Supplementary Information for:

Chapter 2 | Integrative genomics sheds light on the immunogenetics of tuberculosis in cattle

This PDF file includes:

Supplementary Note 2.1

Supplementary Note 2.2

Supplementary Figure S2.1

Supplementary Figure S2.2

Supplementary Figure S2.3

Supplementary Figure S2.4

Supplementary Figure S2.5

Supplementary Figure S2.6

Supplementary Figure S2.7

Supplementary Figure S2.8

Supplementary Figure S2.9

Supplementary Figure S2.10

Supplementary Figure S2.11

Supplementary Figure S2.12

Supplementary Figure S2.13

Supplementary Note 2.1

Genomic and transcriptomic data generation

Genomic DNA was extracted and purified for each animal (bTB+ and bTB−) using 200 µl of whole blood and the QIAamp® DNA Blood Mini Kit (Qiagen) following the manufacturer's recommendations. Purified DNA was quantified using a NanoDrop™ One microvolume spectrophotometer (Thermo Fisher Scientific) and stored at −20°C. The extracted DNA was used for SNP genotyping using the Axiom™ Genome-Wide BOS 1 Bovine Array (Thermo Fisher Scientific), which assays 648,315 SNPs across the bovine genome. SNP array genotyping was performed by an external provider (IdentiGEN/MSD Animal Health).

Total RNA preparation was performed for each animal (bTB+ and bTB−) using the Tempus™ Spin RNA Isolation Kit (Thermo Fisher Scientific). The Tempus™ tubes were inverted to mix the contents, which were then poured into sterile 50 ml conical tubes and total RNA was isolated and purified according to the manufacturer's instructions. Following total RNA purification, an Agilent 2100 Bioanalyzer and the RNA 6000 Nano LabChip kit (Agilent Technologies, Inc.) were used to quality check and generate RNA integrity number (RIN) values for all samples, which were then stored at −80°C. Total RNA samples (bTB+ and bTB−) were used with the NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (New England Biolabs) to construct sequence barcode-indexed RNA-seq libraries, which were then used for 150 bp paired-end sequencing on the NovaSeq™ 6000 Sequencing System (Illumina, Inc.). RNA-seq library preparation and sequencing were performed by an external provider (Novogene).

Supplementary Note 2.2

Variant remapping and strand flipping

Genomic coordinates of variants were updated from UMD3.1 to the ARS-UCD1.2 (Rosen *et al.* 2020) bovine genome assembly using liftOver implemented in the R package rtracklayer v.1.54.0 (Lawrence *et al.* 2009) and only SNPs were retained that matched the positions in a file of the Affymetrix Axiom™ Genome-Wide BOS-1 array coordinates updated to ARS-UCD1.2 available from the United States Department of Agriculture National Animal Genome Research Program (USDA-NAGRP) data repository (www.animalgenome.org/repository/cattle/UMC_bovine_coordinates). Due to potential strand flipping issues, all remaining palindromic SNPs were removed. Raw SNP genotype files with updated coordinates were then filtered using PLINK v1.90b6.25 to remove animals with a genotype call rate < 0.95 and to restrict analysis to autosomal SNPs with a call rate > 0.95.

Strand mismatches or miscoding of alleles can significantly reduce imputation performance (Verma *et al.* 2014) and statistical power in the context of genomic association studies (Winkler *et al.* 2014). To identify the alleles and associated genotypes that needed to be flipped, original reference and alternative allele pairs for each SNP were first determined using the Axiom™ Genome-Wide BOS-1 Array master annotation file (www.thermofisher.com/order/catalog/product/sec/assets?url=TFS-Assets/LSG/Support-Files/Axiom_GW_Bos_SNP_1-na35-annot-csv.zip) due to issues associated with the encodings of the major and minor alleles in PLINK. Following this, the reference and alternative allele pairs of the filtered VCF file were compared to the USDA-NAGRP BOS-1 ARS-UCD1.2 reference allele file (www.animalgenome.org/repository/cattle/UMC_bovine_coordinates). If the NAGRP reference allele matched the alternative allele or the complement of the alternative allele in the filtered VCF file, the SNP and associated genotypes were flipped. Of the remapped variants, 100,888 SNPs (18.23%) were flipped due to ambiguity associated with reference and alternative allele strands. By following guidelines detailed in Winkler *et al.* (2014), to verify that the strand flipping procedure was implemented correctly, the ARS-UCD1.2 reference allele frequencies of the target set were compared to 56 European animals derived from a whole genome sequence (WGS) Global Reference Panel (Dutta *et al.* 2020) at matched genomic loci. A total of 236,325 SNPs overlapped between the two groups, and we observed a significant positive correlation between the allele frequencies present in both groups (Spearman correlation (ρ) = 0.98; $P < 2.2 \times 10^{-16}$) (**Supplementary Figure S2.4**). A total of 150 SNPs present in the target and European WGS reference set were identified as potentially harbouring spurious allele frequency patterns

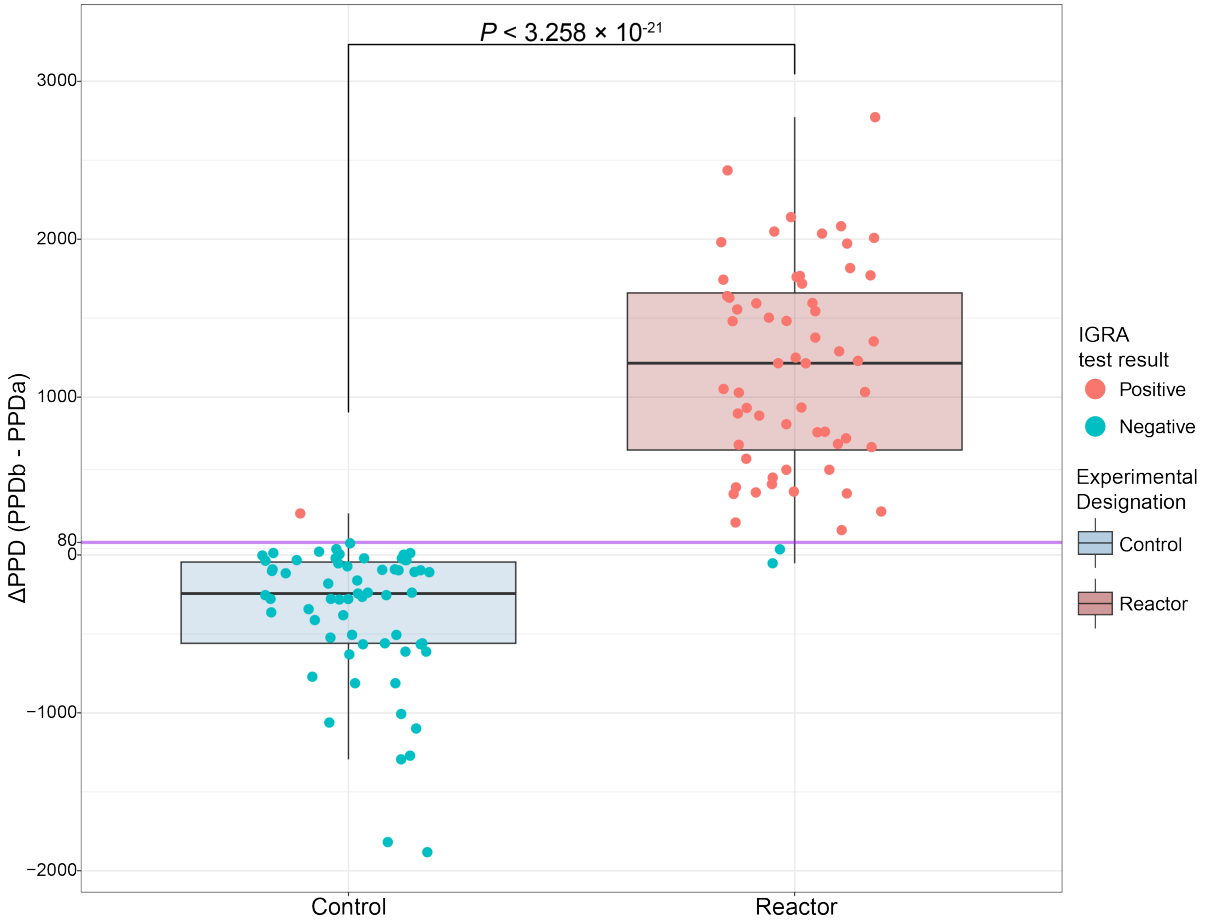
through substantial deviation from the line of parity. These SNPs were removed from the target set and WGS Global Reference Panel. Hence, the number of SNPs remaining in the target set and the WGS Global Reference Panel prior to imputation was 553,369 and 10,282,037, respectively.

Genome-wide SNP imputation

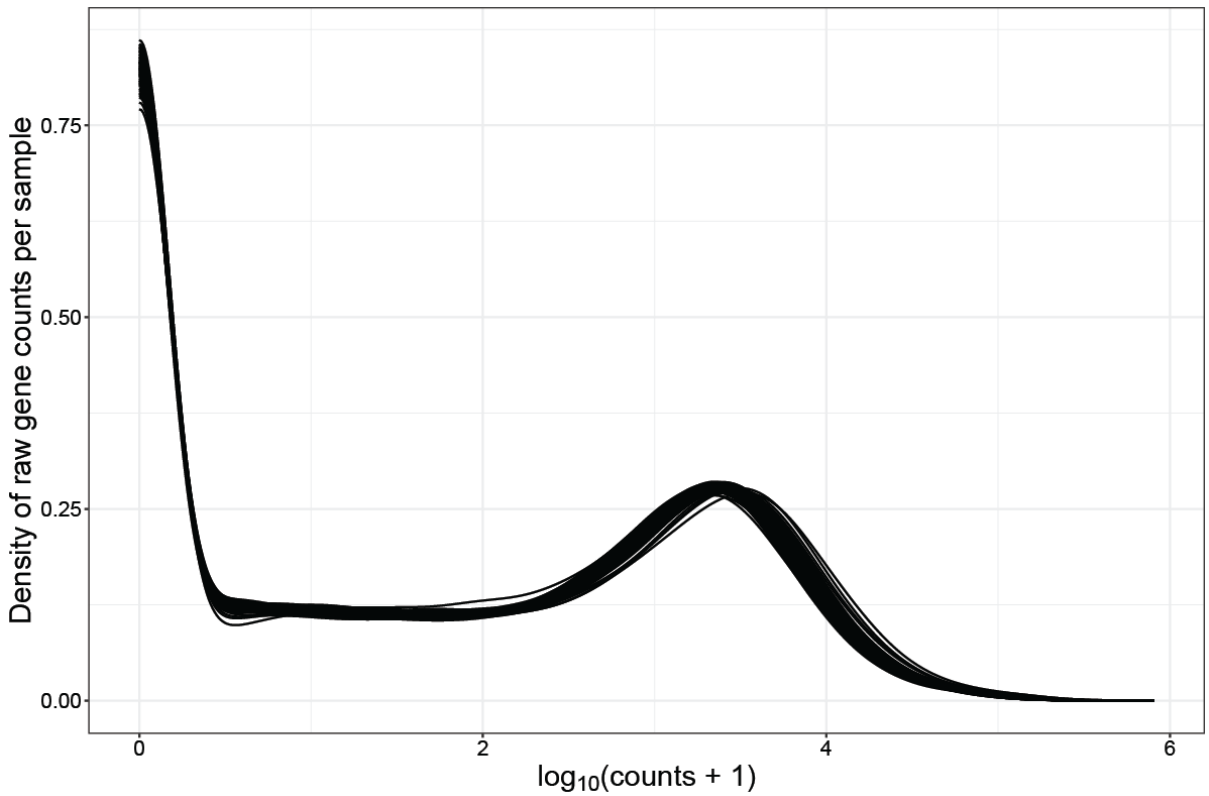
For the imputation of the SNP-array data up to WGS scale, a global cattle reference panel was used, which comprised a total of 10,282,187 SNPs derived from n = 287 distinct animals spanning a diverse range of breeds and geographic locations (55 populations: 13 European, 12 African, 28 Asian, and two Middle Eastern). Details of how the original WGS data were processed is provided by Dutta *et al.* (2020) and the quality control procedure implemented on the data set is described by Riggio *et al*. (2022). Briefly, Illumina WGS data for n = 427 animals representing global cattle breeds were aligned to the UCD-ARS1.2 genome. Animals and SNPs with a high proportion of missingness (> 25%), SNPs with a poor genotype quality (GQ) (< 25) and highly related individuals (relatedness value estimated from VCFtools v.0.1.15 (Danecek *et al.* 2011) with *-relatedness2* (Manichaikul *et al.* 2010) > 0.0625) were removed leaving a total of n = 287 animals with 10,282,187 variants. The 150 genotyped SNPs with spurious allele frequency patterns were also removed from the WGS reference set prior to imputation.

The VCFtools package was used to split the reference and target data sets into 29 VCF files representing the 29 bovine autosomes. Beagle v.5.4 (Browning *et al.* 2021) was then used to phase the reference and target data sets separately and Minimac3 v.2.0.1 (Das *et al.* 2016) was used with default settings to generate the reference haplotype m3.vcf files. Following this, Minimac4 v.1.03 (Das *et al.* 2016) was used with default parameters to impute the target genotype data set up to WGS scale and Bcftools v.1.7 (Danecek *et al.* 2021) was used to concatenate all 29 imputed VCF files, which resulted in a master imputed data set consisting of all n = 123 animals with genotypes for 10,282,037 SNPs.

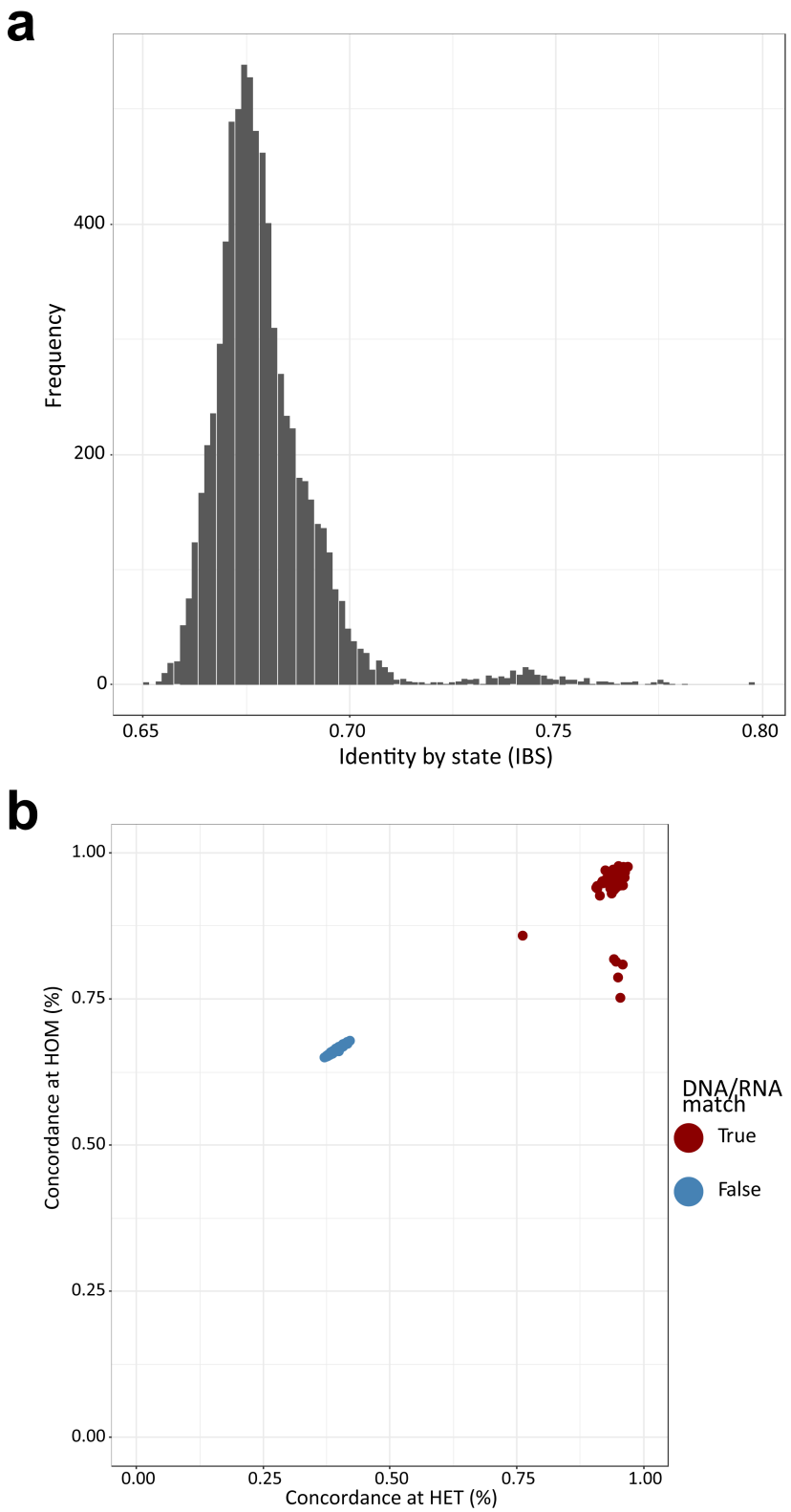
Supplementary Figures



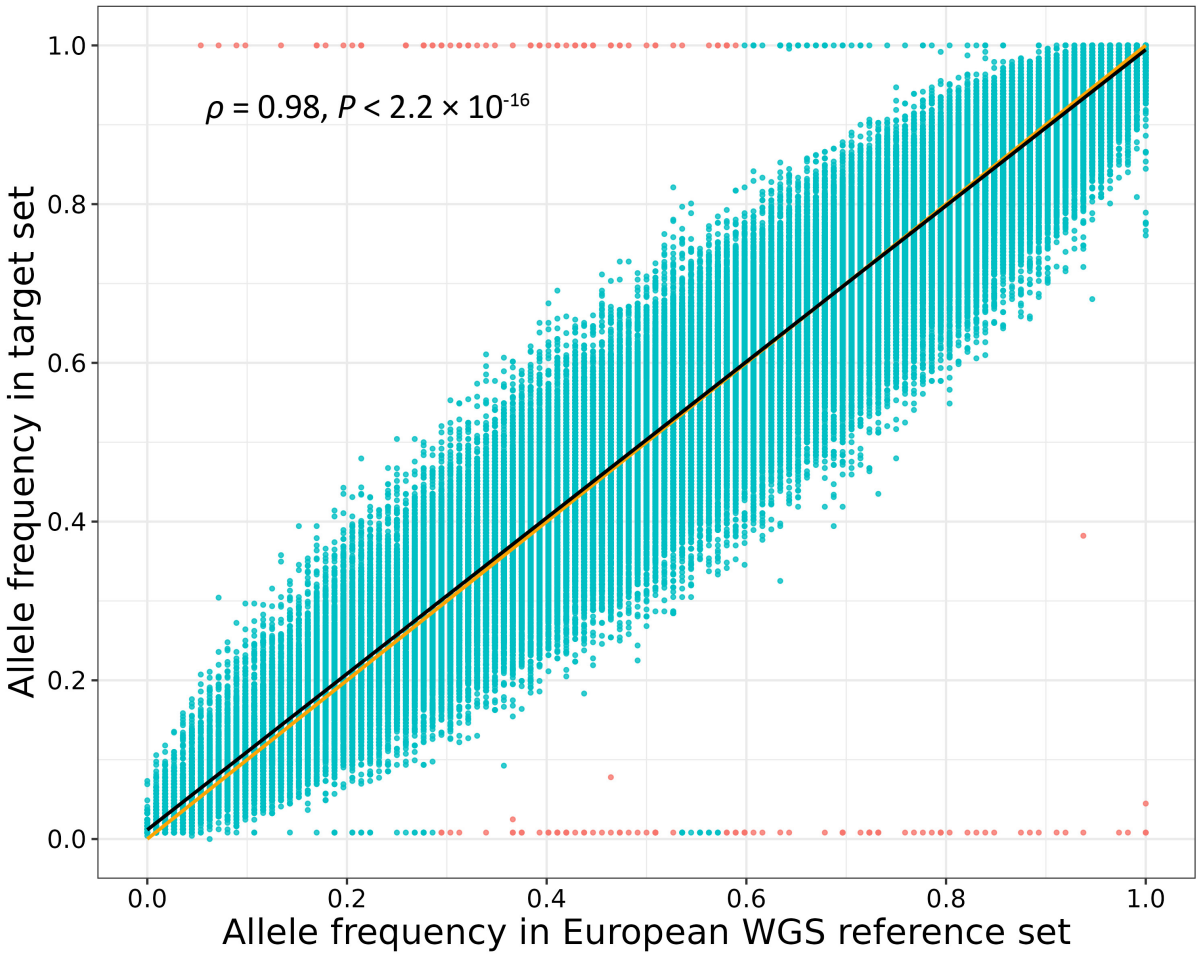
Supplementary Figure S2.1: IFN-γ release assay (IGRA) test results for all (*n* = 123) animals in the present study. Interferon-gamma (IFN-γ) release assay (IGRA) test results for animals designated as control (bTB−; *n* = 63) or reactor (bTB+; *n* = 60) animals. The *y*-axis denotes the Δ purified protein derivative (ΔPPD) value, calculated as PPD-bovine (PPDb) IFN-γ value minus the PPD-avian (PPDa) IFN-γ value. The purple line indicates the threshold for determining whether an animal is positive or negative for the IGRA test. Horizontal lines inside the boxes show the medians, Box bounds show the lower quartile (Q1, the 25th percentile) and the upper quartile (Q3, the 75th percentile).



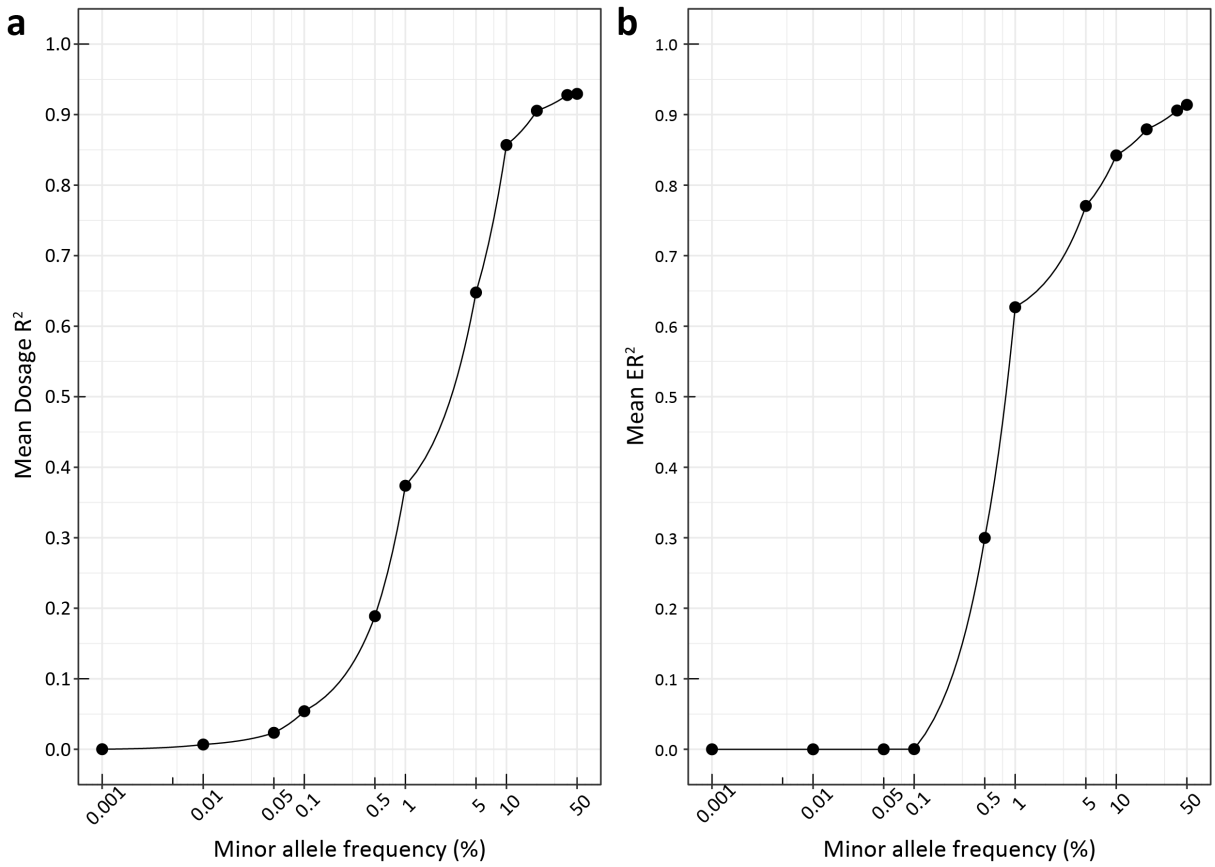
Supplementary Figure S2.2: Transcriptomics data quality control. Density of $\log_{10}$ gene counts per library (*n* = 123) prior to filtering of lowly expressed genes.



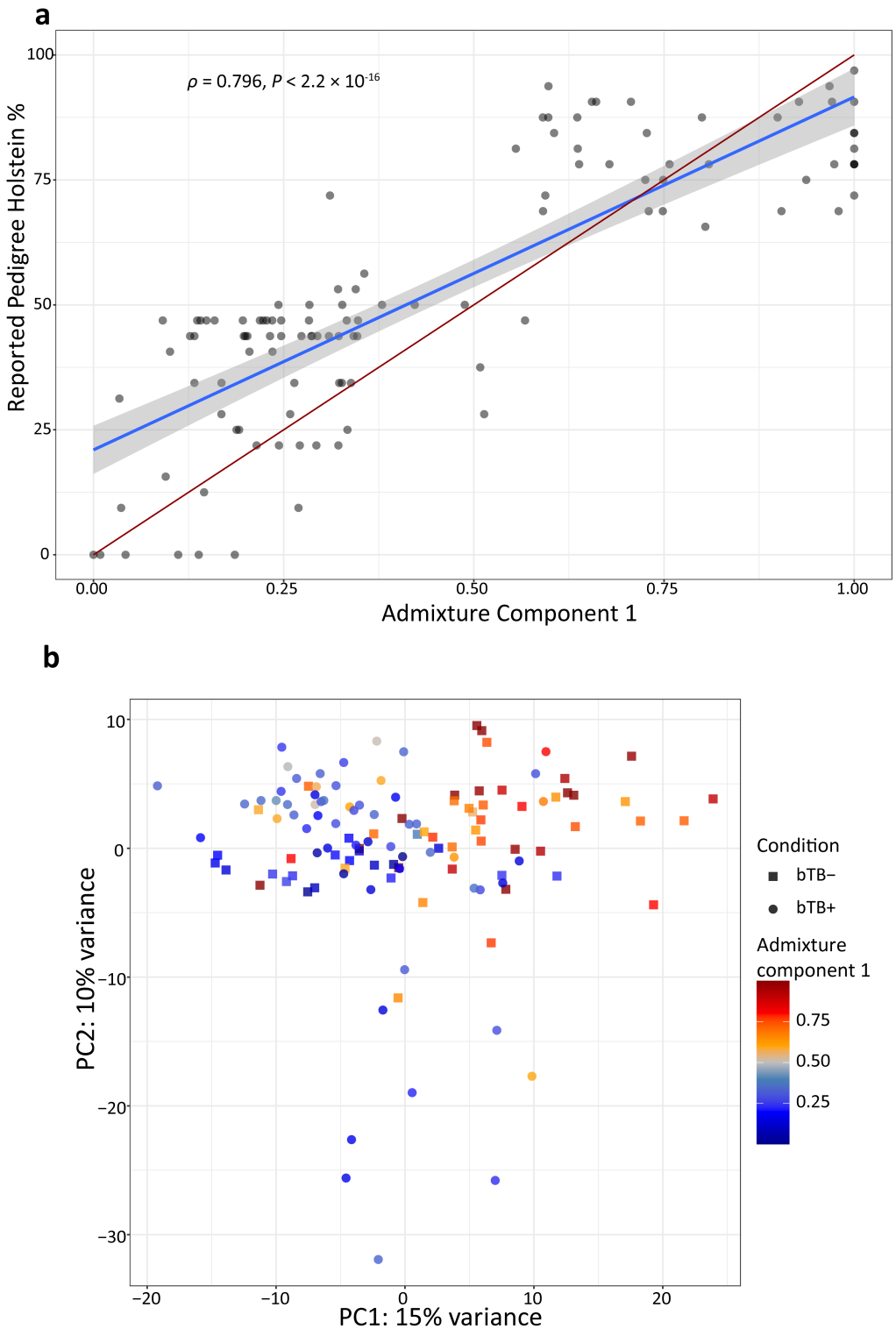
Supplementary Figure S2.3: Sample duplication and DNA/RNA match assessment. (**a**) Identity by state (IBS) values from PLINK (Purcell *et al.* 2007) for all pairwise comparisons among the *n* = 123 animals in the current study using the pruned SNP data prior to imputation. (**b**) Match BAM to VCF (MBV) (Fort *et al.* 2017) outputs showing allelic consistency for heterozygous (*x*-axis) and homozygous (*y*-axis) variants, between genotype and RNA-seq sequences. Red dots indicate proper identity match of RNA-sequencing with expected genotypes. As a control, we show consistency levels for non-matching samples between biological sample T064 and all other samples (blue dots).



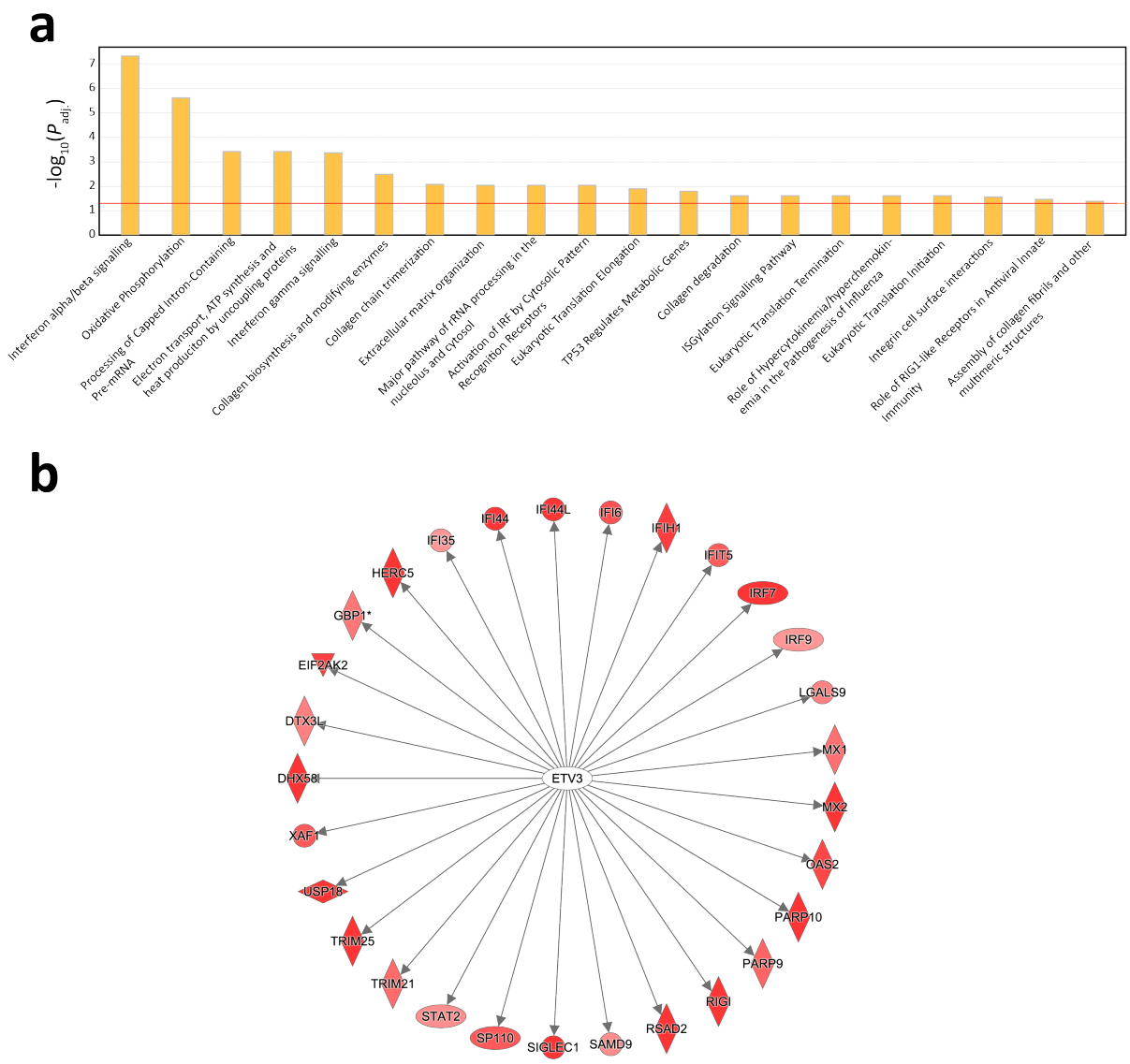
Supplementary Figure S2.4: Allele frequencies between the study population and European *B. taurus* cattle from the WGS reference panel. The observed (study-specific; *n* = 123) allele frequencies reported on the *y*-axis are plotted against allele frequencies derived from *n* = 56 European WGS animals on the *x*-axis at 236,325 matched genomic loci. Orange line denotes the line of parity. Black line denotes line of best fit. Red points indicate variants removed from the study population and Global Reference Panel prior to imputation.



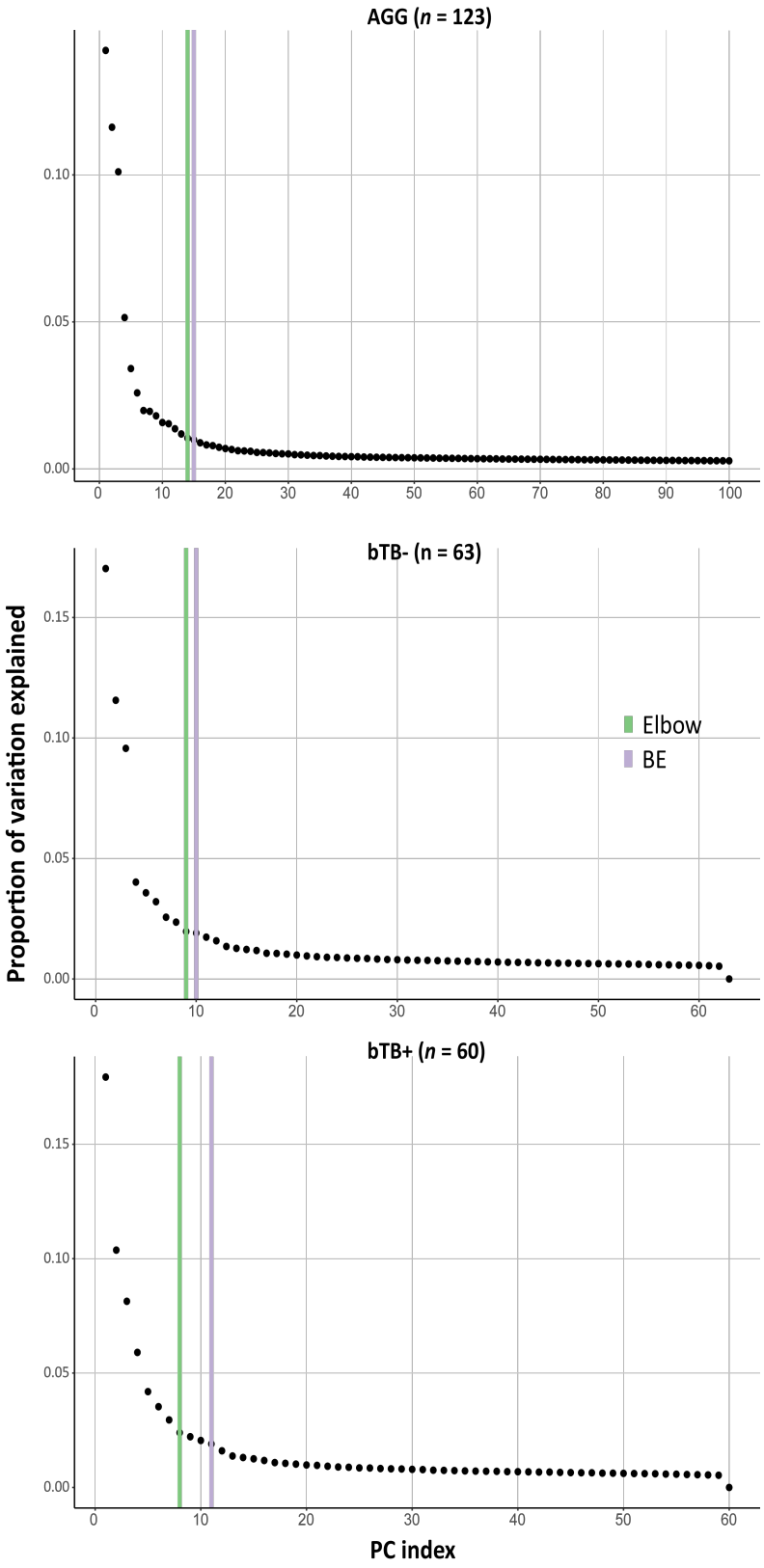
Supplementary Figure S2.5: Imputation Performance. (**a**) Imputation accuracy (Dosage R^2) from Minimac4 (Das *et al.* 2016) for all variants (genotyped and ungenotyped) when imputed up to WGS level using the Global Reference Panel. (**b**) Imputation accuracy measured using the empirical R^2 (ER^2) metric from Minimac4 of genotyped variants when imputed up to WGS level using the Global Reference Panel.



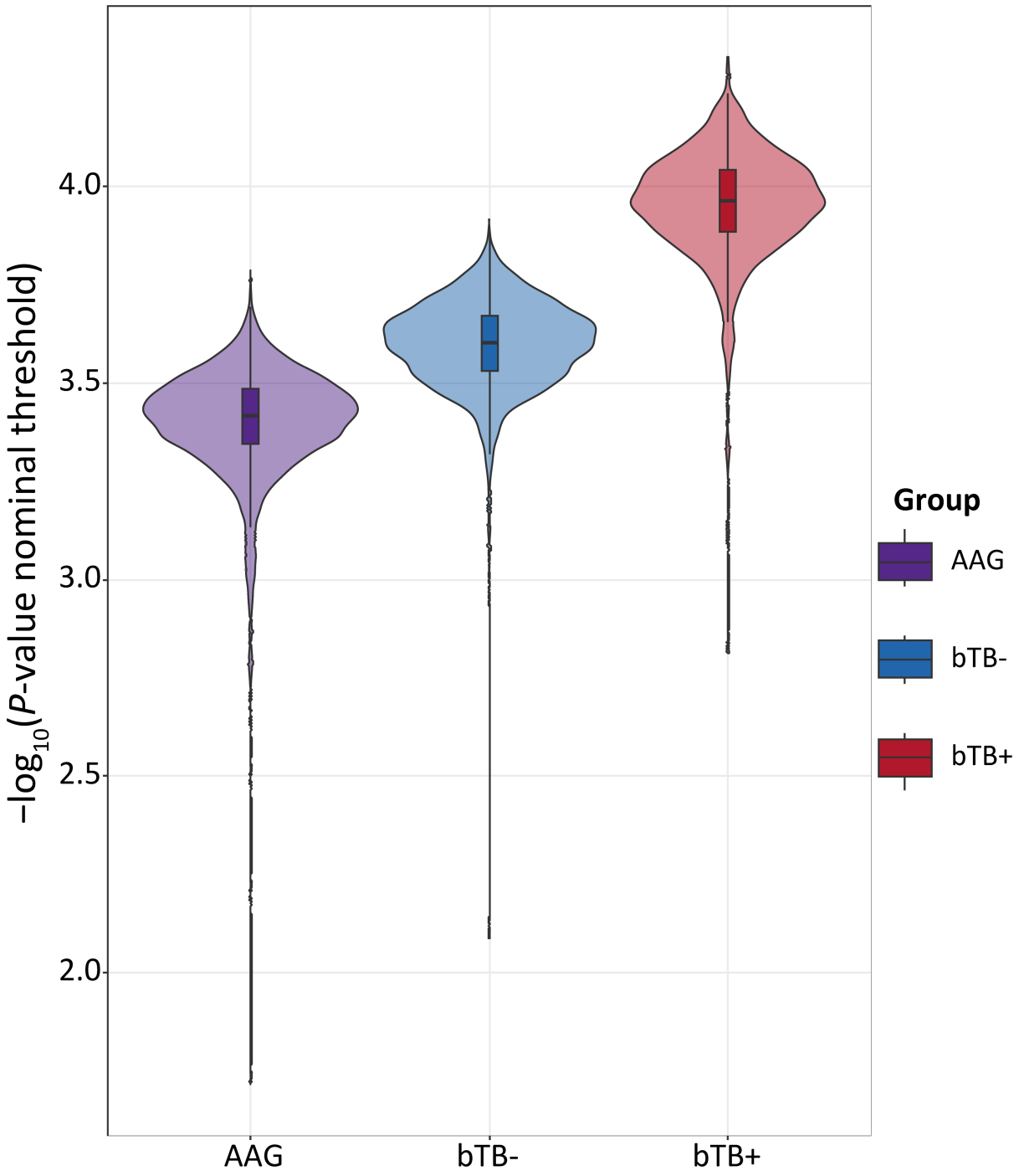
Supplementary Figure S2.6: Admixture analysis and its impact on the transcriptomics principal component analysis (PCA). (**a**) Spearman correlation (ρ) between reported pedigree Holstein % and Admixture component 1 from Admixture analysis (Alexander & Lange 2011) of 34,272 pruned SNPs. Red line indicates line of parity, blue line indicates line of best fit. (**b**) Principal component analysis (PCA) of top 1,500 most variable genes for all n = 123 animals after variance stabilising transformation using DESeq2 (Love *et al.* 2014). Principal components PC1 and PC2 are plotted. Animal data points are shaped based on their experimental condition and coloured according to their corresponding Admixture component 1 value.



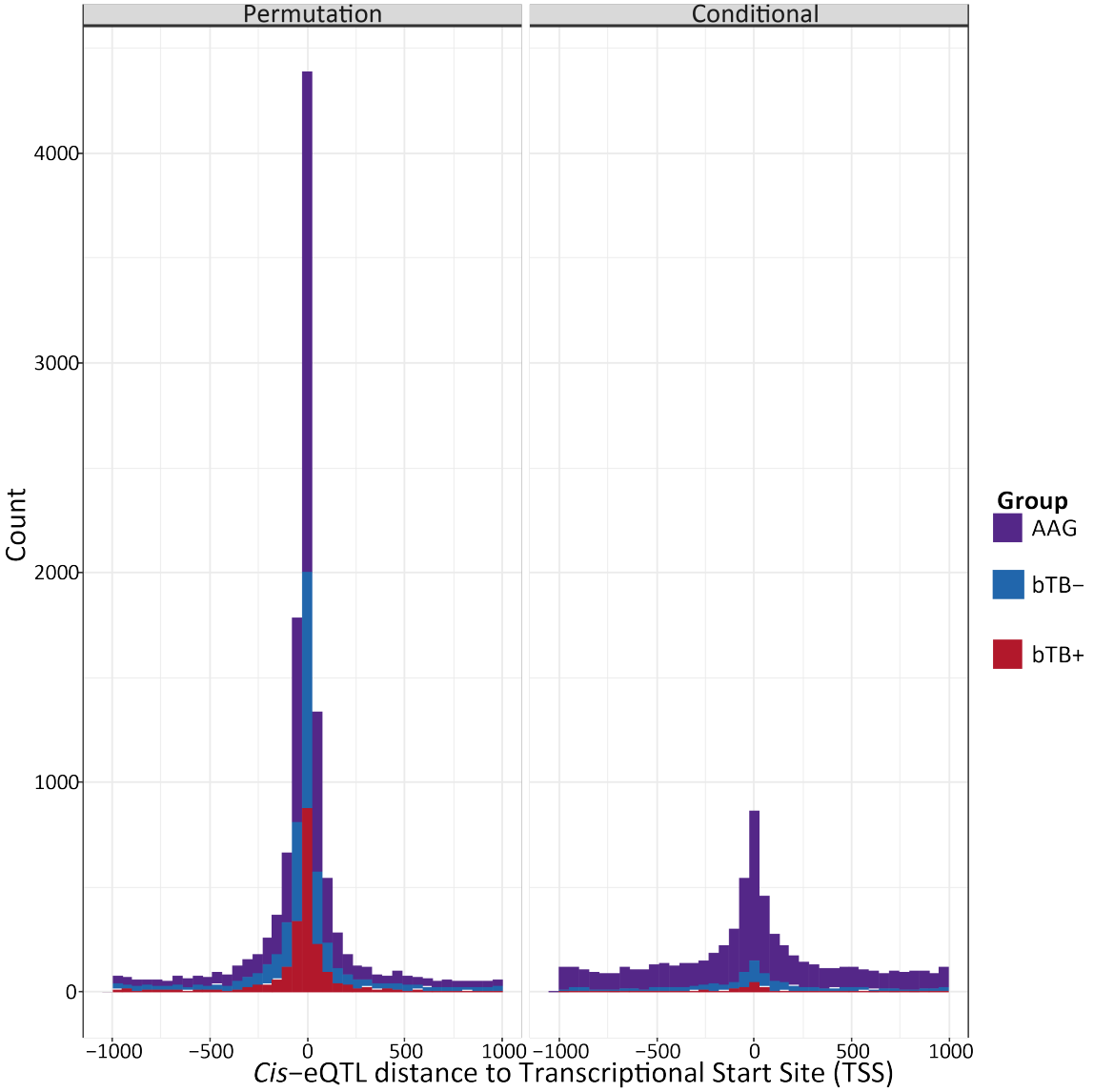
Supplementary Figure S2.7: IPA® enrichment and upstream regulator results for differentially expressed genes. (**a**) Bar plot of significantly enriched (FDR $P_{adj.} < 0.05$) Ingenuity® Pathway Analysis (IPA) pathways for highly significant (FDR $P_{adj.} < 0.01$) differentially expressed genes used as input. Bars are ordered corresponding to their $-\log_{10} P_{adj.}$ value. (**b**) Top upstream biological regulator (ETV3) predicted to be upregulated by IPA based on the highly significant DE genes used as input.



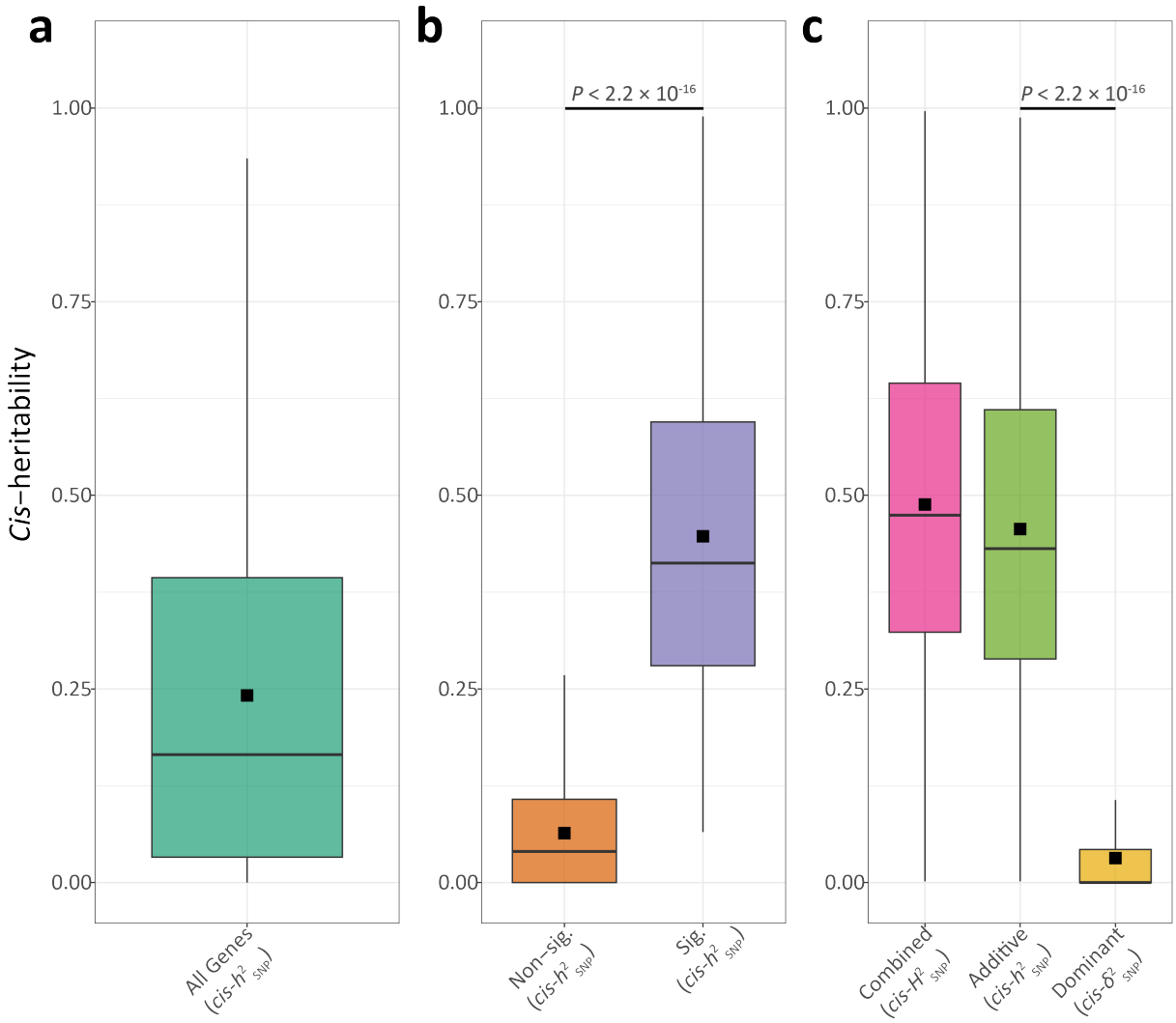
Supplementary Figure S2.8: Transcriptomics principal components (PCs) used in the eQTL analysis. Scree plot showing the proportion of variation explained by each of the transcriptomics PCs identified using the PCAForQTL R package (Zhou *et al.* 2022) across the all animals group (AGG), the control group (bTB−), and the reactor group (bTB+), respectively. Green lines correspond to the number of transcriptomic PCs inferred by the elbow method. Purple line denotes the number of PCs inferred by the Buja and Eyuboglu (BE) algorithm (Buja & Eyuboglu 1992).



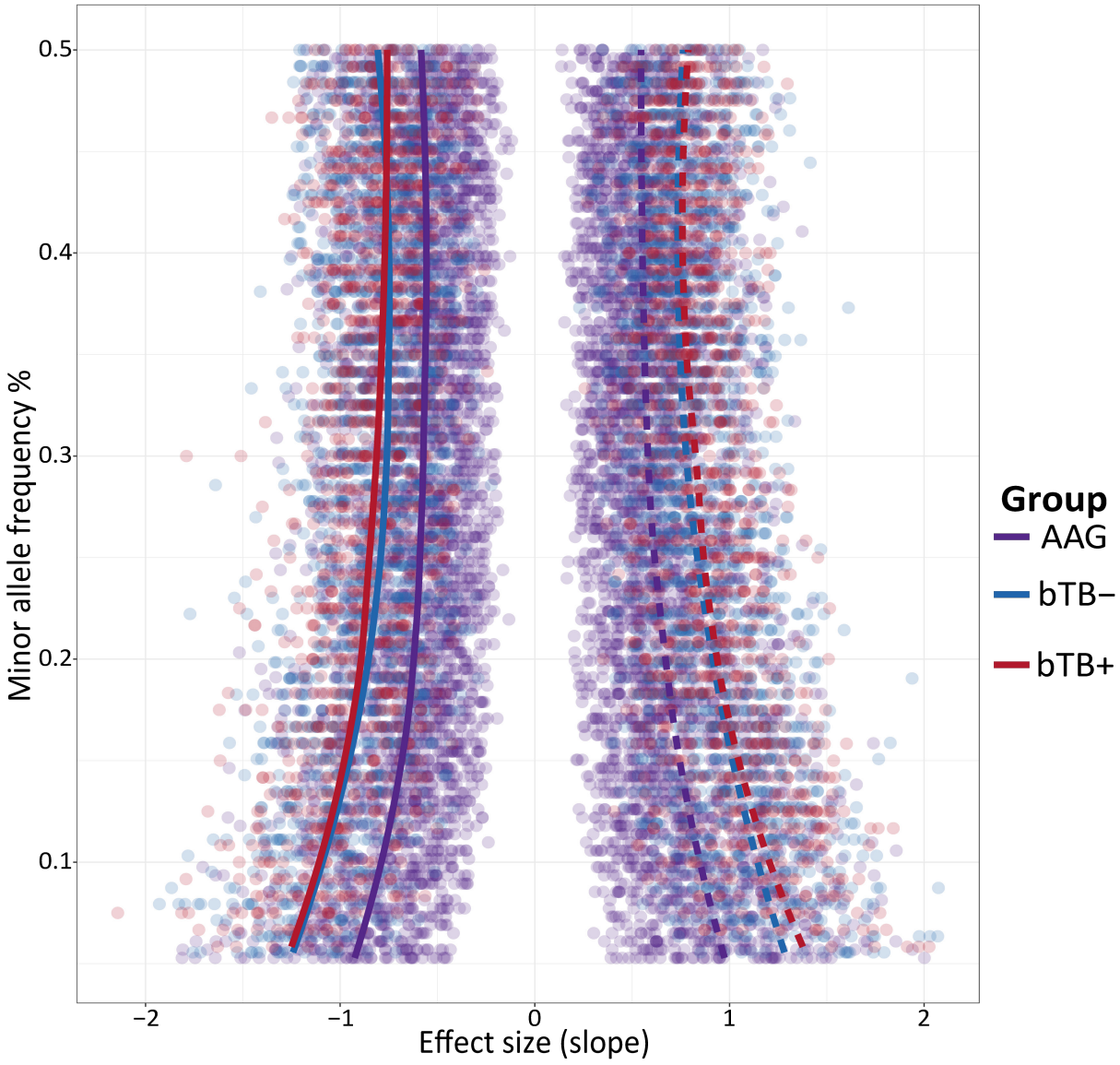
Supplementary Figure S2.9: Nominal P-value thresholds within each group for the determination of *cis*-eVariants. Violin plots showing the distribution of nominal *P*-value threshold per *cis*-eGene for bTB− (control), bTB+ (reactor), and AAG (all animal) groups, respectively.



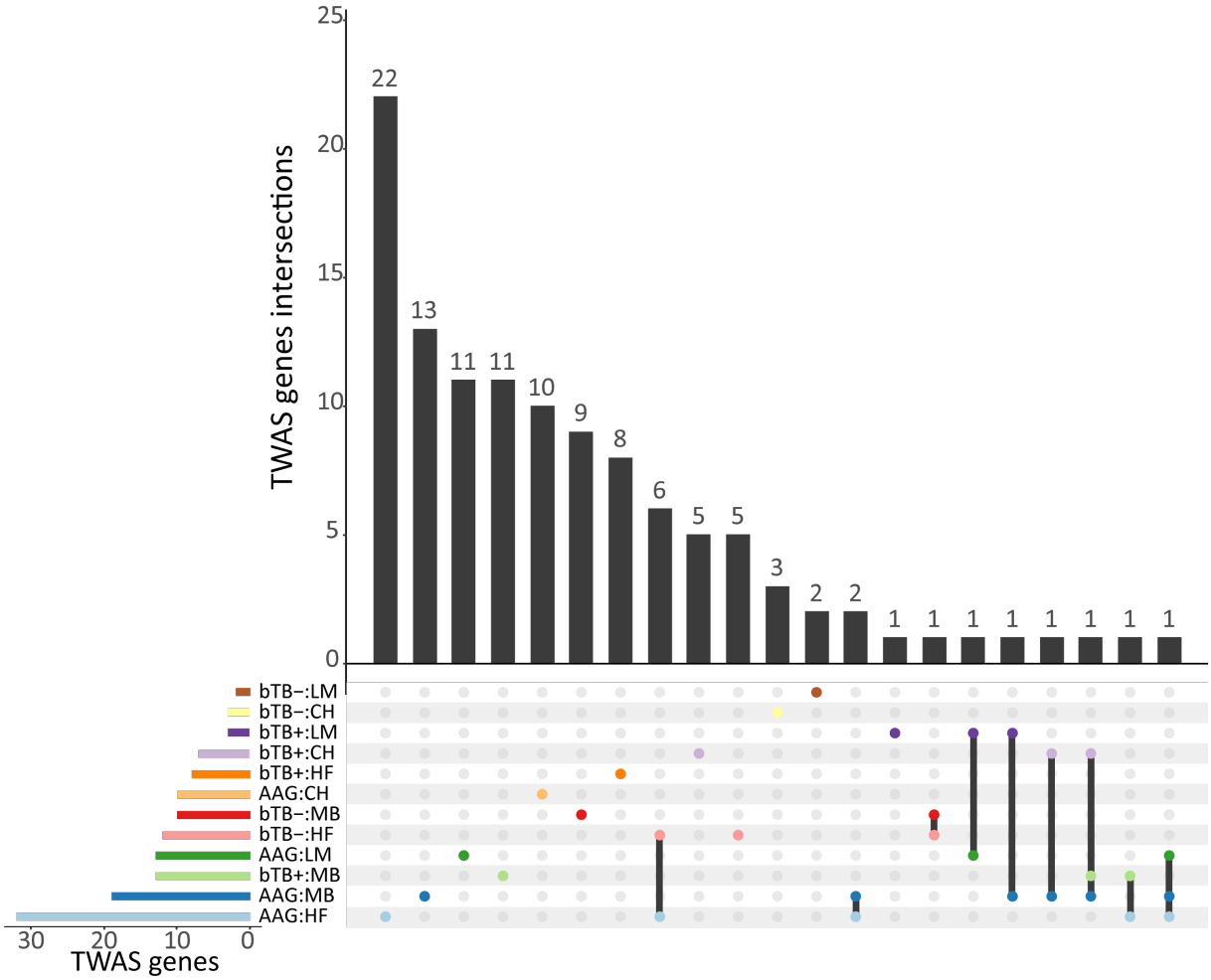
Supplementary Figure S2.10: Distance to transcriptional start site of *cis*-eQTLs. Comparison of the distances to transcriptional start site of all cis-eQTLs identified via the permutation and conditional analysis across the control (bTB−), reactor (bTB+), and combined all animals (AAG) cohorts.



Supplementary Figure S2.11: *Cis*-heritability of gene expression is predominantly governed by additive genetic variance. a Distribution of narrow-sense *cis*-heritability (*cis*-h^2) estimates for all genes (n = 14,563) tested in the *cis*-eQTL analysis in the all animal group (AAG) cohort and for which a GRM could be constructed for. **b** Distribution of *cis*-h^2 values for genes not considered significantly heritable (Non-sig.; likelihood ratio test (LRT) $P_{adj.}$ > 0.05) (n = 7,806) and for those considered significantly heritable (Sig.; LRT $P_{adj.}$ < 0.05) (n = 6,757). **c** Distribution of broad-sense heritability (H^2_{SNP}) estimates for 5,863 significantly heritable genes (dark-green) with this estimate decomposed into the proportion explained by additive variance (light-green) and dominance variance (δ^2_{SNP}; orange). The box plots in **a, b** and **c** cover the interquartile range with the median line denoted at the centre and mean being denoted by the black square box. The whiskers extend to the most extreme data point that is no more than 1.5 × IQR from the edge of the box. *P*-values in figures **b** and **c** are inferred from the Wilcoxon rank-sum test between the heritability estimates within each group



Supplementary Figure S2.12: Visualisation of *cis*-eQTL effect size and relationship to minor allele frequency (MAF). Solid trend lines indicate *cis*-eQTLs where the effect size was negative and dashed lines indicate those where the effect size was positive and were generated using the *geom_smooth* function in R setting the method to "*loess*" in ggplot2.



Supplementary Figure S2.13: Overlap of significant TWAS genes across all 12 TWAS groups. Upset plot showing the number and intersection of significant transcriptome-wide association study (TWAS) genes (Bonferroni $P_{adj.} < 0.05$ and $P < 0.05$ post-permutation) across the bTB−, the bTB+, and AAG reference panels identified using the Charolais (CH), Limousin (LM), Holstein-Friesian (HF), and Multi-breed (MB) GWAS data sets.

References

Alexander D.H. & Lange K. (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* 12, 246.

Browning B.L., Tian X., Zhou Y. & Browning S.R. (2021) Fast two-stage phasing of large-scale sequence data. *Am J Hum Genet* 108, 1880-90.

Buja A. & Eyuboglu N. (1992) Remarks on Parallel Analysis. *Multivariate Behav Res* 27, 509-40.

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R. & Genomes Project Analysis G. (2011) The variant call format and VCFtools. *Bioinformatics* 27, 2156-8.

Danecek P., Bonfield J.K., Liddle J., Marshall J., Ohan V., Pollard M.O., Whitwham A., Keane T., McCarthy S.A., Davies R.M. & Li H. (2021) Twelve years of SAMtools and BCFtools. *Gigascience* 10.

Das S., Forer L., Schonherr S., Sidore C., Locke A.E., Kwong A., Vrieze S.I., Chew E.Y., Levy S., McGue M., Schlessinger D., Stambolian D., Loh P.R., Iacono W.G., Swaroop A., Scott L.J., Cucca F., Kronenberg F., Boehnke M., Abecasis G.R. & Fuchsberger C. (2016) Next-generation genotype imputation service and methods. *Nat Genet* 48, 1284-7.

Dutta P., Talenti A., Young R., Jayaraman S., Callaby R., Jadhav S.K., Dhanikachalam V., Manikandan M., Biswa B.B., Low W.Y., Williams J.L., Cook E., Toye P., Wall E., Djikeng A., Marshall K., Archibald A.L., Gokhale S., Kumar S., Hume D.A. & Prendergast J.G.D. (2020) Whole genome analysis of water buffalo and global cattle breeds highlights convergent signatures of domestication. *Nat Commun* 11, 4739.

Fort A., Panousis N.I., Garieri M., Antonarakis S.E., Lappalainen T., Dermitzakis E.T. & Delaneau O. (2017) MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets. *Bioinformatics* 33, 1895-7.

Lawrence M., Gentleman R. & Carey V. (2009) rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* 25, 1841-2.

Love M.I., Huber W. & Anders S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550.

Manichaikul A., Mychaleckyj J.C., Rich S.S., Daly K., Sale M. & Chen W.M. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867-73.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., de Bakker P.I., Daly M.J. & Sham P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81, 559-75.

Riggio V., Tijjani A., Callaby R., Talenti A., Wragg D., Obishakin E.T., Ezeasor C., Jongejan F., Ogo N.I., Aboagye-Antwi F., Toure A., Nzalawahej J., Diallo B., Missohou A., Belem A.M.G., Djikeng A., Juleff N., Fourie J., Labuschagne M., Madder M., Marshall K., Prendergast J.G.D. & Morrison L.J. (2022) Assessment of genotyping array performance for genome-wide association studies and imputation in African cattle. *Genet Sel Evol* 54, 58.

Rosen B.D., Bickhart D.M., Schnabel R.D., Koren S., Elsik C.G., Tseng E., Rowan T.N., Low W.Y., Zimin A., Couldrey C., Hall R., Li W., Rhie A., Ghurye J., McKay S.D., Thibaud-Nissen F., Hoffman J., Murdoch B.M., Snelling W.M., McDaneld T.G., Hammond J.A., Schwartz J.C., Nandolo W., Hagen D.E., Dreischer C., Schultheiss S.J., Schroeder S.G., Phillippy A.M., Cole J.B., Van Tassell C.P., Liu G., Smith T.P.L. & Medrano J.F. (2020) De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* 9.

Verma S.S., de Andrade M., Tromp G., Kuivaniemi H., Pugh E., Namjou-Khales B., Mukherjee S., Jarvik G.P., Kottyan L.C., Burt A., Bradford Y., Armstrong G.D., Derr K., Crawford D.C., Haines J.L., Li R., Crosslin D. & Ritchie M.D. (2014) Imputation and quality control steps for combining multiple genome-wide datasets. *Front Genet* 5, 370.

Winkler T.W., Day F.R., Croteau-Chonka D.C., Wood A.R., Locke A.E., Magi R., Ferreira T., Fall T., Graff M., Justice A.E., Luan J., Gustafsson S., Randall J.C., Vedantam S., Workalemahu T., Kilpelainen T.O., Scherag A., Esko T., Kutalik Z., Heid I.M., Loos R.J. & Genetic Investigation of Anthropometric Traits C. (2014) Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* 9, 1192-212.

Zhou H.J., Li L., Li Y., Li W. & Li J.J. (2022) PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol* 23, 210.