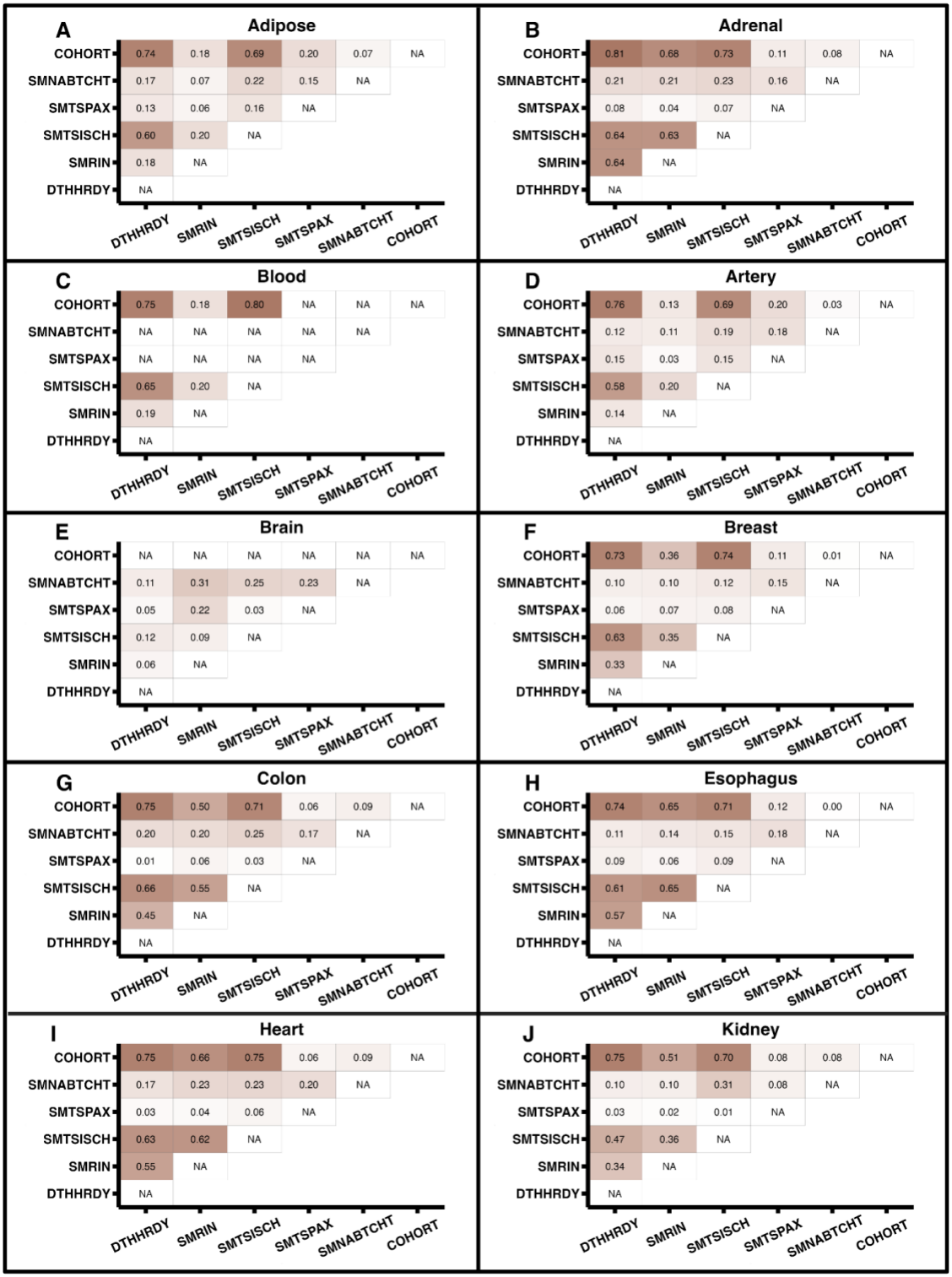
Supplementary Information

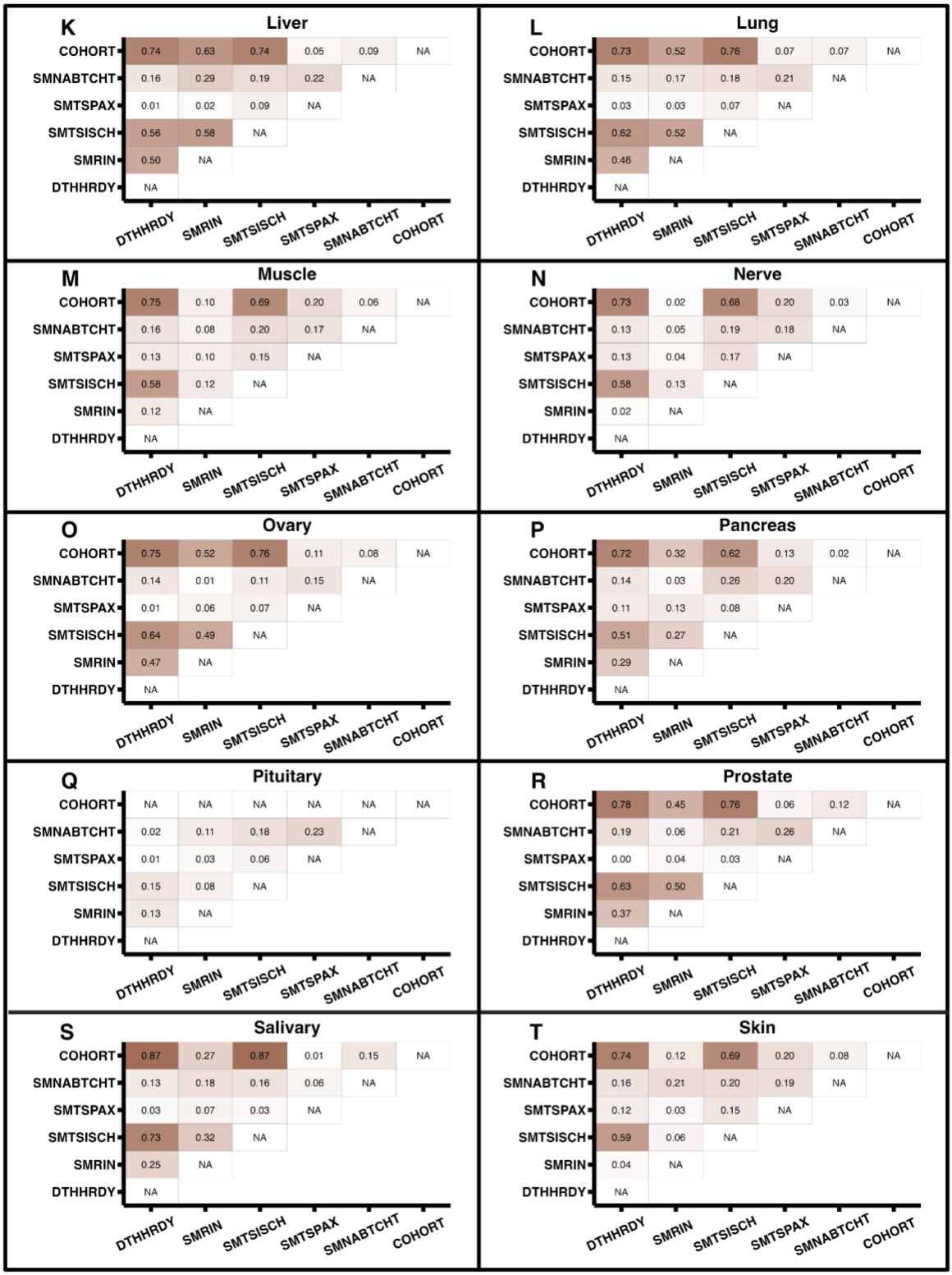
Switch-like Gene Expression Modulates Disease Risk

Supplementary Note 1: Correlation between confounders

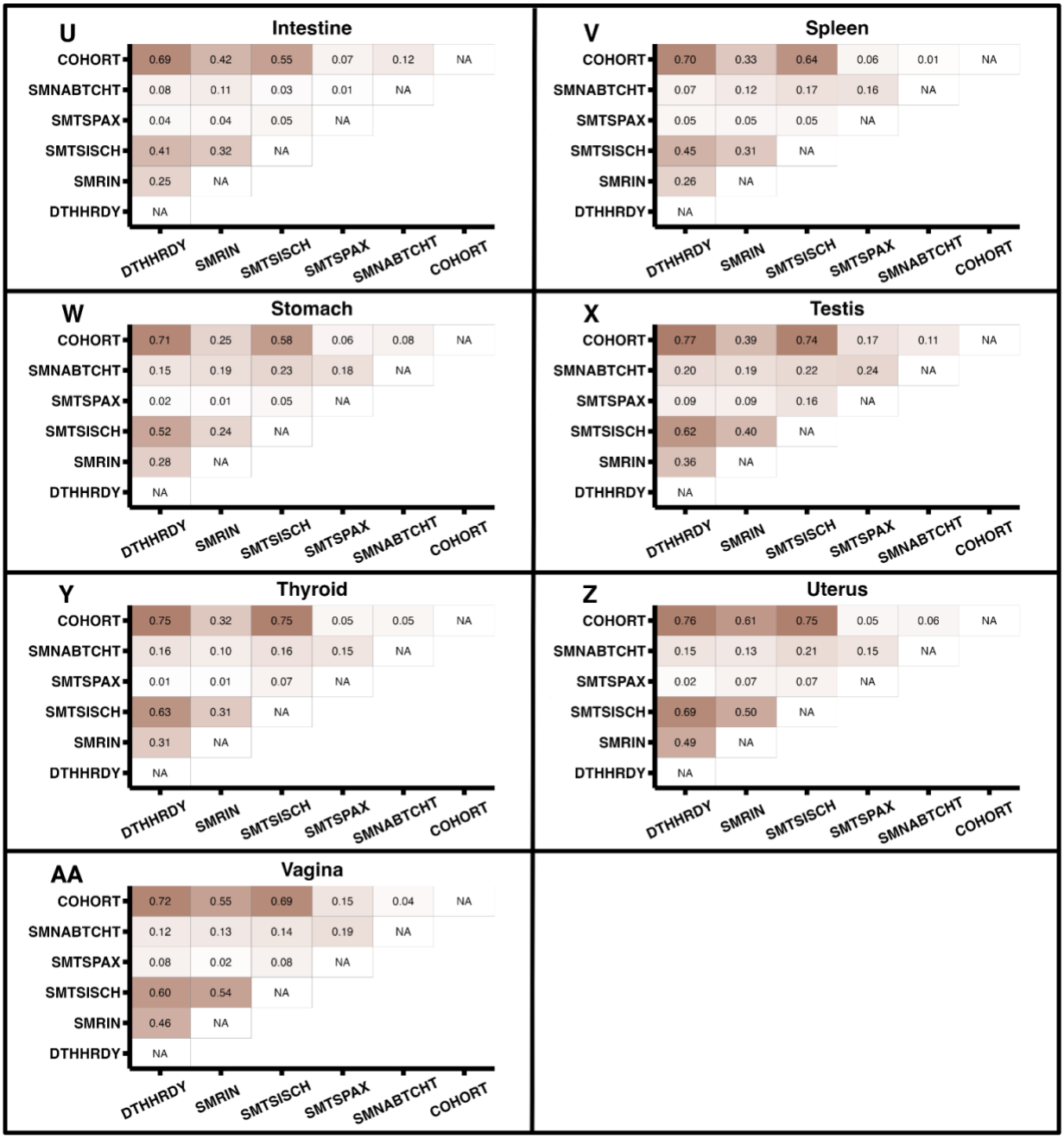
Supplementary Figure 1 shows the pairwise correlation between confounders. The figure indicates that the confounders are not independent. Therefore, we used PCA in our GAMs to address multicollinearity. Notably, across tissues, the Hardy death score (DTHHRDY), Cohort type (organ donor vs. postmortem), ischemic time (SMTSISCH), and RNA integrity number (SMRIN) show strong intercorrelations. In addition, **Supplementary Figure 3C** indicates that these four confounders, in contrast to the two other confounders, are highly correlated with raw expression levels, underscoring how the nature and timing of death affect RNA integrity and contribute to bimodality.



Supplementary Figure 1. Absolute value of Pearson's correlation coefficient, r, for each pair of confounders across samples in each tissue. Panels **A** to **AA** show results for one tissue each. **A.** Adipose. **B.** Adrenal. **C.** Blood. **D.** Artery. **E.** Brain. **F.** Breast. **G.** Colon. **H.** Esophagus. **I.** Heart. **J.** Kidney. **K.** Liver. **L.** Lung. **M.** Muscle. **N.** Nerve. **O.** Ovary. **P.** Pancreas. **Q.** Pituitary. **R.** Prostate. **S.** Salivary. **T.** Skin. **U.** Intestine. **V.** Spleen. **W.** Stomach. **X.** Testis. **Y.** Thyroid. **Z.** Uterus. **AA.** Vagina. Source data are provided as a Source Data file.



Supplementary Figure 1 continued.



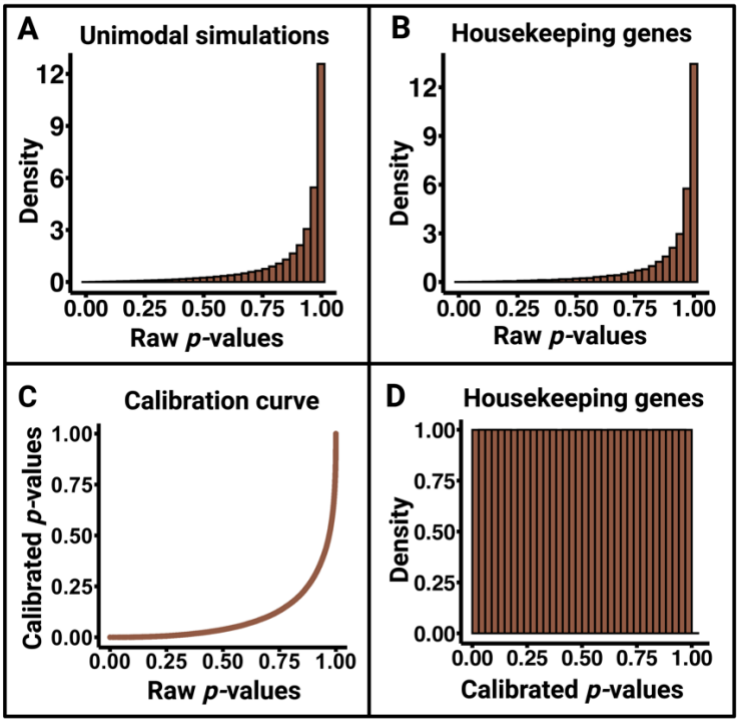
Supplementary Figure 1 continued.

Supplementary Note 2: p-value recalibration for the dip test

We evaluated the calibration of the dip test using simulated and empirical data. First, for each gene–tissue pair, we computed the mean and standard deviation of

log(TPM+1) across individuals, and then used these values as the parameters of the normal distribution to generate the same number of data points as the number of individuals available for that tissue. Under the assumption of normality, the simulated data are expected to have a unimodal distribution; hence, if the dip test were well-calibrated, the p-values obtained from the dip test on these simulated data would be uniformly distributed on the interval [0,1]. However, as shown in **Supplementary Figure 2A**, the p-value distribution is heavily biased toward 1.

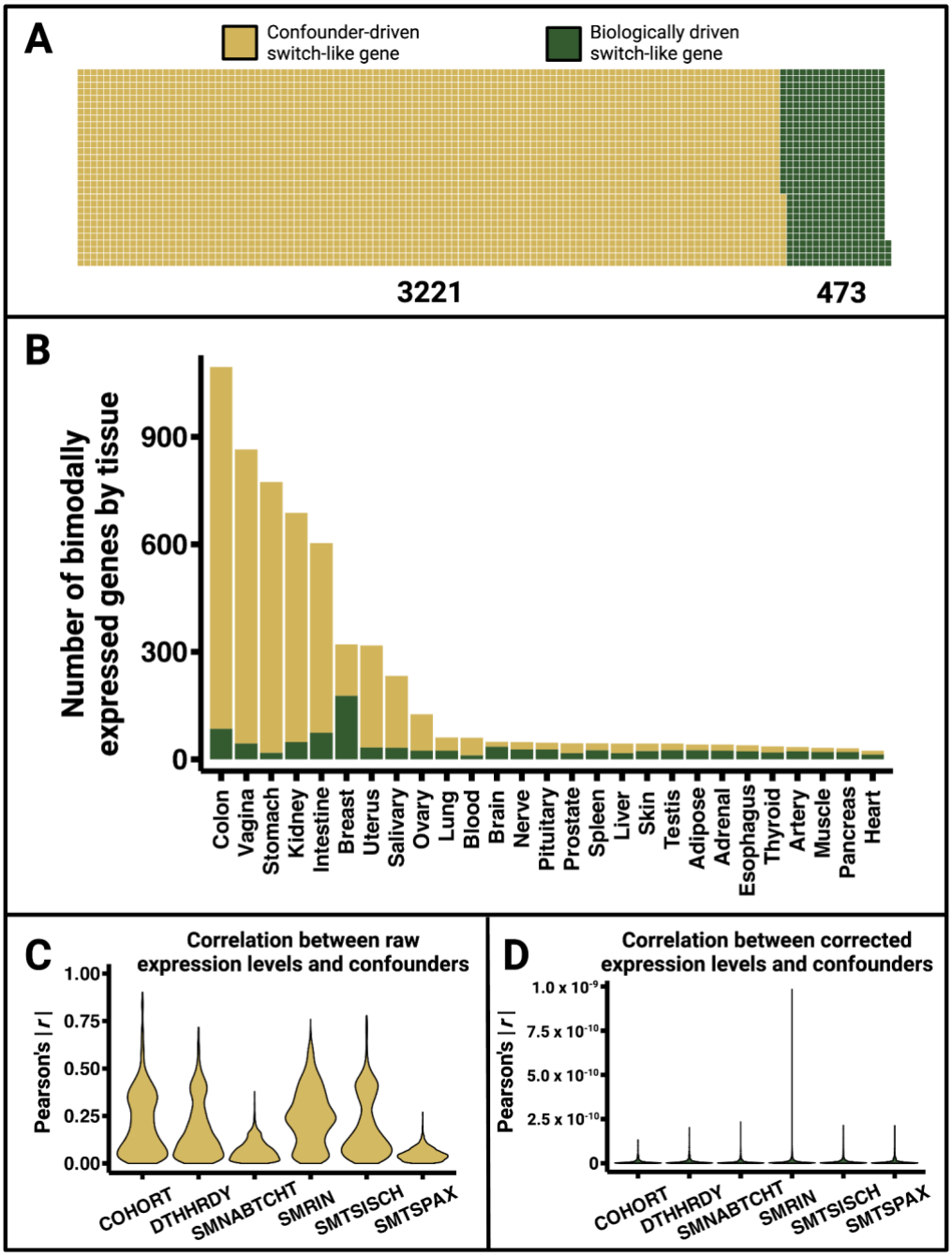
To further confirm this finding, we focused on a set of housekeeping genes obtained from the Housekeeping Transcript Atlas [1]. Housekeeping genes are characterized by low expression variability across biological contexts. We therefore expect them to be unimodally distributed in all tissues across samples. Consistent with the simulated data, dip-test p-values for housekeeping gene-tissue pairs were heavily skewed toward 1 (**Supplementary Figure 2B**), indicating that the dip test is overly conservative. To correct for this bias, we performed an empirical null p-value calibration as described in **Methods**. The resulting calibration curve is presented in **Supplementary Figure 2C**. Applying the calibration curve to the p-values from housekeeping gene-tissue pairs yields uniformly distributed calibrated p-values shown in **Supplementary Figure 2D**.



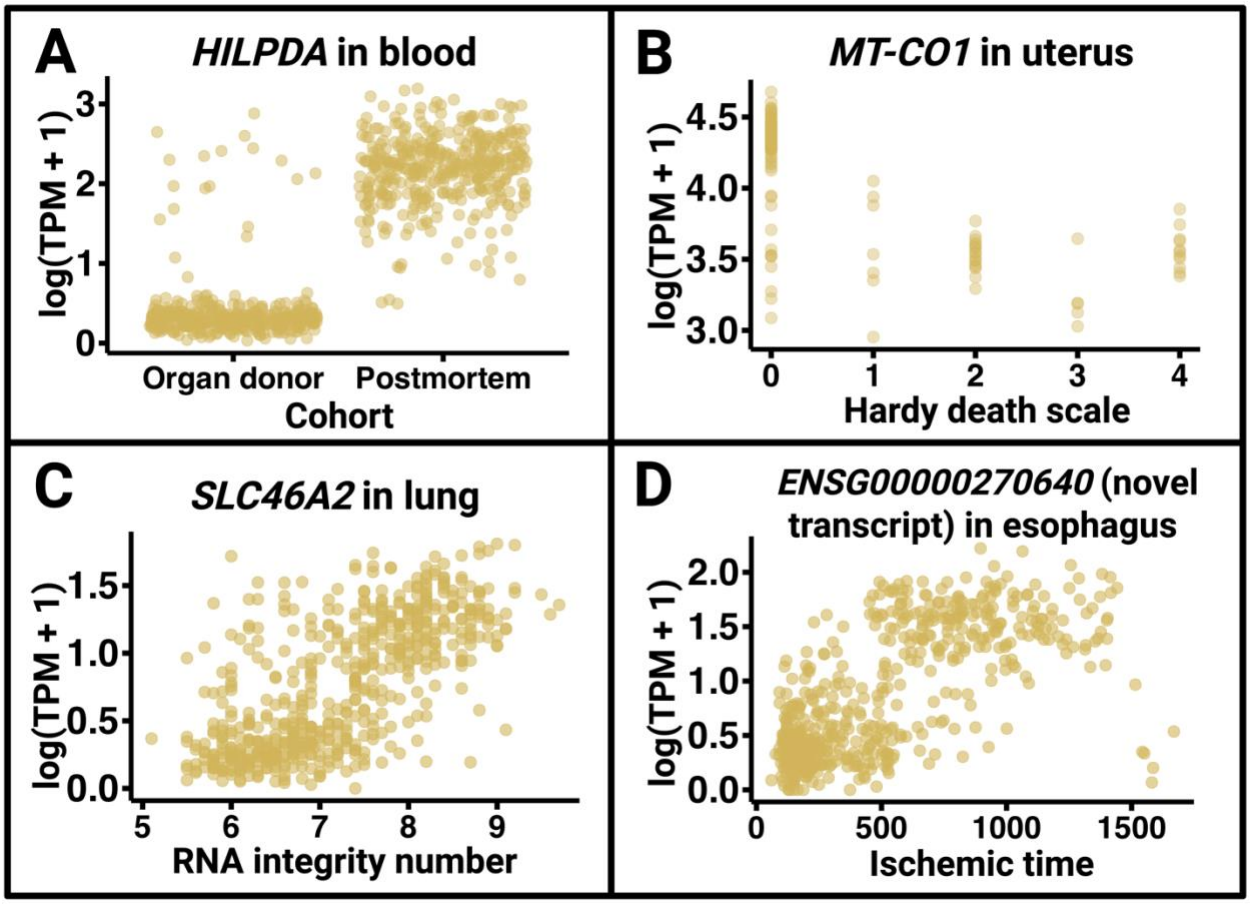
Supplementary Figure 2. p-value recalibration for the dip test. A. Distribution of raw p-values from dip tests performed on simulated unimodal data under the assumption of the normal distribution. **B.** Distribution of raw p-values from dip tests performed on the distribution of confounder-corrected expression levels across samples for housekeeping gene-tissue pairs. Note that the p-values are biased toward 1. **C.** Calibration curve. **D.** Null calibrated p-value distribution.

Supplementary Note 3: Effect of confounders on bimodality

In the **Methods** section, we noted that we performed two rounds of dip tests on the expression data: one before confounder correction and one after. We did this two-round testing to determine how frequently confounders give rise to bimodal expression that disappears once expression levels are corrected for confounders. Of the 3694 switch-like genes identified from calibrated dip tests on uncorrected data, only 15% (473) retain their bimodality in at least one tissue after confounder correction (**Supplementary Figure 3A**). Certain tissues (colon, vagina, stomach, kidney, and intestine) are especially prone to confounder-driven bimodal expression (**Supplementary Figure 3B**). We further illustrate the distributions of confounder-to-expression correlations for switch-like gene–tissue pairs before and after correction (**Supplementary Figures 3C and D**), with **Supplementary Figure 3D** indicating that our confounder-correction procedure has performed effectively. Finally, **Supplementary Figure 4** provides examples of confounders that induce bimodality in uncorrected gene expression. Finally, we clarify the interpretation of higher TPM levels of *HILPDA* in blood from postmortem samples compared to organ donor samples (**Supplementary Figure 4A**). This difference does not indicate that *HILPDA* is "switched on" in deceased individuals. Rather, because TPM represents transcript abundance, the observed elevation in *HILPDA* TPM levels in postmortem samples suggests that after death, its transcripts decay more slowly than those of other genes, resulting in higher relative abundance. This phenomenon has been described previously [2,3].



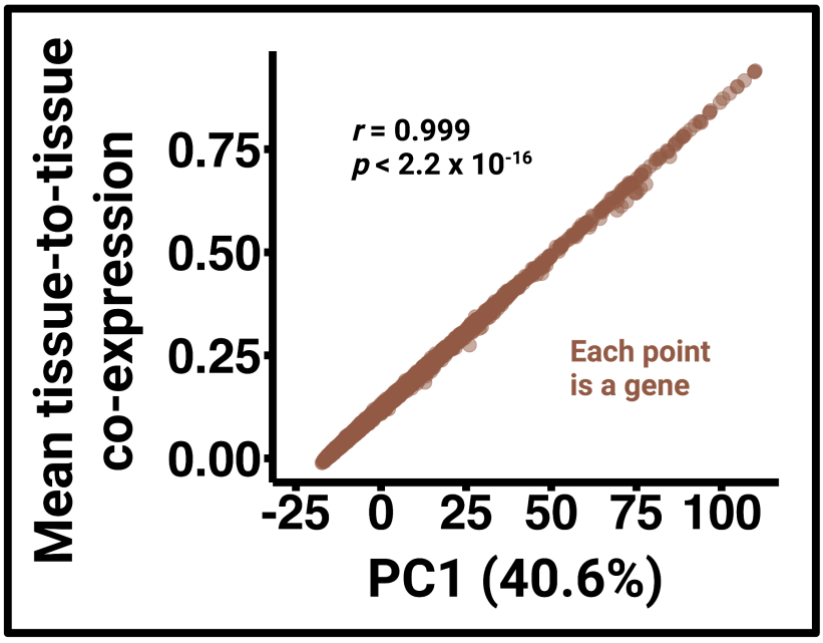
Supplementary Figure 3. Effect of confounders on bimodal gene expression. A. A waffle plot showing the number of switch-like genes identified by dip tests performed on log(TPM + 1) distribution of gene-tissue pairs across samples. Green squares represent the genes that sustain their bimodal expression in at least one tissue after confounder correction. Yellow represents the genes that lost their bimodal expression after correction. **B.** A stacked bar plot showing the relative fractions of genes that lose (yellow) versus sustain (green) their bimodal expression in each tissue upon confounder correction. **C.** Distribution of the absolute value of Pearson's *r* between the raw expression level distribution of gene-tissue pairs and each confounder. **D.** Distribution of the absolute value of Pearson's r between the confounder-corrected expression level distribution of gene-tissue pairs and each confounder. Source data are provided as a Source Data file.



Supplementary Figure 4. Examples of confounders leading to bimodal gene expression. A. Cohort-driven bimodally distributed (switch-like) gene expression of *HILPDA* across individuals in blood. **B.** Switch-like expression of *MT-CO1* in the uterus driven by the nature of death (Hardy death scale). **C.** Switch-like expression of *SLC46A2* in lungs driven by different RNA quality (RNA integrity number) across individuals. **D.** Switch-like expression of ENSG00000270640 in the esophagus driven by postmortem interval (ischemic time). Source data are provided as a Source Data file.

Supplementary Note 4: Principal component analysis on tissue-to-tissue co-expression vectors

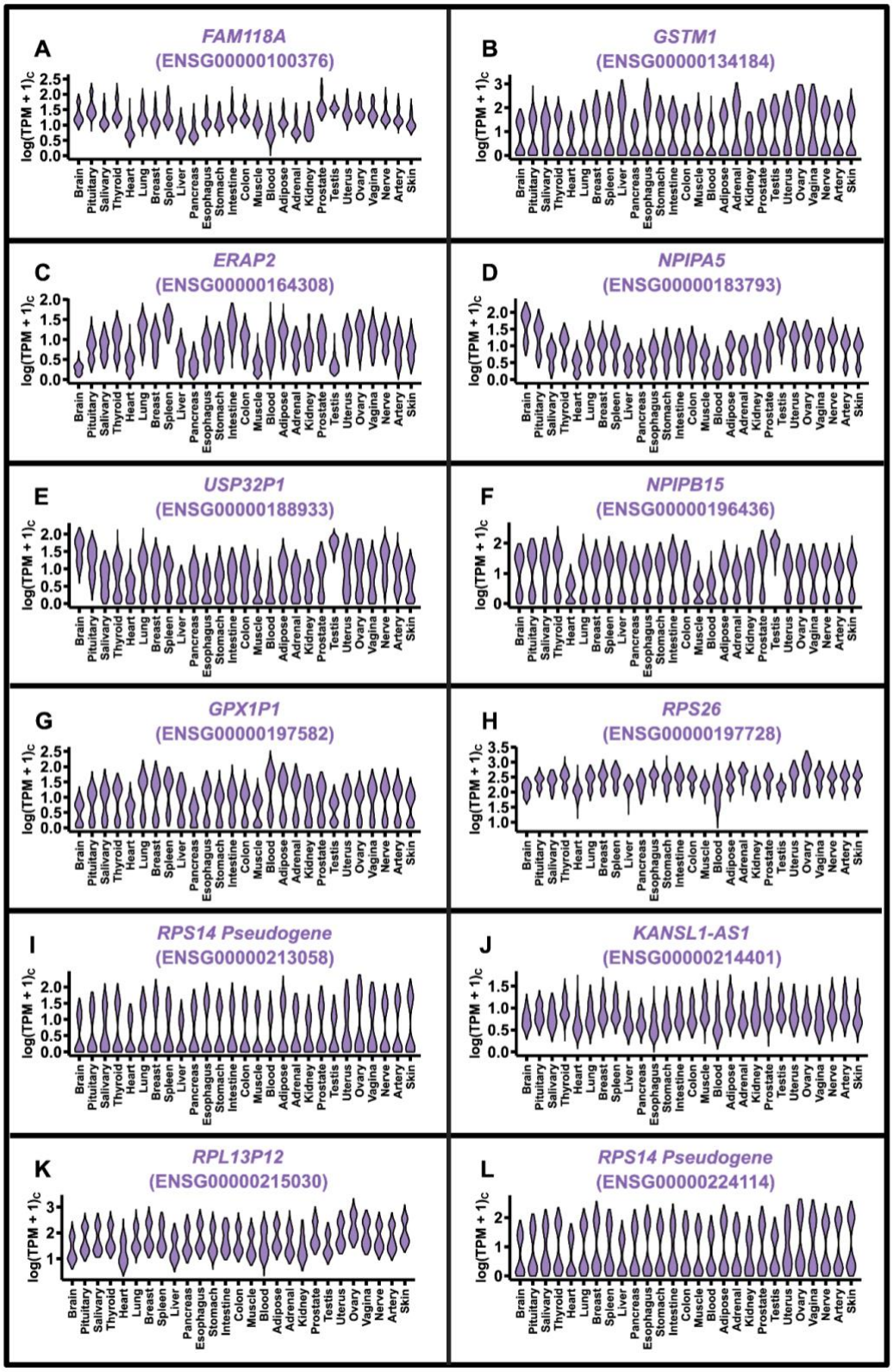
We applied a principal component analysis to the 19,121 vectors of tissue-to-tissue co-expression, one vector for each gene. We find that PC1 (**Figure 2A**), explaining 40.6% of the variation, is nearly perfectly correlated with mean tissue-to-tissue co-expression across tissue-tissue pairs (r = 0.999, p-value < 2.2 x 10^{-16}; **Supplementary Figure 5**). This result indicates that 40.6% of the variation in the tissue-to-tissue co-expression of genes is primarily explained by the mean tissue-to-tissue co-expression of genes.



Supplementary Figure 5. The mean tissue-to-tissue co-expression of genes shows a near-perfect correlation with PC1.

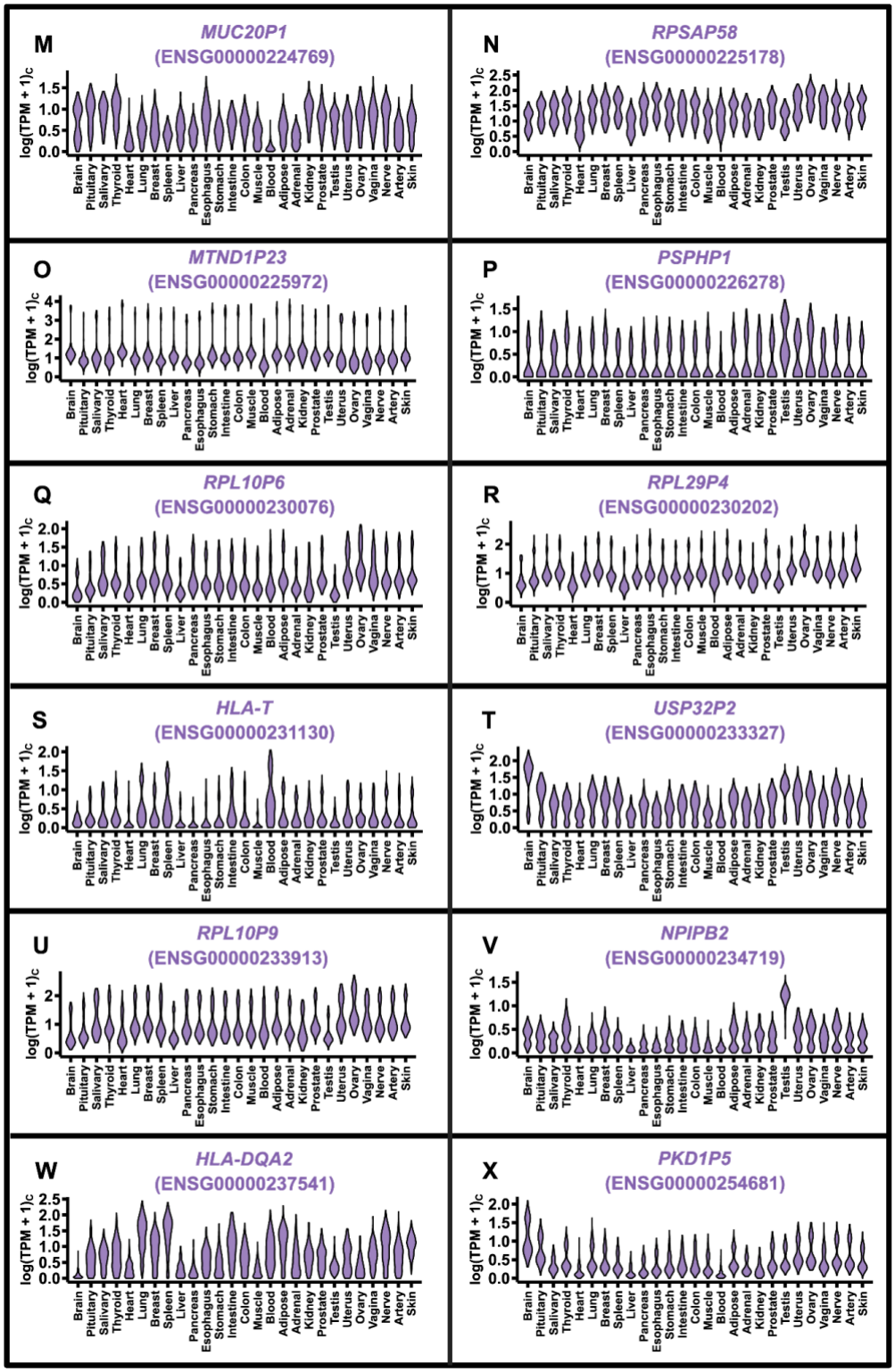
Supplementary Note 5: Tissue-specific switch-like genes

We divided switch-like genes into three clusters in the space spanned by the first two principal components (**Figure 2A**). While we said that genes in cluster 1 **(Figure 2A-B)** are tissue-specific switch-like genes, manual inspection reveals this is not true for all genes in cluster 1. In particular, the transcript ENSG00000273906, coming from chr Y, was labeled cluster 1 by hierarchical clustering even though it is universally switch-like in tissues common to both sexes. Indeed, we removed all chr-Y genes from our analyses of genuine cluster-1 genes. Other cluster-1 genes, bimodally expressed in a large number of tissues, lie on the autosomes. For example, *CLPS*, *CELA3A*, and multiple members of the *PRSS* gene family (*PRSS1*, *PRSS2*, *PRSS3*, *PRSS8*, *PRSS16*, and *PRSS22*), despite having low overall tissue-to-tissue co-expression, are bimodally expressed in a large number of tissues.

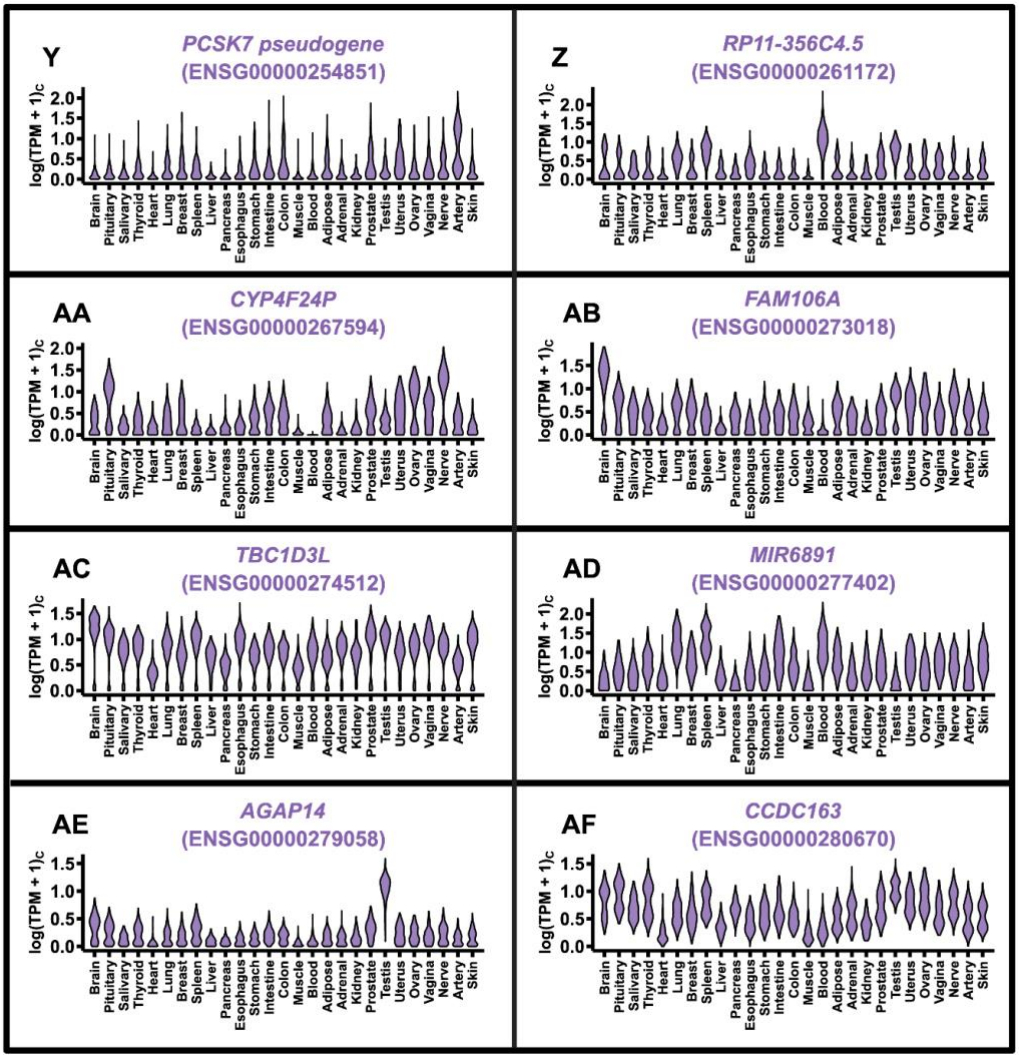


Supplementary Figure 6. Violin plots for expression level distributions of switch-like genes in cluster 2A. Panels **A** to **AF** show results for one cluster-2A gene each. Source data are provided as a Source Data file. **A.** *FAM118A* (ENSG00000100376). **B.** *GSTM1* (ENSG00000134184). **C.** *ERAP2* (ENSG00000164308). **D.** *NPIPA5* (ENSG00000183793). **E.** *USP32P1* (ENSG00000188933). **F.**

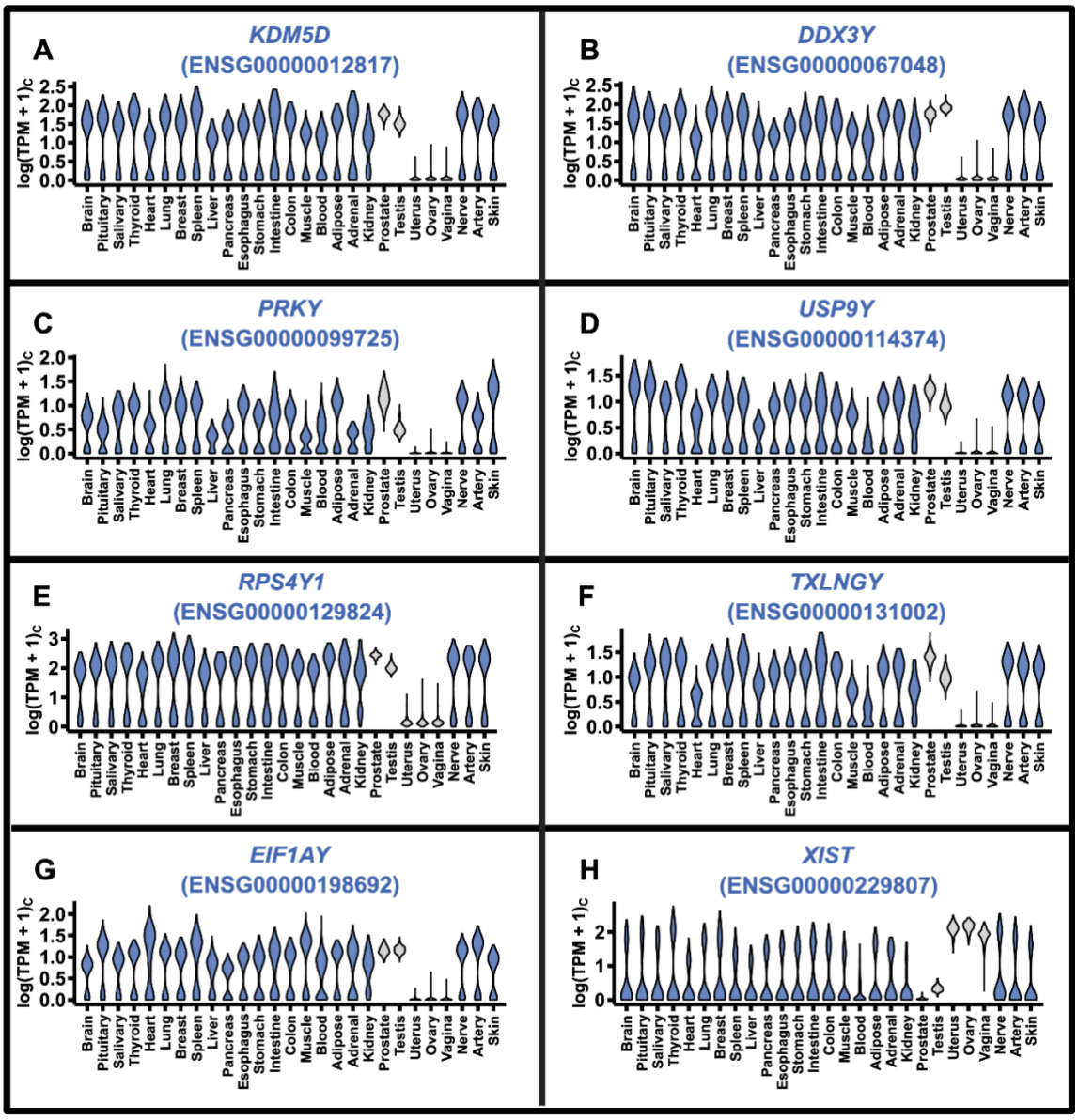
NPIPB15 (ENSG00000196436). **G.** *GPX1P1* (ENSG00000197582). **H.** *RPS26* (ENSG00000197728). **I.** *RPS14* pseudogene (ENSG00000213058). **J.** *KANSL1-AS1* (ENSG00000214401). **K.** *RPL13P12* (ENSG00000215030). **L.** *RPS14* pseudogene (ENSG00000224114). **M.** *MUC20P1* (ENSG00000224769). **N**. *RPSAP58* (ENSG00000225178). **O.** *MTND1P23* (ENSG00000225972). **P.** *PSPHP1* (ENSG00000226278). **Q.** *RPL10P6* (ENSG00000230076). **R.** *RPL29P4* (ENSG00000230202). **S.** *HLA-T* (ENSG00000231130). **T.** *USP32P2* (ENSG00000233327). **U.** *RPL10P9* (ENSG00000233913). **V.** *NPIPB2* (ENSG00000234719). **W.** *HLA-DQA2* (ENSG00000237541). **X.** *PKD1P5* (ENSG00000254681). **Y.** *PCSK7* pseudogene (ENSG00000254851). **Z.** *RP11-356C4.5* (ENSG00000261172). **AA.** *CYP4F24P* (ENSG00000267594). **AB.** *FAM106A* (ENSG00000273018). **AC.** *TBC1D3L* (ENSG00000274512). **AD.** *MIR6891* (ENSG00000277402). **AE.** *AGAP14* (ENSG00000279058). **AF.** *CCDC163* (ENSG00000280670)**.**



Supplementary Figure 6 continued.



Supplementary Figure 6 continued.



Supplementary Figure 7. Violin plots for expression level distributions of switch-like genes in cluster 2B. Panels **A** to **H** show results for one cluster-2B gene each. **A.** *KDM5D*. **B.** *DDX3Y*. **C.** *PRKY*. **D.** *USP9Y*. **E.** *RPS4Y1*. **F.** *TXLNGY*. **G.** *EIF1AY*. **H.** *XIST*. Source data are provided as a Source Data file.

Supplementary Note 6: Universally switch-like genes and their biomedical implications

The violin plots for the expression level distributions for all cluster-2A and cluster-2B switch-like genes not shown in the main text are present in **Supplementary Figure 6** and **Supplementary Figure 7**, respectively. In the main text, we discussed the *USP32P2* and *FAM106A*. Here, we discuss some other interesting examples of universally switch-like genes.

Firstly, a common ~20kb whole-gene deletion (esv3587154) of the *GSTM1* gene [4,5] is associated with bladder cancer in humans [6]. *GSTM1* is bimodally expressed across individuals in all tissues (**Supplementary Figure 6B**) that we analyzed, as well as across multiple tumor types [7], with different expression peaks corresponding to differential prognoses among patients. These findings suggest a compelling hypothesis: the common deletion of *GSTM1*, maintained either by drift or balancing selection [8], has no significant effect on the health of non-cancerous individuals; however, it could have significant implications for prognosis once certain types of tumors develop. Therefore, screening patients with certain tumor types for the *GSTM1* deletion could significantly advance our ability to predict the course of tumor progression in an individualized manner. However, unlike the case of *GSTM1* (along with *USP32P2* and *FAM106A*), in which structural variation gives rise to bimodal expression, the high copy number variation at the *PGA3* [9] and amylase [10–12] loci did not have the same effect in the GTEx data. In particular, *PGA3* and amylase genes were not identified as switch-like genes. This result is consistent with a previous study suggesting that the copy number of *PGA3* seems to have no impact on *PGA3* expression in cancer samples [13]. Similarly, it has been shown that copy number variation at *AMY1* is only a minor contributor to the gene's expression variation [14]. Nevertheless, when copy number variation involves a large number of alleles (as in the case of *PGA3* and *AMY1*) rather than just two alleles (as with *GSTM1*), it is possible for the copy number to influence gene expression without producing a clearly multimodal distribution across individuals.

Secondly, *ERAP2* (**Supplementary Figure 6C**) exhibits universally switch-like expression due to a single eQTL (rs2248374). One variant produces a full-length transcript, while the other results in a truncated transcript that undergoes nonsense-mediated decay, thereby creating a bimodal expression pattern [15]. *ERAP2* is involved in trimming antigens within cells so that they can be presented on the cell surface, subsequently activating immune responses. It has been further hypothesized that the coexistence of both transcript variants represents an evolutionary trade-off, balancing the risk of autoimmune diseases with protection against *Yersinia pestis*, the causative agent of the Black Death [16].

Thirdly, the bimodality of *NPIPA5* (**Supplementary Figure 6D**), too, can be explained by a single eQTL. The T allele of the SNV rs3198697 is associated with *NPIPA5* being switched on across tissues, while the C allele is associated with the gene being switched off. *NPIPA5* has been reported as one of the top differentially expressed genes among patients with multiple sclerosis in both blood and brain [17]. Moreover, this study [17] showed that this gene is co-expressed in blood and brain. Here, we have shown that this gene is switch-like and that the co-expression of *NPIPA5* is not restricted to blood and brain but extends to all pairs of tissues.

Fourthly, genes that are bimodally expressed across multiple tissues raise an evolutionary paradox. Typically, genes with a wide expression breadth (i.e., expression across a large number of tissues) affect fitness and are thus constrained at both the sequence and expression level [18–21]. However, despite having a high expression breadth, universally switch-like genes are not conserved at the expression level. This could imply different health consequences for individuals with an “off” versus an “on” state of the genes. For example, the universally switch-like *RPS14* pseudogene (ENSG00000213058; **Supplementary Figure 6I**) is upregulated in the peripheral blood mononuclear cells of patients with ankylosing spondylitis [22], eutopic endometrium in endometriosis patients [23], and peripheral blood mononuclear cells of multiple sclerosis patients [24]. In contrast, it is downregulated in the peripheral blood mononuclear cells of Sjögren's syndrome patients [25]. These results suggest that this gene being switched on versus off may predispose individuals to certain diseases while protecting them against others. This balance between susceptibility and protection could explain why both high-expression and low-expression states are maintained in the population at comparable frequencies.

Lastly, a single eQTL can explain the bimodality of a member of the PKD1 gene family in cluster 2A, *PKD1P5* (**Supplementary Figure 6X**). For *PKD1P5*, the C allele of the SNV rs201525245 is associated with the gene being switched on, while the G allele is associated with the gene being switched off.

Supplementary Note 7: Conceptual issues regarding bimodal expression distributions driven by genetic polymorphisms

In the main text, we claimed that genetic polymorphisms drive the bimodal expression of universally switch-like genes in cluster 2A. For a polymorphism with two alleles (A and a), there are three possible genotypes (aa, Aa, and AA). Since each of the three genotypes can lead to three different expression levels, we expect the expression distribution of a cluster-2A gene to have three modes. This leads to the question: Why do we not see trimodal, as opposed to bimodal, expression distributions for genes in cluster 2A? To answer this question, we develop the following frameworks. Let us assume that a genetic polymorphism exists with two alleles, A and a, with frequencies p_A and $(1-p_A)$, respectively. The three genotypes for this polymorphism, aa, Aa, and AA, lead to three different expression states (TPM levels) for the gene with averages μ_{aa}, μ_{Aa}, and μ_{AA}, respectively. Let us also assume that the Hardy-Weinberg equilibrium holds for this locus. Then, the frequency of $aa = (1-p_A)^2$, the frequency of $Aa = 2p_A(1-p_A)$, and the frequency of $AA = {p_A}^2$. We assume that $\mu_{aa} \leq \mu_{Aa} \leq \mu_{AA}$. Next, we define a dominance coefficient $0 \leq \alpha \leq 1$ by

$$\mu_{Aa} = \mu_{aa} + (\mu_{AA} - \mu_{aa})\alpha. \quad (1)$$

If we define the ratio R by

$$R = \frac{\mu_{AA}}{\mu_{aa}}, \quad (2)$$

then, we obtain

$$\mu_{Aa} = \mu_{aa}(1 - \alpha + R\alpha), \quad (3)$$

and

$$\mu_{AA} = R\mu_{aa}. \quad (4)$$

We can then divide individuals into three groups depending on their genotypes. Let us assume that the coefficient of variation (CV) of expression is the same for each genotypic group. Then, we can model the TPM value of this gene in a given individual as a normal random variable with:

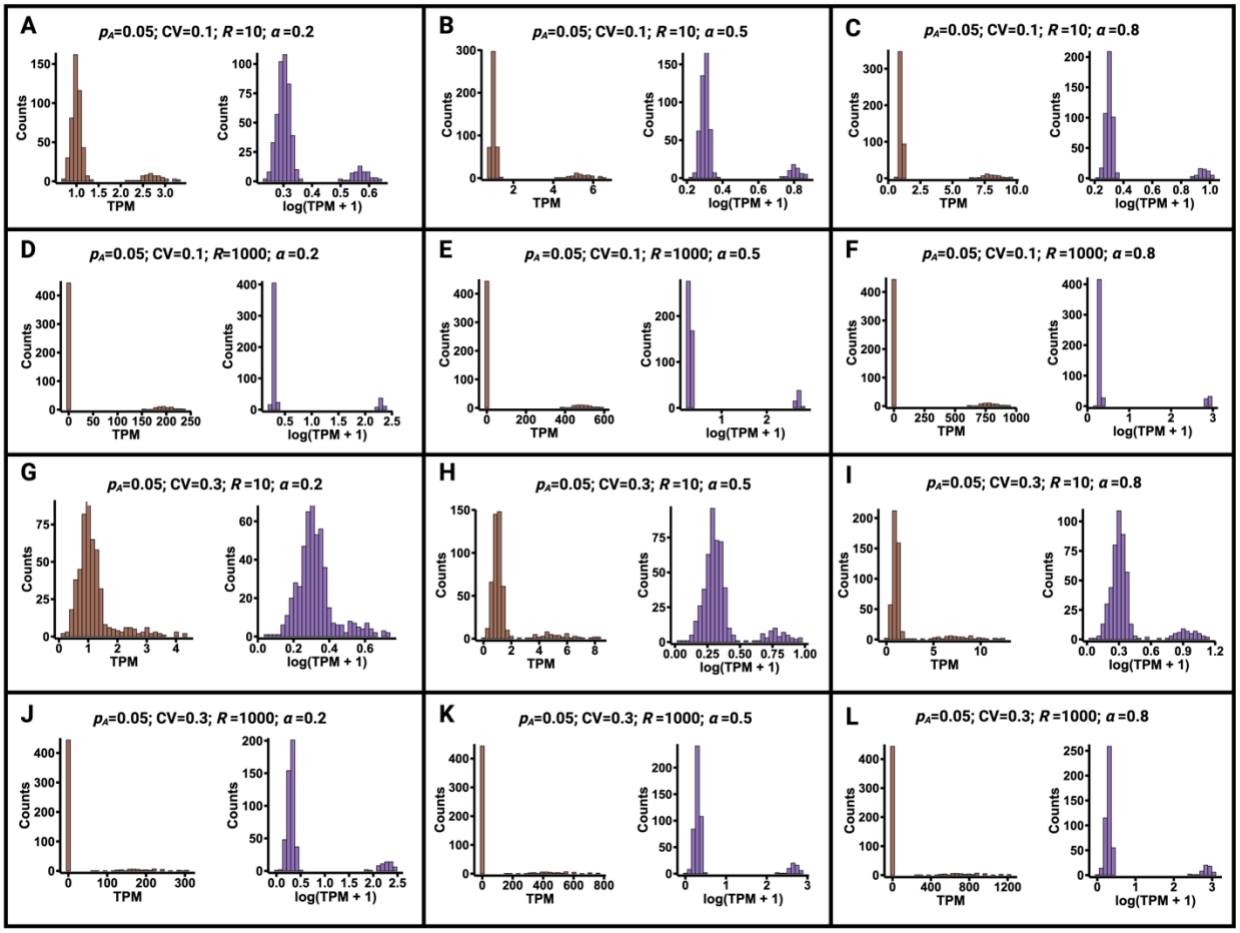
1) mean = μ_{aa} and standard deviation = CV x μ_{aa} if the genotype is aa;
2) mean = μ_{Aa} and standard deviation = CV x μ_{Aa} if the genotype is Aa; and
3) mean = μ_{AA} and standard deviation = CV x μ_{AA} if the genotype is AA.

The value of μ_{aa} is irrelevant for gauging the effect of polymorphisms on the shape of the expression level distributions. Therefore, we set $\mu_{aa} = 1$.

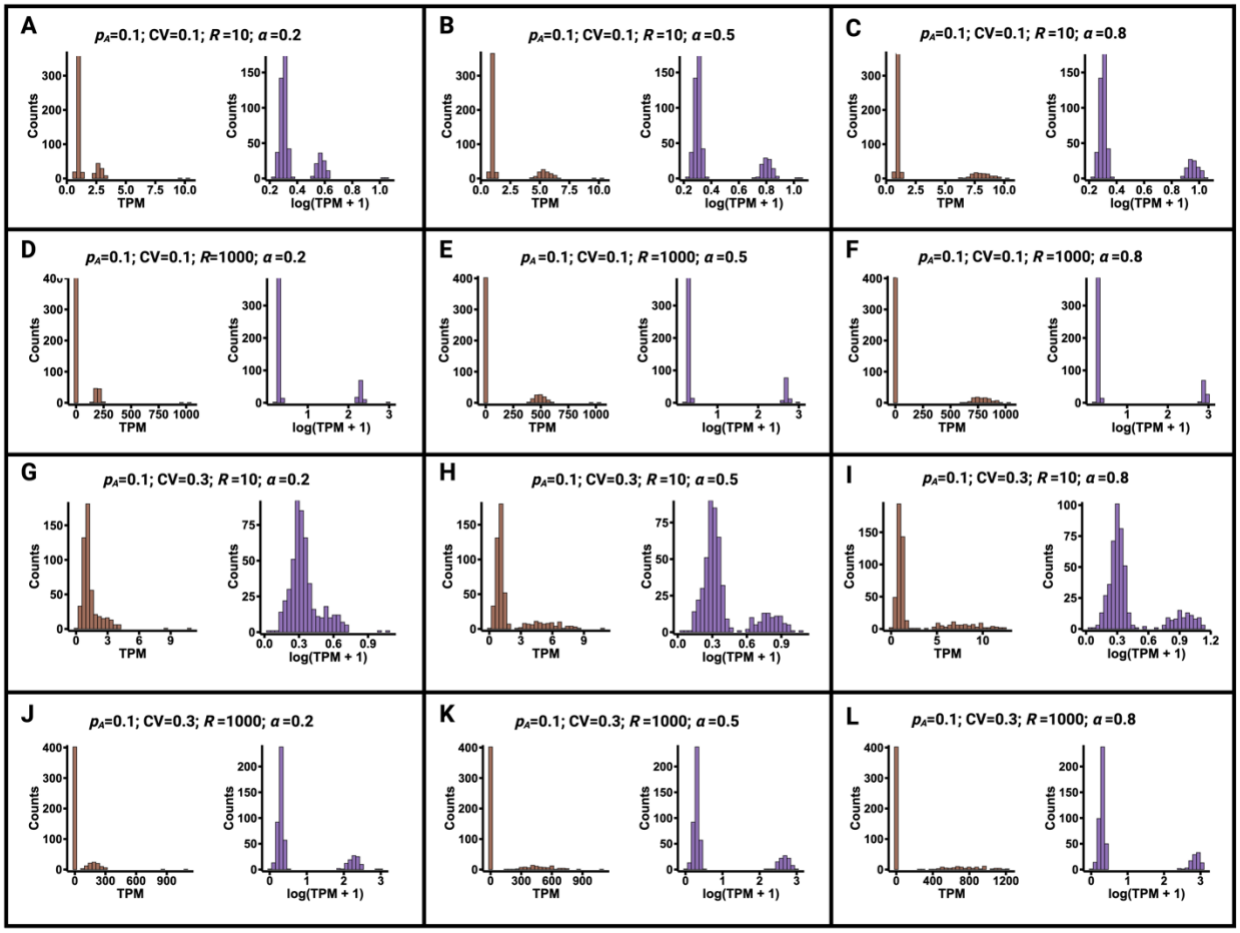
Under these mathematical assumptions, we performed simulations using 36 distinct models. These models vary by four parameters: $p_A \in \{0.05, 0.1, 0.5\}$, CV $\in \{0.1,$

0.3}, $R \in \{10, 1000\}$, and $\alpha \in \{0.2, 0.5, 0.8\}$. For each model, defined by a unique combination of the values of these four parameters, we performed a two-step sampling procedure. First, we obtained a random sample of 500 genotypes, based on p_A and the Hardy-Weinberg equilibrium. Next, for each of the 500 genotypes sampled, we sample a TPM value from the normal distribution corresponding to that genotype. Thus, for each of the 36 models, we simulated 500 TPM values. We present these values as histograms with and without log transformation. The results for p_A = 0.05, p_A = 0.1, and p_A = 0.5 are shown in **Supplementary Figure 8**, **Supplementary Figure 9**, and **Supplementary Figure 10**, respectively**.** These simulations help us answer the question we first asked: Why do we not see a trimodal distribution if a genetic polymorphism drives expression-level variability in a gene?

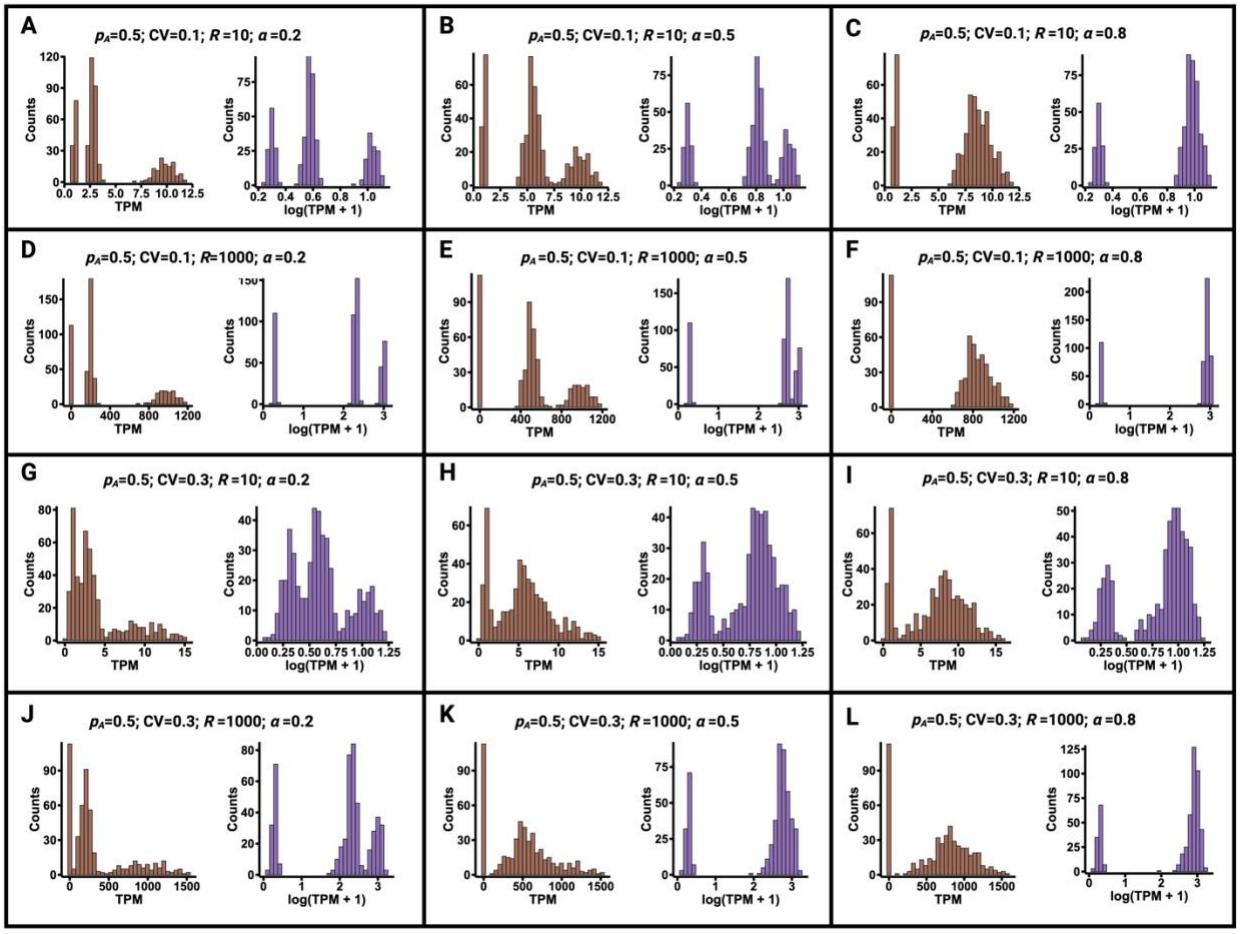
Firstly, even when the minor allele (*A*) frequency is not low (e.g., 10%), the frequency of the genotype *AA* is still quite low (e.g., 1%). Therefore, the third peak is not always conspicuously visible. We see this in all models with p_A = 0.05 and p_A = 0.1 (**Supplementary Figures 8** and **9**), regardless of CV, R, and α values. At higher allele frequencies (e.g., 50%), the effect of the remaining parameters becomes more apparent. **Supplementary Figure 10** shows that a higher dominance coefficient α makes the expression level distribution more bimodal. By contrast, a lower dominance coefficient α makes the expression level distribution more trimodal. The lack of observed trimodality in the GTEx data may suggest that expression levels of switch-like genes tend to be more dominant than additive with regard to causal genetic polymorphisms. Secondly, greater variation (CV) in the data can also obscure the third peak. For example, comparing **Supplementary Figure 10B** to **Supplementary Figure 10H**, we find that increasing the CV can change the distribution from trimodal to bimodal when the other parameters are held constant. However, R does not seem to affect whether the expression level distribution is bimodal or trimodal.



Supplementary Figure 8. TPM simulations for a hypothetical gene whose expression is driven by a genetic polymorphism with an allele frequency (p_A) of 5%. Panels **A** to **L** show results for one model (defined by a unique set of parameters) each. **A.** p_A = 0.05, CV = 0.1, R = 10, and α = 0.2. **B.** p_A = 0.05, CV = 0.1, R = 10, and α = 0.5. **C.** p_A = 0.05, CV = 0.1, R = 10, and α = 0.8. **D.** p_A = 0.05, CV = 0.1, R = 1000, and α = 0.2. **E.** p_A = 0.05, CV = 0.1, R = 1000, and α = 0.5. **F.** p_A = 0.05, CV = 0.1, R = 1000, and α = 0.8. p_A = 0.05, CV = 0.1, R = 10, and α = 0.2. **G.** p_A = 0.05, CV = 0.3, R = 10, and α = 0.2. **H.** p_A = 0.05, CV = 0.3, R = 10, and α = 0.5. **I.** p_A = 0.05, CV = 0.3, R = 10, and α = 0.8. **J.** p_A = 0.05, CV = 0.3, R = 1000, and α = 0.2. **K.** p_A = 0.05, CV = 0.3, R = 1000, and α = 0.5. **L.** p_A = 0.05, CV = 0.3, R = 1000, and α = 0.8. Source data are provided as a Source Data file.



Supplementary Figure 9. TPM simulations for a hypothetical gene whose expression is driven by a genetic polymorphism with an allele frequency (p_A) of 10%. Panels **A** to **L** show results for one model (defined by a unique set of parameters) each. **A.** p_A = 0.1, CV = 0.1, R = 10, and α = 0.2. **B.** p_A = 0.1, CV = 0.1, R = 10, and α = 0.5. **C.** p_A = 0.1, CV = 0.1, R = 10, and α = 0.8. **D.** p_A = 0.1, CV = 0.1, R = 1000, and α = 0.2. **E.** p_A = 0.1, CV = 0.1, R = 1000, and α = 0.5. **F.** p_A = 0.1, CV = 0.1, R = 1000, and α = 0.8. p_A = 0.05, CV = 0.1, R = 10, and α = 0.2. **G.** p_A = 0.1, CV = 0.3, R = 10, and α = 0.2. **H.** p_A = 0.1, CV = 0.3, R = 10, and α = 0.5. **I.** p_A = 0.1, CV = 0.3, R = 10, and α = 0.8. **J.** p_A = 0.1, CV = 0.3, R = 1000, and α = 0.2. **K.** p_A = 0.1, CV = 0.3, R = 1000, and α = 0.5. **L.** p_A = 0.1, CV = 0.3, R = 1000, and α = 0.8. Source data are provided as a Source Data file.



Supplementary Figure 10. TPM simulations for a hypothetical gene whose expression is driven by a genetic polymorphism with an allele frequency (p_A) of 50%. Panels **A** to **L** show results for one model (defined by a unique set of parameters) each. **A.** p_A = 0.5, CV = 0.1, R = 10, and α = 0.2. **B.** p_A = 0.5, CV = 0.1, R = 10, and α = 0.5. **C.** p_A = 0.5, CV = 0.1, R = 10, and α = 0.8. **D.** p_A = 0.5, CV = 0.1, R = 1000, and α = 0.2. **E.** p_A = 0.5, CV = 0.1, R = 1000, and α = 0.5. **F.** p_A = 0.5, CV = 0.1, R = 1000, and α = 0.8. p_A = 0.05, CV = 0.1, R = 10, and α = 0.2. **G.** p_A = 0.5, CV = 0.3, R = 10, and α = 0.2. **H.** p_A = 0.5, CV = 0.3, R = 10, and α = 0.5. **I.** p_A = 0.5, CV = 0.3, R = 10, and α = 0.8. **J.** p_A = 0.5, CV = 0.3, R = 1000, and α = 0.2. **K.** p_A = 0.5, CV = 0.3, R = 1000, and α = 0.5. **L.** p_A = 0.5, CV = 0.3, R = 1000, and α = 0.8. Source data are provided as a Source Data file.

Supplementary Table 1. A list of tissues used in this study, and the number of individuals for each tissue.

GTEx tissue name	Tissue name used in figures	Number of samples
adipose_subcutaneous	Adipose	663
adrenal_gland	Adrenal	258
whole_blood	Blood	755
artery_tibial	Artery	663
brain_cerebellum	Brain	241
breast_mammary_tissue	Breast	459
colon_transverse	Colon	406
esophagus_mucosa	Esophagus	555
heart_left_ventricle	Heart	432
kidney_cortex	Kidney	85
liver	Liver	226
lung	Lung	578
muscle_skeletal	Muscle	803
nerve_tibial	Nerve	619
ovary	Ovary	180
pancreas	Pancreas	328
pituitary	Pituitary	283
prostate	Prostate	245
minor_salivary_gland	Salivary	162
skin_sun_exposed_lower_leg	Skin	701
small_intestine_terminal_ileum	Intestine	187
spleen	Spleen	241
stomach	Stomach	359
testis	Testis	361
thyroid	Thyroid	653
uterus	Uterus	142
vagina	Vagina	156

Supplementary references

1. Hounkpe, B. W., Chenou, F., de Lima, F. & De Paula, E. V. HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res.* **49**, D947–D955 (2021).
2. Pozhitkov, A. E. *et al.* Tracing the dynamics of gene transcripts after organismal death. *Open Biol.* **7**, (2017).
3. Gilad, Y. *An Intuitive Primer on Effective Functional Genomics Study Design*. (Chicago, Illinois, 2025).
4. Saitou, M. & Gokcumen, O. An Evolutionary Perspective on the Impact of Genomic Copy Number Variation on Human Health. *J. Mol. Evol.* **88**, 104–119 (2020).
5. Saitou, M., Satta, Y., Gokcumen, O. & Ishida, T. Complex evolution of the GSTM gene family involves sharing of GSTM1 deletion polymorphism in humans and chimpanzees. *BMC Genomics* **19**, 293 (2018).
6. Rothman, N. *et al.* A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat. Genet.* **42**, 978–984 (2010).
7. Justino, J. R., Reis, C. F. dos, Fonseca, A. L., Souza, S. J. de & Stransky, B. An integrated approach to identify bimodal genes associated with prognosis in câncer. *Genet. Mol. Biol.* **44**, e20210109 (2021).
8. Aqil, A., Speidel, L., Pavlidis, P. & Gokcumen, O. Balancing selection on genomic deletion polymorphisms in humans. *eLife* **12**, (2023).
9. Otto, M., Zheng, Y., Grablowitz, P. & Wiehe, T. Detecting adaptive changes in gene copy number distribution accompanying the human out-of-Africa expansion. *Hum. Genome Var.* **11**, 37 (2024).
10. Yilmaz, F. *et al.* Reconstruction of the human amylase locus reveals ancient duplications seeding modern-day variation. *Science* **386**, eadn0609 (2024).
11. Bolognini, D. *et al.* Recurrent evolution and selection shape structural diversity at the

amylase locus. *Nature* **634**, 617–625 (2024).

12. Scheer, K. *et al.* Adaptive increase of amylase gene copy number in Peruvians driven by potato-rich diets. *bioRxiv* (2025) doi:10.1101/2025.03.25.644684.
13. Shen, S., Li, H., Liu, J., Sun, L. & Yuan, Y. The panoramic picture of pepsinogen gene family with pan-cancer. *Cancer Med.* **9**, 9064–9080 (2020).
14. Carpenter, D., Mitchell, L. M. & Armour, J. A. L. Copy number variation of human AMY1 is a minor contributor to variation in salivary amylase expression and activity. *Hum. Genomics* **11**, 2 (2017).
15. Andrés, A. M. *et al.* Balancing selection maintains a form of *ERAP2* that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genet.* **6**, e1001157 (2010).
16. Klunk, J. *et al.* Evolution of immune genes is associated with the Black Death. *Nature* (2022) doi:10.1038/s41586-022-05349-x.
17. Islam, T. *et al.* Detection of multiple sclerosis using blood and brain cells transcript profiles: Insights from comprehensive bioinformatics approach. *Informatics in Medicine Unlocked* **16**, 100201 (2019).
18. Duret, L. & Mouchiroud, D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17**, 68–74 (2000).
19. Zhang, L. & Li, W.-H. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.* **21**, 236–239 (2004).
20. Park, J., Xu, K., Park, T. & Yi, S. V. What are the determinants of gene expression levels and breadths in the human genome? *Hum. Mol. Genet.* **21**, 46–56 (2011).
21. Khan, Z. *et al.* Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* **342**, 1100–1104 (2013).
22. Li, C., Qu, W. & Yang, X. Comprehensive lncRNA and mRNA profiles in peripheral blood

mononuclear cells derived from ankylosing spondylitis patients by RNA-sequencing analysis. *Medicine* **101**, e27477 (2022).

23. von Grothusen, C. Endometrial receptivity and regeneration in health and disease: Molecular, cellular and clinical perspectives. (Doctoral dissertation, Karolinska Institutet, Stockholm, Sweden, 2022).
24. Almsned, F. M. Understanding the Genetic Nature of Multiple Sclerosis Using Next-Generation Sequencing Genomic Analysis Methods. (Doctoral dissertation, George Mason University, VA, United States, 2020).
25. Chery, G. Understanding Sjögren's Syndrome as a Systemic Autoimmune Disorder. (Master's dissertation, State University of New York at Albany, NY, United States, 2022).