GENE-SWitCH

The regulatory GENomE of SWine and CHicken: functional annotation during development

Deliverable D4.2
Significant associations detected by eQTL analysis in pigs

Deliverable leader: IRTA

Authors: Daniel CRESPO-PIAZUELO (IRTA) and Maria BALLESTER (IRTA)

Version: 1.0

Due date of deliverable (as in DoA) :	M24
Actual submission date:	M27

Dissemination level:

PU Public	**X**
CO Confidential, only for members of the consortium (including the Commission Services)	

Research and Innovation Action, SFS-30-2018-2019-2020 Agri-Aqua Labs
Duration of the project: 01 July 2019 – 30 June 2023, 48 months

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817998

Table of contents

1 Summary 3
2 Introduction 4
3 Results 4
3.1 Significant associations detected by eQTL analysis in duodenum 5
3.2 Significant associations detected by eQTL analysis in liver 5
3.3 Significant associations detected by eQTL analysis in muscle 5
3.4 Comparison of the significant associations found between tissues 6
4 Conclusion 8
5 Deviations or delays 8
6 Acknowledgements 8
7 References 8
8 Annexes 9
9 Glossary 9

1 Summary

- **Objectives:** The genetic variants that are significantly associated with a trait of interest are called quantitative trait loci (QTLs), and they are named expression-QTLs (eQTLs) if the association is found between the variant and the expression of a gene. In the GENE-SWitCH project, we aim to identify eQTLs in three porcine tissues (i.e., small intestine (duodenum), liver and muscle).

- **Method:** Duodenum, liver and muscle samples were obtained at slaughter from 300 pigs from three different breeds (n=100 Duroc, n=100 Landrace and n=100 Large White) and RNA was extracted using spin column-based kit and sequenced on the Illumina NovaSeq6000 platform. In addition, the same platform (NovaSeq6000) was used to sequence the whole genome of these 300 pigs, obtaining 44,127,400 genetic variants with GATK/4.1.8.0. A total of 25,315,878 polymorphisms were kept after filtering out those with more than 10% missing genotype data, and those with a minor allele frequency below 5%. RNA sequences were mapped against the reference genome Sscrofa11.1 with STAR/v2.5.3a, counts were quantified by RSEM/1.3.0 and normalized by TMM (trimmed mean of M-values). Genes with low expression (counts per million below 10/minimum library size in millions) and those that did not show expression in at least 5% of the animals were removed. Expression genome wide association studies (eGWAS) were conducted between the filtered polymorphisms and the normalized expression data using the fastGWA tool from GCTA/1.93.2, following a model where sex and breed were included as fixed effects. Bonferroni correction was applied afterwards considering only associations with an adjusted *p*-value <0.05 to be significant.

- **Main Results:** eGWAS were conducted for the normalized expression of 16,753 genes in duodenum, 15,710 genes in liver, and 13,887 genes in muscle. In duodenum, a total of 4,802,045 significant associations were found between 2,813,177 polymorphisms and the expression of 6,551 genes. In liver, 7,024,941 significant associations were found between 4,187,249 polymorphisms and the expression of 7,433 genes. In muscle, 8,099,604 significant associations were found between 4,814,732 polymorphisms and the expression of 7,496 genes. The genomic position of the significantly associated polymorphisms was also studied to assess if they were *cis*-regulatory elements (regulating a gene located at 1Mb or less from them). Out of the total number of significantly associated polymorphisms, the proportion of *cis*-regulatory elements was the highest across the three tissues (duodenum: 80.3%; liver: 89.7%; and muscle: 68.2%).

- **Teams involved:**

  Animal Breeding and Genetics Program, IRTA

  Hendrix Genetics Research Technology & Services B.V.

  Hypor B.V.

  IFIP-Institut du porc and Alliance R&D

  INRAE GABI

2 Introduction

In the last 10 years, the reduction in costs of genotyping and whole genome sequencing made genomic selection a powerful tool for animal breeding. This tool takes advantage of genetic variations (polymorphisms) that are present in the genome of an individual and makes it different from the rest. The idea is that the differences in the phenotypes of interest (e.g., carcass weight, coat colour, or levels of white blood cells) are driven by these genetic variants, and as such can be transmitted to the next generation. Sometimes, a polymorphism falls inside a gene region that modifies the protein that is encoded by this gene, making it clearly visible on the phenotype. However, more than 90% of polymorphisms that are associated with the traits of interest are not located on the coding regions of a gene, but within intergenic and intronic regions. These regions of the genome can regulate gene expression, which can be considered as "intermediate phenotypes", expected to be more closely related to genetic information than conventionally measured trait phenotypes. The genetic variants that are significantly associated with a trait of interest are called quantitative trait loci (QTLs), and they are named expression-QTLs (eQTLs) if the association is found between the polymorphism and the expression of a gene. In addition, we talk about *cis*-eQTL if the variant is associated to the expression levels of a nearby gene, and *trans*-eQTL if the variant is far away from the gene or even in another chromosome. The study of eQTLs can be performed with genetic markers located across all the genome using expression genome wide association studies (eGWAS). This method has the advantage that can be easily parallelizable, as the associations are performed individually, reducing computing times considerably. The identification of these genetic variants associated with gene expression levels will contribute to shed light on the relationship between genetic variation and end-trait phenotypes.

3 Results

After sequencing the whole genome of the 300 pigs, 44,127,400 genetic variants were obtained. A total of 25,315,878 polymorphisms were kept after filtering out those with more than 10% missing genotype data, and those with a minor allele frequency below 5% (**Table 1**). Regarding the type of genetic variation, the polymorphisms were classified as SNPs (74.92%), deletions (11.20%) and insertions (13.88%).

Table 1. Distribution per chromosome of the 25,315,878 polymorphisms detected by whole genome sequencing after the filtering step.

Chromosome	No. of polymorphisms
1	2,274,009
2	1,598,944
3	1,481,069
4	1,419,464
5	1,140,001
6	1,843,185
7	1,361,758
8	1,689,350
9	1,474,232
10	1,139,451
11	1,053,896
12	888,676
13	1,890,576
14	1,586,643

15	1,398,896
16	1,082,114
17	883,858
18	665,364
X	443,425
Y	728
MT	239

After the RNA-seq mapping and quantification processes, eGWAS were conducted for the normalized expression of 16,753 genes in duodenum, 15,710 genes in liver, and 13,887 genes in muscle. Regarding these three normalized datasets, 18,097 genes were expressed at least in one tissue, whereas the three tissues had in common the expression of 12,892 genes.

3.1 Significant associations detected by eQTL analysis in duodenum

In duodenum, 4,802,045 significant associations were found between 2,813,177 polymorphisms and the expression of 6,551 genes. A total of 23,403 polymorphisms were associated with the expression of 10 or more genes and were considered hotspot regulatory elements. Regarding their genomic position, 2,260,300 polymorphisms (80.3%) were annotated as *cis*-regulatory elements, as they were located at 1Mb or less than their associated gene. The most significantly associated variant (adj.*p*-value=4.55×10^{-216}) was located on an intergenic region, close to the novel gene (ENSSSCG00000051495) that encodes a lncRNA, whose expression was being *cis*-regulated by this variant. From the 2,813,177 significantly associated polymorphisms, 849,810 were novel variants (30.2%) and the remaining 1,963,367 were existing variants (69.8%), already described in the Ensembl database.

3.2 Significant associations detected by eQTL analysis in liver

In liver, 7,024,941 significant associations were found between 4,187,249 polymorphisms and the expression of 7,433 genes. A total of 24,216 polymorphisms were associated with the expression of 10 or more genes and were considered hotspot regulatory elements. Regarding their genomic position, 3,756,049 polymorphisms (89.7%) were annotated as *cis*-regulatory elements, as they were located at 1Mb or less than their associated gene. The most significantly associated variant (adj.*p*-value=9.99×10^{-300}) was also associated with the expression levels of the aforementioned novel gene (ENSSSCG00000051495) and was located on the 3'-UTR region of the 6-Phosphofructo-2-Kinase/Fructose-2,6-Biphosphatase 1 (*PFKFB1*) gene and on an intronic region of the trophinin (*TRO*) gene. From the 4,187,249 significantly associated polymorphisms, 1,254,419 were novel variants (30.0%) and the remaining 2,932,830 were existing variants (70.0%).

3.3 Significant associations detected by eQTL analysis in muscle

In muscle, 8,099,604 significant associations were found between 4,814,732 polymorphisms and the expression of 7,496 genes. A total of 33,569 polymorphisms were associated with the expression of 10 or more genes and were considered hotspot regulatory elements. Regarding their genomic position, 3,283,835 polymorphisms (68.2%) were annotated as *cis*-regulatory elements, as they were located at 1Mb or less than their associated gene. The most significantly associated variant (adj.*p*-value=2.66×10^{-178}) was located on an intronic region of the SRSF Protein Kinase 3 (*SRPK3*) gene, which encodes a protein associated to muscle development. From the 4,814,732 significantly associated variants, 1,438,442 were novel variants (29.9%) and the remaining 3,376,290 were existing variants (70.1%).

3.4 Comparison of the significant associations found between tissues

A total of 19,926,590 significant associations were found for the expression of 6,551 genes in duodenum, 7,433 genes in liver, and 7,496 genes in muscle. From these genes with significant associations, 1,730 were found overlapping the three tissues (**Figure 1**). Muscle had the highest number of unique genes with significant associations (2,688), followed by liver (2,349), whereas duodenum had the lowest number (2,179). Duodenum and muscle were the tissues that shared the lowest number of genes with significant associations (1,183), while the highest number of shared genes with significant associations was found between liver and muscle (1,895).

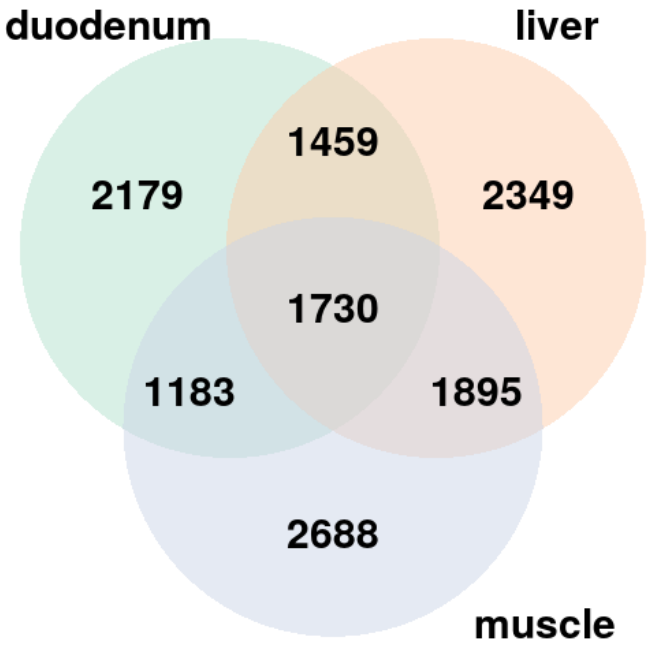

Figure 1. Venn diagram representing the overlap in genes with significant associations between the three studied tissues (duodenum, liver, and muscle).

Regarding the distribution of significant associations across *Sus scrofa* chromosomes (SSC) for the three tissues (**Figure 2**), SSC6, SSC12 and SSC14 had the highest number of them. Because duodenum had the lowest number of significant associations, the number of significant associations found in liver and muscle was higher in almost all the chromosomes. However, liver had higher number of significant associations than muscle in few instances (SSC3, SSC10 and SSC14).

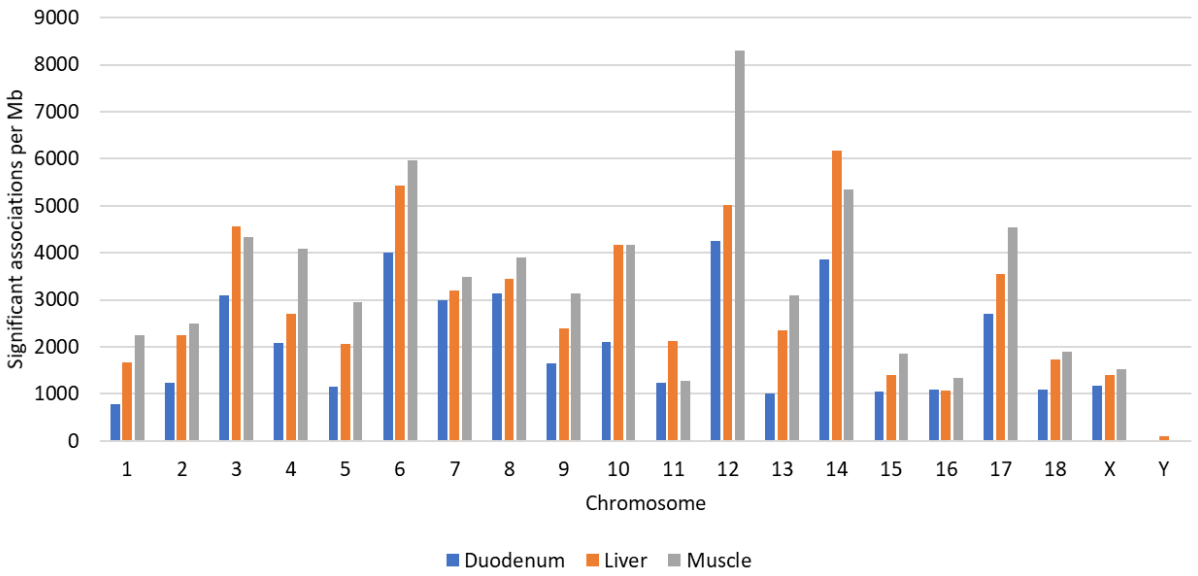

Figure 2. Number of significant associated variants per Mb across all chromosomes in the three studied tissues (duodenum, liver, and muscle).

Variant effect predictor (VEP) tool was used to evaluate the position of the significantly associated polymorphisms over genomic regions and predict their possible consequences (**Table 2**). No clear differences were observed among the three tissues, although it is worth to note that most of the significantly associated polymorphisms (65%) were located within intronic regions, while the number of significantly associated polymorphisms located on intergenic regions was lower (13-14%).

Table 2. List of the predicted consequences by the VEP tool regarding the genomic position and the alternative alleles of the significant polymorphisms found in the three studied tissues (duodenum, liver, and muscle).

Consequences (all)	**Duodenum**	**Liver**	**Muscle**
intron_variant:	65%	65%	65%
intergenic_variant:	13%	14%	14%
downstream_gene_variant:	8%	8%	8%
upstream_gene_variant:	8%	8%	7%
non_coding_transcript_variant:	3%	3%	3%
3_prime_UTR_variant:	1%	1%	1%
non_coding_transcript_exon_variant:	1%	1%	1%
synonymous_variant:	1%	1%	1%
5_prime_UTR_variant:	<1%	<1%	<1%
others:	<1%	<1%	<1%

Coding consequences	**Duodenum**	**Liver**	**Muscle**
synonymous_variant:	61%	62%	62%
missense_variant:	33%	32%	32%
frameshift_variant:	4%	4%	4%
inframe_insertion:	1%	1%	1%

inframe_deletion:	1%	1%	1%
others:	<1%	<1%	<1%

The datasets containing all the significant associations detected by eQTL analysis in the three tissues will be released in a scientific publication on an open access journal.

4 Conclusion

A total of 25,315,878 polymorphisms were identified by whole genome sequencing of 300 pigs representing three different breeds. From the same animals, 900 transcriptomes were produced using mRNA from 3 different tissues (duodenum, liver and skeletal muscle). Combining these datasets, eGWAS was performed and 19,926,590 eQTLs were obtained in duodenum, liver, and muscle after performing more than 1.17×10^{12} combinations, which took 68.2 days of CPU time. These key regulatory elements identified in our results may be included in new predictive models to increase the accuracy of genomic predictions, speeding up the rate of genetic improvement of economically important traits in pigs.

5 Deviations or delays

This deliverable was originally scheduled for submission in M24, and it was delayed to M27 due to the temporary closure of the sequencing centre (CNAG, Barcelona) during the start of the COVID19 pandemic in March 2020.

6 Acknowledgements

Not applicable.

7 References

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1), 323. https://doi.org/10.1186/1471-2105-12-323

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

8 Annexes

Not applicable.

9 Glossary

eGWAS: expression genome wide association studies

eQTLs: expression-QTLs

lncRNA: long non-coding RNA

QTLs: quantitative trait loci

SNP: single nucleotide polymorphism

SSC: *Sus scrofa* chromosome