Deciphering genetic variants from whole genome affecting duodenum, liver and muscle transcriptomes in pigs

D. Crespo-Piazuelo[1*], O. González-Rodríguez[1], M. Mongellaz[2], H. Acloque[2], M.-J. Mercat[3], M.C.A.M. Bink[4], A.E. Huisman[4], Y. Ramayo-Caldas[1], J.P. Sánchez[1] and M. Ballester[1]

[1] Animal Breeding and Genetics Program, IRTA, Torre Marimon, 08140 Caldes de Montbui, Spain; [2] INRAE GABI, Domaine de Vilvert, 78350 Jouy-en-Josas, France; [3] IFIP-Institut du porc and Alliance R&D, La Motte au Vicomte, 35651 Le Rheu, France; [4] Hendrix Genetics, P.O. Box 114, 5830 AC Boxmeer, the Netherlands; [*]daniel.crespo@irta.cat

Abstract

With the aim of deciphering genetic variants affecting gene expression levels of duodenum, liver and muscle, the DNA and RNA of 300 pigs from three different breeds (Duroc, Landrace and Large White) was extracted and sequenced. After filtering and normalizing both the genomic and the transcriptomic datasets, expression genome-wide association studies (eGWAS) were conducted between 25,315,878 polymorphisms and the expression of 16,753 genes in duodenum, 15,710 genes in liver, and 13,887 genes in muscle. In total, more than 1.17×10^{12} combinations were performed, obtaining 19,926,590 significantly associated polymorphisms, which were grouped in 148,639 expression quantitative trait locus (eQTL) regions. Half of the polymorphisms were located in close proximity to their associated gene, but only 7,773 *cis*-eQTL regions were defined, the remaining being *trans*-eQTLs regions. Out of the cis-eQTL regions, 18 shared the same top polymorphism among the three tissues; thus, they were considered the strongest candidates for gene expression regulation.

Introduction

Rather than the usual measured phenotypes, gene expression acts as an "intermediate phenotype", which is expected to be closer to the genomic information than a conventional end-trait phenotype (Stranger *et al.*, 2007). In this regard, genetic variants affecting gene-expression phenotypes are an important source of phenotypic variation. Over the last decade, expression genome-wide association studies (eGWAS) have become one of the most used methods to study the association of the polymorphisms distributed across the genome and the expression levels of a gene. At the time, this method used a limited dataset of genetic markers, but nowadays all the genetic variants found in the genomes of a population can be included, thanks in particular to the progress made to improve computing efficiency. In addition, RNA sequencing approaches have become cheaper and can be applied to larger datasets. This is crucial to understand the relationship between gene regulation mechanisms and the traits of interest, since more than 90% of the phenotype-associated polymorphisms are located within intergenic and intronic regions, rather than on the coding regions of a gene (Maurano *et al.*, 2012). When a location in the genome is found in association with the expression of a certain gene, an expression quantitative trait locus (eQTL) region is defined. Classically, *cis*-eQTL regions have been defined at 1Mb of the gene analysed; otherwise, they are considered *trans*-eQTL regions. In this study we will aim to decipher potential genetic variants that could modify gene expression levels in duodenum, liver and muscle.

Materials & Methods

Duodenum, liver and muscle samples were collected at slaughter from 300 pigs of three different breeds (n=100 Duroc, n=100 Landrace and n=100 Large White). Genomic DNA was extracted from blood (Duroc and Landrace) and liver (Large White) samples using NucleoSpin Blood kit (Macherey-Nagel, Düren, Germany). RNA from muscle samples was extracted using RiboPure™ RNA Purification Kit (Invitrogen, Carlsbad, CA, USA) and RNeasy Fibrous Tissue Mini Kit (Qiagen, Hilden, Germany), and RNA from liver and duodenum samples was extracted using a chemagic™ 360 instrument with RNA Tissue10 Kit H96 (PerkinElmer, Baesweiler, Germany). The whole genome of the 300 pigs and the transcriptome from the three tissues (duodenum, liver and muscle; n=900) were paired-end (2×150bp) sequenced in an Illumina NovaSeq6000 platform (Illumina, San Diego, CA, USA). DNA sequences were mapped against the reference genome (Sscrofa11.1 assembly) with BWA-MEM/0.7.17 (H. Li, 2013) and the genetic variants were obtained with GATK/4.1.8.0 HaplotypeCaller (McKenna *et al.*, 2010). Then, the genetic variants were filtered (minor allele frequency below 5% and/or more than 10% missing genotype data) for downstream analyses.

RNA sequences were mapped against the reference genome (Sscrofa11.1 assembly) with STAR/v2.5.3a (Dobin *et al.*, 2013), counts were quantified by RSEM/1.3.0 (B. Li and Dewey, 2011) 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. Thereafter, eGWAS were conducted between the filtered polymorphisms and the normalized expression data using the fastGWA tool from GCTA/1.93.2 (Yang *et al.*, 2011) using the following model:

$$y_{hijk} = sex_{hj} + breed_{hk} + u_{hi} + s_{il} \cdot a_{hl} + e_{hijk}$$

where y_{hijk} corresponds to the expression of the hth gene in the ith individual of sex j and belonging to the kth breed; sex_{hj} corresponds to the jth sex effect (two levels); $breed_{hk}$ to the kth breed effect (three levels); u_{hi} is the infinitesimal genetic effect of the individual i, with $\mathbf{u}\sim\text{MVN}(\mathbf{0},\mathbf{G}\cdot\sigma^2_u)$, where $\mathbf{G}$ is the genomic relationship matrix calculated using the filtered autosomal polymorphisms and σ^2_u is the additive genetic variance to be estimated; s_{il} is the genotype (coded as 0, 1 or 2) for the lth polymorphism; and a_{hl} is the allele substitution effect of the lth polymorphism on the expression level of the hth gene; finally, e_{hijk} is the residual term. Bonferroni correction was applied to assess the statistical significance for the association studies at genome-wide level with the p.adjust function of R. Only those associations with an adjusted *p*-value<0.05 were considered significant.

Results

The whole genome sequence of the 300 pigs resulted in 44,127,400 genetic variants, which after the filtering step, got reduced to a total of 25,315,878 polymorphisms that were kept for the association analyses. Regarding the type of genetic variant, the polymorphisms were classified as SNPs (74.92%), insertions (13.88%) and deletions (11.20%). After normalizing and filtering the three transcriptomic datasets, 12,892 genes were expressed in all the three tissues, whereas 18,097 genes were expressed in total. For each tissue, eGWAS were conducted for 16,753 genes in duodenum, 15,710 genes in liver, and 13,887 genes in muscle. The eGWAS resulted in a total of 19,926,590 significant associations among the three tissues, which are detailed in Table 1. From all the significantly associated polymorphisms in the three tissues, 29.9-30.2% were novel variants and the remaining 69.8-70.1% were already described in the Ensembl database. In this regard, a polymorphism could be associated with the expression of more than one gene. Thus, in duodenum, 2,813,177 unique polymorphisms were associated with the expression of 6,551 genes; in liver, 4,187,249 unique polymorphisms were associated with the expression of 7,433 genes, and in muscle, 4,814,732 polymorphisms were associated with the expression of 7,496 genes. Out of these genes with any significant association 1,730 were expressed in the three tissues.

Table 1. Number of significantly associated variants and eQTL regions per tissue.

Tissue	Significantly associated variants		eQTL regions	
	No.	*cis*-regulatory elements[1]	No.	*cis*-eQTL regions[2]
Duodenum	4,802,045	2,260,300 (47.1%)	53,236	1,866 (3.5%)
Liver	7,024,941	3,756,049 (53.5%)	36,835	2,704 (7.3%)
Muscle	8,099,604	4,461,375 (55.1%)	58,568	3,203 (5.5%)
Total	19,926,590	10,477,724 (52.6%)	148,639	7,773 (5.2%)

[1] A *cis*-regulatory element was considered if the polymorphism was located at a maximum distance of 1Mb from the associated gene.

[2] A *cis*-eQTL region was considered if the gene was located within it.

An eQTL region was defined ±1Mb from a significantly associated polymorphism, which were merged into the same eQTL region if they intersected. Thus, 148,639 eQTL regions were defined (Table 1). However, this number of eQTL regions should be carefully considered as 43.52-59.82% were defined exclusively by a single polymorphism: 23,167 in duodenum, 21,828 in liver, and 35,033 in muscle. In proportion, the ratio of *cis*-eQTL regions defined by a single polymorphism was much lower (5.62-7.61%): 142 in duodenum, 152 in liver, and 191 muscle. For each eQTL region, most of the significant polymorphisms were located within 1Mb of the top polymorphism of the region (Figure 1).

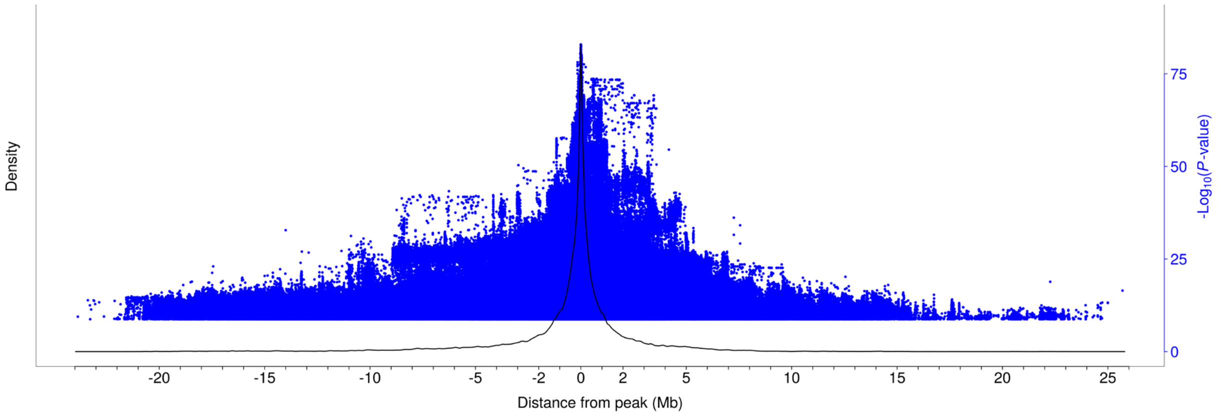

Figure 1. Density plot of the distance of the significant polymorphisms associated with liver gene expression to the peak of their eQTL region (sexual and mitochondrial chromosomes excluded). The significance of the polymorphisms is provided in blue dots.

Regarding the location of the polymorphisms in respect to the expression of their associated gene, 47.1-55.3% of the polymorphisms were closer than 1Mb to their gene. Despite so, only ~5% of the 148,639 eQTL regions defined were in *cis* (Table 1). When comparing the top polymorphisms of the eQTL regions found in all the three tissues, 332 genetic variants where shared between the three of them, being 18 of them cis-regulatory elements.

Discussion

In this study, more than 1.17×10^{12} combinations were performed through eGWAS in duodenum, liver, and muscle, obtaining 19,926,590 associated polymorphisms. These polymorphisms were grouped in 148,639 eQTL regions, but despite half of the polymorphisms being located in close proximity to their associated gene, only ~5% of the

eQTLs regions were in *cis*. A possible explanation of this result is due to the fact that most of the *trans*-eQTL regions were defined by a single polymorphism. Hence, to avoid such artifacts, filters to define eQTL regions with 2 or more polymorphisms may be required.
Considering the distribution of the *p*-values due to linkage disequilibrium, the definition of *cis* elements if they were located at less than 1Mb from their associated gene seemed appropriate, as they were rarely surpassing the 2Mb. However, this definition may change depending on the structure of the population used.
One of the top 18 *cis*-regulatory elements of the eQTL regions found in common in the three tissues was a polymorphism located upstream the *SUPT3H* gene, a gene already identified by our group as a potential key regulator for health-related traits in pigs (Crespo-Piazuelo *et al.*, 2021). Five out of the 18 *cis*-regulatory elements were located in long non-coding RNAs (lncRNAs). Of interest was the polymorphism located upstream the *ENSSSCG00000041410* gene, which encodes for a lncRNA that was positively correlated in all the three tissues (*r*=0.590-0.732) with the expression of *TRIM39*, a nearby gene involved in the immune response and cell cycle progression (Suzuki *et al.*, 2016; Zhang *et al.*, 2012), and that has been already described as participating in the same processes regulated by lncRNAs (Zhou *et al.*, 2019). In conclusion, the identification of genetic variants associated with gene expression levels in the three tissues will contribute to shed light on the molecular mechanisms of regulatory variations to shape end-trait phenotypes.

Acknowledgements: This project is part of GENE-SWitCH (https://www.gene-switch.eu) and has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement n° 817998. It is also part of EuroFAANG (https://eurofaang.eu), a synergy of five Horizon 2020 projects that share the common goal to discover links between genotype to phenotype in farmed animals and meet global Functional Annotation of ANimal Genomes (FAANG) objectives.

References

Crespo-Piazuelo D., Ramayo-Caldas Y., González-Rodríguez O., Pascual M., *et al.* (2021). Front. Immunol. 12:784978. https://doi.org/10.3389/fimmu.2021.784978

Dobin A., Davis C.A., Schlesinger F., Drenkow J., *et al.* (2013). 29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635

Li B., and Dewey C.N. (2011). BMC Bioinformatics 12(1):323. https://doi.org/10.1186/1471-2105-12-323

Li H. (2013). https://arxiv.org/abs/1303.3997

Maurano M.T., Humbert R., Rynes E., Thurman R.E., *et al.* (2012). Science 337(6099):1190–1195. https://doi.org/10.1126/science.1222794

McKenna A., Hanna M., Banks E., Sivachenko A., *et al.* (2010). Genome Res. 20(9):1297–1303. https://doi.org/10.1101/gr.107524.110

Stranger B.E., Nica A.C., Forrest M.S., Dimas A., *et al.* (2007). Nat. Genet. 39(10):1217–1224. https://doi.org/10.1038/ng2142

Suzuki M., Watanabe M., Nakamaru Y., Takagi D., *et al.* (2016). Cell. Mol. Life Sci. 73(5):1085–1101. https://doi.org/10.1007/s00018-015-2040-x

Yang J., Lee S.H., Goddard M.E., and Visscher P.M. (2011). Am. J. Hum. Genet. 88(1):76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Zhang L., Mei Y., Fu N. -y., Guan L., *et al.* (2012). Proc. Natl. Acad. Sci. 109(51):20937–20942. https://doi.org/10.1073/pnas.1214156110

Zhou Y., He L., Liu X.-D., Guan H., *et al.* (2019). Biomed Res. Int. 2019:6305065. https://doi.org/10.1155/2019/6305065