Mapping of *cis*-regulatory variants by differential allelic expression analysis identifies candidate risk variants and target genes of 27 breast cancer risk loci

Joana M. Xavier *et al*.

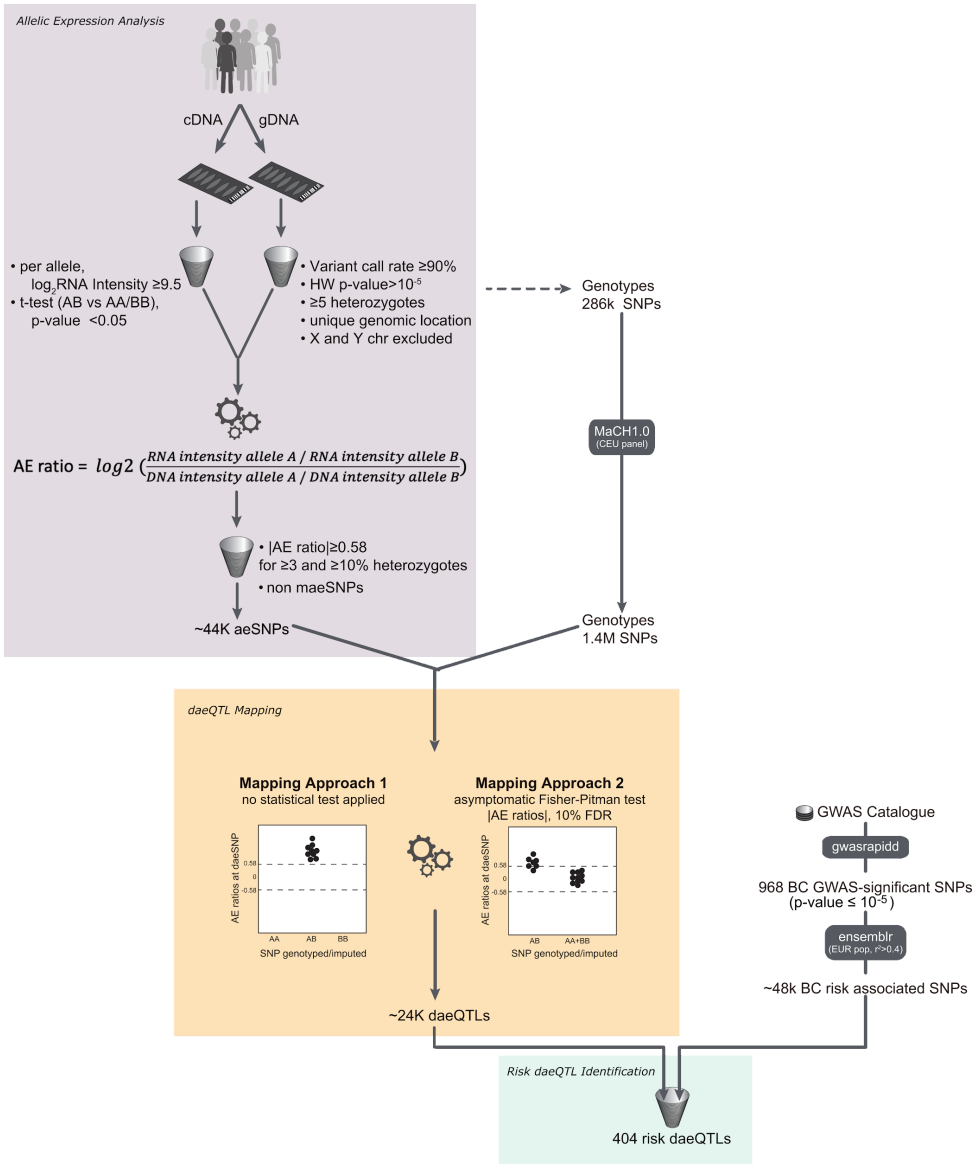


Figure S1: Schematic representation of the methodology workflow to identify daeSNPs, daeQTLs and risk daeQTLs. The workflow is divided into five steps: (i) allelic expression analysis (purple); (ii) genotype imputation of SNPs utilizing MACH; (iii) daeQTL mapping (orange); (iv) breast cancer risk associated variants and proxies download; and (v) risk daeQTL identification (green). Legend: HW - Hardy-Weinberg; AE ratio - allelic expression ratio; maeSNP - mono-allelic expressed SNP; aeSNP - allelic expressed SNP.

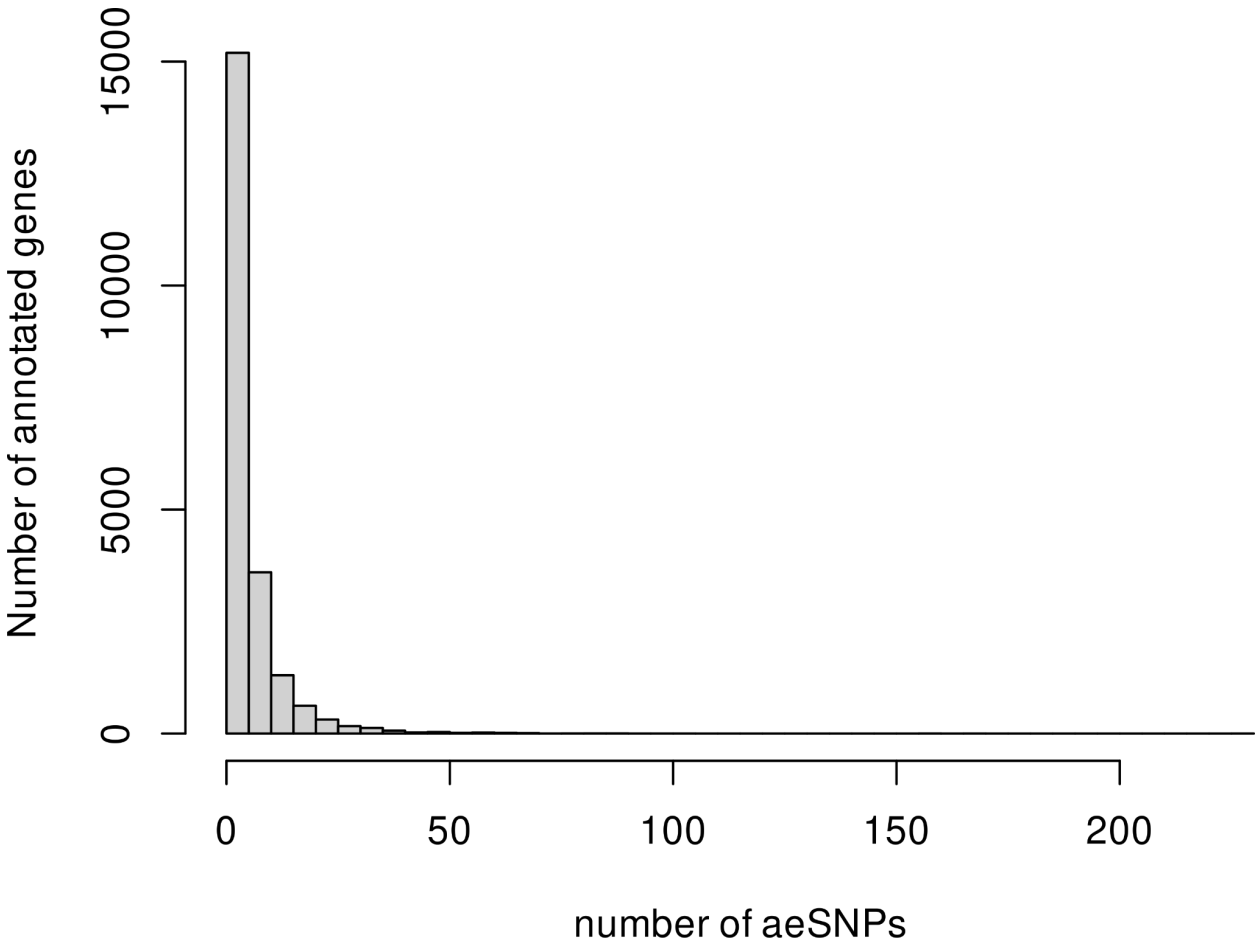


Figure S2: Histogram of the number of aeSNPs per gene across 21,527 annotated Ensembl genes.

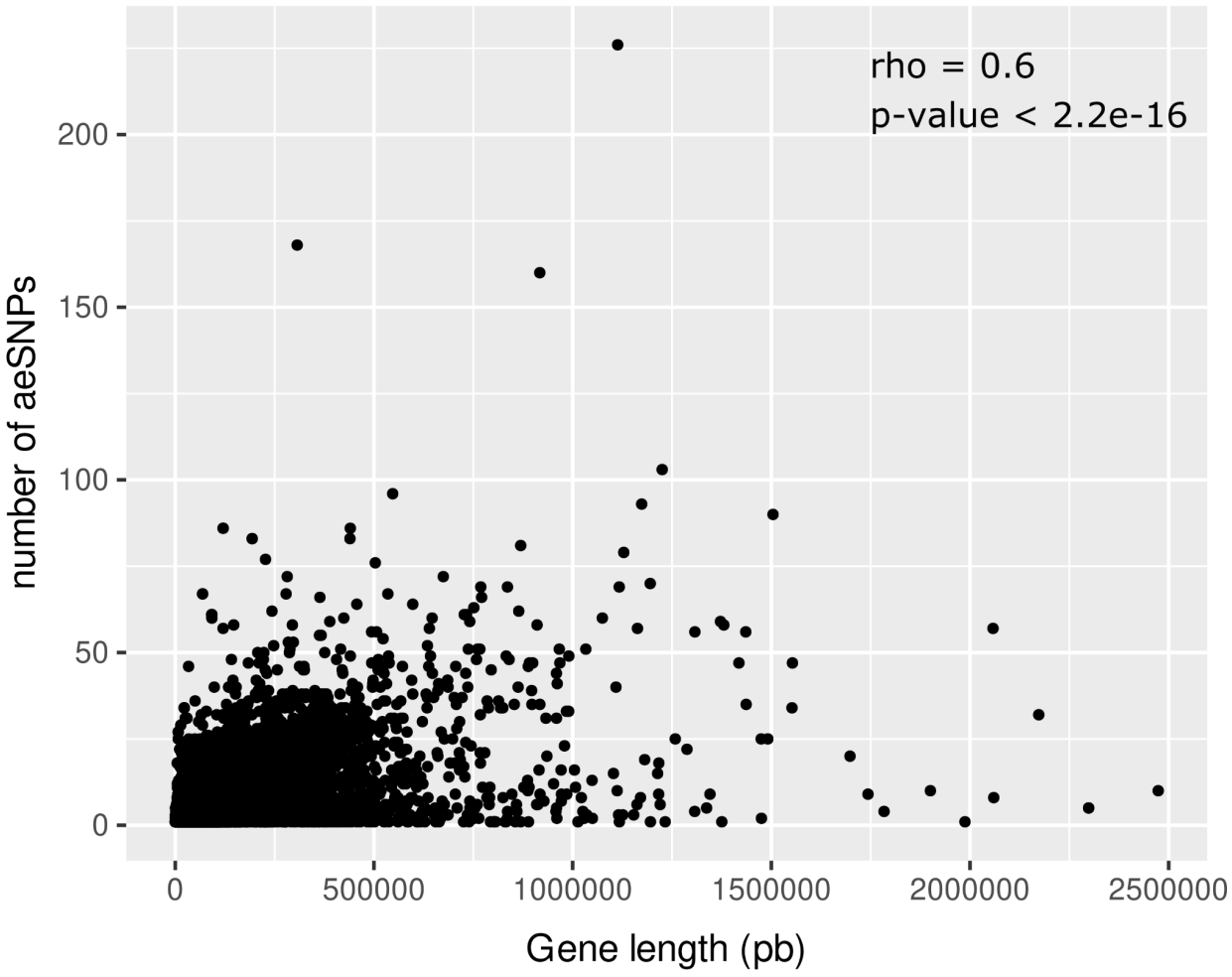


Figure S3: Distribution of the number of aeSNPs analysed according to gene size. Rho and p-value correspond to a Pearson's correlation analysis.

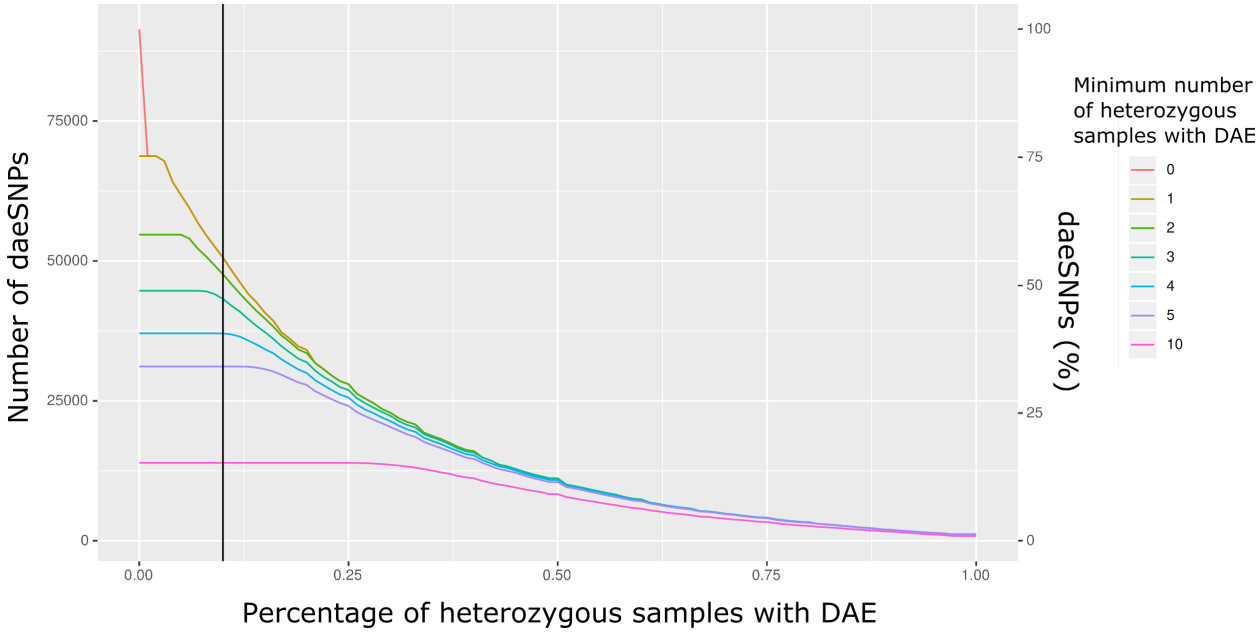


Figure S4: Criteria and thresholds to define daeSNPs. The number of daeSNPs identified (y-axis) varies according to the threshold used for percentage (x-axis) and minimum number of heterozygous individuals (coloured differently) with DAE. aeSNPs were defined as daeSNPs when at least 3 samples and 10% of the heterozygotes showed |AE values| ≥ 0.58.

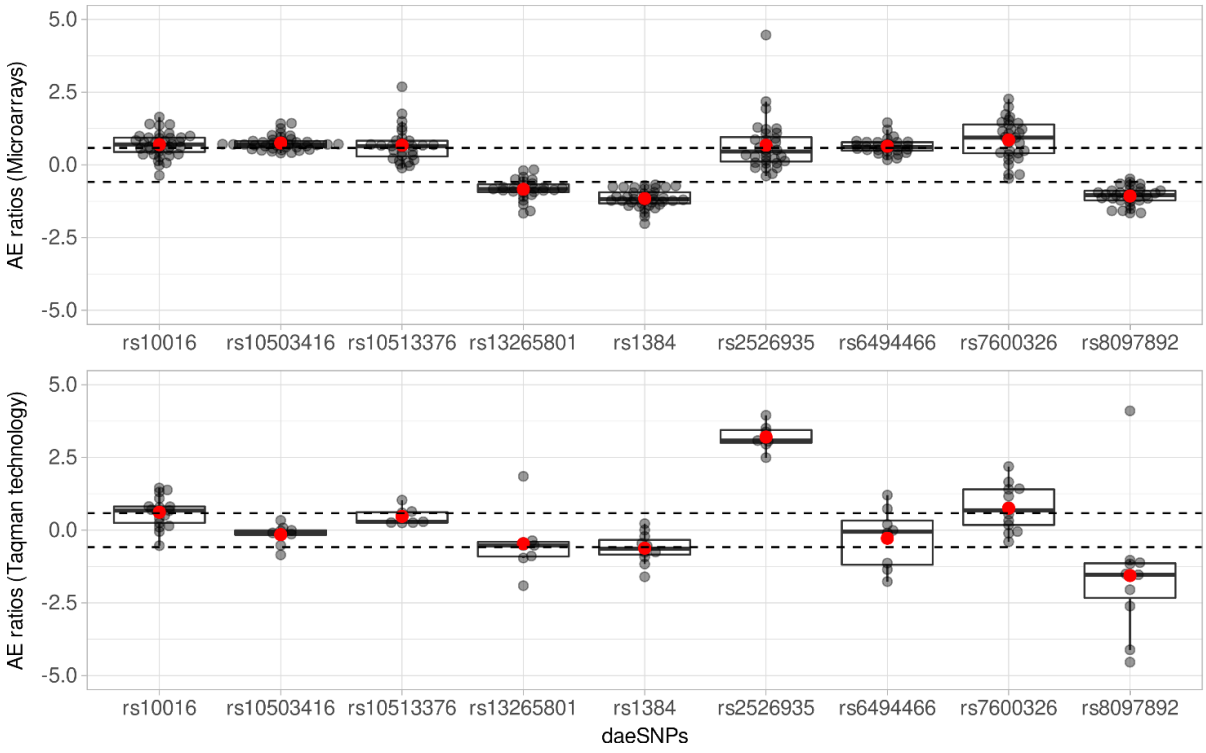


Figure S5: Validation of nine daeSNPs by Taqman PCR technology. AE ratios for nine daeSNPs measured by Microarrays (above) and Taqman PCR technology (below) are represented on the y-axis, and each data point in grey corresponds to a heterozygous individual for the SNP indicated on the x-axis. The red points correspond to the medium value of AE ratios at each snp. Fisher's exact test for concordance between number of validated daeSNP showed a p-value > 0.05.

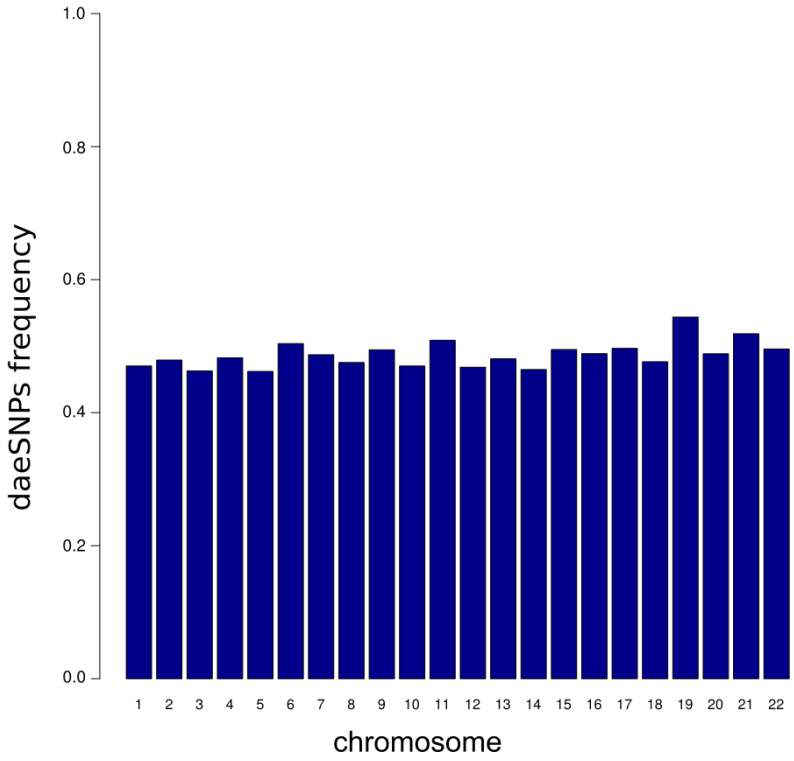


Figure S6: Percentage of daeSNPs across the human autosomal chromosomes. daeSNP frequency is defined by the number of daeSNPs divided by the total number of aeSNPs analysed.

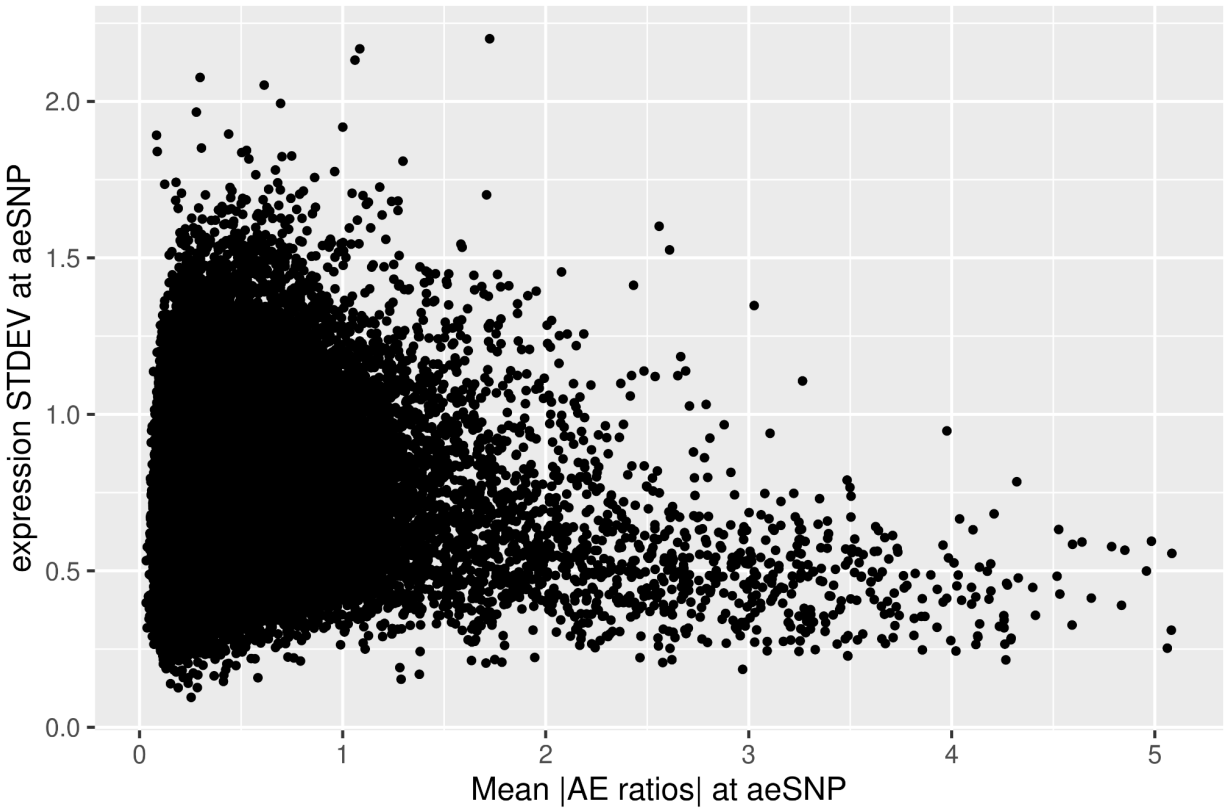
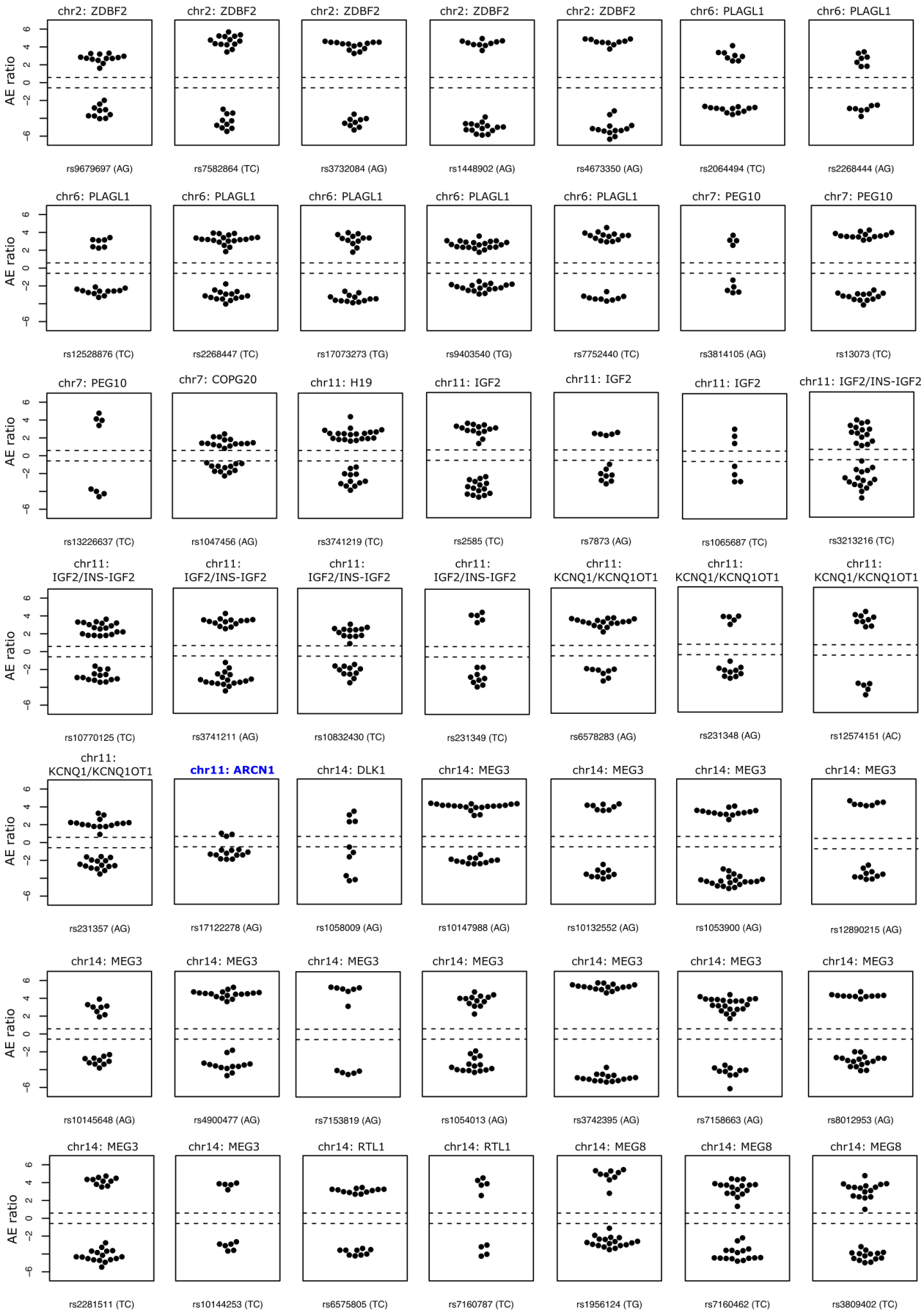


Figure S7: Distribution of the mean of the absolute value of AE ratios at each aeSNP according to the expression standard deviation, in microarray data.


chr2: ZDBF2
chr2: ZDBF2
chr2: ZDBF2
chr2: ZDBF2
chr2: ZDBF2
chr6: PLAGL1
chr6: PLAGL1
AE ratio
rs9679697 (AG)
rs7582864 (TC)
rs3732084 (AG)
rs1448902 (AG)
rs4673350 (AG)
rs2064494 (TC)
rs2268444 (AG)
chr6: PLAGL1
chr6: PLAGL1
chr6: PLAGL1
chr6: PLAGL1
chr6: PLAGL1
chr7: PEG10
chr7: PEG10
AE ratio
rs12528876 (TC)
rs2268447 (TC)
rs17073273 (TG)
rs9403540 (TG)
rs7752440 (TC)
rs3814105 (AG)
rs13073 (TC)
chr7: PEG10
chr7: COPG20
chr11: H19
chr11: IGF2
chr11: IGF2
chr11: IGF2
chr11: IGF2/INS-IGF2
AE ratio
rs13226637 (TC)
rs1047456 (AG)
rs3741219 (TC)
rs2585 (TC)
rs7873 (AG)
rs1065687 (TC)
rs3213216 (TC)
chr11:
IGF2/INS-IGF2
chr11:
IGF2/INS-IGF2
chr11:
IGF2/INS-IGF2
chr11:
IGF2/INS-IGF2
chr11:
KCNQ1/KCNQ1OT1
chr11:
KCNQ1/KCNQ1OT1
chr11:
KCNQ1/KCNQ1OT1
AE ratio
rs10770125 (TC)
rs3741211 (AG)
rs10832430 (TC)
rs231349 (TC)
rs6578283 (AG)
rs231348 (AG)
rs12574151 (AC)
chr11:
KCNQ1/KCNQ1OT1
chr11: ARCN1
chr14: DLK1
chr14: MEG3
chr14: MEG3
chr14: MEG3
chr14: MEG3
AE ratio
rs231357 (AG)
rs17122278 (AG)
rs1058009 (AG)
rs10147988 (AG)
rs10132552 (AG)
rs1053900 (AG)
rs12890215 (AG)
chr14: MEG3
chr14: MEG3
chr14: MEG3
chr14: MEG3
chr14: MEG3
chr14: MEG3
chr14: MEG3
AE ratio
rs10145648 (AG)
rs4900477 (AG)
rs7153819 (AG)
rs1054013 (AG)
rs3742395 (AG)
rs7158663 (AG)
rs8012953 (AG)
chr14: MEG3
chr14: MEG3
chr14: RTL1
chr14: RTL1
chr14: MEG8
chr14: MEG8
chr14: MEG8
AE ratio
rs2281511 (TC)
rs10144253 (TC)
rs6575805 (TC)
rs7160787 (TC)
rs1956124 (TG)
rs7160462 (TC)
rs3809402 (TC)

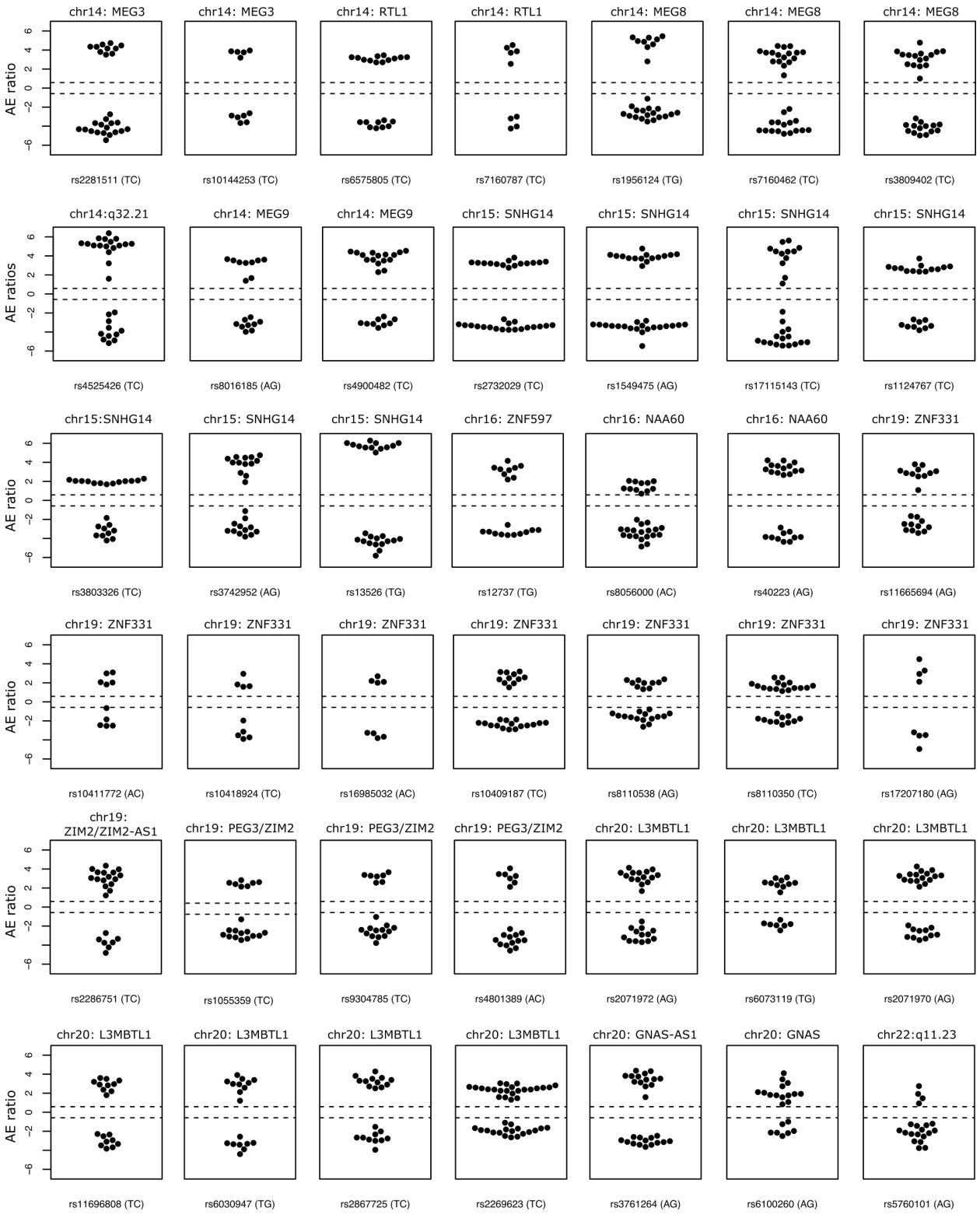

chr14: MEG3
chr14: MEG3
chr14: RTL1
chr14: RTL1
chr14: MEG8
chr14: MEG8
chr14: MEG8
AE ratio
rs2281511 (TC)
rs10144253 (TC)
rs6575805 (TC)
rs7160787 (TC)
rs1956124 (TG)
rs7160462 (TC)
rs3809402 (TC)
chr14:q32.21
chr14: MEG9
chr14: MEG9
chr15: SNHG14
chr15: SNHG14
chr15: SNHG14
chr15: SNHG14
AE ratios
rs4525426 (TC)
rs8016185 (AG)
rs4900482 (TC)
rs2732029 (TC)
rs1549475 (AG)
rs17115143 (TC)
rs1124767 (TC)
chr15:SNHG14
chr15: SNHG14
chr15: SNHG14
chr16: ZNF597
chr16: NAA60
chr16: NAA60
chr19: ZNF331
AE ratio
rs3803326 (TC)
rs3742952 (AG)
rs13526 (TG)
rs12737 (TG)
rs8056000 (AC)
rs40223 (AG)
rs11665694 (AG)
chr19: ZNF331
chr19: ZNF331
chr19: ZNF331
chr19: ZNF331
chr19: ZNF331
chr19: ZNF331
chr19: ZNF331
AE ratio
rs10411772 (AC)
rs10418924 (TC)
rs16985032 (AC)
rs10409187 (TC)
rs8110538 (AG)
rs8110350 (TC)
rs17207180 (AG)
chr19:
ZIM2/ZIM2-AS1
chr19: PEG3/ZIM2
chr19: PEG3/ZIM2
chr19: PEG3/ZIM2
chr20: L3MBTL1
chr20: L3MBTL1
chr20: L3MBTL1
AE ratio
rs2286751 (TC)
rs1055359 (TC)
rs9304785 (TC)
rs4801389 (AC)
rs2071972 (AG)
rs6073119 (TG)
rs2071970 (AG)
chr20: L3MBTL1
chr20: L3MBTL1
chr20: L3MBTL1
chr20: L3MBTL1
chr20: GNAS-AS1
chr20: GNAS
chr22:q11.23
AE ratio
rs11696808 (TC)
rs6030947 (TG)
rs2867725 (TC)
rs2269623 (TC)
rs3761264 (AG)
rs6100260 (AG)
rs5760101 (AG)

Figure S8: AE ratios for 84 maeSNPs. In each plot, data points correspond to heterozygous individuals for the SNPs indicated below with the respective AE ratios given by the y-axis. For each SNP the two alleles are shown in the order used for calculating the AE ratios. The dashed lines correspond to the threshold defined to call DAE (AE ratios ≤ -0.58 or AE ratios ≥ 0.58). The gene *ARCN1*, suggested as a candidate novel mono-allelic expressed protein-coding gene is highlighted in blue.

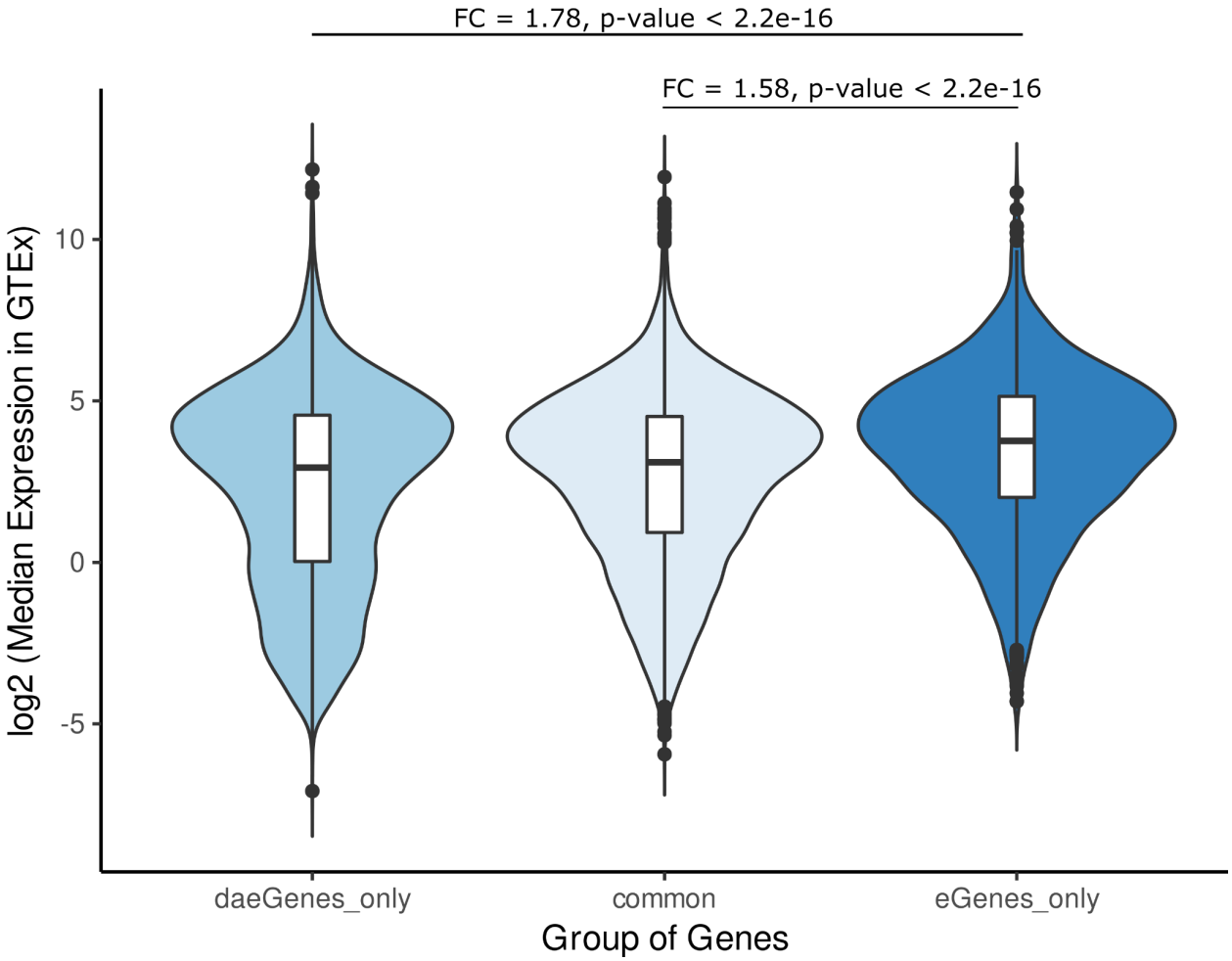


Figure S9: Comparison of total gene expression levels between daeGenes and eGenes. It is shown the fold-change (FC) and the p-values from a two-sample Wilcoxon test between daeGenes_only and eGenes_only, and between common (genes that are both daeGenes and eGenes) and eGenes_only.

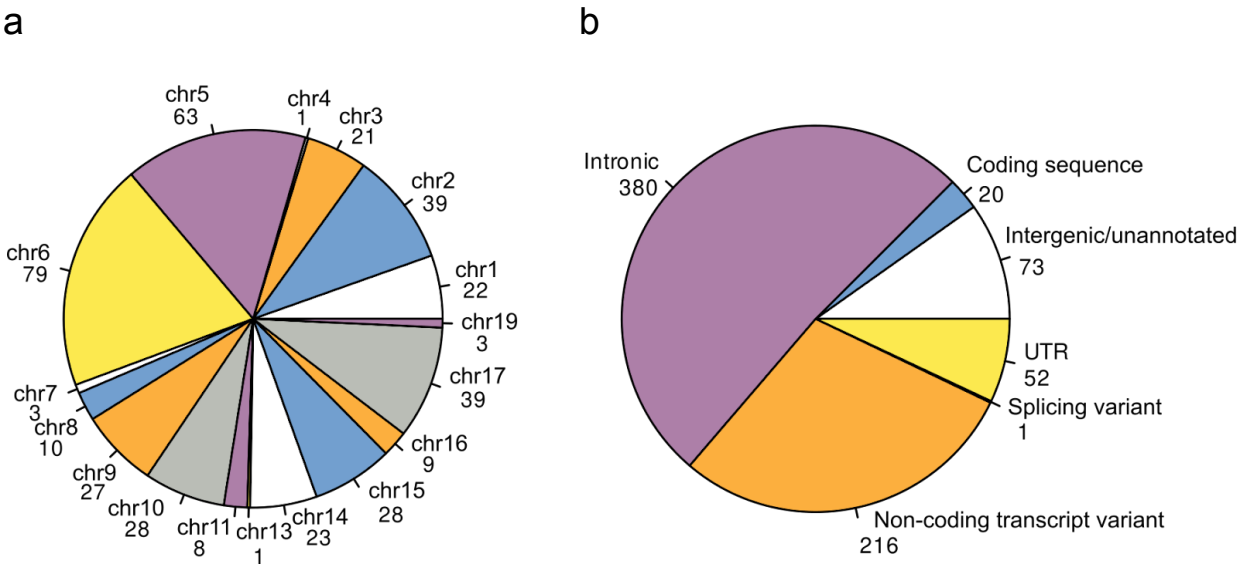


Figure S10: Distribution of the risk-daeQTLs according to their genomic location and consequence types. a) Distribution according to chromosome location. b) Distribution according to transcript consequence types. Variants can have more than one consequence type due to co-localization to different transcripts.

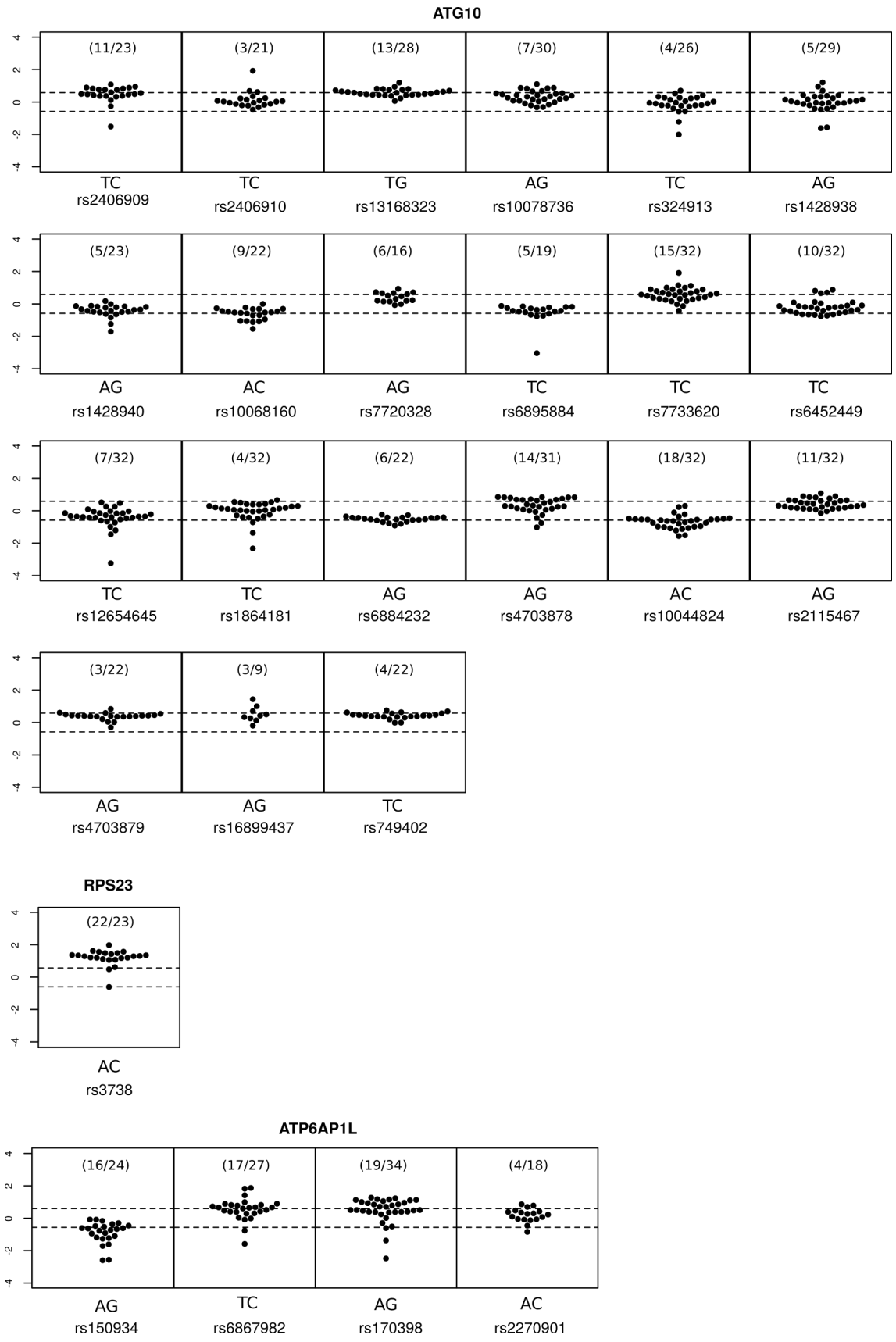

ATG10
(11/23)
(3/21)
(13/28)
(7/30)
(4/26)
(5/29)
TC rs2406909
TC rs2406910
TG rs13168323
AG rs10078736
TC rs324913
AG rs1428938
(5/23)
(9/22)
(6/16)
(5/19)
(15/32)
(10/32)
AG rs1428940
AC rs10068160
AG rs7720328
TC rs6895884
TC rs7733620
TC rs6452449
(7/32)
(4/32)
(6/22)
(14/31)
(18/32)
(11/32)
TC rs12654645
TC rs1864181
AG rs6884232
AG rs4703878
AC rs10044824
AG rs2115467
(3/22)
(3/9)
(4/22)
AG rs4703879
AG rs16899437
TC rs749402
RPS23
(22/23)
AC rs3738
ATP6AP1L
(16/24)
(17/27)
(19/34)
(4/18)
AG rs150934
TC rs6867982
AG rs170398
AC rs2270901

Figure S11: AE ratios at the daeSNPs located at the *ATG10*, *RPS23* and *ATP6AP1L* genes. In each plot, the data points correspond to the heterozygous individuals indicated below the graphs, with the alleles for each SNP indicated above the rsID, and the respective AE ratios are given by the y-axis. The dashed lines correspond to the threshold defined to call DAE (AE ratios ≤ -0.58 or AE ratios ≥ 0.58).

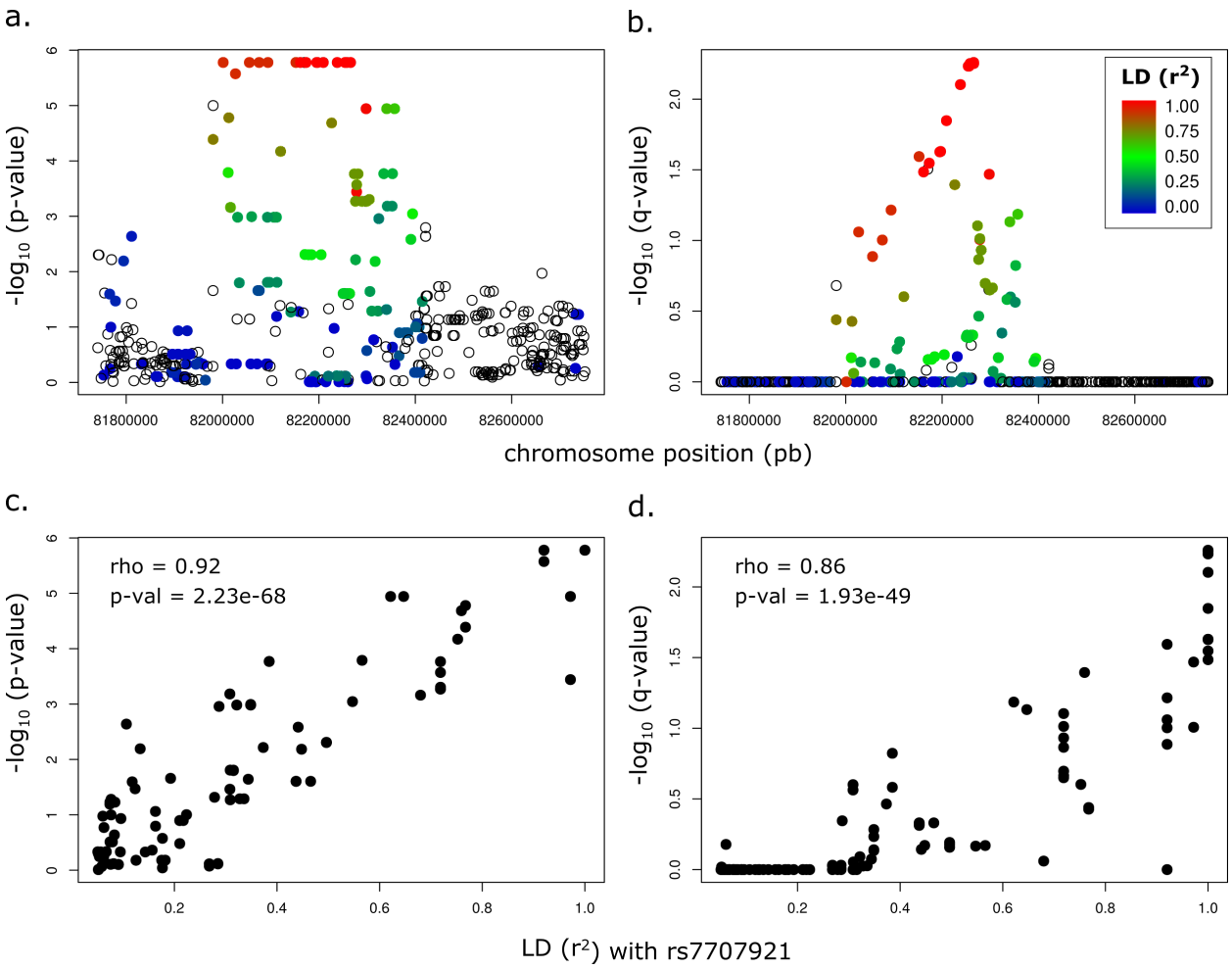


Figure S12: Summary of daeQTL mapping for four *ATG10* daeSNPs in normal breast tissue samples. *ATG10* daeSNPs analysed were: rs7733620, rs2115467, rs4703878 and rs10044824. a) and b) $-\log_{10}$(p-value) and $-\log_{10}$(q-values) of the daeQTL analysis, respectively, plotted according to the chromosome 5 position (hg38) give in the x-axis. Data points are coloured according to the LD (r^2) between the corresponding variant and the GWAS lead SNP rs7707921. c) and d) Correlation between the daeQTL analysis -log10(p-value) and -log10(q-values), respectively, and the r^2 between the corresponding variant and the GWAS lead SNP rs7707921.

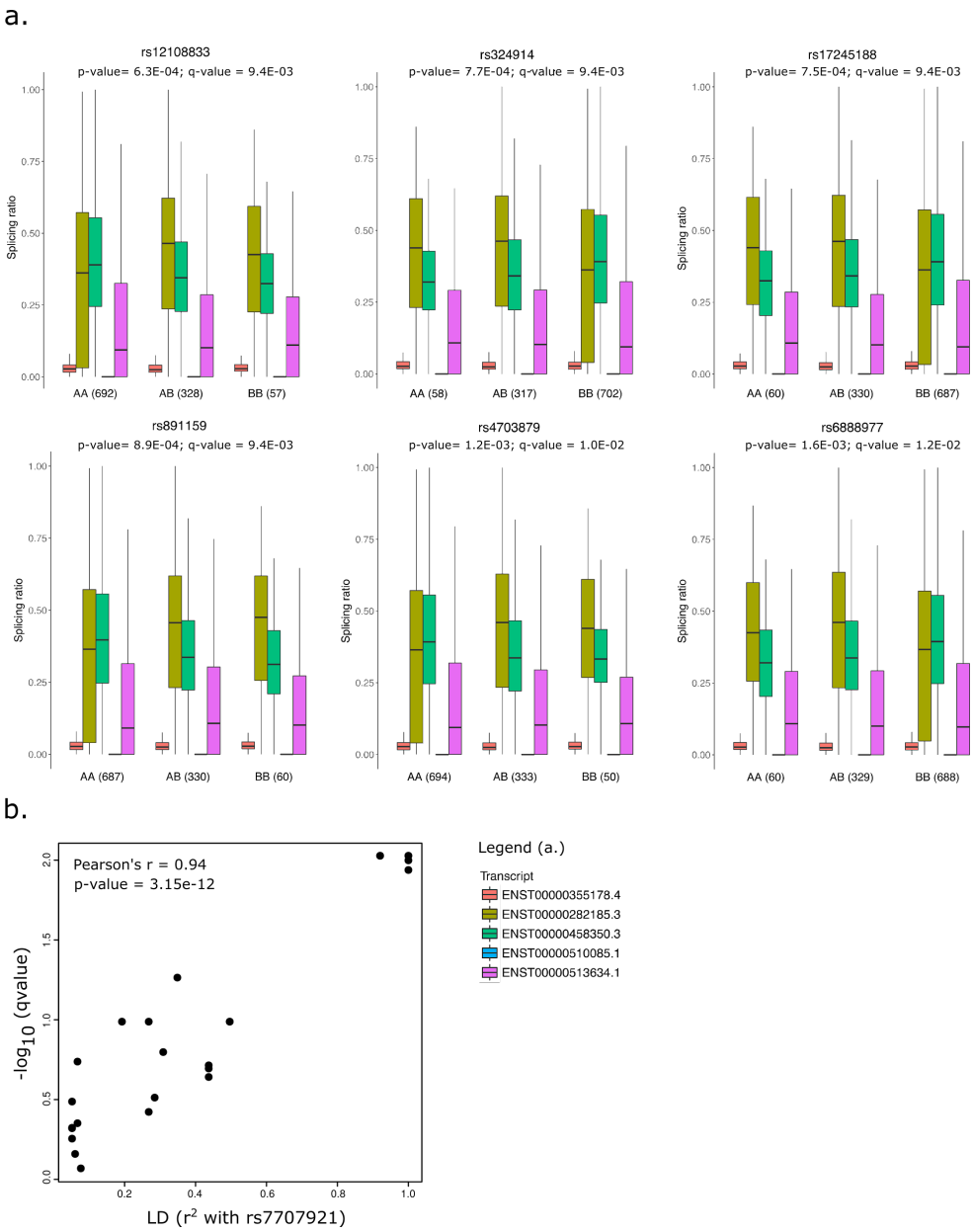


Figure S13: Variants at the 5q14.1 locus associated with alternative transcrioption of *ATG10*. a) sQTL analysis in primary solid tumours from the TCGA-BRCA project for the six significant sQTL variants identified for *ATG10*. Both the unadjusted p-value and the 5% FDR corrected p-value. b) Correlation between -$\log_{10}$(q-value) (y-axis) and the linkage disequilibrium (LD) r^2 value between each corresponding variant and rs7707921 (x-axis). The Pearson's correlation coefficient and p-value are shown.

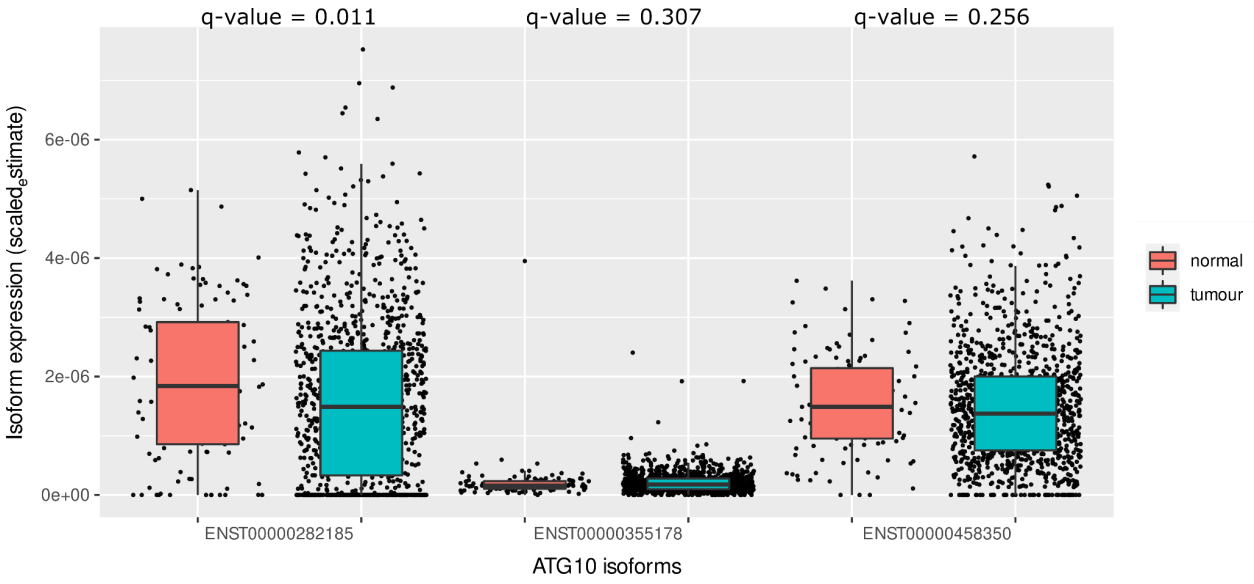


Figure S14: *ATG10* isoforms are differentially expressed between normal-matched tissue (n=113) and breast tumour samples (n=1102). Levels of expression of three *ATG10* isoforms are indicated on the y-axis from the TCGA-BRCA project. Pairwise comparisons were performed using a Mann-Whitney-Wilcoxon test and corrected q-values are presented.

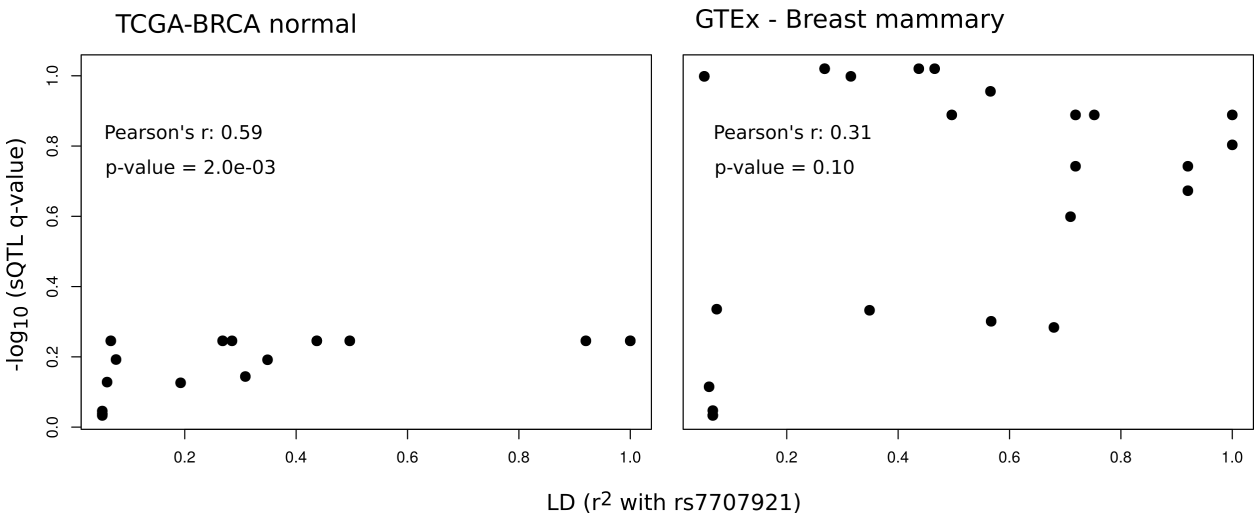


Figure S15: Correlation analysis of sQTL q-values and LD with GWAS lead SNP for TCGA-BRCA normal matched dataset and GTEx mammary tissue. Pearson's coefficient and p-value are shown for the correlation analysis.

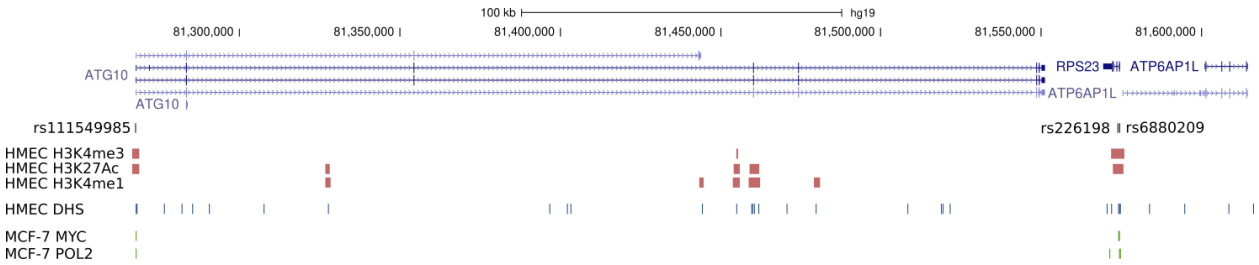


Figure S16: Genomic and epigenomic landscape of the 5q14.1-14.2 locus harbouring the candidate causal variants affecting the binding of transcription factors. Top to bottom tracks show chromosomal position according to hg19 assembly, RefSeq genes, three candidate risk rSNPs (rs111549985, rs226198 and rs6880209), histone modifications peaks in human mammary epithelial cells (HMEC), DNaseI hypersensitivity sites (DHS) in HMEC, RNA Polymerase II (POL2) and MYC binding in MCF-7 cells.

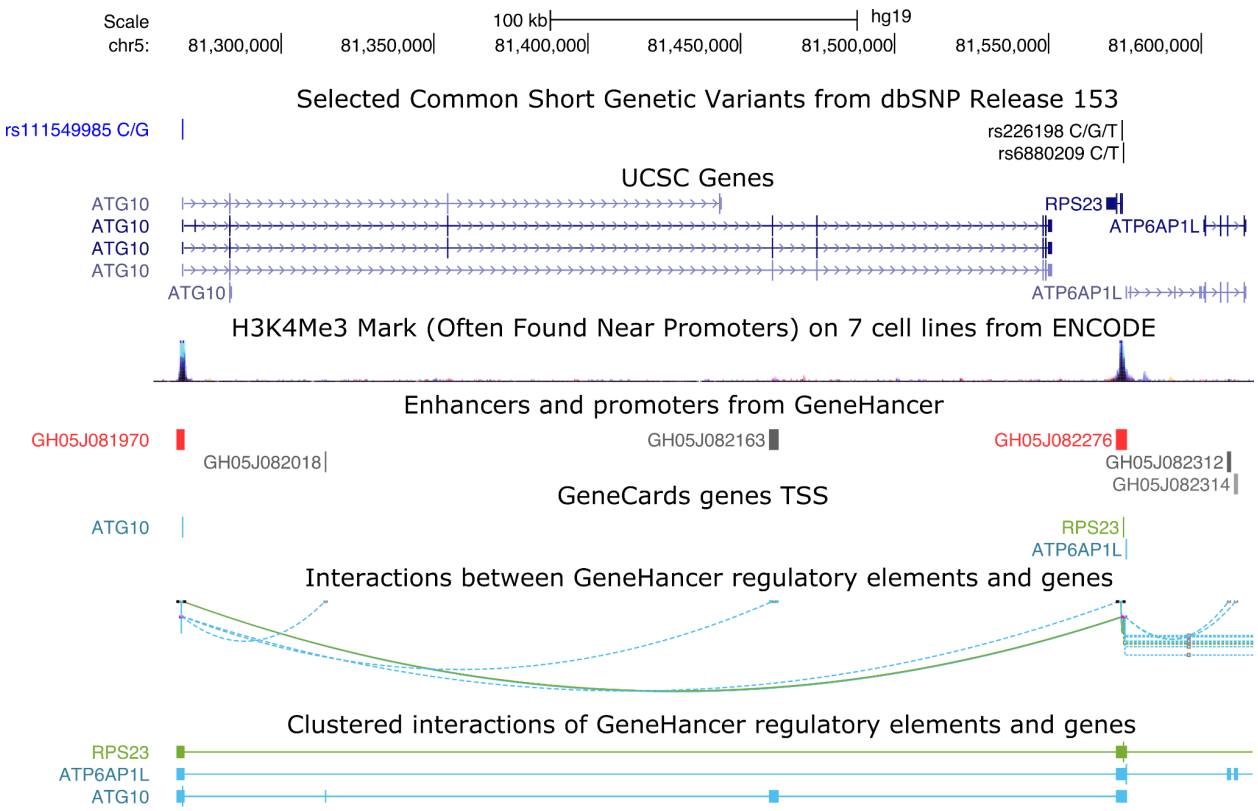


Figure S17: Genomic interactions between GeneHancer regulatory elements and genes at the 5q14.1-14.2 locus. Tracks from top to bottom show chromosomal position according to hg19 assembly, three candidate risk rSNPs, RefSeq genes, H3K4Me3 marks, and GeneHance analysis tracks. rs226198 and rs6880209 are located in *RPS23* and *ATP6AP1L* transcription start sites (TSS) and are over a GeneHancer Promoter/Enhancer (GH05J082276) predicted to interact with an *ATG10* GeneHancer Promoter/Enhancer (GH05J081970) where rs111549985 is located.

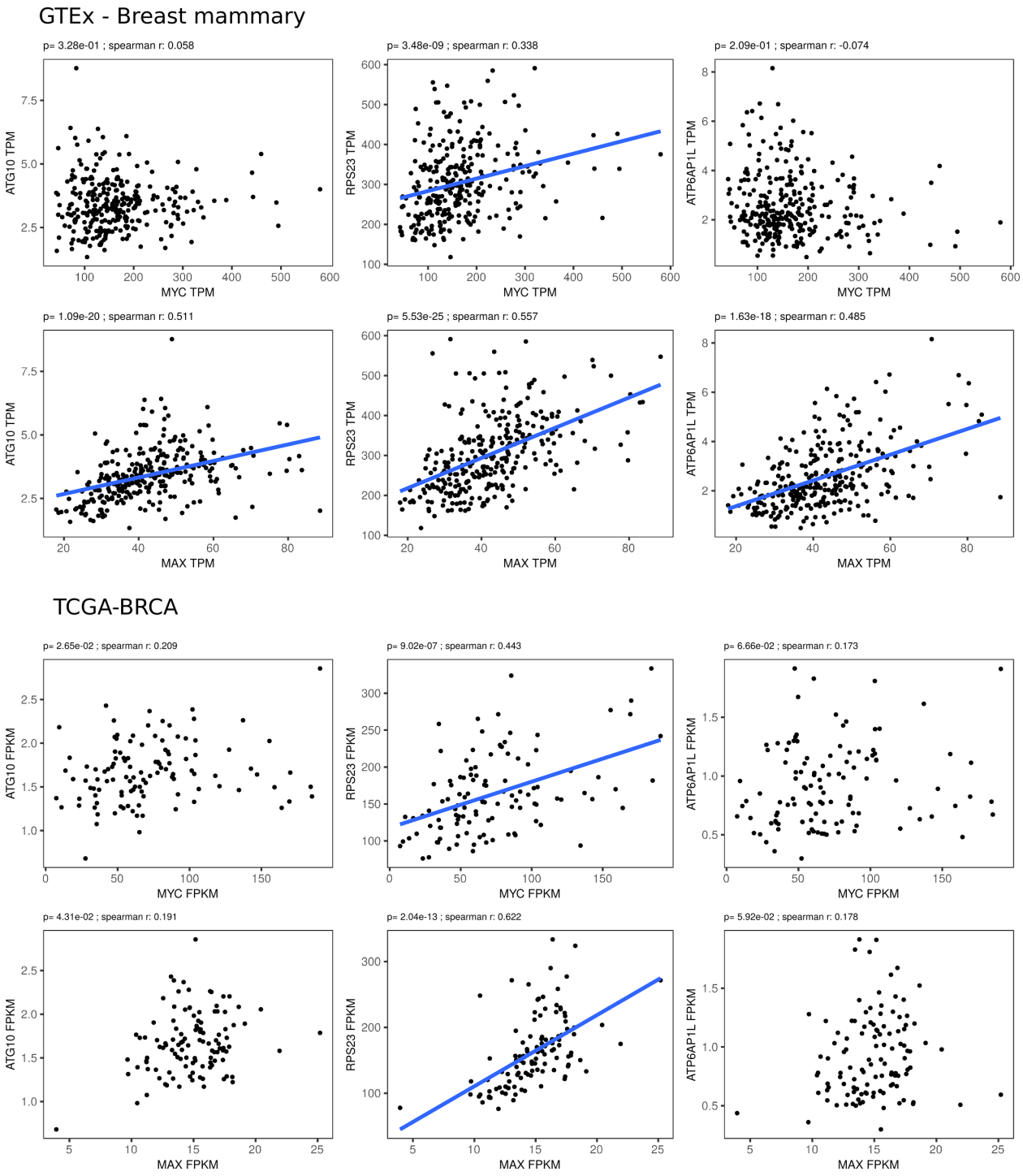


Figure S18: Correlation analysis between the gene expression of *ATG10*, *RPS23* and *ATP6AP1L* target genes and the gene expression of MYC and MAX transcription factors. The blue lines represent the significant ($p \leq 0.05$) linear regression for Spearman correlations with r greater than 0.3.

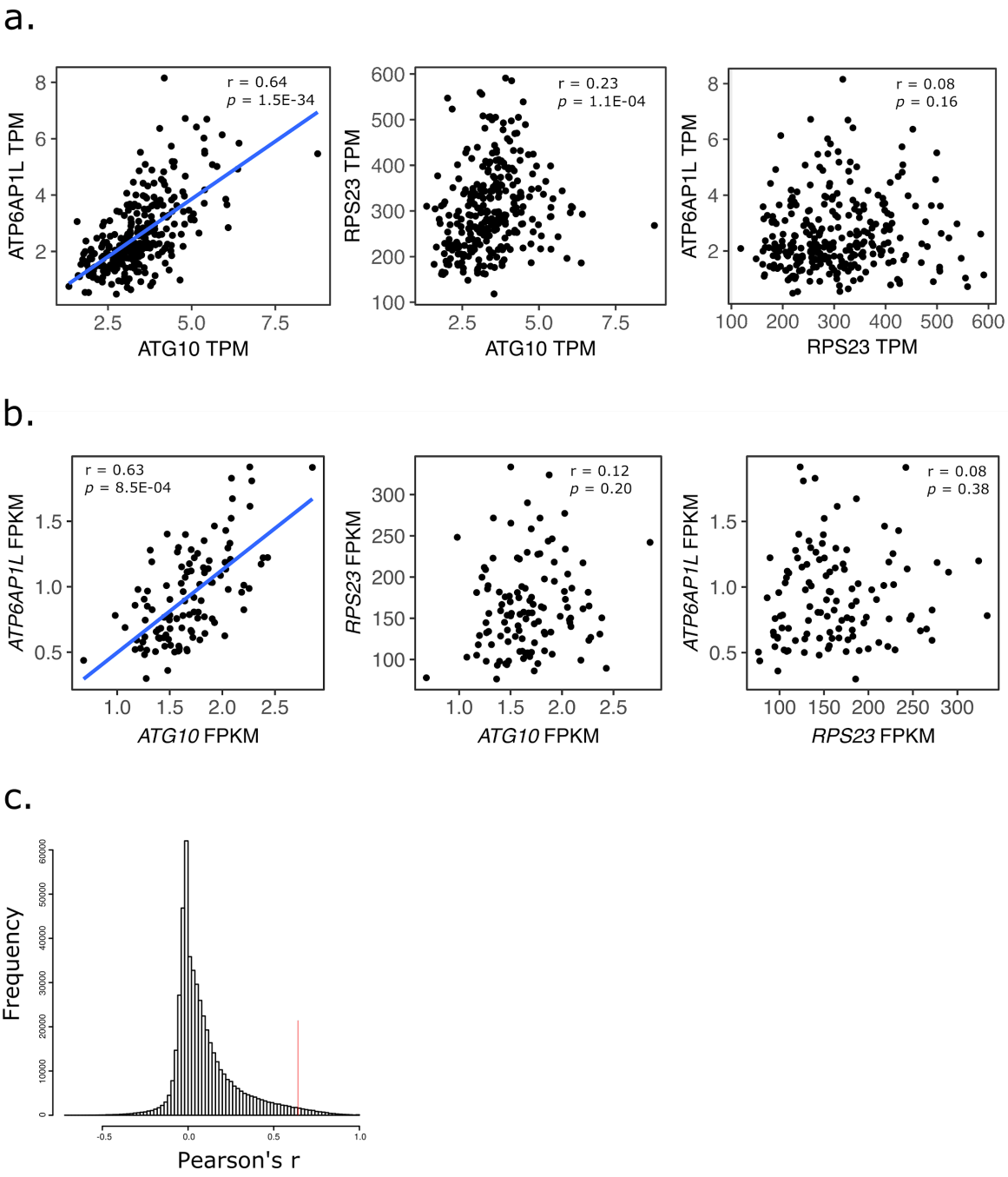


Figure S19: **Pairwise correlation analysis of the expression of the *ATG10*, *RPS23* and *ATP6AP1L* genes.** Data was analysed from a) GTEx breast mammary tissue and b) The Cancer Genome Atlas (*TCGA*) Breast invasive carcinoma (BRCA) Solid Tissue Normal. Plots show Pearson's correlation coefficient (r) and the corresponding p-value. The blue lines represent the significant ($p \leq 0.05$) linear regression for Pearson's correlations with r greater than 0.3. c) Histogram of the rank of pairwise gene expression correlation across 500,000 pairs of genes from the GTEx dataset. Red line indicates the location of an r equals to 0.64.

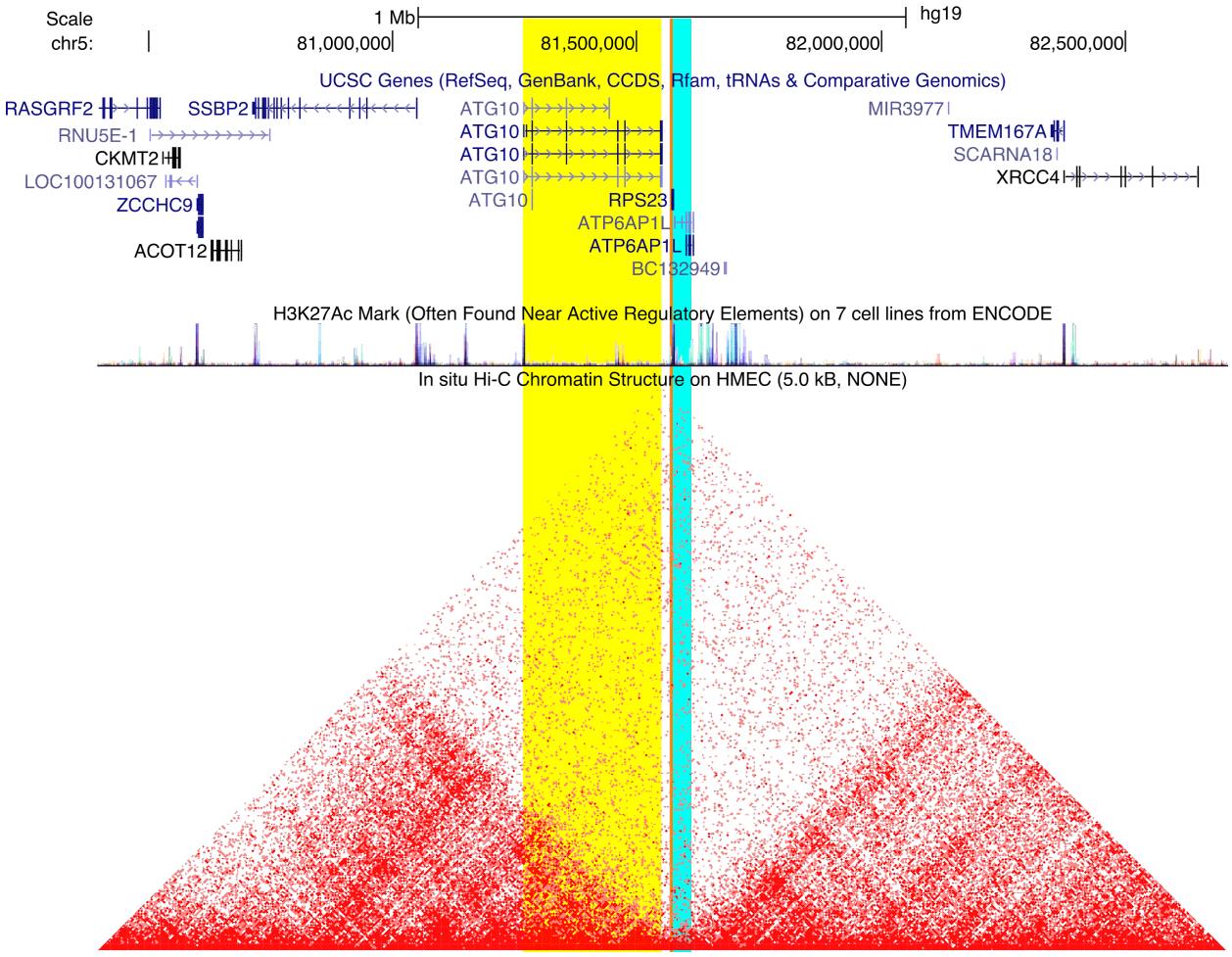


Figure S20: Genome browser visualization of two distinct topologically associating domains (TADs) at the 5q14.1 locus. *ATG10*, *RPS23* and *ATP6APL1L* genes are highlighted in yellow, orange and light blue, respectively. Red dots represent regions where DNA sequences preferentially contact each other, retrieved from Rao et al. 2014 (Hi-C track data available at the UCSC genome browser hg19).

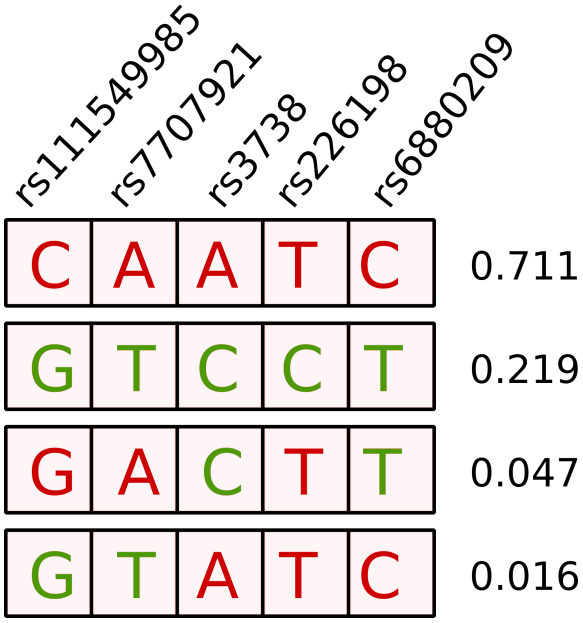


Figure S21: Haplotype analysis for candidate risk regulatory variants in the 5q14.1 risk locus. The candidate risk regulatory variants major alleles are shown in red and the minor alleles in green. The frequency of each haplotype in the 64 normal breast tissue samples analysed for DAE is shown on the right.

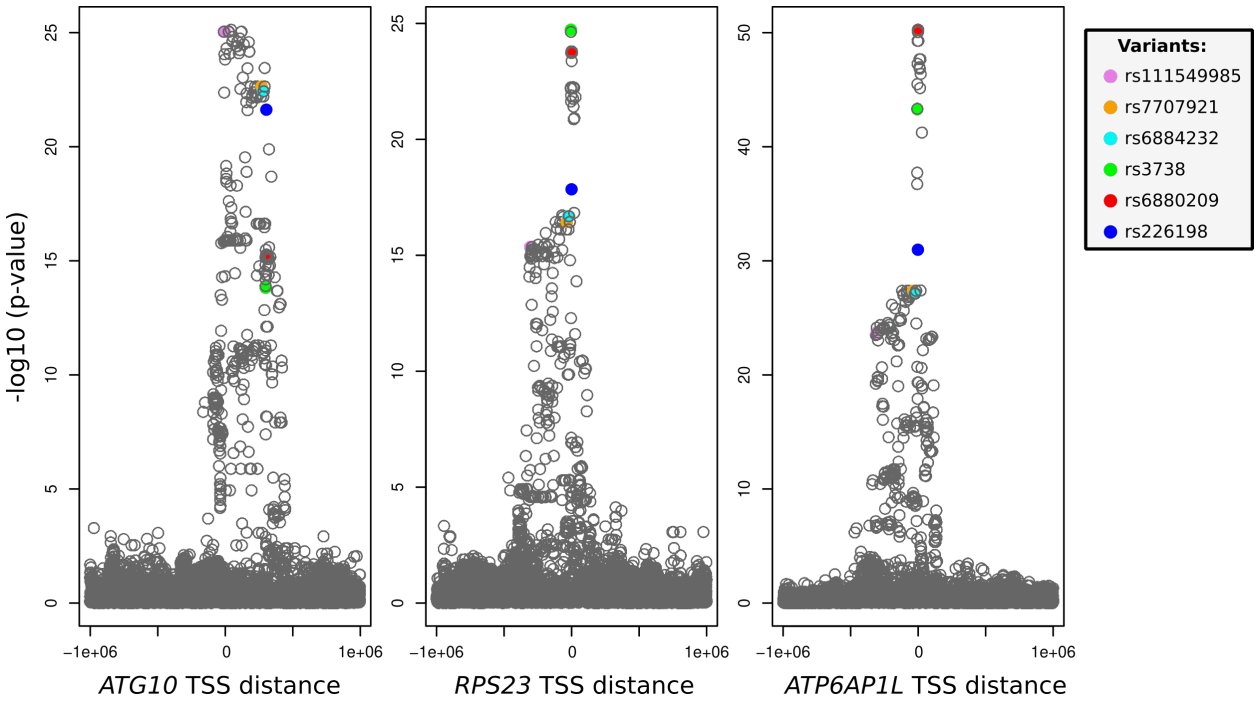


Figure S22: Summary of eQTL analysis for *ATG10*, *RPS23* and *ATP6AP1L* genes. The Y-axis represents the -$\log_{10}$(p-value) and the X-axis corresponds to the genomic distance between the variants and the transcription start site (TSS) of the corresponding gene. Data was retrieved from the Genotype-Tissue Expression (GTEx) project (http://www.gtexportal.org/home/).