Supplementary materials: Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human

Barbara Piasecka[1,2,3], Marc Robinson-Rechavi[1,3,*], Sven Bergmann[2,3,*]
1 Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland
2 Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland
3 Swiss Institute of Bioinformatics, Lausanne, Switzerland
∗ **E-mail: Sven.Bergmann@unil.ch; Marc.Robinson-Rechavi@unil.ch**

Theoretical analysis

Here we prove that for data transformed into z-scores the Pearson's correlation coefficient and the squared Euclidean distance are linearly dependent.
Let $\mathbf{x}$ be a vector of expression values of a given gene measured in n conditions. The z-score vector of $\mathbf{x}$ is calculated as follows:

$$\mathbf{z_x} = \frac{\mathbf{x} - \bar{x}}{s_\mathbf{x}} = \frac{\tilde{\mathbf{x}}}{s_\mathbf{x}}, \tag{1}$$

where $\bar{x}$ is a mean and $s_\mathbf{x}$ is a standard deviation of gene expression values. $s_\mathbf{x}$ is defined as:

$$s_\mathbf{x} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

Euclidean norm of vector $\mathbf{x}$ is defined as:

$$||\mathbf{x}|| = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{3}$$

To normalize vector $\tilde{\mathbf{x}}$ we divide it by its Euclidean norm. We refer to it as z-like normalization of vector $\mathbf{x}$ and denote it as $\tilde{\mathbf{z}}_\mathbf{x}$:

$$\tilde{\mathbf{z}}_\mathbf{x} = \frac{\tilde{\mathbf{x}}}{||\tilde{\mathbf{x}}||} = \frac{\tilde{\mathbf{x}}}{\sqrt{\sum_{i=1}^{n} \tilde{x_i}^2}} = \frac{\mathbf{x} - \bar{x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{\mathbf{x} - \bar{x}}{\sqrt{n}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{1}{\sqrt{n}}\mathbf{z_x} \tag{4}$$

So the z-score vector can be represented as:

$$\mathbf{z_x} = \sqrt{n}\tilde{\mathbf{z}}_\mathbf{x} \tag{5}$$

The Pearson's correlation coefficient between vectors x and y is defined as:

$$r_\mathbf{xy} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_\mathbf{x} s_\mathbf{y}} = \frac{1}{n}\mathbf{z_x}\mathbf{z_y} = \tilde{\mathbf{z}}_\mathbf{x}\tilde{\mathbf{z}}_\mathbf{y} \tag{6}$$

The Euclidan distance between vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$d_\mathbf{xy} = ||\mathbf{x} - \mathbf{y}|| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{7}$$

The squared Euclidean distance between $\tilde{\mathbf{z}}_\mathbf{x}$ and $\tilde{\mathbf{z}}_\mathbf{y}$ can be expressed as follows:

$$d^2_{\tilde{\mathbf{z}}_\mathbf{x}\tilde{\mathbf{z}}_\mathbf{y}} = ||\tilde{\mathbf{z}}_\mathbf{x} - \tilde{\mathbf{z}}_\mathbf{y}||^2 = ||\tilde{\mathbf{z}}_\mathbf{x}||^2 + ||\tilde{\mathbf{z}}_\mathbf{y}||^2 - 2\tilde{\mathbf{z}}_\mathbf{x}\tilde{\mathbf{z}}_\mathbf{y} \tag{8}$$

Replacing $\tilde{\mathbf{z}}_\mathbf{x}\tilde{\mathbf{z}}_\mathbf{y}$ by equation 6 we obtain:

$$\begin{aligned} d^2_{\tilde{\mathbf{z}}_\mathbf{x}\tilde{\mathbf{z}}_\mathbf{y}} &= ||\tilde{\mathbf{z}}_\mathbf{x} - \tilde{\mathbf{z}}_\mathbf{y}||^2 = ||\tilde{\mathbf{z}}_\mathbf{x}||^2 + ||\tilde{\mathbf{z}}_\mathbf{y}||^2 - 2r_{\mathbf{xy}} \\ &= 1 + 1 - 2r_{\mathbf{xy}} \\ &= 2 - 2r_{\mathbf{xy}} \\ &= 2(1 - r_{\mathbf{xy}}) \end{aligned} \tag{9}$$

We see from equation 9 that in case of z-like normalization there is a linear dependence between the squared Euclidean distance (d^2) and the Pearson's distance ($1 - r$), namely:

$$d^2_{\tilde{\mathbf{z}}_\mathbf{x}\tilde{\mathbf{z}}_\mathbf{y}} = 2(1 - r_{\mathbf{xy}}) \tag{10}$$

Numerical analysis

Dependence between Pearson's and Euclidean distances for different normalizations

In order to illustrate the strength of dependence between the Pearson's and the Euclidean distances for different data normalization modes, we calculated them on the human-mouse gene expression data from the GNF atlas (Su *et al.* 2004). The data consisted of 27 homologous organ groups (table S1) and 8,942 one-to-one orthologous gene pairs. For every gene expression profile, within both species, we applied Manhattan, Euclidean and z-like normalization followed by the Pearson's correlation coefficient and the Euclidean distance calculation. As predicted by eq. 9, r and d^2 were linearly dependent for z-like normalization (figure S1C). In contrast, for Manhattan and Euclidean normalizations the correlation between r and d^2 was low (figure S1A-B).

Dependence between Euclidean distance and expression specificity for different normalizations

In order to explain the dependence between the distance measure for different normalization and the expression specificity of compared genes we used simulated data:

1. two genes with uniform expression levels across 50 conditions (figure S2A)
2. two genes with expression specific to a single condition (figure S2B)

We calculated d^E_E, d^Z_E and r for both pairs of genes. Intuitively, one might assume that d_Es (or r) for two pairs of genes with conserved expression pattern should be similar and close to zero. However, because of the mathematical properties of the transformation used (subtracting the mean, dividing by the euclidean norm), measured distances are different for different normalizations and expression specificity. Note that d^2_E can be estimated by summing squared distances (in the y-direction) from the points (representing the expression values of two genes for a given condition) to the identity line $y = x$. Simply put, the further away from the identity line the points are, the higher the d_E between two expression profile is. d^E_E is lower for uniformly expressed genes (figure S2C) than for specifically expressed genes (figure S2D). In contrast, d^Z_E is very high for uniformly expressed genes (figure S2E) and very low for specifically expressed genes (figure S2F). The reason is that $d^Z_E = \sqrt{2(1-r)}$ is a decreasing function of r, and r is around 0 for

uniformly expressed genes (figure S2E), and high for specifically expressed genes (figure S2F). In the first case, r reflects mainly the noise of the measurement. In the second case, the value of r is driven by the single outlier.

We illustrate the correlation between the distance and expression specificity using replicated data, both from mouse and human. We considered only the genes with both replicates in the same τ-group. We calculated the correlation between the distance of two replicates and their mean τ value. As expected, the correlation was positive for d_E^M and d_E^E, and negative for d_E^Z (figure S3).

Analysis of mouse gene expression data set

As mentioned in the main text, we performed our study also on mouse gene expression data set from the GNF atlas (Su *et al.* 2004). We divided mouse probe set pairs into three τ-groups of similar size (table S2). Similarly to the τ distribution in the human data (figure S4A-B), in case of the mouse data τ distribution is not uniform (figure S4C-D). Therefore, the first two τ-groups consisted mostly of broadly expressed genes ($\tau < 0.425$) and the third τ-group contained genes with more specific patterns of expression ($\tau > 0.425$). We measured Euclidean distances for probe set pairs within every τ-group and we found that values of d_E^M and d_E^E were significantly lower for broadly expressed genes than for organ-specific genes (figure S5A-B). On the contrary, values of d_E^Z were significantly higher for broadly expressed genes than for organ specific genes (figure S5C). The same conclusion is reached when the genes are divided into three τ-groups with balanced τ distribution (figure S6 for human data, and S7 for mouse data).

We also generated two sets of random mouse probe set pairs. One by random permutation of the two data set, second by the procedure of τ-uniform sampling. Again, we found that overrepresentation of broadly expressed genes caused underestimation of d_E^M and d_E^E between randomly permuted pairs and overestimation of d_E^Z between replicates (figure S8). d_E^M and d_E^E for τ-uniform random pairs seemed to estimate better the expected level of expression divergence under neutral evolution (figure S8).

For the details of this analysis and its conclusions, please refer to the main text.

Results of the comparative study of human and mouse gene expression with d_E^M

As mentioned in the main text, we performed the comparative study of human and mouse gene expression to demonstrate the effect of our approach. Here, we present the results of analogous analysis, but with d_E^M as a distance measure. We selected 8,942 one-to-one orthologous gene pairs from the human and mouse data sets (Su *et al.* 2004). We created two sets of random gene pairs, using both random permutation and the procedure of τ-uniform sampling, and we calculated the Euclidean distance (d_E^M) for orthologous gene pairs and for both sets of random pairs. If the d_E^M value for human-mouse orthologous gene pairs is smaller than the 5th percentile of d_E^M for randomly paired genes, there is some evidence that the expression evolution of this pair has been constrained (Liao and Zhang 2006). Using randomly permuted gene pairs did not provide clear evidence for constrained evolution (figure S9A). Only 8% of orthologous pairs were identified to have a conserved expression pattern, which was close to the random expectation of 5%. In contrast, using τ-uniform random pairs, 29% of orthologous genes were identified to have conserved expression (figure S9).

Comparison with the study of Liao and Zhang (2006a)

The number of detected genes with conserved expression pattern may seem surprisingly low in comparison to Liao and Zhang (2006), who reported that as much as 84% of genes showed conserved expression between human and mouse. We would like to draw the reader's attention to the fact that Liao and Zhang (2006) used two different metrics to calculate the distance between orthologous genes and between randomly paired genes - the so called net distance and the euclidean distance, respectively. The net distance was defined as $D = d - (d_h + d_m)/2$, where d was the Euclidean distance between mouse and human expression profiles, d_h was the Euclidean distance between the expression profiles of two randomly picked probe sets for the human gene, and d_m was the Euclidean distance between the expression profiles of two randomly picked probe sets for the mouse gene. We believe that the incompatibility of the distances calculated over orthologous and random gene pairs caused the overestimation of expression conservation. The authors argued that they applied the correction only for orthologous genes, because "randomly paired genes should have no expression similarity; thus, the Euclidean distance do not require correction" (page 535 in Liao and Zhang (2006)). Note that this implies that authors assumed in advance, that orthologous genes should be similar and because of that they corrected only their distance. Thus, the method applied by Liao and Zhang (2006) implies by default a difference between distributions of distance values for orthologous genes and randomly paired genes. Here, we demonstrate that for two sets of randomly paired genes, which by definition should display a similar rate of divergence, the net distance is significantly lower than the Euclidean distance (figure S10). Consequently, the method used in Liao and Zhang (2006) clearly introduces a bias towards overestimation of conservation.

References

Domazet-Lošo, T. and Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, **468**(7325), 815–8.

Liao, B.-Y. and Zhang, J. (2006). Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol*, **23**(3), 530–540.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, **101**(16), 6062–6067.

Supplementary Tables and Figures

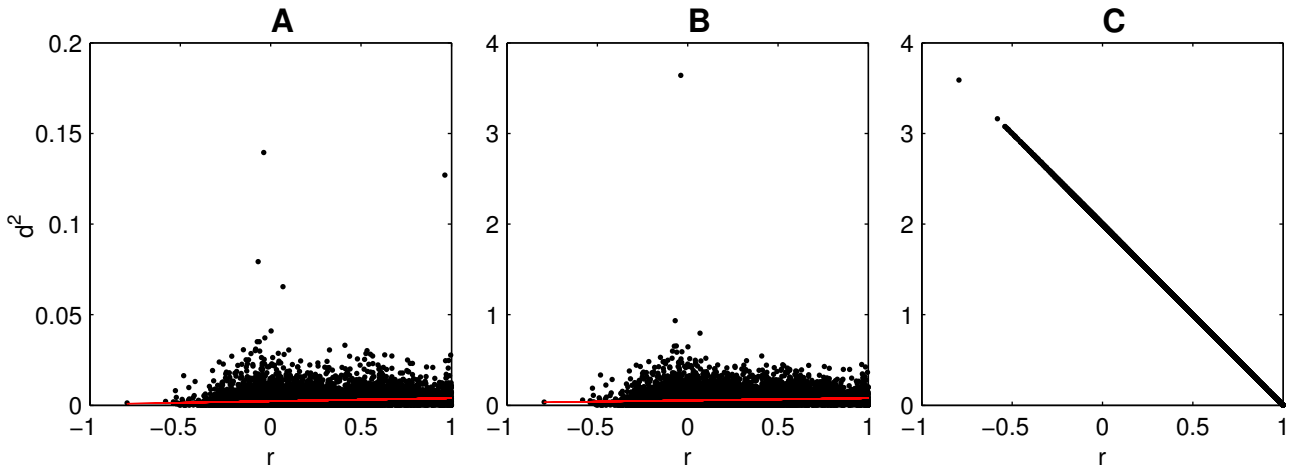

Figure S1. Interdependence between r and d depends on data normalization mode. Both for Manhattan (A) and Euclidean (B) normalizations the correlation between r and d^2 is low (0.12 and 0.09, respectively). (C) For z-like normalization there is linear dependence between r and d^2.

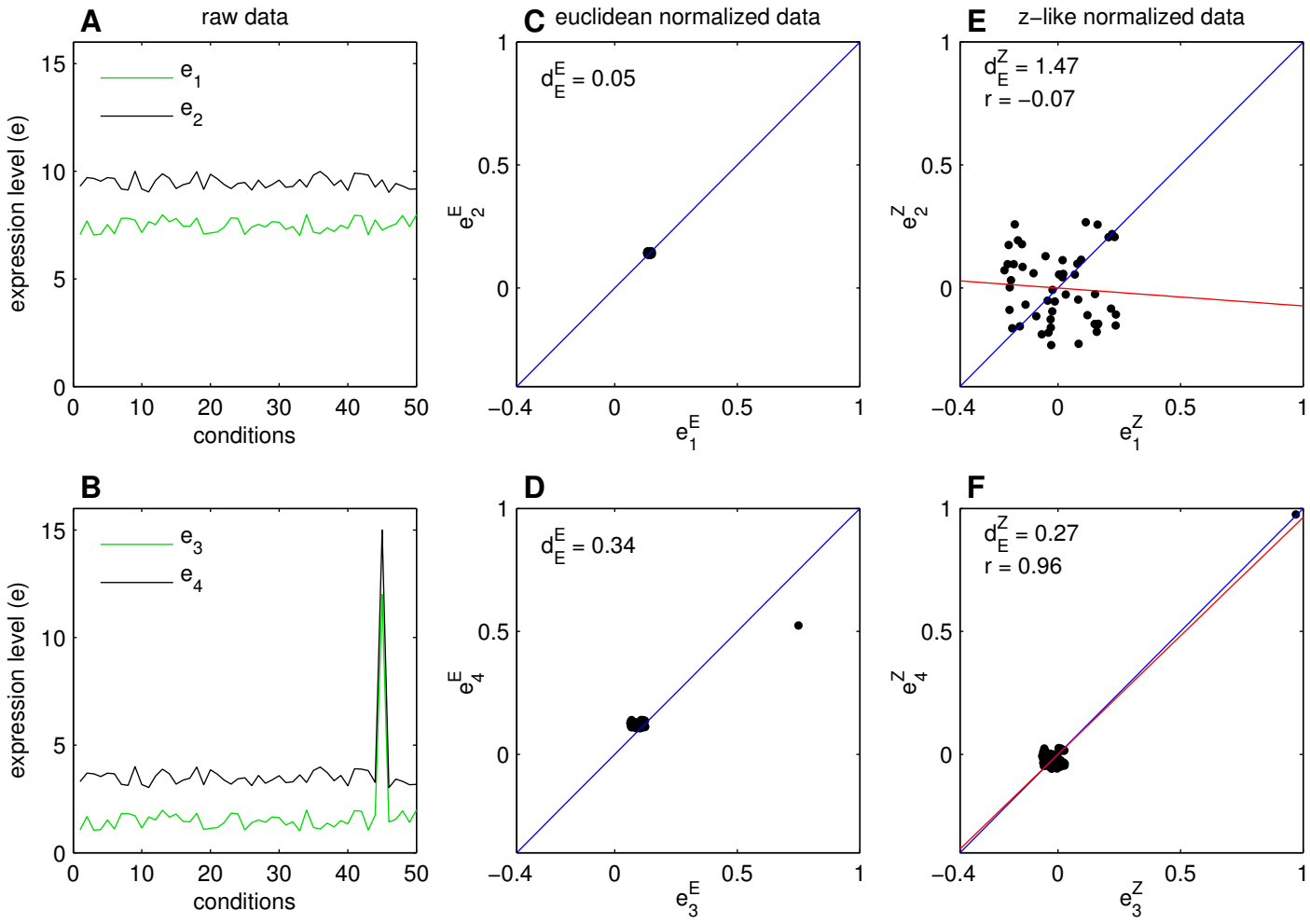

Figure S2. Euclidean distance between two genes with conserved expression patterns (A, B) depends on data normalization mode and specificity of genes expression. For Euclidean normalization the distance is lower for genes expressed over all conditions (C) than for specifically expressed genes (D). For z-like normalization the distance is higher for genes expressed over all conditions (E) than for specifically expressed genes (F). Regression line is plotted in red. Identity line ($y = x$) is plotted in blue. Note that d_E^2 can be estimated by summing squared distances (in the y-direction) from the points to the blue line.

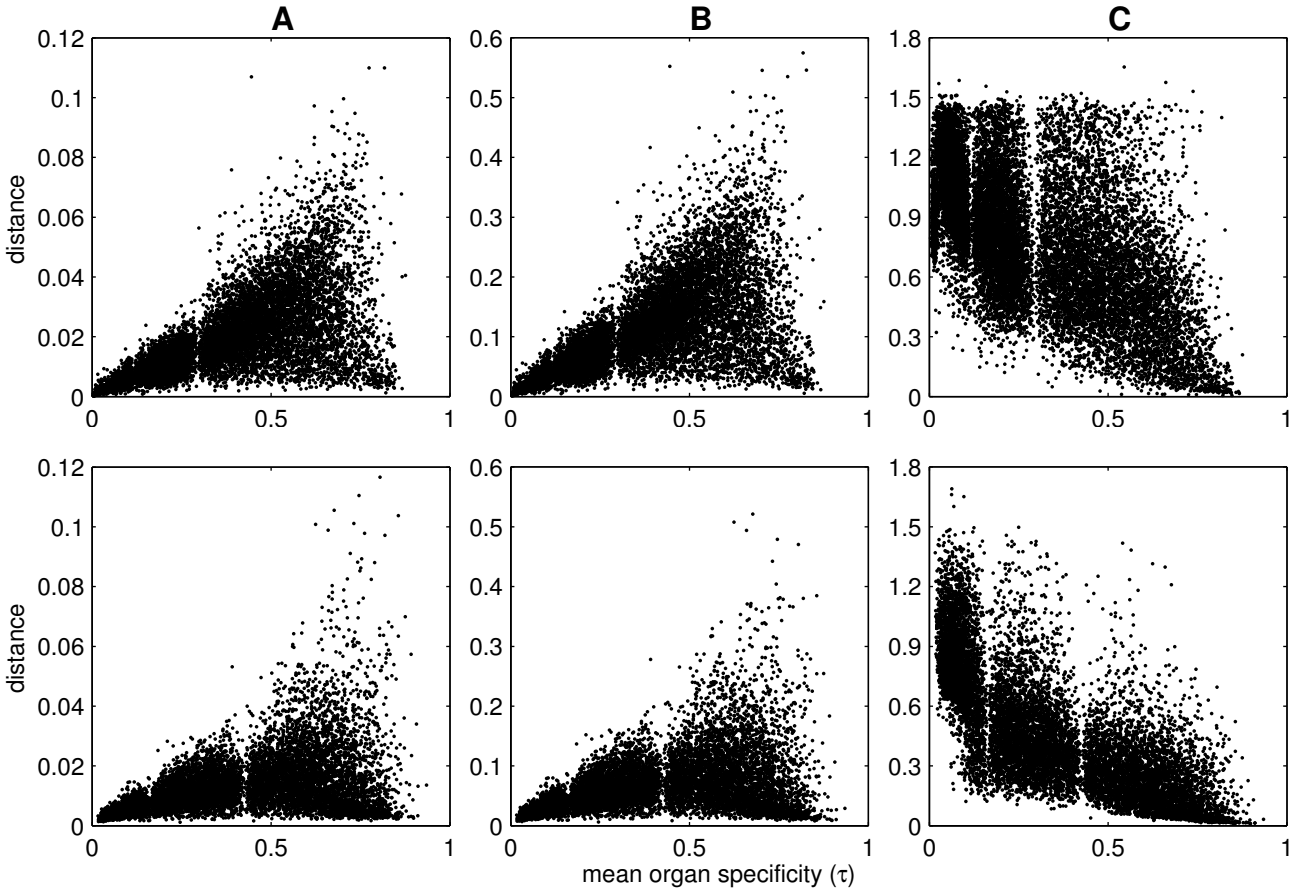

Figure S3. Euclidean distance between replicates (simulating genes with conserved expression patterns) depends on data normalization and specificity of genes expression. For Manhattan and Euclidean normalization (A, B) the distance is positively correlated with the expression specificity, whereas for z-like normalization (C) this correlation is negative. Top: human replicates. Spearman correlation coefficients: for d_E^M : 0.89, for d_E^E : 0.88, for d_E^Z : −0.56. Bottom: mouse replicates. Spearman correlation coefficients: for d_E^M : 0.68, for d_E^E : 0.64, for d_E^Z : −0.81.

Table S1. List of homologous organ groups (HOGs) and their corresponding organs (sample names) in mouse and human

HOG id	HOG name	sample name (human)	sample name (mouse)
HOG:0001141	adrenalgland_interrenalgland	3AJZ02022645_adrenal_gland.CEL	MGJZ030211002Aadrenalgland.CEL
		1BJZ02022646_adrenalgland.CEL	MGJZ030211002Badrenalgland.CEL
		1BJZ02022761_adrenalgland.CEL	
		3AJZ02022762_adrenal_gland.CEL	
HOG:0001277	amygdala	1BMH02022705_Brain_Amygdala.CEL	MGMH030211062Aamygdala.CEL
		1BMH02022105_Brain_Amygdala.CEL	MGMH030211062Bamygdala.CEL
		3AMH02022205_Brain_Amygdala.CEL	
		3AMH02022605_Brain_Amygdala.CEL	
HOG:0001218	bonemarrow	3AJZ02022865_bone_marrow.CEL	MGMH030312056Abonemarrow.CEL
		3AJZ02030474_bonemarrow.CEL	MGMH030312056Bbonemarrow.CEL
		1BJZ02022866_bonemarrow.CEL	
		1BJZ02030473_bonemarrow.CEL	
HOG:0000024	cerebellum	3AJZ02022764_cerebellum.CEL	MGJZ030207007Acerebellum.CEL
		1BJZ02022763_cerebellum.CEL	MGJZ030207007Bcerebellum.CEL
		1BJZ02051612_CerebellumPeduncles.CEL	
		1BJZ02022648_cerebellum.CEL	
		3AJZ02061902_CerebellumPeduncles.CEL	
		1BJZ02051611_CerebellumPeduncles.CEL	
		3AJZ02022647_cerebellum.CEL	
		3AJZ02051604_CerebellumPeduncles.CEL	
HOG:0000722	cerebralcortex	3AJZ02060407_CingulateCortex.CEL	MGJZ030212008Bcerebralcortex.CEL
		3AJZ02053109_PrefrontalCortex.CEL	MGMH030312008Acortex.CEL
		3AJZ02060508_CingulateCortex.CEL	
		1BJZ02060505_CingulateCortex.CEL	
		1BJZ02060408_CingulateCortex.CEL	
		3AJZ02053111_PrefrontalCortex.CEL	
		1BJZ02053112_PrefrontalCortex.CEL	
		1BJZ02053110_PrefrontalCortex.CEL	

HOG:0000222	dorsalrootganglion	3ARS02080736e_DRG.CEL	MGJZ030312065Bdorsalrootganglion.CEL
		3ARS02080736f_DRG.CEL	MGJZ030312065Adorsalrootganglion.CEL
		1BRS02081536e_DRG.CEL	
		3AJZ02081478a_Superior_Cervical_Ganglion.CEL	
		1BJZ02081478a_Superior_Cervical_Ganglion.CEL	
		3AJZ02081478b_Superior_Cervical_Ganglion.CEL	
		1BJZ02081478b_Superior_Cervical_Ganglion.CEL	
		1BRS02081536f_DRG.CEL	
HOG:0000276	heart	1BRS02080872a_atrioventricular_node.CEL	MGJZ030207054Bheart.CEL
		3ARS02080772a_atrioventricular_node.CEL	MGJZ030207054Aheart.CEL
		3AJZ02021909_HEART.CEL	
		1BRS02080872b_atrioventricular_node.CEL	
		1BMH02022808_Heart.CEL	
		1BMH02022802_Heart.CEL	
		3AMH02030702_Heart.CEL	
		3ARS02080772b_atrioventricular_node.CEL	
HOG:0000143	hypophysis	3AJZ02022867_Pituitary.CEL	MGJZ030228080Bpituitary.CEL
		1BJZ02061809_pituitary.CEL	MGJZ030228080Apituitary.CEL
		1BJZ02061909_pituitary.CEL	
		3AJZ02030476_pituitary.CEL	
HOG:0000179	hypothalamus	3AJZ02060506_Hypothalamus.CEL	MGMH030212053Ahypothalamus.CEL
		1BJZ02060406_Hypothalamus.CEL	MGMH030402094Bpreoptic.CEL
		3AJZ02061907_Hypothalamus.CEL	MGMH030401094Apreoptic.CEL
		1BJZ02072563_Hypothalamus.CEL	MGMH030212053Bhypothalamus.CEL
HOG:0000257	liver	3AJZ02021915__LIVER.CEL	MGJZ030211018Aliver.CEL
		1BMH02022812_Liver.CEL	MGJZ030211018Bliver.CEL
		1BMH02022806_Liver.CEL	
		3AMH02030706_Liver.CEL	
HOG:0000310	lung_swimbladder	3AJW02021805_lung.CEL	MGMH030211019Blung.CEL
		3AMH02030705_Lung.CEL	MGMH030211019Alung.CEL
		1BMH02022805_Lung.CEL	
		1BMH02022811_Lung.CEL	
HOG:0001273	lymphnode	3AJZ02022115__lymph_node.CEL	MGJZ030207020Alymphnode.CEL
		3AJZ02022232__lymph_node.CEL	MGJZ030207020Blymphnode.CEL
		1BJZ02022116_lymphnode.CEL	
		1BJZ02022231_lymphnode.CEL	

HOG:0000039	metanephros	3AJZ02022760_kidney.CEL	MGMH030312016Akidney.CEL
		3AJZ02022643_kidney.CEL	MGJZ030211016Bkidney.CEL
		1BJZ02030684_KIDNEY.CEL	
		1BJZ02022644_kidney.CEL	
HOG:0000033	olfactorybulb	1BJZ02060511_OlfactoryBulb.CEL	MGJZ030212070Bolfactorybulb.CEL
		3AJZ02072561_OlfactoryBulb.CEL	MGJZ030212070Aolfactorybulb.CEL
		3AJZ02060514_OlfactoryBulb.CEL	
		1BJZ02060414_OlfactoryBulb.CEL	
HOG:0000251	ovary	3AJZ02052302_Ovary.CEL	MGMH030402089Boocyte.CEL
		3AJZ02050806_Ovary.CEL	MGMH030228089Aoocyte.CEL
		1BJZ02052208_Ovary.CEL	MGMH030312023Aovary.CEL
		1BJZ02050813_Ovary.CEL	MGMH030312023Bovary.CEL
HOG:0000050	pancreas	1BMH02022702_Pancreas.CEL	MGMH030212060Apancreas.CEL
		1BJZ02060501_Pancreas.CEL	MGMH030212060Bpancreas.CEL
		3AJZ02060401_Pancreas.CEL	
		3AJZ02060502_Pancreas.CEL	
HOG:0001266	placenta	3ARS0207263HB_PLACENTA.CEL	MGJZ030312066Aplacenta.CEL
		1BRS0207253HB_PLACENTA.CEL	MGJZ030312066Bplacenta.CEL
		1BRS0207253IB_PLACENTA.CEL	
		3ARS0207253IA_PLACENTA.CEL	
HOG:0001261	prostate	1BMH02022804_Prostate.CEL	MGJZ030212025Aprostate.CEL
		3AMH02030704_Prostate.CEL	MGJZ030212025Bprostate.CEL
		1BMH02030601_Prostate.CEL	
		3AJZ02021911__PROSTATE.CEL	
HOG:0000376	salivarygland	1BJZ02041227_salivarygland.CEL	MGJZ030212027Asalivarygland.CEL
		3AJZ02041226_salivarygland.CEL	MGJZ030212027Bsalivarygland.CEL
		3AJZ02040823_salivarygland.CEL	
		1BJZ02040822_salivarygland.CEL	
HOG:0000319	skeletalmuscle	1BJZ02083092b_Skeletal_Muscle_Psoas.CEL	MGMH030312028Bskeletalmuscle.CEL
		3AJZ02083092a_Skeletal_Muscle_Psoas.CEL	MGMH030312028Askeletalmuscle.CEL
		1BJZ02083092a_Skeletal_Muscle_Psoas.CEL	
		3AJZ02083092b_Skeletal_Muscle_Psoas.CEL	
HOG:0000601	spinalcord	3AJZ02022107___spinal_cord.CEL	MGMH030212058Bspinalcordupper.CEL
		1BJZ02022223_spinalcord.CEL	MGMH030212058Aspinalcordupper.CEL
		1BJZ02022108_spinalcord.CEL	MGMH030212057Aspinalcordlower.CEL
		3AJZ02022224___spinal_cord.CEL	MGMH030212057Bspinalcordlower.CEL

HOG:0000252	testis	3AJZ02052114_TestiSeminiferousTubule.CEL	MGJZ030207135Btestis.CEL
		3AJZ02051707_Testi_GermCell.CEL	MGJZ030207135Atestis.CEL
		3AJZ02051709_Testi_Intersitial.CEL	
		3AJZ02052108_Testi-GermCell.CEL	
		1BJZ02051708_Testi_GermCell.CEL	
		3AJZ02051711_Testi_LeydigCell.CEL	
		1BJZ02052111_TestiLeydigCell.CEL	
		3AJZ02051713_Testi_SeminiferousTubule.CEL	
		3AJZ02052112_TestiLeydigCell.CEL	
		3AJZ02052305_TestiIntersitial.CEL	
		1BJZ02022636_testis.CEL	
		1BJZ02052113_TestiSeminiferousTubule.CEL	
		1BJZ02052107_TestiGermCell.CEL	
		1BJZ02052109_TestiIntersitial.CEL	
		1BJZ02051712_Testi_LeydigCell.CEL	
		1BJZ02022751_testis.CEL	
		3AJZ02022752_testis.CEL	
		3AJZ02022635_testis.CEL	
		1BJZ02051714_Testi_SeminiferousTubule.CEL	
		1BJZ02051710_Testi_Intersitial.CEL	
HOG:0000253	thymus	3AJZ02031411_thymus.CEL	MGMH030311036Bthymus.CEL
		1BJZ02022642_thymus.CEL	MGMH030311036Athymus.CEL
		3AJZ02022758_thymus.CEL	
		1BJZ02022757_thymus.CEL	
HOG:0000418	thyroid	1BMH02022704_Thyroid.CEL	MGMH030212037Athyroid.CEL
		3AMH02022204_Thyroid.CEL	MGMH030212037Bthyroid.CEL
		3AMH02022604_Thyroid.CEL	
		1BMH02022104_thyroid.CEL	
HOG:0000419	tongue	3AJZ02082987B_TONGUE.CEL	MGMH030311038Btongueepidermis.CEL
		1BJZ022987A_TONGUE.CEL	MGMH030311038Atongueepidermis.CEL
		3AJZ02082987A_TONGUE.CEL	
		1BJZ022987B_TONGUE.CEL	
HOG:0000371	trachea	3AJZ02022639_trachea.CEL	MGMH030311039Btrachea.CEL
		1BJZ02022755_trachea.CEL	MGMH030311039Atrachea.CEL
		3AJZ02022756_trachea.CEL	
		1BJZ02022640_trachea.CEL	

HOG:0001137	uterus	3AMH02030703_Uterus.CEL 1BJZ02083089b_Uterus_Corpus.CEL 3AJZ02083089b_Uterus_Corpus.CEL 1BMH02022809_Uterus.CEL 3AJZ02083089a_Uterus_Corpus.CEL 1BMH02022803_Uterus.CEL 1BJZ02083089a_Uterus_Corpus.CEL 3AJZ02021913__UTERUS.CEL	MGJZ030207041Auterus.CEL MGJZ030207041Buterus.CEL

Table S2. Composition of three τ-groups of mouse probe set (ps) pairs

	Organ-specificity (τ)	Number of mouse ps pairs
τ-group 1	$0.011 \leq \tau \leq 0.163$	4442
τ-group 2	$0.163 < \tau \leq 0.425$	4041
τ-group 3	$0.425 < \tau \leq 0.942$	4603

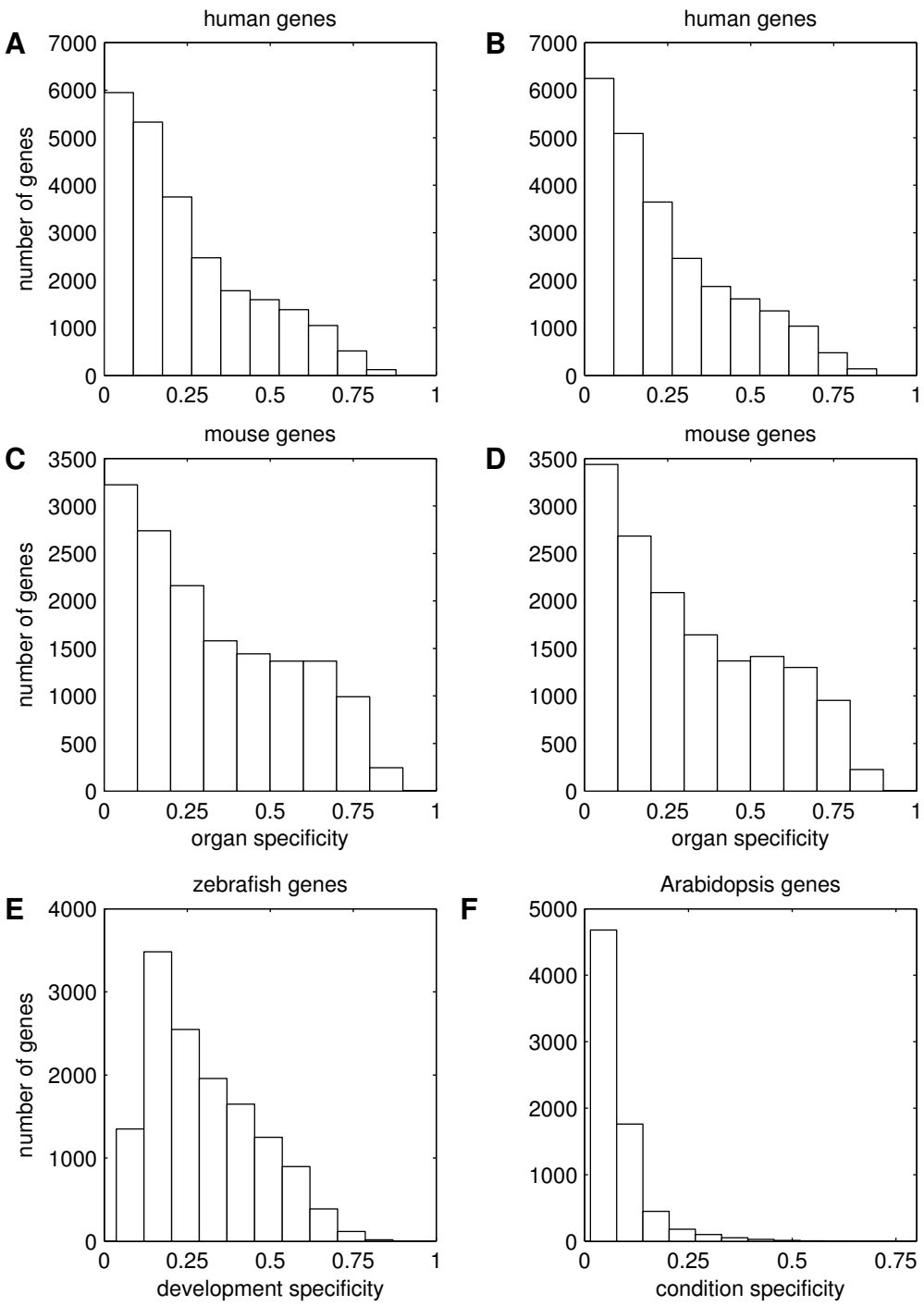

Figure S4. τ distribution is not uniform in the real data. (A, B) τ distribution for human replicates. (C, D) τ distribution for mouse replicates. (E) τ distribution for zebrafish genes expressed during the ontogeny (Domazet-Lošo and Tautz 2010). (F) τ distribution for Arabidopsis genes expressed in different light conditions (NASC 2007, GEO accession number GSE5617).

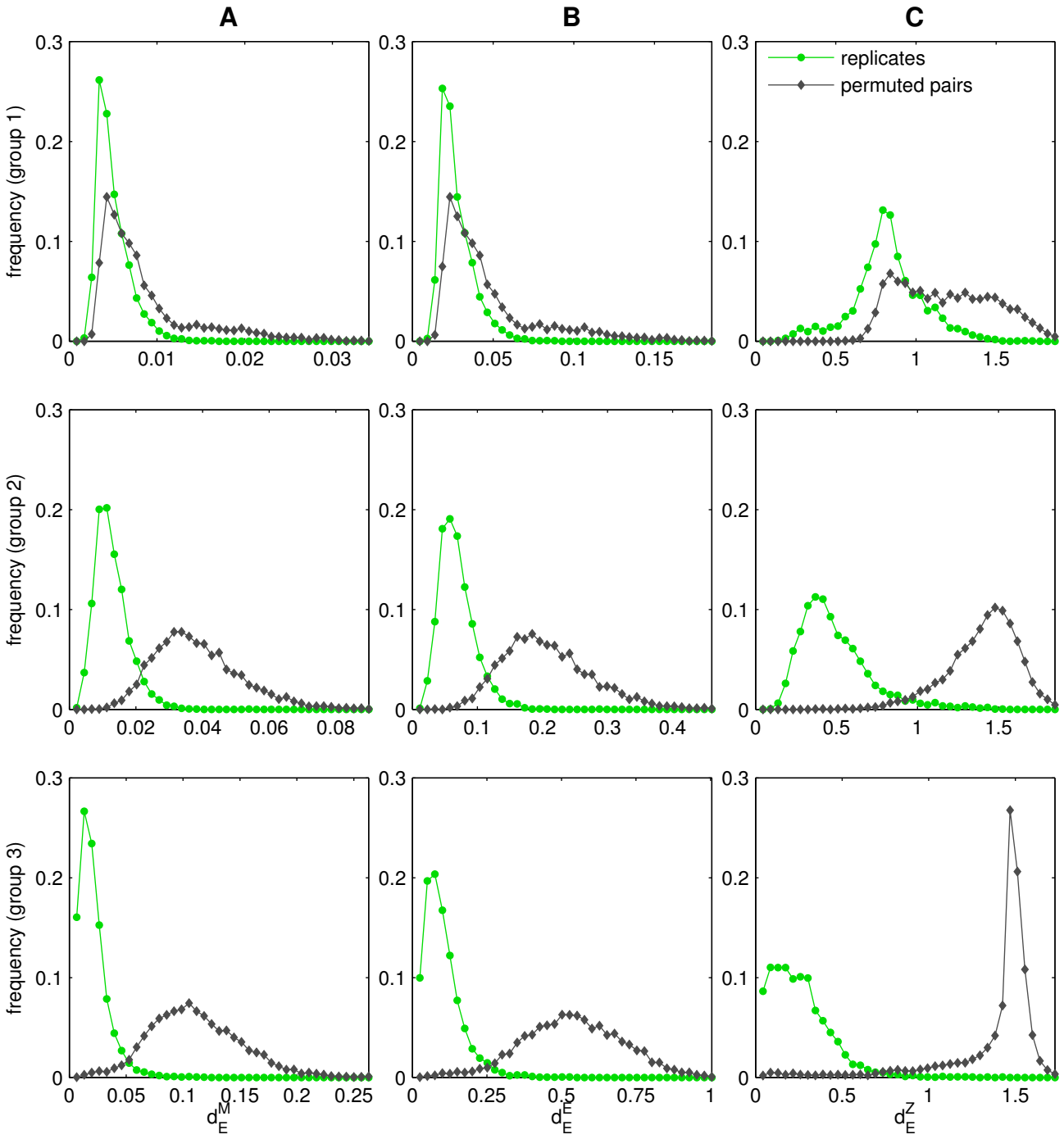

Figure S5. The distribution of expression similarity between mouse replicates depends on their organ specificity. (A) d_E^M and (B) d_E^E are significantly lower for broadly expressed genes (group 1) than for organ specific genes (group 3). For randomly permuted pairs of genes d_E^M and d_E^E also differ between the three τ-groups. They are significantly lower for random pairs in group 1 than in group 3. (C) d_E^Z is significantly higher for broadly expressed genes (group 1) than for organ specific genes (group 3). d_E^Z for randomly permuted pairs is high in all three groups even in the first τ-group, where random pairs consist of two broadly expressed genes (this is a consequence of low r for uniformly expressed genes) Note that scale of x-axis differs strongly between graphs.

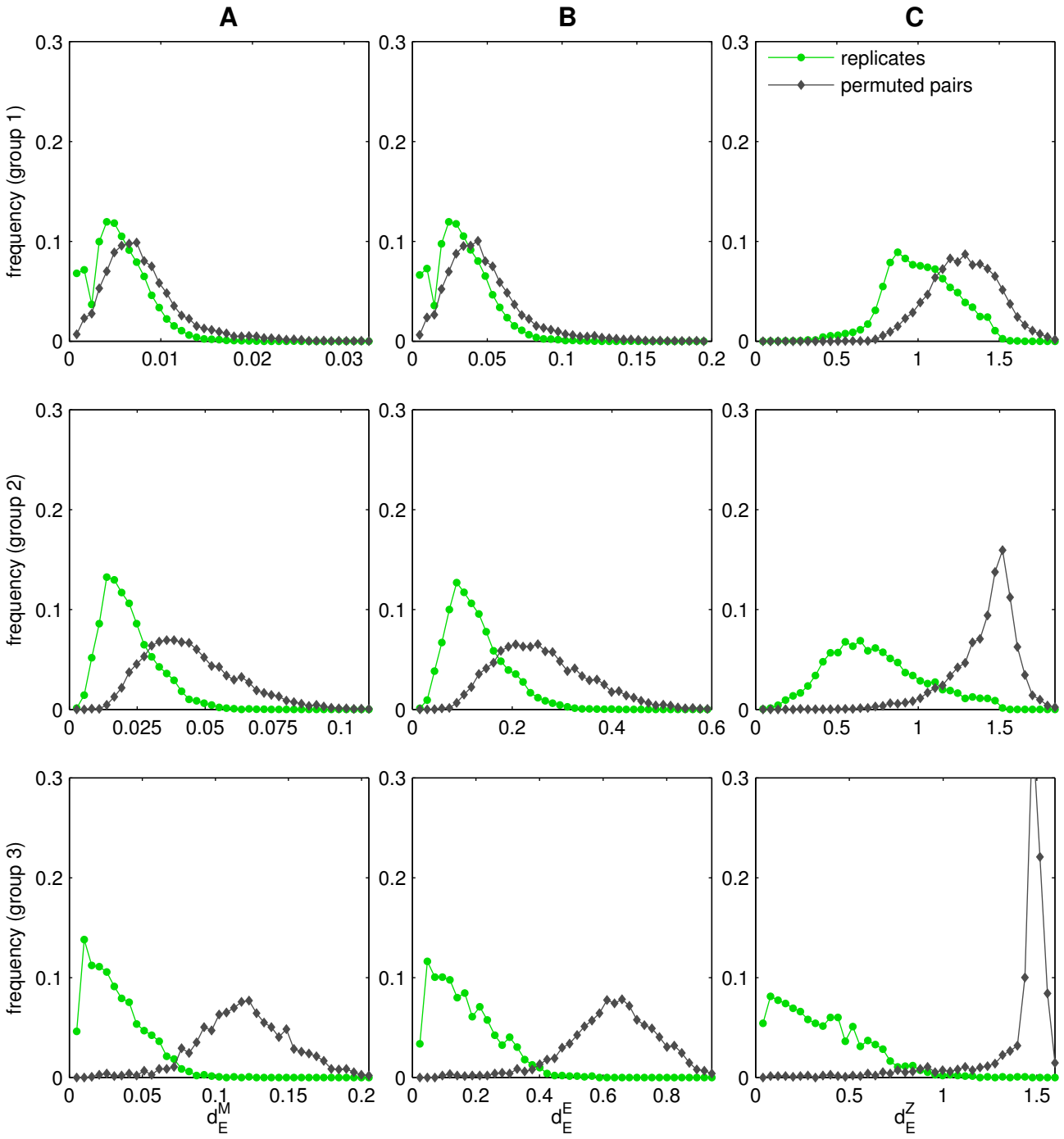

Figure S6. The distribution of expression similarity between human replicates depends on their organ specificity. Presented 3 groups of gene pairs have balanced τ distribution. Group 1: τ: 0 - 0.2, 10723 gene pairs; Group 2: τ: 0.2 - 0.6, 7551 gene pairs; Group 3: τ: 0.6 - 1, 1514 gene pairs. For the explanation of the figure please refer to figure S5.

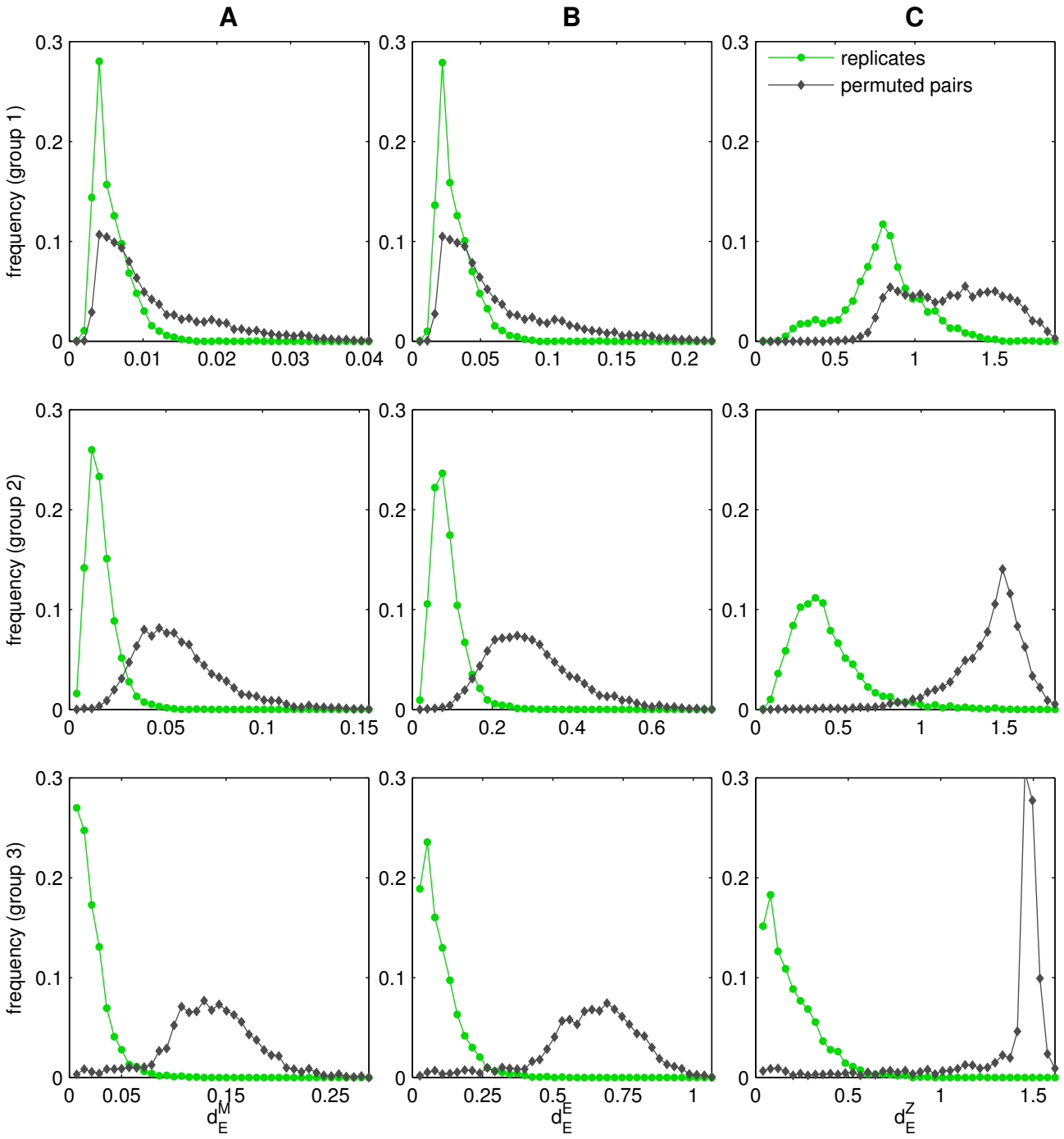

Figure S7. The distribution of expression similarity between mouse replicates depends on their organ specificity. Presented 3 groups of gene pairs have balanced τ distribution. Group 1: τ: 0 - 0.2, 5424 gene pairs; Group 2: τ: 0.2 - 0.6, 5688 gene pairs; Group 3: τ: 0.6 - 1, 2303 gene pairs. For the explanation of the figure please refer to figure S5.

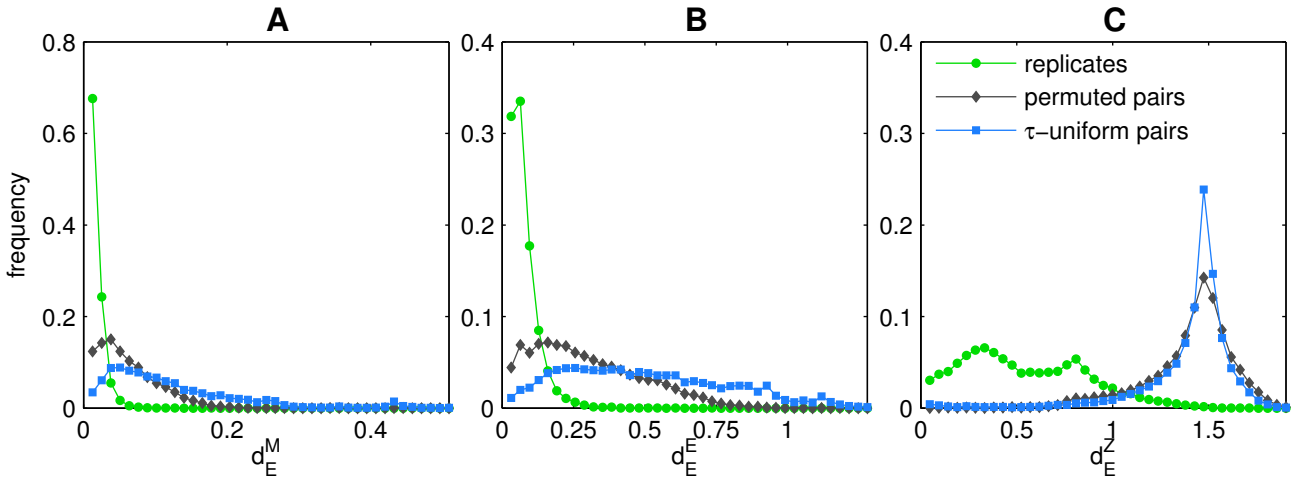

Figure S8. Overrepresentation of broadly expressed mouse genes causes underestimation of the conservation of expression when randomly permuted pairs are used to approximate the neutral evolution rate. (A, B) For noticeable number of randomly permuted pairs the distances (d_E^M and d_E^E) are small, indistinguishable from the distances for replicates. (C) d_E^Z is high both for permuted gene pairs and for the group of replicates. (A, B) For τ-uniform random pairs d_E^E and d_E^M are higher, which is more consistent with the assumption about neutral evolution from Jordan et al. (2005). (C) distribution of d_E^Z does not change with the new random pairs set.

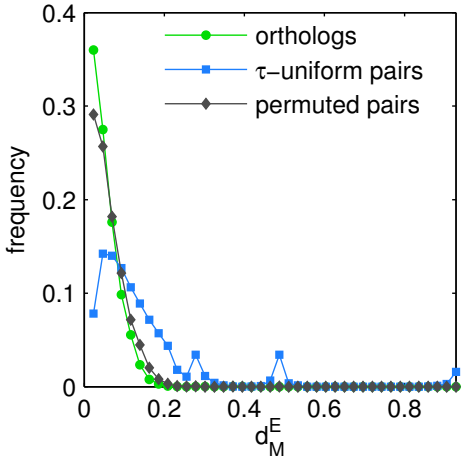

Figure S9. The choice of the randomization method changes the conclusions about gene expression evolution between mouse and human. There is no clear evidence for constrained evolution if we compare the distribution of d_E^M for orthologous (green) and randomly permuted gene pairs (grey). Whereas, comparison of d_E^M distribution for orthologous (green) and τ-uniform random pairs (blue) suggest that expression evolution is far from neutral.

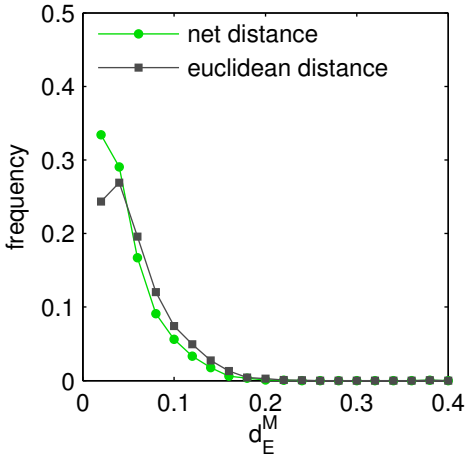

Figure S10. One-sided correction of the Euclidean distance lead to different distributions of distance values for two sets of randomly paired genes. Using 3,193 human-mouse orthologous gene pairs (all human genes covered by multiple probe sets), we generated two sets of randomly permuted gene pairs. For the first set (simulating the set of genes with non-conserved expression profiles) we calculated the net distance, for the second set (used to estimate neutral evolution) we calculated the Euclidean distance. Because both sets were "equally random", one should not expect any differences between them. However, as much as 20% of gene pairs from the first random set (green) was detected to be more conserved than gene pairs from the second random set (grey).