RNA extraction and sequencing

RNA extraction and sequencing was performed at a commercial facility (Eurofins, Germany). Briefly, total RNA was extracted from the testis tissue of the analysis population by RNeasy Mini Kit (QIAGEN) following instructions of the manufacturer. Concentration of RNA was measured by Nanodrop® 2000 (ThermoScientific, Massachusetts, USA) and quality was evaluated with a Bioanalyzer (Agilent, California, USA). Sequencing libraries were prepared from 400 ng RNA using either the TruSeq stranded mRNA (Illumina, San Diego, USA) kit following manufacturer's instructions or the TruSeq total stranded RNA (Illumina) kit following manufacturer's instructions according to RNA quality. The samples were sequenced with an Illumina HiSeq 2500 (Illumina, San Diego, USA), which amounted to a theoretical 20 million reads per sample and demultiplexed into FastQ-files by CASAVA®-software (Illumina, San Diego, USA).

Quality control and read counting

Quality control (QC) of RNA-Seq reads was conducted with FastQC (v. 0.11.3) (3). Reads were trimmed for known Illumina TruSeq adapter sequences using the software CutAdapt (v. 1.8.1) (9). Poor reads were trimmed by the software Trimmomatic (v. 0.33) (4) using default parameters. The trimmed reads were then mapped to the *Sus scrofa* reference genome (10.2, version 87) obtained from Ensembl (2) by the STAR aligner (v. 020201) (5) using default parameters. Post-mapping QC was performed with Qualimap (11). Post-mapping QC showed a mean of 12.23 million reads successfully mapped to the reference genome. The mean number of reads aligned (± SD) was 18.26 (± 12.05) million reads. The genomic origin of the reads had a

mean (± SD) of 50.58% (± 4.62%) exonic, 15.59% (± 1.91%) intronic and 33.83% (± 6.14%) intergenic. The mapped reads were counted to each gene by HTSeq (v. 0.6.0) (1) using default parameters. All subsequent statistical analysis were performed in R (v. 3.1.0) (13). To remove genes with very low expression levels, only genes with a mean count of more than five were included in the gene count matrices. This filtering approach was consistent with previous eQTL analysis in the porcine genome (6) and consistent with previous warrants to obtain the highest statistical power (7). The final number of genes used in the study was 14,277. Normalisation of gene counts was performed by the voom variance-stabilization function implemented in the R package limma (v. 3.30.3) (14) using the "scale" normalisation method. Furthermore, sample quality weights were used to increase comparability between samples (8). All genes without gene symbols were denoted in the manuscript with their *Sus scrofa* Ensembl identifier code.

Genotyping and filtering

Genotyping was performed by a commercial facility (Neogen, Lincoln, NE, USA). Briefly, genomic DNA was extracted from loin muscle by DNA Minikit (QIAGEN) following manufacturer instructions and genotyped using the GeneSeek Genomic Profiler (GGP) Porcine Bead Chip 80K (Neogen). For filtering of genotype data, PLINK (12) (v. 1.90b3.44) was used. The variants were included in the genotype data when they achieved a call rate above 0.95, a minor allele frequency (MAF) of above 0.05 and were in Hardy-Weinberg equilibrium ($P < 1 \times 10^{-5}$). In order to remove variants, which were in linkage disequilibrium (LD), LD-based variant pruning was performed with a window size of 50 Kb, a step size of 5 Kb and an r^2 threshold of 0.8. Finally, the variants were converted to their genomic coordinates and Reference SNP cluster ID (rs) accession numbers, which was obtained from SNPchiMp v. 3 database (10). After

running the filtering pipeline, the genotype data comprised a total of 68,516 SNPs with a total genotyping rate of 0.943578. Due to missing genotype data, 4,701 variants were removed. Furthermore, 6,322 SNPs were removed due to Hardy-Weinberg exact test filtering and 6,646 SNPs were removed due to minor allele threshold. Linkage disequilibrium (LD) based variant pruning removed 19,858 variants due to high LD ($r^2 \geq 0.8$) and a final subset of 28,959 SNPs was used for the analysis in this study.

Validation of top eQTL by RT-qPCR

Reverse transcriptase quantitative real-time PCR (RT-qPCR) was performed on total RNA from the 32 testis samples, which had been subjected to RNA-Seq and eQTL analysis. The total RNA was isolated by identical protocol for the isolation of total RNA for RNA-Seq. cDNA was synthesised from 1 µg of total RNA by the RevertAid First Strand cDNA Synthesis Kit (ThermoFischer, Slangerup, Denmark) with slight modifications in volumes from manufacturer's protocol: 5 µl RT M-MLv RT buffer, 0.8 µl RNase inhibitor and 1.3 µl dNTP 10 mM mix. Gene specific primers for RT-qPCR were designed using the Primer3 primer designer tool (15). The housekeeping gene *RPL4* was used for relative quantification. Information on the primers is available in Supplementary file 7. Each run was performed in a 96-well plate comprising samples and no template controls and conducted with the following program: pre-incubation for 95°C for 5 min and 45 cycles of quantification at 95°C for 10 s/60°C for 10 s/72°C for 20 s and finally, melting curve analysis of 95°C for 5 s/65°C for 1 min and cooling at 40°C for 30 s with the SYBR® Green Lightcycler© 480 II system (Roche). PCR efficiencies were calculated by a standard curve of eight dilutions and were always above 90%. In order to statistically validate the top eQTL gene, a Pearson correlation between delta RNA-Seq expression divided with

expression of reference gene *RPL4* compared with delta CT values from the RT-qPCR procedure and expressed as 1 / (*CYP1A2* / *RPL4*). To confirm the eQTL relationship, an ANOVA test was performed between the RT-qPCR gene expression and the genotypes of the top eQTL SNP.

References

1. **Anders S, Pyl PT, and Huber W**. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* btu638, 2014.
2. **Andrew Yates, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juettemann, Stephen Keenan, Ilias Lavidas, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Michael Nuhn, Anne Parker, Mateus Patricio, Miguel Pignatelli, Matthew Rahtz, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Ewan Birney, Jennifer Harrow, Matthieu Muffato, Emily Perry, Magali Ruffier, Giulietta Spudich, Stephen J. Trevanion, Fiona Cunningham, Bronwen L. Aken, Daniel R. Zerbino, and Flicek P**. Ensembl 2016. *Nucleic Acids Res* D710-D716: 2016.
3. **Andrews S**. FastQC: a quality control tool for high throughput sequence data http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
4. **Bolger AM, Lohse M, and Usadel B**. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* btu170, 2014.
5. **Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR**. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15-21, 2013.
6. **Kogelman LJ, Zhernakova DV, Westra H-J, Cirera S, Fredholm M, Franke L, and Kadarmideen HN**. An integrative systems genetics approach reveals potential causal genes and pathways related to obesity. *Genome Med* 7: 105, 2015.
7. **Łabaj PP, Leparc GG, Linggi BE, Markillie LM, Wiley HS, and Kreil DP**. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics* 27: i383-i391, 2011.
8. **Law CW, Chen Y, Shi W, and Smyth GK**. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15: R29, 2014.
9. **Martin M**. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal* 17: 2011.
10. **Nicolazzi EL, Caprera A, Nazzicari N, Cozzi P, Strozzi F, Lawley C, Pirani A, Soans C, Brew F, and Jorjani H**. SNPchiMp v. 3: integrating and standardizing single nucleotide polymorphism data for livestock species. *BMC Genomics* 16: 283, 2015.
11. **Okonechnikov K, Conesa A, and García-Alcalde F**. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 32: 292-294, 2016.

12. **Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, and Daly MJ**. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559-575, 2007.
13. **R Core Team**. R: A language and environment for statistical computing R Foundation for Statistical Computing. http://www.R-project.org/.
14. **Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK**. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* e47, 2015.
15. **Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, and Rozen SG**. Primer3—new capabilities and interfaces. *Nucleic Acids Res* 40: e115-e115, 2012.