

Supplementary materials

Re-analysis of the Asian samples of Yang et al.¹⁵ – Supplementary methods

In this analysis we aimed to reproduce the results published by Yang et al. in 2018. To do so, we used only the Asian samples. We also excluded the continent ID from the model as all animals were from Asia.

In accordance with the applied analysis workflow, we defined an *expressed gene* as a gene that contained minimum ten reads in at least two individual samples. The minimum number of samples was reduced to two as compared to the main study, where this parameter was three; this was necessary as the smallest cluster included only two samples in this reduced experiment (n=2 in the dog cluster). With this definition, we found 12,435 expressed genes in our samples.

Supplementary data 1: the complete list of significant gene ontology terms from the main analysis.

Supplementary tables

Table S1 Hemoglobin-related genes from Harrington et al.¹⁴ and their canine orthologs together with genomic information about the genes. The dashed line separates the active and archived canine genes in Ensembl's annotation database.

Chr ID	Gene start	Gene end	Strand	Dog gene ID	Human gene ID	Human gene abbreviation	Human gene name	Gene type	Ortholog type	Dog gene's status
6	40324263	40326351	Reverse	ENSCAFG00000032615	ENSG00000188536	HBA2	hemoglobin subunit alpha 2	Protein coding	Many to many	Active
6	40324263	40326351	Reverse	ENSCAFG00000032615	ENSG00000206172	HBA1	hemoglobin subunit alpha 1	Protein coding	Many to many	Active
6	40330683	40332438	Reverse	ENSCAFG00000029224	ENSG00000206177	HBM	hemoglobin subunit mu	Protein coding	One to one	Active
6	40342562	40343504	Reverse	ENSCAFG00000028569	ENSG00000130656	HBZ	hemoglobin subunit zeta	Protein coding	One to many	Active
21	28181347	28205138	Reverse	ENSCAFG00000030286	ENSG00000223609	HBD	hemoglobin subunit delta	Protein coding	Many to many	Active
21	28181347	28205138	Reverse	ENSCAFG00000030286	ENSG00000244734	HBB	hemoglobin subunit beta	Protein coding	Many to many	Active
6	40326459	40329857	Reverse	ENSCAFG00000029904	ENSG00000188536	HBA2	hemoglobin subunit alpha 2	Protein coding	Many to many	Archived
6	40326459	40329857	Reverse	ENSCAFG00000029904	ENSG00000206172	HBA1	hemoglobin subunit alpha 1	Protein coding	Many to many	Archived
6	40331793	40332696	Reverse	ENSCAFG00000031055	ENSG00000130656	HBZ	hemoglobin subunit zeta	Protein coding	One to many	Archived
21	28179119	28180299	Reverse	ENSCAFG00000029518	ENSG00000223609	HBD	hemoglobin subunit delta	Protein coding	Many to many	Archived
21	28179119	28180299	Reverse	ENSCAFG00000029518	ENSG00000244734	HBB	hemoglobin subunit beta	Protein coding	Many to many	Archived
21	28193272	28194670	Reverse	ENSCAFG00000024181	ENSG00000196565	HBG2	hemoglobin subunit gamma 2	Protein coding	One to many	Archived
21	28193272	28194670	Reverse	ENSCAFG00000024181	ENSG00000213934	HBG1	hemoglobin subunit gamma 1	Protein coding	One to many	Archived

Table S2 Biotype classification of the significantly differentially expressed genes. Only the number of genes per biotype is shown; the biotype descriptions are quoted from Ensembl's website.

Ensembl biotype	Biotype description ¹	Gene count	Note
Protein coding	Gene/transcript that contains an open reading frame (ORF)	1712	
lncRNA	A non-coding gene/transcript >200bp in length	36	A subset of the <i>Processed transcript</i> biotype
Pseudogene	A gene that has homology to known protein-coding genes but contains a frameshift and/or stop codon(s) which disrupts the ORF. Thought to have arisen through duplication followed by loss of function	25	
Processed pseudogene	Pseudogene that lacks introns and is thought to arise from reverse transcription of mRNA followed by reinsertion of DNA into the genome	10	A subset of the <i>pseudogene</i> biotype
IG V gene	Variable chain immunoglobulin gene that undergoes somatic recombination before transcription	11	A subset of the <i>IG gene</i> biotype
miRNA	A small RNA (~22bp) that silences the expression of target mRNA	5	A subset of the <i>ncRNA</i> biotype
Mt rRNA	transfer RNA	2	A subset of the <i>ncRNA</i> biotype
Mt tRNA	transfer RNA	2	A subset of the <i>ncRNA</i> biotype
snoRNA	Small RNA molecules that are found in the cell nucleolus and are involved in the post-transcriptional modification of other RNAs	2	A subset of the <i>ncRNA</i> biotype
miscRNA	Miscellaneous RNA. A non-coding RNA that cannot be classified	1	A subset of the <i>ncRNA</i> biotype
snRNA	Small RNA molecules that are found in the cell nucleus and are involved in the processing of pre messenger RNAs	1	A subset of the <i>ncRNA</i> biotype

1: sources – Ensembl (25), GenCode (32) and Vega (33)

Supplementary figures

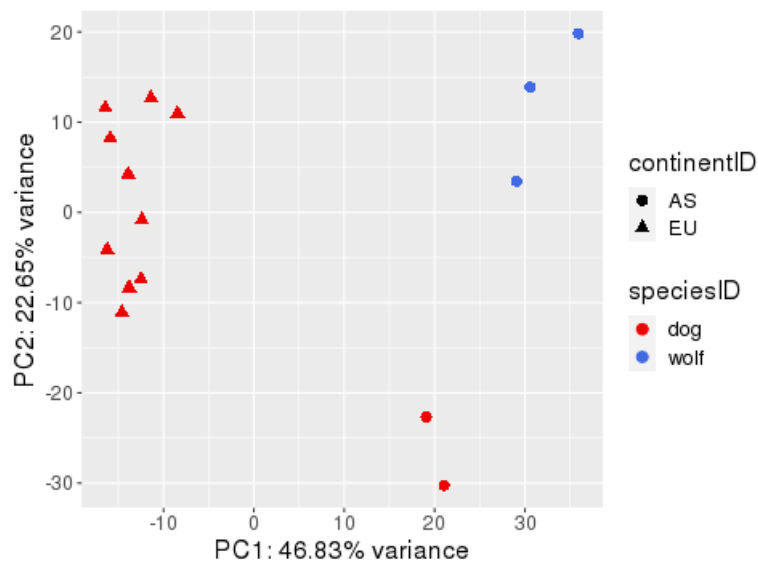


Figure S1: Principal component analysis of the raw read counts after regularized-logarithm transformation (first two PCs are shown). The amount of variance explained by the first two components is also indicated. Marker shapes indicate the origin of the samples, marker colors indicate the species.

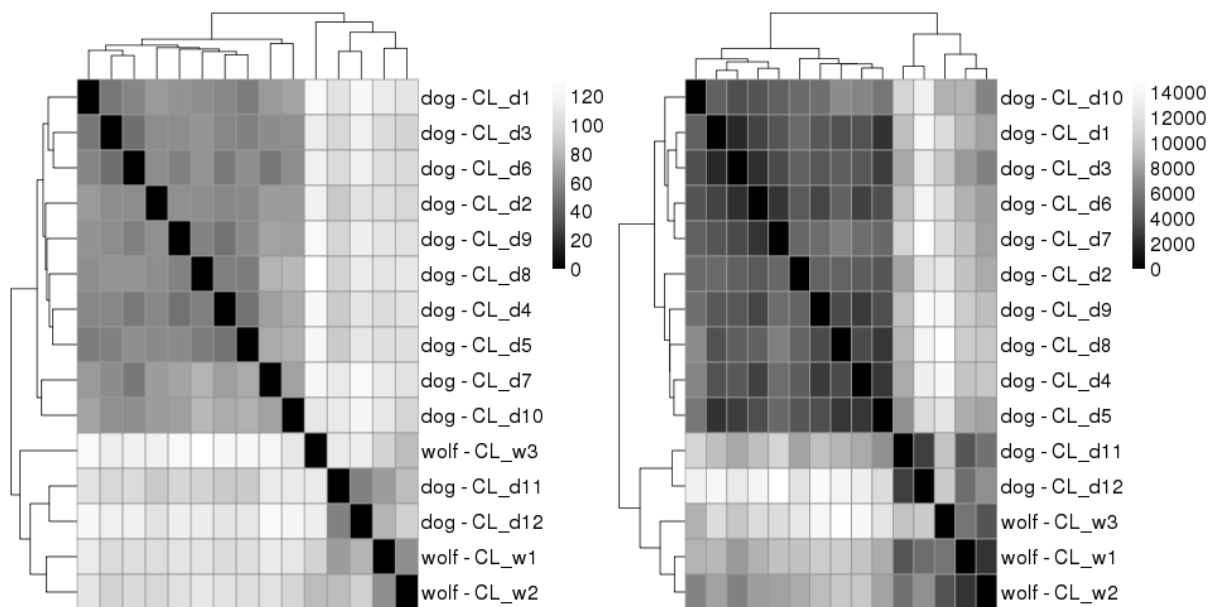


Figure S2: Euclidean (left) and Poisson (right) distances between the samples based on the read counts after regularized-logarithm transformation. Clustering of the test subjects is also shown at the margins of the figures.

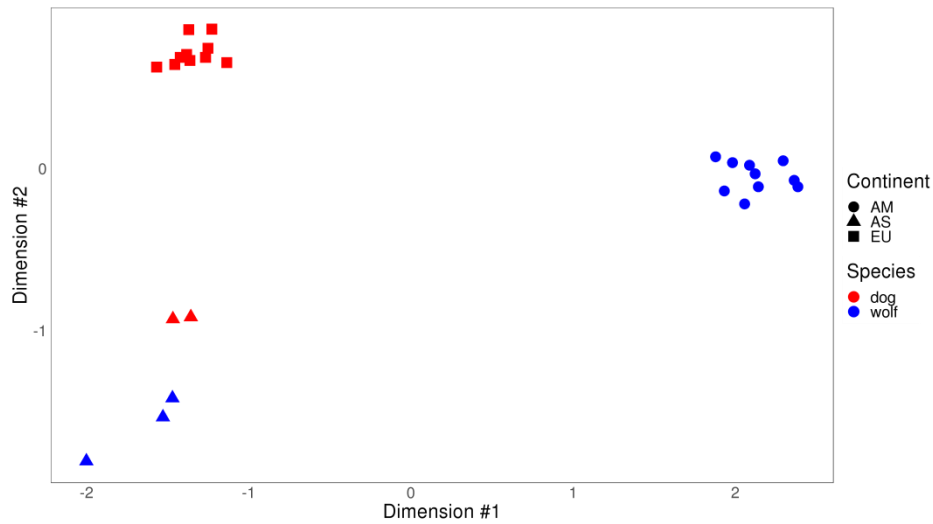


Figure S3: Multidimensional scaling of 25 canids from three continents: Europe ($n=10$)¹⁶, Asia ($n=5$)¹⁵ and the Americas ($n=10$)¹⁸.

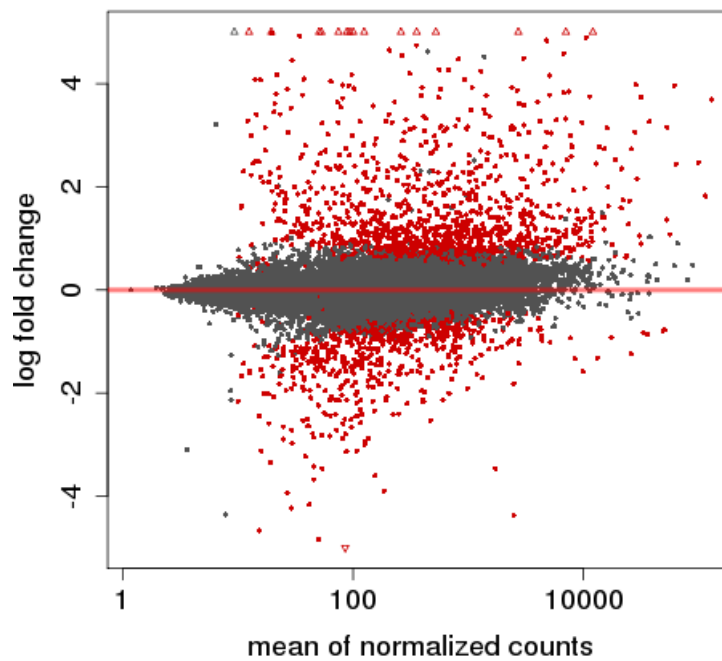


Figure S4: An MA-plot showing the mean normalized read counts of each gene (“A”; x axis) vs $\log_2$ fold change (“M”; y axis). Significant genes ($n=1807$) are highlighted in red. Genes with a higher than 5, or lower than $-5 \log_2$ fold change value are indicated with triangles.

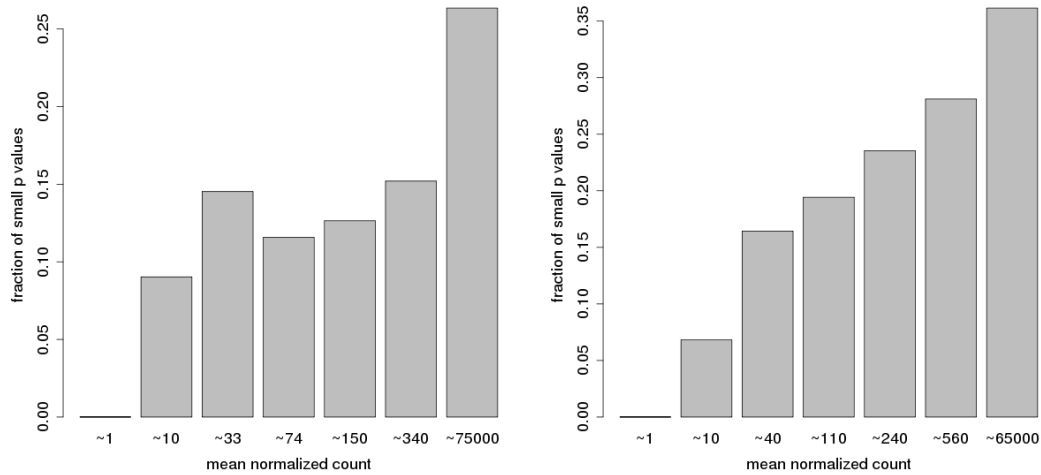


Figure S5: Proportion of low p-values ($p < 0.05$) among the tested genes divided into 6 quantiles, each containing $\sim 16.67\%$ of the genes. Genes were divided into quantiles based on their normalized read count. The same plot is shown for the re-analysis of the data published by Yang et al.¹⁵ (left) and for the analysis including our additional ten pet border collie dog samples (right).

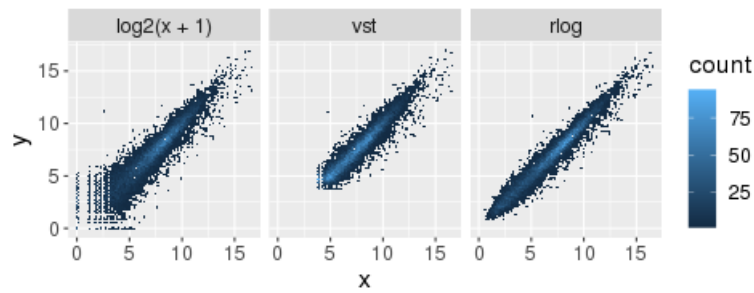


Figure S6: The effect of the different transformations on the raw read counts in the first two European dogs (CL_d1 and CL_d2).