560. Large-scale cis-eQTL analysis of gene expression in blood of young healthy pigs using PigGTEx

L.M. Kramer[1*], J. Teng[2], K.S. Lim[1], Y. Gao[3,4], H. Yin[5], L. Bai[5], G.E. Liu[3], Z. Zhang[2], L. Fang[6], G.S. Plastow[7], C.K. Tuggle[1] and J.C.M. Dekkers[1]

[1]Iowa State University, 806 Stange Rd, Ames, IA 50011, USA; [2]South China Agricultural University, Guangzhou, 510642 Guangdong, China; [3]Animal Genomics and Improvement Laboratory Beltsville Agricultural Research Center, USDA, Beltsville, MD 20705, USA; [4]University of Maryland, College Park, MD 20742, USA; [5]Agricultural Genome Institute at Shenzhen, Chinese Academy of Agricultural Sciences, 518120 Shenzhen, China; [6]University of Edinburgh, Edinburgh, EH4 2XU, United Kingdom; [7]University of Alberta, Edmonton, AB, Canada; lmkramer@iastate.edu

Abstract

RNA sequencing and high-density genotyping or whole genome sequencing can be used in tandem to identify genomic regions associated with the control of gene expression, also known as expression quantitative trait loci (eQTL) mapping. Whole blood is a convenient tissue for many purposes due to its ease of collection and relation to health. Whole blood RNA-seq data and high-density SNP array genotypes were generated on ~1,600 healthy nursery pigs as part of a study of resilience to a polymicrobial disease challenge. This data was used for genotype imputation to whole genome sequence with the reference panel and bioinformatics pipeline developed by the PigGTEx consortium to identify cis-eQTL. Cis-eQTL were typically associated with expression of one or two genes. Chromosome 12 was enriched for cis-eQTL. The results provide insights into the control of gene expression in whole blood, allowing further study of its relationship with disease resilience.

Introduction

Selection and breeding of livestock species thrives on understanding the genetic control of economically important phenotypes. Extensive research is ongoing to better understand the genetic links between phenotypes and genetic markers through intermediate phenotypes, such as gene expression. This research utilizes various tissues to identify these intermediate phenotypes, with one of the most common tissues being blood due to its ease of collection and its relevance to health. As a tissue, blood is interesting as it interacts with other tissues across the organism, which allows it to be used for the detection of abnormalities or changes within the body, such as the onset of disease, or a change in behavior or physiological status. Blood's interconnectedness to other tissues in an organism also provides difficulties, as gene expression levels in blood can be influenced by external factors such as stress, cell composition, and temporal transcriptional differences. This makes it of utmost importance to characterize these influential factors to enhance the informativeness of blood as a tissue and its use to address a multitude of other research questions. In addition to improving blood as a source of gene expression data, it is also important to link these data to economic traits in livestock such as disease resistance/resilience, response to environmental changes, or even just production traits. The use of gene expression data and dense marker genotypes throughout the entire genome allows for eQTL mapping to better understand regions of genomic regulatory control of transcriptomic variation and response to external stressors. Identification of eQTL may indicate hotspots that are indicative of specialized regions of overarching genetic control. The purpose of this study is to capitalize on a uniquely large dataset with both genome-wide SNP genotypes and RNA-seq data to identify cis-eQTL for gene expression in whole blood of young healthy pigs for their potential use to predict future trait phenotypes for the animal.

R.F. Veerkamp and Y. de Haas (eds)

DOI: 10.3920/978-90-8686-940-4_560, © L.M. Kramer *et al.* 2022

Materials & methods

Animals and samples. Blood samples were collected in Tempus tubes from ~1,600 healthy Landrace × Yorkshire nursery barrows during the quarantine phase of the natural disease challenge model described by Putz *et al.* (2018) and Cheng *et al.* (2020). Piglets were brought to the nursery in batches of 60 or 75 from one of seven different breeding companies. All pigs were genotyped for approximately 650k markers across the genome.

RNA-seq. RNA-seq data were obtained as per Lim *et al.* (2021). RNA-seq libraries were generated using QuantSeq 3' mRNA Library Prep Kit FWD for Illumina and the RNA Removal Solution Globin Block (Lexogen, Austria) (Lim *et al.* 2019). QuantSeq libraries from 96 samples were multiplexed and sequenced with single end 50bp using Illumina HiSeq 3000 Sequencing Systems (Illumina, USA) for the first 887 samples from Lim *et al.* (2021) and with single-end 100 bp using NovaSeq 6000 Sequencing Systems (Illumina, USA) for an additional 707 samples. QuantSeq reads were trimmed and filtered using BBDuk. Read quality control was performed using FASTQC and mapped using STAR 2.5.3a. Mapped reads were removed if they had a zero count for more than 80% of the samples or were mapped to globin genes, *HBA* and *HBB*. The remaining reads were normalized using EdgeR (Robinson *et al.* 2010).

PigGTEx. PigGTEx is a part of the FarmGTEx project, aiming to develop a comprehensive catalog of regulatory variants in the pig transcriptome. The eQTL mapping pipeline that was used was developed by PigGTEx as a derivative of the cattle GTEx (Liu *et al.* 2021), and was designed to impute genotypes up to a whole genome sequence reference and then perform eQTL mapping using the imputed genotypes and RNA-seq data, using the following softwares: Java, Plink (Chang *et al.* 2015; Purcell and Chang), Beagle (Browning *et al.* 2018), ConformGT, Bcftools (Li, 2011), and TensorQTL (Taylor-Weiner *et al.* 2019). Imputation to whole genome sequence was performed on a chromosome basis using ~1,600 external pigs with whole genome sequence data based on the reference genome Susscrofa11.1. eQTL mapping was performed using TensorQTL by utilizing gene expression data corrected for read counts following Lim *et al.* (2021) and the imputed genotypes. Additionally, PEER estimation (Stegle *et al.* 2012) was used to obtain ten factors from the expression data and PCA was used to obtain the first ten principle components from the imputed genotypes to utilize as covariates in eQTL mapping. Cis-eQTL were defined as eQTL that were within one megabase of the gene whose expression is being evaluated. Multiple testing correction of identified cis-eQTL was done using a q-value calculation based on the approach of Nettleton *et al.* (2006).

Results & discussion

A set of cis-eQTL was obtained for each chromosome for our pig population. From this, we were able to identify the total number of eQTL after q-value control for each chromosome (Table 1), as well as the distribution of the number of genes whose expression is associated with each cis-eQTL. Chromosome six had the largest number of unique cis-eQTL after filtering (268,389), while chromosome 16 had the smallest number of cis-eQTL (51,688). The number of detected cis-eQTL per chromosome generally decreased with chromosome size, except for chromosome 12, which had many unique cis-eQTL, despite its smaller physical size. The largest number of genes whose expression was significantly associated with a given cis-eQTL by chromosome ranged from 12 for chromosome 16 to 53 for chromosome 7. Most cis-eQTL, however, had less than 10 associated genes, with the median number of genes associated with an eQTL equal to five and most eQTL only being associated with one or two genes (Figure 1). The total number of unique genes with a significant cis-eQTL per chromosome ranged from 179 for chromosome 11 to 1,053 for chromosome 6. Again, chromosome 12 had many unique genes for its size, which suggests that chromosome 12 has an abundance of cis-eQTL associated with gene expression in whole blood.

Table 1. Total number of unique cis-eQTL and associated expressed genes (# genes) by chromosome and the maximum number (Max #) of genes whose expression is associated with a single eQTL.

Chromosome	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
# cis-eQTL	247,046	229,560	194,465	150,175	118,033	268,389	154,336	119,889	140,710
# genes	864	952	763	568	569	1,053	661	342	493
Max # genes	34	50	51	44	35	48	53	20	25
Chromosome	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
# cis-eQTL	76,294	52,045	166,380	220,995	225,801	103,632	51,688	95,557	53,619
# genes	235	179	621	685	668	386	194	300	208
Max # genes	17	15	41	29	26	24	12	25	23

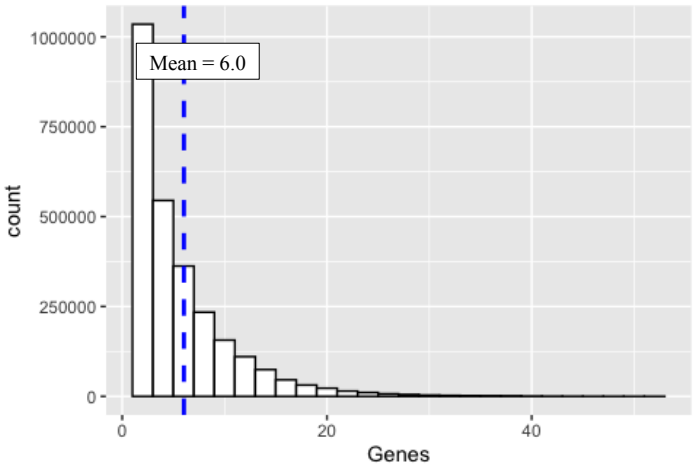

Figure 1. Frequency distribution of the number of genes whose expression is associated with each cis-eQTL.

Future analyses of eQTL for whole blood for this data set will involve an analysis of trans eQTL and of the associated genes and eQTL with disease resilience traits, following Lim *et al.* (2021). Cis- and trans-eQTL analyses after deconvolution of the gene expression data into various cell-type-specific proportions rather than gene expression of whole blood will also be conducted to identify putative regulatory regions associated with gene expression in these individual cell populations, in order to further develop and utilize blood as a tissue for understanding the biological processes behind economically important traits in swine. Through this approach it could become possible to use individual blood component gene expression profiles to identify individuals that are more or less resilient to disease.

Acknowledgements

This work was funded by USDA-NIFA grants # 2017-67007-26144 and #2021-67015-34562, along with Genome Canada and PigGen Canada. The natural disease challenge model was designed and implemented by John Harding (University of Saskatchewan), Michael Dyck (University of Alberta), Frederic Fortin (Centre de Développement du Porc du Québec, Inc.), Jack Dekkers and Graham Plastow together with PigGen Canada.

References

Andrews SFASTQC. *A quality control tool for high throughput sequence data*. 2010 Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Browning B.L., Zhou Y., Browning S.R. (2019) Am J Hum Genet 103(3):338-348 https://doi.org/10.1016/j.ajhg.2018.07.015

Chang C.C., Chow C.C., Tellier L.CAM., Vattikuti S., Purcell S.M., *et al.* (2015) GigaScience 4(1) https://doi.org/10.1186/s13742-015-0047-8

Cheng J., Putz A.M., Harding J.C.S., Dyck M.K., Fortin F., *et al.* (2020) J Anim Sci., 98(8) https://doi.org/10.1093/jas/skaa244

Li H. (2011) Bioinformatics 237(21):3987-93 https://doi.org/10.1093/bioinformatics/btr509

Lim K.S., Dong Q., Moll P., Vitkovska J., Wiktorin G., *et al.* (2019) BMC Genomics, 20(1):1-10. https://doi.org/10.1186/s12864-019-6122-2

Lim K.S., Cheng J., Putz A.M., Dong Q., Bai X., *et al.* (2021) BMC Genomics 22:614 https://doi.org/10.1186/s12864-021-07912-8

Liu S., Gao Y., Canela-Xandri O., Wang S., Yu Y., *et al.* (2021) bioRxiv https://doi.org/10.1101/2020.12.01.406280

Nettleton D., Hwang J.T.G., Caldo R.A., Wise R.P. (2006) JABES 11, 337-356. https://doi.org/10.1198/108571106X129135

Purcell S, Change C. PLINK Available at: www.cog-genomics.org/plink/1.9/

Putz A.M., Harding J.C.S., Dyck M.K., Fortin F., Plastow G.S., *et al.* (2019) Front Genet 9:660 https://doi.org/10.3389/fgene.2018.00660

Robinson M.D., McCarthy D.J., Smyth G.K. (2010) Bioinformatics 26(1):139-140 https://doi.org/10.1093/bioinformatics/btp616

Stegle O., Parts L., Piipari M., Winn J., Durbin R. (2012) Nat Protoc 7(3): 500-7 https://doi.org/10.1038/nprot.2011.457

Taylor-Weiner A., Aguet F., Haradhvala N.J., Gosai S., Anand S., *et al.* (2019) Genome Biology 20(228) https://doi.org/10.1186/s13059-019-1836-7