: Beta coefficient; SE: Standard Error; VgExp: Genetic Variance explained; VpExp: Phenotypic Variance explained;
[1] AX-116265245: Most significant SNP on the GWAS analysis (leadSNP);
[2] 2GCAAG: core haplotype

DISCUSSION

Backfat is one of the most important traits in pork production and it is closely related to the efficiency-related traits of pigs. The composition of lipid in backfat is mainly triacylglycerol (ROONGSITTHICHAI; TUMMARUK, 2014), and it is affected by the amount of fat and feed intake (WOOD et al., 1989). In a previous study, van Son et al. (2017) identified a QTL for fatty acid composition on SSC14 in two out of the pig populations evaluated in this study. In the current study we identified a QTL with strong association with BFE on SSC5 on four out of the five evaluated populations.

The most significant SNP was the same in all four population and explains up to 6% of the genetic variation and 3% of the phenotypic variance of BFE in these four population. In the Duroc population, where this QTL region was not identified, the most significant SNP identified in the other populations is not segregating, being fixed for the G allele, which is associated with increased backfat. This indicates that this region has been heavily selected already in the Duroc population, a breed known for excels on meat quality traits.

Several studies have reported QTLs and candidate genes to possibly identify functional mutations in different chromosomes associated with backfat in pigs (FONTANESI et al., 2012; PANT et al., 2015; ZHU et al., 2014). However, we have identified one strong candidate SNP (leadSNP) located on SSC5:66103958 which is shared across breeds according to the performed fine mapping. Pietrain, Synthetic, and Landrace breeds have high frequency for the A allele, which is associated with less backfat, while Large White has high frequency for the G allele which is associated with more backfat, however, Large White has the lowest p-values among the four breeds.

The leadSNP is present within intron 3 of the *CCND2* gene which is essential for control of secreting beta-cells and growth of pancreatic islet (KUSHNER et al., 2005; LU et al., 2009). Nevertheless, the core haplotype is present in a region taking *CCND2* gene and LOC106507534 (a non-conding RNA (ncRNA) - https://www.ncbi.nlm.nih.gov/gene/106507534). Using the human genome annotation, the *CCND2* gene region is located on chromosome 12 and *LOC106507534* (pig annotation) seems to be a long non-coding RNA (lncRNA) named *CCND2-AS1* (human annotation). This lncRNA is associated with regulation of Wnt/b-catenin signaling which directly regulates cell proliferation, cell polarity and cell fate determination during embryo development and tissue homeostasis (LOGAN; NUSSE, 2004; ZHANG et al., 2017).

For fine mapping the core haplotype region we used both animals which were genotyped and had the whole genome sequenced. From the core region with seven SNPs (~45kb) we narrowed down to a region with three variants (~6kb) with high allele frequency across

breeds. Two variants (SSC5:66097445 and SSC5:66099282) are within intron 4 and the variant SSC5:66103958 is located within intron 3 of *CCND2* gene. Le et al., (2017) has identified the *CCND2* gene as a possible candidate for back conformation in Landrace.

Since we have identified the three candidate variants, we searched for regulatory elements using ChIP-Seq and RNA-Seq data. Using ChiP-Seq data from pig alveolar macrophages cultured for 2 and 6 hours we have found H3K27me3 and H3K4me3 marks on the region of the three variants, indicating that there are histone modifications during the embryo development (ERHARDT et al., 2003; VAN DER HEIJDEN et al., 2005; WANG; DEY, 2006; ZHANG et al., 2009) which are associated with gene repression and activation, respectively (RINGROSE; PARO, 2004; SCHUETTENGRUBER et al., 2007). We have not found any epigenetic mark around the variants using data from muscle and liver. In addition, we used RNA-Seq data from backfat and muscle tissues in which *CCND2* gene was expected to be highly expressed (PÉREZ-MONTARELO et al., 2014), however, the expression level in these tissues was lower than the expected. Using RNA-Seq from fetuses and embryos we identified high expression for the CCND2 gene. High expression of *CCND2* is associated with adipose tissue development and differentiation (VAITTINEN et al., 2011).

Since 1964 is already known that histones are modified post-translationally (ALLFREY; FAULKNER; MIRSKY, 1964) and there exist many types of histone modification, such as acetylation, phosphorylation, methylation and others, however, histone methylation does not alter the overall charge of the molecule (XHEMALCE; DAWSON; BANNISTER, 2011). In this study we have used histone methylation marks that occur mainly, but not only, on the side chains of lysine residuals and it can be mono-, di- or tri-methylated, meaning that the lysine methyltransferase can catalyze one, two or three methyl groups on the histone tails during the methylation and may have different functional consequences. H3K4me3 and H3K27me3 are histone marks trimethylated and, according to BERNSTEIN et al., (2006), they may be important for Embryonic Stem cell differentiation and self-renew. Furthermore, these marks are active and repressive, respectively, in the regulation of developmental important genes on stem cells during pre-natal stage in mammals (BOGLIOTTI; ROSS, 2012; LI, 2002). We hypothesize that the expression of *CCND2* gene has been regulated during the pre-natal development stage and these results suggest that the causal variant might be located in the region of SSC5:66097445-66103958 which is regulating the *CCND2* gene.

Due to the low LD between the leadSNP and the other variants in the core haplotype, we hypothesized that the Lead SNP could be the causal variant itself. Haplotype-based association analysis has shown that the association with the core haplotype is lower than the association with the leadSNP, it may confirm our hypothesis that the leadSNP is the causal variant itself.

Although the association found on chromosome five is high, the genetic variant explained

is low, up to 6%. Duroc population was the only one which did not identify the QTL, due to the leadSNP is already fixed for the G allele, which is related to increasing backfat. It is interesting because Duroc has been selected for meat quality and it may indicate that this SNP or gene can have some influence in other traits related to backfat and meat quality.

It is also interesting to note that breeds which have the lowest average backfat, such as Pietrain and Landrace, the G allele is still segregating at high frequency. Finally, as expected, Large-White population has the highest backfat and the highest frequency for the G allele. However, the Synthetic breed has the second higher average backfat and shows the lowest frequency of the G allele.

CCND2 gene is somehow involved and is a strong candidate gene for backfat deposition in pig, nevertheless, more specific data such as ChiP-Seq and RNA-Seq from embryonic stage tissue from the same population is needed to have a better conclusion and better understanding how this gene is involved in this process. Furthermore, we could demonstrate that the impact of epigenetic and genetic coding in embryonic or fetal life can greatly impact production traits in adult animals.

CONCLUSION

In thisstudy we identified a strong QTL on SSC5 associated with backfat in four out of five pig breeds. The QTL was not identified in Duroc's breed due to that SNP is not segregating anymore in this population, seems to be already fixed. The leadSNP, AX-116265245, explaining up to 6 percent of genetic variance is high associated with backfat on the other breeds. Fine mapping of this region on SSC5 showed that there are three candidate variants associated with decreasing backfat located in intronic regions of the *CCND2* gene and the leadSNP may be the causal mutation for this association. Moreover, *CCND2* gene has shown to be high expressed during embryo development and it seems to have some influence on expressing the phenotype in the adult life.

REFERENCES

ALLFREY, G.; FAULKNER, R.; MIRSKY, A. E. ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS. **Biochemistry**, v. 315, n. 1938, p. 786–794, 1964.

BERNSTEIN, B. E. et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. **Cell**, v. 125, n. 2, p. 315–326, 2006.

BOGLIOTTI, Y. S.; ROSS, P. J. Mechanisms of histone H3 lysine 27 trimethylation remodeling during early mammalian development. **Epigenetics**, v. 7, n. 9, p. 976–981, 7 set. 2012.

BOSI, P.; RUSSO, V. The production of the heavy pig for high quality processed products. **Italian Journal of Animal Science**, v. 3, n. 4, p. 309–321, jan. 2004

BROWNING, S. R.; BROWNING, B. L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. **The American Journal of Human Genetics**, v. 81, n. 5, p. 1084–1097, nov. 2007.

CHEN, J. N. et al. Distribution of H-FABP and ACSL4 gene polymorphisms and their associations with intramuscular fat content and backfat thickness in different pig populations. **Genetics and Molecular Research**, v. 13, n. 3, p. 6759–6772, 2014.

DO, D. N. et al. Genome-wide association and systems genetic analyses of residual feed intake, daily feed consumption, backfat and weight gain in pigs. **BMC Genetics**, v. 15, n. 1, p. 1–15, 2014.

ERHARDT, S. et al. Consequences of the depletion of zygotic and embryonic enhancer of zeste 2 during preimplantation mouse development. **Development**, v. 130, n. 18, p. 4235–4248, 2003.

FONTANESI, L. et al. A genome wide association study for backfat thickness in Italian Large White pigs highlights new regions affecting fat deposition including neuronal genes. **BMC Genomics**, v. 13, n. 1, 2012.

FOWLER, K. E. et al. Genome wide analysis reveals single nucleotide polymorphisms associated with fatness and putative novel copy number variants in three pig breeds. **BMC Genomics**, v. 14, n. 1, p. 1, 2013.

GARRISON, E.; MARTH, G. Haplotype-based variant detection from short-read sequencing. 17 jul. 2012.

GILMOUR, A R. et al. ASReml user guide release 3.0. **VSN International Ltd**, p. 275, 2009.

HANSON, C. et al. Influence of diet and obesity on COPD development and outcomes. **International Journal of COPD**, v. 9, p. 723–733, 2014.

KUMAR, P.; HENIKOFF, S.; NG, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. **Nature Protocols**, v. 4, n. 7, p. 1073–1081, 25 jul. 2009.

KUSHNER, J. A. et al. Cyclins D2 and D1 Are Essential for Postnatal Pancreatic -Cell Growth. **Molecular and Cellular Biology**, v. 25, n. 9, p. 3752–3762, 2005.

LE, T. H. et al. Genome-wide association study for conformation traits in three Danish pig breeds. **Genetics Selection Evolution**, v. 49, n. 1, p. 1–12, 2017.

LI, E. Chromatin modification and epigenetic reprogramming in mammalian development. **Nature Reviews Genetics**, v. 3, n. 9, p. 662–673, set. 2002.

LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078–2079, 15 ago. 2009.

LI, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 16 mar. 2013.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754–1760, 2009.

LOGAN, C. Y.; NUSSE, R. the Wnt Signaling Pathway in Development and Disease. **Annual Review of Cell and Developmental Biology**, v. 20, n. 1, p. 781–810, 2004.

LONERGAN, S. M. et al. Selection for lean growth efficiency in Duroc pigs influences pork quality. **Journal of Animal Science**, v. 79, n. 8, p. 2075, 2001.

LU, M. H. et al. Cyclin D2 protein stability is regulated in pancreatic β-cells. **Molecular Endocrinology**, v. 23, n. 11, p. 1865–1875, 2009.

MCLAREN, W. et al. The Ensembl Variant Effect Predictor. **Genome Biology**, v. 17, n. 1, p. 122, 6 dez. 2016.

MERKS, J. W. M. One century of genetic changes in pigs and the future needs. **BSAP Occasional Publication**, v. 27, p. 8–19, 27 fev. 2000.

OKUMURA, N. et al. Genomic regions affecting backfat thickness and cannon bone circumference identified by genome-wide association study in a Duroc pig population. **Animal Genetics**, v. 44, n. 4, p. 454–457, ago. 2013.

PANT, S. D. et al. Comparative analyses of QTLs influencing obesity and metabolic phenotypes in pigs and humans. **PLoS ONE**, v. 10, n. 9, p. 1–22, 2015.

PÉREZ-MONTARELO, D. et al. Identification of genes regulating growth and fatness traits in pig through hypothalamic transcriptome analysis. **Physiological Genomics**, v. 46, n. 6, p. 195–206, 15 mar. 2014.

RAJENDER RAO, K.; LAL, N.; GIRIDHARAN, N. V. **Genetic & epigenetic approach to obesityIndian Journal of Medical Research**Indian Council of Medical Research, , 1 nov. 2014.

RINGROSE, L.; PARO, R. Epigenetic Regulation of Cellular Memory by the Polycomb and Trithorax Group Proteins. **Annual Review of Genetics**, v. 38, n. 1, p. 413–443, dez. 2004.

ROONGSITTHICHAI, A.; TUMMARUK, P. Importance of backfat thickness to reproductive performance in female pigs. **Thai Journal of Veterinary Medicine**, v. 44, n. 2, p. 171–178, 2014.

SARGOLZAEI, M.; CHESNAIS, J. P.; SCHENKEL, F. S. A new approach for efficient genotype imputation using information from relatives. **BMC Genomics**, v. 15, n. 1, 2014.

SCHUETTENGRUBER, B. et al. Genome Regulation by Polycomb and Trithorax Proteins. **Cell**, v. 128, n. 4, p. 735–745, fev. 2007.

TRAPNELL, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated ranscripts and isoform switching during cell differentiation. **Nature Biotechnology**, v. 28, n. 5, p. 511–515, 2 maio 2010.

VAITTINEN, M. et al. Microfibrillar-associated protein 5 is linked with markers of obesity-related extracellular matrix remodeling and inflammation. **Nutrition & Diabetes**, v. 1, n. 8, p. e15–e15, 2011.

VAN DER HEIJDEN, G. W. et al. Asymmetry in Histone H3 variants and lysine methylation between paternal and maternal chromatin of the early mouse zygote. **Mechanisms of Development**, v. 122, n. 9, p. 1008–1022, 2005.

WANG, H.; DEY, S. K. Roadmap to embryo implantation: Clues from mouse models. **Nature Reviews Genetics**, v. 7, n. 3, p. 185–199, 2006.

WANG, S.; PENG, D. Q.; YI, Y. The unsolved mystery of apoA-I recycling in adipocyte. **Lipids in Health and Disease**, v. 15, n. 1, p. 1–8, 2016.

WARR, A. et al. An improved pig reference genome sequence to enable pig genetics and genomics research. **bioRxiv**, p. 668921, 2019.

WESTESSON, O.; SKINNER, M.; HOLMES, I. Visualizing next-generation sequencing data with JBrowse. **Briefings in Bioinformatics**, v. 14, n. 2, p. 172–177, 1 mar. 2013.

WOOD, J. D. et al. Backfat composition in pigs: Differences between fat thickness groups and sexes. **Livestock Production Science**, v. 22, n. 3–4, p. 351–362, 1989.

WOOD, J. D. et al. Fat deposition, fatty acid composition and meat quality: A review. **Meat Science**, v. 78, n. 4, p. 343–358, abr. 2008.

XHEMALCE, B.; DAWSON, M. A.; BANNISTER, A. J. Histone Modifications. In: **Encyclopedia of Molecular Cell Biology and Molecular Medicine**. Weinheim, Germany: Wiley-

VCH Verlag GmbH & Co. KGaA, 2011.

YANG, J. et al. GCTA: A Tool for Genome-wide Complex Trait Analysis. **The American Journal of Human Genetics**, v. 88, n. 1, p. 76–82, jan. 2011.

YANG, J. et al. Advantages and pitfalls in the application of mixed-model association methods. **Nature Genetics**, v. 46, n. 2, p. 100–106, fev. 2014.

ZHANG, H. et al. Highly expressed lncRNA CCND2-AS1 promotes glioma cell proliferation through Wnt/β-catenin signaling. **Biochemical and Biophysical Research Communications**, v. 482, n. 4, p. 1219–1225, 2017.

ZHANG, M. et al. Defective chromatin structure in somatic cell cloned mouse embryos. **Journal of Biological Chemistry**, v. 284, n. 37, p. 24981–24987, 2009.

ZHANG, Y. et al. Model-based Analysis of ChIP-Seq (MACS). **Genome Biology**, v. 9, n. 9, p. R137, 2008.

ZHU, D. et al. Genome-wide association study of the backfat thickness trait in two pig populations. **Frontiers of Agricultural Science and Engineering**, v. 1, n. 2, p. 91–95, 2014.

GENERAL CONCLUSIONS

In the current research we have shown the applicability of Next Generation Sequencing (NGS) data in Animal Science, presented different possibilities to use for these types of data and how to do it, as well as the benefits of its use in genetic breeding studies. NGS has been used widely in the agriculture field, especially in Animal Breeding. Along three chapters we have shown its application at development of a tool for transcriptome analysis and, analyzing quantitative and health traits in pigs and chickens, contributing for a better understanding of the molecular genetics of backfat in the pig and bacterial chondronecrosis with osteomyelitis in birds.

The second chapter in this thesis has described the development of a pipeline for NGS analysis. BAQCOM is mostly written in R language and its usage is easy and simple. Performing quality control, mapping and counting reads using the best available software. In multithread processing and multi samples, the pipeline can facilitate the command line for those who are not comfortable with the UNIX environment. Moreover, BAQCOM has shown to be efficient and fast in its analysis. Although the main purpose of BAQCOM is to analyze RNA-Seq data, it can be used for other type of NGS data such as whole genome sequence (WGS) and metagenomics as well. Additionally, there is the flexibility to integrate other genomic tools in the future, such as fine-mapping analysis.

In the first published paper (third chapter), entitled "RNA-seq reveals downregulated osteochondral genes potentially related to tibia bacterial chondronecrosis with osteomyelitis in broilers" we found, using RNA-Seq, 16 differentially expressed genes in the tibia growth plate transcriptome that are directly responsible for bone and cartilage formation, which were downregulated in Bacterial Chondronecrosis with Osteomyelitis (BCO)-affected broilers. According to our data, the lack of ossification might be the main cause of BCO in broilers, and the bacterial process seems to be a secondary condition. Moreover, our results highlight the pathogenesis of BCO, and show that to pursue prevention and control of such condition, breeding strategies have to focus on the improvement of ossification and cartilage formation. Moreover, we conclude that the use of RNA-Seq data is suitable for discover the molecular cause of diseases.

In the second paper (fourth chapter), we have identified a QTL on SSC5 with strong association with backfat in four pig breeds. Using genotype data, we have found a high significant SNP which have led us to fine map the QTL region using WGS, RNA-Seq and ChIP-Seq

data finding three candidate variants associated with decreasing backfat located in intronic regions of the CCND2 gene. Using ChIP-Seq data we were led to believe that there are an epigenetic modifications during the embryonic stage with histone marks such as H3K27me3 and H3K4me3 in the studied genomic region. Indeed, using RNA-Seq of embryos and fetuses we could testify that there is high expression of CCND2 gene in the same genomic region.

In Animal Breeding there is a long history of working with genomic selection using only genotyping and phenotyping data, trying to identify quantitative trait loci and their application. It might be because animal breeding has put more focus on statistical approaches and prediction models and less focus on molecular data. However, they should be working together, filling in the gaps of each approach. In this thesis, we have used genotyping, phenotyping, transcriptome, ChIP-Sequence and WGS data, integrating all together to narrow down a QTL region. These results and the proposed tool in the current thesis are valuable for a better understanding of the genetic architecture and to improve the livestock breeds in the future.