Taming of the wild: a new method for cross study RNA-seq analysis

Diana Lobo (diana.lobo@cibio.up.pt)
Universidade do Porto Centro de Investigacao em Biodiversidade e Recursos Geneticos
https://orcid.org/0000-0001-6988-9993

Raquel Godinho
Universidade do Porto Centro de Investigacao em Biodiversidade e Recursos Geneticos;Departamento de Biologia, Faculdade de Ciencias, Universidade do Porto, Rua do Compo Alegre

John Archer
Universidade do Porto Centro de Investigacao em Biodiversidade e Recursos Geneticos

Methodology article

Keywords: canids, domestication, intra-condition variation, gene expression, RNA-seq, software, transcriptomics

Posted Date: May 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-27674/v1

Abstract

Background

In the last decades, the evolution of RNA-Seq has yielded archived datasets that possess the potential for providing unprecedented inter-study insight into transcriptome evolution, once background noise has been reduced. Here we present a method to quantify intra-condition variation and to remove reference-based transcripts associated with highly variable read counts, prior to differential expression analysis. The method utilizes variation within pairwise distances between normalized read counts for each transcript across all included samples of a given condition. As a case study, we demonstrate our approach at an inter and intra-study level using RNA-seq data from brain samples of dogs, wolves, and two strains of fox (aggressive and tame) prior to performing differential expression analysis to identify common genes associated with tame behaviour.

Results

By applying our method, the distribution of the gene-wise dispersion estimates improved and the number of outliers detected in differential expression analysis decreased. Several genes that initially were differentially expressed in the non-filtered datasets were removed due to high intra-condition variation. Additionally, by optimizing the detection of differentially expressed transcripts, the overall number increased between dogs vs wolves and tame vs aggressive foxes when compared to the non-filtered datasets. Using these filtered sets, we found common over expressed genes in dogs and tame foxes, including those involved in brain development, neurotransmission and immunity, factors known to be involved in domestication.

Conclusions

We presented a method to quantify and remove intra-condition variation from RNA-seq count data and demonstrate its usage in improving the distribution of gene-wise dispersion estimates and ultimately, reduce the number of false positives in differential gene expression analysis. We provide the method as a freely available tool, to aid studies using RNA-seq to calculate and characterize the variation present within data prior to perform differential expression analysis. Additionally, we identify candidate genes involved with selection for tameness, which seems to have played a crucial role in the canine domestication.

Background

The advance of RNA-seq technology [1] has revolutionized gene expression analysis by allowing a rapid hi-resolution view of gene expression under varying conditions, compartments or timepoints. In a typical RNA-seq experiment, gene expression profiles are estimated for each sample using a metric based upon

the number of reads associated with each transcript within a reference transcriptome. Expression profiles are compared to identify differentially expressed genes (DEGs) [2]. A challenge arises due to sources of variation within gene expression profiles that are independent of, or partially overlapping with, the condition of interest [3]. At an intra-study level, the inclusion of biological and technical replicates can be applied to reduce the effects of such noise [4].

At times, replicates may not always be possible due to cost or difficulty in obtaining samples. As an alternative, the incorporation of RNA-seq data from the rapidly growing repertoire of published works could complement the number of effective biological replicates associated with a given condition [5]. A hurdle is in accounting for the inherent variability of the data [6], which is amplified at an inter-study level as there is little control over the sample environments or experimental setups. For a given experiment, between two conditions, differential expression tools generally compute a p-value for each gene, based on the overall distribution of normalized read counts, that reflect the likelihood of that gene being differentially expressed. As intra-condition variation increases, the ability to decipher differential expression patterns decreases. Several methods have been proposed for data normalization and bias removal on RNA-seq data including, EDASeq [7], RUV2 [8], sva [3] and PEER [9]. However, when these methods are compared, highly variable results are observed, with the number of false positives of DEGs being increased [10]. No consensus exists on the best approach to apply.

Here, we provide a simple method for removing transcripts associated with high levels of intra-condition read count variation and thus results in a subset of the data that contains reduced noise. The method is to be applied to data prior to differential gene expression analysis, following the grouping of datasets into conditions of interest. We assume that, within a single condition, if a given transcript possesses large amounts of variation across normalized read count values, then an accurate expression pattern cannot be determined relative to the condition itself, independently of the source of variation. When this variation is large, it suggests that the condition-associated datasets are discordant with each other and thus comparisons to other conditions to identify DEGs will not yield meaningful results. The metric our method uses is the variation present within the pairwise differences of normalized read counts between datasets associated with a specific condition across each transcript of the reference set. Within a condition, the range of these values across all transcripts indicates the compatibility of the data for usage within a differential expression analysis.

As a case study, we demonstrate our approach at an inter and intra-study level using multiple published RNA-seq datasets from brain samples of dogs and wolves derived from several studies, and to brain samples of foxes from domestication experiments derived from the same study. Animal domestication has resulted in a shared set of common behavioural traits, including reduced fear of humans, diminished aggression and altered tendencies of exploration [11]. Despite the genetic basis underlying such behaviours being unclear, it has been shown that selection for tameness leads to common phenotypic changes in domesticated species, for example in rats [12], red foxes [13, 14] and red junglefowl [15]. Among canids, two illustrative examples of human-defined behavioural shifts have been studied through RNA-sEq. First, domestic dogs, that present marked behaviour differences from wolves, their wild

ancestors, have evolved unique social cognitive capabilities [16–18]. Second, a lineage of tame red foxes recently discovered to have originated in fur farms in Canada [19], which resulted from deliberated selection against fear and aggression over several generations of cross-breeding [20–22]. An inter-study comparison of expression patterns between RNA-seq data for wolves and dogs as well as for aggressive and tame foxes would provide insights on the gene dynamics involved in the evolution of tameness.

Within this study we initially explore the effects of removing transcripts associated with several levels of variation relative to inter and intra-study datasets and compare patterns of gene expression between i) wolves and dogs and ii) aggressive and tame foxes to uncover the genes associated with behavioural traits involved in both domestication events.

Results

Alignment of RNA-seq data

Available RNA-seq data from the brain of dogs, wolves, aggressive and tame foxes (Supplementary Table 1) was used. The alignment of the 44 samples against the dog reference transcriptome (26,107 transcripts) revealed a mapping success of 60% and 58% for dogs and wolves, respectively (Supplementary Fig. 1), as expected due to their recent divergence (~ 27 kya) [23]. A similar portion of reads failing to map against the transcriptome (~ 40%) has been previously reported for dog brain samples [24] and is most likely due to (i) novel genes (ii) regions that are not translated despite being transcribed, (iii) contamination with genomic DNA and (iv) uncharacterized chimeras within reference resulting from assembly errors. For the fox samples, an average of 50% of reads mapped to the dog reference transcriptome (Supplementary Fig. 1). The lower percentage of mapping for foxes was expected due to an increased genetic divergence to dogs (~ 10 mya) [25] together with the other aforementioned factors.

Intra-condition Variation

To remove individual transcripts associated with high amounts of variation prior to differential gene expression analysis, we developed a method that utilizes the intra-condition variation within pairwise differences of normalized read counts between samples for each transcript (Fig. 1; Methods section). The method handles raw count data which is then normalized considering the length of each transcript and the sum of all counts for the corresponding sample. We grouped samples across two contrasts, wolves (n = 6) vs dogs (n = 10) and aggressive vs tame foxes (n = 12 for each condition), according to the groups that were posteriorly used for differential gene expression analysis. Across all transcripts, the mean intra-condition variation observed between wolves and dogs was not significantly different (Wilcoxon-test, p-value < 0.198, Supplementary Fig. 2a), while between aggressive and tame foxes a significant difference was found (Wilcoxon-test, p-value < $2.2e^{-16}$, Supplementary Fig. 2a). In the latter, tame fox samples exhibited a higher number of transcripts associated with increased variability, mostly due to five samples

that differentiate from the remaining ones in the PCA plot of normalized count values (axis PC1 explained 80% of the variance, Supplementary Fig. 2b).

Based on the combined variance distribution for each of the contrasts, transcripts were removed from the reference according to a series of threshold values defined by the percentiles associated with the distribution of variation (e.g. variance score at the 95th percentile correspond to a removal of the 5% of most variable transcripts) (Fig. 2, Supplementary Table 2; Methods section). For each level, only the transcripts that pass the filter were maintained in the newly created output files of each sample, for downstream analysis. Initially, for wolves and dogs, 184 transcripts associated with the 99th percentile of variance and above were removed, while for the aggressive vs tame foxes, 235 transcripts were removed. Overall, the number of transcripts removed was higher among intra-study samples compared to samples combined from different studies, suggesting higher discordance between fox samples.

Differential Gene Expression Analysis

Levels of gene expression between conditions of each contrast (wolves vs dogs and aggressive vs tame foxes) were assessed using raw read count data from non-filtered and filtered data on DESeq2 [2]. To evaluate the effect of removing variation on dispersion estimates calculated by DESeq2, we used non-filtered and all the different filtered-level datasets, independently, to calculate the estimates of gene-wise dispersions and to perform a regression analysis over the mean of normalized counts. The regression analysis revealed a better fitting for the contrast wolves vs dogs (Figure 3), that displayed a high correlation ($r^2 > 0.7$) and a low deviation of the residuals (root mean square error – RMSE) around the line of best fit. By removing only 1% of the transcripts with high intra-condition variation, the correlation between both variables improved, the RMSE decreased and the number of outliers, recognize by DESeq2 as the points with extremely high dispersion values that cannot be shrunken towards the fit curve, has decreased (Supplementary Figure 3, Supplementary Table 3). The removal of the top 10% variable transcripts led to an increase of the r^2 to 0.82, to less 109 outliers and to a decrease in the number of transcripts with over-dispersion (variance > mean) (Figure 3). For the fox contrast, the linear regression did not fit so well the correlation between both variables ($r^2 = 0.49$), where an elevated number of transcripts presented over-dispersion (Figure 3). In this case, the shrinkage was more extensive (Supplementary Figure 3) to include more points scattered around the fit curve. Nevertheless, the same improvement in the r^2, RMSE, and the number of outliers, was observed after removing intra-condition variation (Figure 3, Supplementary Figure 3, Supplementary Table 3).

Regarding the differential gene expression analysis prior to filtering, 430 differentially expressed transcripts (DETs) were identified between wolves and dogs. Of those, 259 were over expressed in dogs while 171 were under expressed (Supplementary Table 4). Between aggressive and tame foxes, 651 DETs were observed, of which, 532 and 119 were over and under expressed, respectively (Supplementary Table 4). Post filtering, within the first ten steps of size one from the 99^{th} to the 90^{th} percentiles, the number of DETs identified, peaks at the 97^{th} (n=430; over=255, under=175) and the 95^{th} percentiles (n=730;

over=607, under=123) in dogs and tame foxes (Figure 4), respectively. These peaks suggest that the removal of the 3% (n= 854) and 5% (n=1940) of transcripts associated with the highest levels of intra-condition variation optimizes the detection of DETs. These filtered datasets were selected as inputs for the gene annotation.

Individual Gene Identification And Gene Families

Following the annotation of DETs within each contrast, using the non-filtered (Supplementary Table 5 and Supplementary Table 6) and filtered datasets at 3% and 5% thresholds (Supplementary Table 7 and Supplementary Table 8), six and 43 annotated genes originally over expressed in the non-filtered datasets of dogs and tame foxes, respectively, were removed in the filtered datasets due to high intra-condition variation. Similarly, seven different under expressed annotated genes were removed in dogs and tame foxes. Of those, in dogs, all seven genes were transcripts removed whereas in tame foxes, two were removed, while the other five, although kept in the reference, were no longer significantly differentially expressed ($p > 0.05$).

Between the filtered data, 21 gene families, containing 50 genes, were observed to be simultaneously over expressed within dogs and tame foxes (Table 1). Of these 50 genes, 19 were exclusive to dogs while 24 were exclusive to tame foxes. The remaining seven genes (RGR, CHRNA5, SQLE, ARHGAP25, ITGA7, MYO7A and TRIB2), belonging to seven different families, were common to both dogs and tame foxes. Additionally, three gene families, containing four genes, were found to be simultaneously under expressed (Table 2). Two of these genes (STMND1 and OASL) were shared between dogs and tame foxes while the other two were unique to each condition. The same analysis performed on the non-filtered datasets revealed similar results (Supplementary Table 9), however, the RGR gene family, which included a shared gene between dogs and tame foxes, was lost.

Table 1

List of the gene families that were simultaneously over expressed in dogs (DG) and tame foxes (TF) between the filtered datasets. The number and the name of the genes that composed each family and the species they are present (shared or exclusively to dogs/tame foxes) are presented with the corresponding value of log2Fold change in brackets. When more than one variant for a specific gene was present, all the log2FC values were reported.

Gene Family	Group	Number of OE	Gene name and log2FC value
Retinal G protein-coupled receptor	**Shared**	1	**RGR** (2.10 in DG, 0.78 in TF)
Cholinergic receptor nicotinic alpha	**Shared**	1	**CHRNA5** (1.1 in DG, 0.4 in TF)
Squalene epoxidase	**Shared**	1	**SQLE** (0.54 in DG, 0.31 in TF)
Rho GTPase activating protein	**Shared**	1	**ARHGAP25** (0.86 in DG, 0.72 in TF)
	TF	2	ARHGAP4 (0.64); ARHGAP30 (0.57)
Integrin alpha subunits	DG	3	ITGA6 (1.25, 1.24); ITGA8 (1.14, 0.90); ITGAX (0.97)
	TF	1	ITGAL (0.73)
	Shared	1	**ITGA7** (0.76 in DG, 0.46 and 0.49 in TF)
Myosin	DG	1	MYO3A (1.12)
	TF	3	MYOZ1 (1.53); MYO1F (0.93); MYO1C (0.47)
	Shared	1	**MYO7A** (0.82 in DG; 0.41 in TF)
Tribbles pseudokinase	TF	2	TRIB1 (0.94); TRIB3 (0.78)
	Shared	1	**TRIB2** (0.61 in DG; 0.2 in TF)
EF hand calcium binding	DG	1	EFCAB1 (2.59)
	TF	1	EFCAB2 (0.46)
Transcription factor	DG	1	TCF23 (2.04)
	TF	1	TCF19 (0.63)
Adhesion G protein-coupled receptors	DG	1	ADGRG6 (1.45)
	TF	1	ADGRG1 (0.57)
Patatin Like Phospholipase Domain	DG	1	PNPLA4 (1.41)
	TF	1	PNPLA7 (0.59)
SRY-box	DG	1	SOX6 (1.26)
	TF	2	SOX17(0.84); SOX10 (0.66)

Gene Family	Group	Number of OE	Gene name and log2FC value
Hyaluronan and proteoglycan link protein	DG	1	HAPLN1 (1.15)
	TF	1	HAPLN3 (0.70)
Serine/threonine kinase	DG	2	STK17A (1.15, 1.14); STK32A (1.10)
	TF	1	STK40 (0.57)
Potassium channels	DG	1	KCTD16 (0.98)
	TF	1	KCTD15 (0.72)
Podocalyxin like	DG	1	PODXL (0.95, 0.84)
	TF	1	PODXL2 (0.70, 0.69, 0.67)
ATP binding cassette subfamily B	DG	1	ABCB1 (0.93)
	TF	1	ABCB9 (0.52)
Zinc finger DHHC-type	DG	1	ZDHHC15 (0.75)
	TF	1	ZDHHC1 (0.70)
Sushi domain	DG	1	SUSD1 (0.68)
	TF	2	SUSD3 (0.79); SUSD6 (0.47)
TBC1 domain family	DG	1	TBC1D5 (0.54)
	TF	1	TBC1D7 (0.27)
Mitogen-activated protein kinase kinase kinases	DG	1	MAP3K5 (0.51)
	TF	1	MAP3K11 (0.76)

Table 2
List of the gene families that were simultaneously under expressed in dogs (DG) and in tame foxes (TF) between the filtered datasets. The number and the name of the genes that composed each family and the species they are present (shared or exclusively to dogs/tame foxes) are presented with the corresponding value of log2Fold change in brackets. When more than one variant for a specific gene was present, all the log2FC values were reported.

Gene Family	Group	Number of UE	Gene name and log2FC value
Stathmin domain	**Shared**	1	**STMND1** (-1.18 in DG, -0.53 in TF)
Oligoadenylate synthetase like	**Shared**	1	**OASL** (-0.41 in DG, -0.52 in TF)
Heat shock protein family B	DG	1	HSPB8 (-0.70)
	TF	1	HSPB11 (-0.32)

Discussion

Despite the existence of several published RNA-seq datasets, the quantification and control of variation present between replicates of a given condition are often overlooked. Here we developed a method to quantify intra-condition variation and to remove reference-based transcripts associated with highly variable read counts, prior to differential expression analysis. By applying our method, we were able to reduce the level of noise that standard differential expression tools are required to accommodate when determining differential expression patterns.

Following the application of our method to intra and inter-study datasets derived from brain samples of dogs, wolves, and foxes, we observed an improvement in the distribution of the gene-wise dispersion estimates used by DESeq2 to determine DEGs, by removing 1% of the transcripts that displayed high intra-condition variation. The correlation between the mean of normalized counts and dispersion estimates per gene improved when intra-condition variation was removed. This is a consequence of reducing the number of transcripts that displayed large differences in variance for the respective mean of normalized counts, or, in other words, have high overdispersion. By removing those from the reference, prior to usage within the differential expression software, the gene-wise estimates of dispersion will be more accurate since they are calculated based on information across all genes by assuming that genes with similar expression levels have similar dispersion [2]. Surprisingly, samples from the fox intra-study case contained a higher number of transcripts associated with high amounts of intra-condition variation. This high discordance between counts of samples from the same condition has resulted in a more spread distribution of dispersion estimates around the fit curve which affected the strength of the shrinkage method. When the detection of outliers is dependent on how far a certain point is from the fit curve, having points that are highly spread will affect the detection of outliers and increase the number of false positives. This may explain the low number of outliers detected in the fox data, even having the highest variation. Our method proves effective to improve the distribution of dispersion estimates prior to differential expression analysis and although we have used DESeq2, its utility covers other methods for differential expression analysis, since they also rely on share information across genes for dispersion estimation (e.g. edgeR [26], BBSeq [27], DSS [28], baySeq [29] and ShrinkBayes [30]).

It was observed that removing the 3% and 5% of the transcripts associated with the highest levels of variation maximized the number of DETs between wolves vs dogs and aggressive vs tame foxes, respectively. More importantly, several genes that were over expressed in dogs and tame foxes in the non-filtered datasets, some at the top of the list, were removed after filtering. This is concerning because it reveals how dependent the final list of DEGs is on the accuracy of gene-wise dispersion estimates. When we compared the results of differential expression analysis obtained in the intra-study case with the original publication [21] we found most of the genes (92%) that they have identified as differentially expressed and verified that some of those genes disappeared after filtering.

Amongst the 50 over expressed genes identified across the 21 shared gene families, seven genes were shared between dogs and tame foxes. Up until now, almost no gene overlap has been observed in gene

expression analyses involving domesticated or tame animals, across three different mammal families (Canidae, Suidae, and Leporidae) [18]. Of the seven shared genes, three main functions related to brain development, neurotransmission, and immune response were identified. These functions have been repeatedly associated with behaviour selection during domestication by different approaches, such as QTL analysis [12, 31, 32], whole-genome sequencing [33–35] and RNA data both using microarrays and RNA-seq [16, 17, 20, 21, 36, 37]. Of these genes, the shared gene ITGA7 belongs to a gene family that is known to play an essential role in the control of neuronal connectivity [38] and the inflammatory response [39]. Other genes from this family, for example, ITGA8, have been previously observed to be over expressed in tame foxes [20], and here we also observed its over expression in dogs. Thus, our novel analysis provides further evidence of the family role in tameness. Similar functions are associated with the shared genes CHRNA5 [40, 41] and TRIB2 [42] from the cholinergic and tribbles family, respectively. Also, we found a shared gene involved in sensing local environmental stimuli, the MYO7A, whose mutation results in loss of hearing and vision [43]. Amongst the three gene families identified as under expressed, we found the shared gene STMND1, which deficiency in the amygdala of mice was connected to a deficiency in innate and learned fear [44], a behaviour that might also have played an important role in domestication.

Conclusions

In the present study, we have presented a method to quantify and remove intra-condition variation from RNA-seq count data and demonstrate its usage in improving the distribution of gene-wise dispersion estimates and ultimately, reduce the number of false positives in differential gene expression analysis. Studies using RNA-seq data should characterize and discuss the variation present in their data prior to performing differential expression analysis, independently of the software, to reduce the number of genes that appear, erroneously, as differentially expressed. Additionally, we demonstrated the use of our method in the identification of candidate genes involved with selection for tameness, which seems to have played a crucial role in the canine domestication.

Methods

RNA-seq data

Available RNA-seq data from the brain of dogs, wolves, and foxes (Supplementary Table 1) was downloaded from the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/) and the European Bioinformatics Institute (EMBL-EBI, https://www.ebi.ac.uk/). A total of 44 samples belonging to five studies [18, 21, 24, 45, 46] were used. We accounted for the available details of all samples including the relative location of the tissue, age, and sex of animals, replicate information and sequencing details (Supplementary Table 1). The sample codes will be referred to throughout the rest of this manuscript.

Read Mapping And Normalization

Reads from each of the 44 samples were mapped to the dog reference transcriptome [46], containing 26,107 annotated contigs (Ensembl CanFam3.1, release 92) using Bowtie2.3.4.1 [47]. The percentage of reads mapped for each sample was calculated and used as an indicator of mapping success. For each sample, read counts for each transcript were obtained using BBMap [48] and the output files were used as input in the differential expression software. For our filtering method, normalization was performed by dividing the count values by the length of the transcript they were associated with and by the sum of all counts within the file for a given sample.

Removing Intra-condition Variation

The method we developed removes individual transcripts associated with high amounts of intra-condition variation within the normalized read counts across samples prior to differential expression analysis (Fig. 1). The method utilizes the intra-condition variation within pairwise differences of normalized read counts between samples for each transcript. Samples were grouped across two contrasts, each with two conditions: wolves vs dogs (wolves n = 6 and dogs n = 10) and aggressive vs tame foxes (n = 12 for each condition). Read counts from technical replicates of “Dog_8” and “Dog_9” were averaged and merged into one file, while read counts from the two biological replicates of “Dog_7” were treated separately.

The input for the method is a set of files, each containing the raw counts that are associated with each transcript of the reference. For each contrast the method implements the following steps:

A. Preprocessing:

Allocate each input dataset of read counts to either condition A or condition B and normalize read counts.

B. Calculating intra-condition variation:

(i) For each transcript, calculate the absolute pairwise differences between normalized read counts across all samples within condition A. From these values calculate the corresponding variance (σ_{An}) where n represents the transcript number. Following this step, the number of variance values are equal to the number of transcripts within the reference.

(ii) Repeat for condition B to obtain values for σ_{Bn}.

C. Filtering:

(i) All the variance scores are placed in ascending order into a single file, regardless of condition.

(ii) Calculate the corresponding percentile values across the full distribution of variance values.

(iii) Use percentiles as thresholds to identify the amount of variation to be removed. For example, use the variance score at the 95^{th} percentile to remove the 5% of transcripts associated with the largest amount of intra-condition variation.

(iv) For each transcript, if either σ_{An} or σ_{Bn} is greater than the threshold, remove that transcript and all associated counts from the input data.

(v) Output the raw read counts associated with the remaining transcripts to a separate file for each dataset.

For each contrast, variance thresholds ranging from the 70th to the 90th percentiles (in steps of five), and from the 91th to the 99th (in steps of one) were explored. Steps of one were used in the latter in order to explore this range containing the transcripts associated with the highest levels of intra-condition variation in more detail.

The method was written in Java and is freely available on the web at: https://sourceforge.net/projects/tvscript/.

Differential gene expression analysis

To perform the differential gene expression analysis we used DESeq2 [2] since it was shown to be one of the most consistent methods to identify DEGs [49]. DESeq2 estimates the gene-wise dispersions and shrink these estimates to generate more accurate estimations of dispersions to model the counts. The dispersions are inversely related to the mean since lower mean counts are more affected by within replicates variation. For each contrast, using as input the non-filtered and all the different filtered-level datasets, independently, estimates of dispersion were calculated and used in a linear regression analysis in relation to the mean of normalized counts using R [50]. The number of DETs between conditions of each contrast (wolves vs dogs and aggressive vs tame foxes) was assessed prior to and post-filtering stages. A PCA was performed, using the *plotPCA* function from DESeq2 with non-filtered normalized count values, to visualize the overall effects of experimental covariates in each contrast.

Gene Annotation And Gene Family Analysis

Independently for each contrast, DETs obtained using the non-filtered and filtered datasets, were annotated to the correspondent gene ID using the R package BioMart [51] against the Ensembl Gene database (version 94). Over expressed genes in dogs and tame foxes were classified into gene families. Families containing genes from both dogs and tame foxes were selected for further analysis given their potential for being involved in the evolution of tame behaviour. Genes within gene families were grouped according to whether they were unique to either dogs or tame foxes or shared between the two. Under expressed genes were treated in the same manner. For a given contrast, a family was only maintained if

all the associated genes agreed in relation to their direction of differential expression (i.e. all genes were either over or under expressed).

Abbreviations

DEGs
differentially expressed genes
DETs
differentially expressed transcripts
Kya
thousand years ago
Mya
million years ago
PCA
principal component analysis
RMSE
root mean square error
RNA-seq
RNA sequencing

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: The datasets analyzed during the current study were already publicly available. Full information about all data sources used in this study is described in supplementary Table 1. The code generated during the current study is freely available at: https://sourceforge.net/projects/tvscript/ under the tab “Code”. A full description on how to implement the method can be found under the tab “Wiki”.

Competing interests: The authors declare that they have no competing interests.

Funding: This work was supported by the Portuguese Foundation for Science and Technology, FCT, projects PTDC/BIA-EVF/2460/2014 and PTDC/BIA-EVL/29115/2017. DL, RG were supported by FCT (PD/BD/132403/2017 to DL, contract under DL57/2016 to RG) and JA was supported by FEDER funds through the Operational Programme for Competitiveness Factors - COMPETE (POCI-01-0145-FEDER-029115). Funding entities played no additional role in the design of the study, analysis, interpretation of data, nor in writing the manuscript.

Authors' contributions: DL, RG and JA designed the study; JA and DL conceived and designed methodology. DL and JA analysed the data; DL, RG and JA wrote the manuscript. All authors gave final approval for publication.

Acknowledgements: Not applicable

References

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.
2. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15.
3. Hansen KD, Wu Z, Irizarry RA, Leek JT. Sequencing technology does not eliminate biological variability. Nat Biotechnol. 2011;29:572–3.
4. Liu Y, Zhou J, White KP. RNA-seq differential expression studies: More sequence or more replication? Bioinformatics. 2014;30:301–4.
5. Rau A, Marot G, Jaffrézic F. Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics. 2014;15.
6. McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, et al. RNA-seq: Technical variability and sampling. BMC Genomics. 2011;12.
7. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-Content Normalization for RNA-Seq Data. BMC Bioinformatics. 2011;12.
8. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012;13:539–52.
9. Stegle O, Parts L, Durbin R, Winn J. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol. 2010;6(5):e1000770.
10. Li S, Labaj PP, Zumbo P, Sykacek P, Shi W, Shi L, et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat Biotechnol. 2014;32:888–95.
11. Wiener P, Wilkinson S. Deciphering the genetic basis of animal domestication. Proc R Soc B Biol Sci. 2011;278:3161–70.
12. Albert FW, Carlborg Ö, Plyusnina I, Besnier F, Hedwig D, Lautenschläger S, et al. Genetic architecture of tameness in a rat model of animal domestication. Genetics. 2009;182:541–54.
13. Trut L. Early Canid Domestication: The Farm-Fox Experiment. Am Sci. 1999;87:160.
14. Trut L, Oskina I, Kharlamova A. Animal evolution during domestication: The domesticated fox as a model. BioEssays. 2009.
15. Bélteky J, Agnvall B, Johnsson M, Wright D, Jensen P. Domestication and tameness: Brain gene expression in red junglefowl selected for less fear of humans suggests effects on reproduction and

immunology. R Soc Open Sci. 2016;3:160033.

16. Li Y, Wang GD, Wang MS, Irwin DM, Wu DD, Zhang YP. Domestication of the dog from the Wolf was promoted by enhanced excitatory synaptic plasticity: A hypothesis. Genome Biol Evol. 2014;6:3115–21.
17. Li Y, Von Holdt BM, Reynolds A, Boyko AR, Wayne RK, Wu DD, et al. Artificial selection on brain-expressed genes during the domestication of dog. Mol Biol Evol. 2013;30:1867–76.
18. Albert FW, Somel M, Carneiro M, Aximu-Petri A, Halbwax M, Thalmann O, et al A Comparison of Brain Gene Expression Levels in Domesticated and Wild Animals. Akey JM, editor. PLoS Genet. 2012;8: e1002962.
19. Lord KA, Larson G, Coppinger RP, Karlsson EK. The History of Farm Foxes Undermines the Animal Domestication Syndrome. Trends Ecol Evol. 2019;35:125–36.
20. Kukekova A, Johnson J, Teiling C, Li L, Oskina I, Kharlamova A, et al. Sequence comparison of prefrontal cortical brain transcriptome from a tame and an aggressive silver fox (Vulpes vulpes). BMC Genomics. 2011;12.
21. Wang X, Pipes L, Trut L, Herbeck Y, Vladimirova A, Gulevich R, et al. Genomic responses to selection for tame/aggressive behaviors in the silver fox (Vulpes vulpes). Proc Natl Acad Sci. 2018;115:10398–403.
22. Hekman J, Johnson J, Edwards W, Vladimirova A, Gulevich R, Ford A, et al. Anterior Pituitary Transcriptome Suggests Differences in ACTH Release in Tame and Aggressive Foxes. G3; Genes|Genomes|Genetics. 2018;8:859–73.
23. Skoglund P, Ersmark E, Palkopoulou E, Dalén L. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. Curr Biol. 2015;25:1515–9.
24. Roy M, Kim N, Kim K, Chung WH, Achawanantakun R, Sun Y, et al. Analysis of the canine brain transcriptome with an emphasis on the hypothalamus and cerebral cortex. Mamm Genome. 2013;24:484–99.
25. Wayne RK, Geffen E, Girman DJ, Koepfli K-P, Lau LM, Marshall CR. Molecular Systematics of the Canidae. Syst Biol. 1997;46:622–53.
26. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
27. Zhou YH, Xia K, Wright FA. A powerful and flexible approach to the analysis of RNA sequence count data. Bioinformatics. 2011;27:2672–8.
28. Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics. 2013;14:232–43.
29. Hardcastle TJ, Kelly KA. BaySeq. Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11.
30. Van De Wiel MA, Leday GGR, Pardo L, Rue H, Van Der Vaart AW, Van Wieringen WN. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics.

2013;14:113–28.

31. Kukekova A, Trut L, Chase K, Kharlamova A, Johnson J, Temnykh S, et al. Mapping loci for fox domestication: Deconstruction/Reconstruction of a behavioral phenotype. Behav Genet. 2011;41:593–606.
32. Wirén A, Wright D, Jensen P. Domestication-related variation in social preferences in chickens is affected by genotype on a growth QTL. Genes, Brain Behav. 2013;12:330–7.
33. Carneiro M, Rubin CJ, Palma F, Di, Albert FW, Alföldi J, Barrio AM, et al. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. Science. 2014;345:1074–9.
34. Freedman AH, Schweizer RM, Ortega-Del Vecchyo D, Han E, Davis BW, Gronau I, et al. Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs. PLoS Genet. 2016;12:e1005851.
35. Kukekova A, Johnson J, Xiang X-Y, Feng S-H, Liu S, Rando H, et al. The red fox genome assembly identifies genomic regions associated with tame and aggressive behaviors. Nat Ecol Evol Springer US. 2018;2:1479–91.
36. Saetre P, Lindberg J, Ellegren H, Vila C, Jazin E, Leonard JA, et al. From wild wolf to domestic dog: Gene expression changes in the brain. Mol Brain Res. 2004;126:198–206.
37. Heyne HO, Lautenschläger S, Nelson R, Besnier F, Rotival M, Cagan A, et al. Genetic influences on brain gene expression in rats selected for tameness and aggression. Genetics. 2014;198:1277–90.
38. Lilja J, Ivaska J. Integrin activity in neuronal connectivity. J Cell Sci. 2018;131.
39. González-Amaro R, Sánchez-Madrid F. Cell adhesion molecules: selectins and integrins. Crit Rev Immunol. 1999;19:389–429.
40. Winterer G, Mittelstrass K, Giegling I, Lamina C, Fehr C, Brenner H, et al. Risk gene variants for nicotine dependence in the CHRNA5-CHRNA3-CHRNB4 cluster are associated with cognitive performance. Am J Med Genet Part B Neuropsychiatr Genet. 2010;153:1448–58.
41. Zhang H, Kranzler HR, Poling J, Gruen JR, Gelernter J. Cognitive flexibility is associated with KIBRA variant and modulated by recent tobacco use. Neuropsychopharmacology. 2009;34:2508–16.
42. Eyers PA, Keeshan K, Kannan N. Tribbles in the 21st Century: The Evolving Roles of Tribbles Pseudokinases in Biology and Disease. Trends Cell Biol. 2017;27:284–98.
43. Miller KA, Williams LH, Rose E, Kuiper M, Dahl HHM, Manji SSM. Inner Ear Morphology Is Perturbed in Two Novel Mouse Models of Recessive Deafness. PLoS One. 2012;7:e51284.
44. Martel G, Nishi A, Shumyatsky GP. Stathmin reveals dissociable roles of the basolateral amygdala in parental and social behaviors. Proc Natl Acad Sci U S A. 2008;105:14620–5.
45. Fushan AA, Turanov AA, Lee SG, Kim EB, Lobanov AV, Yim SH, et al. Gene expression defines natural changes in mammalian lifespan. Aging Cell. 2015;14:352–65.
46. Hoeppner MP, Lundquist A, Pirun M, Meadows JRS, Zamani N, Johnson J, et al. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. PLoS One. 2014;9:91172.

47. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
48. Bushnell B. BBMap: A Fast, Accurate, Splice-Aware Aligner. Conf 9th Annu Genomics Energy Environ Meet. 2014.
49. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. PLoS One. 2017;12:e0190152.
50. R Development Core Team. R: A language and environment for statistical computing. Austria: Vienna; 2017.
51. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21:3439–40.

Figures

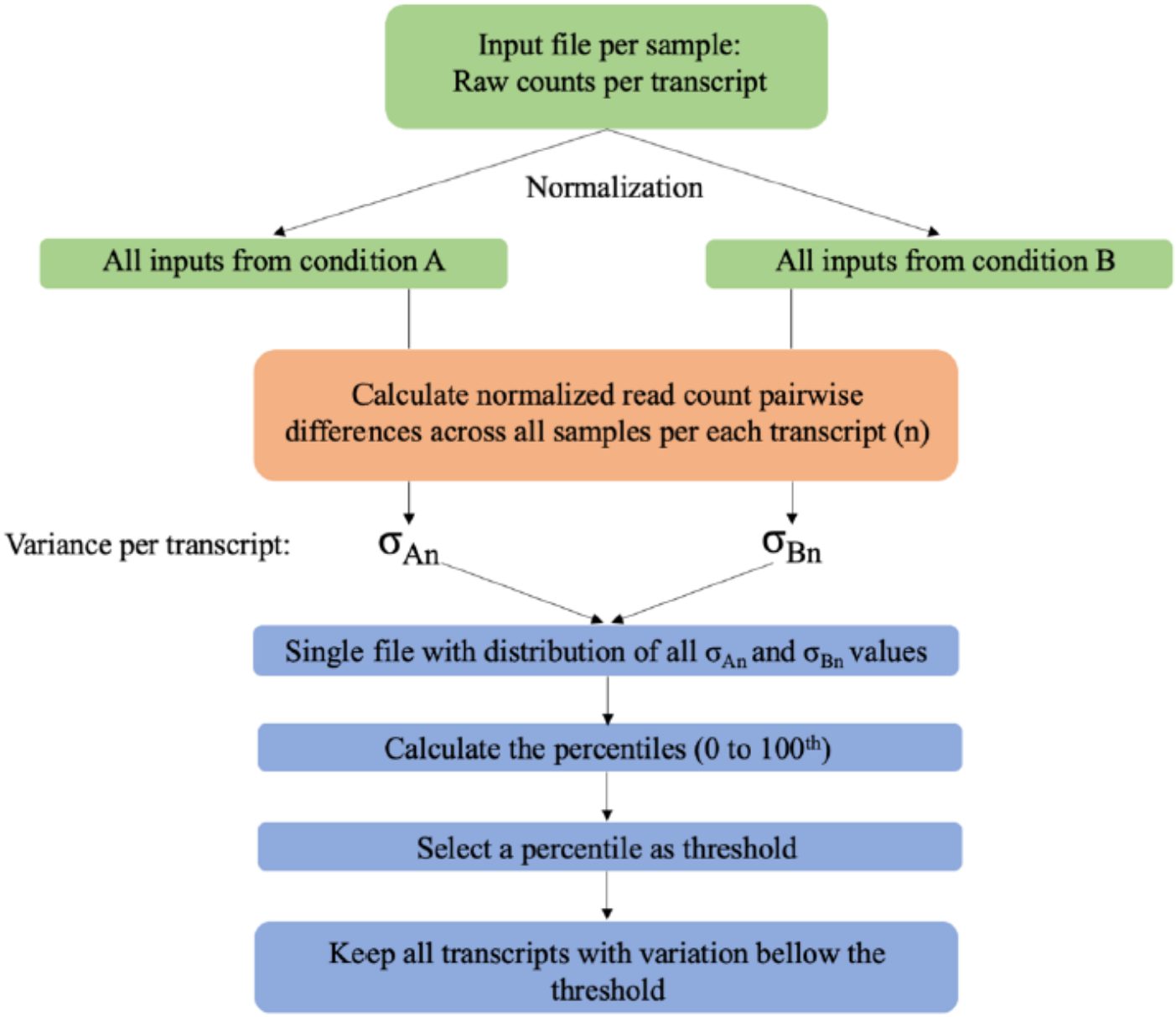

Figure 1

The method developed to filter individual transcripts associated with high amounts of intra-condition variation within expression profiles across samples of both conditions for each contrast.

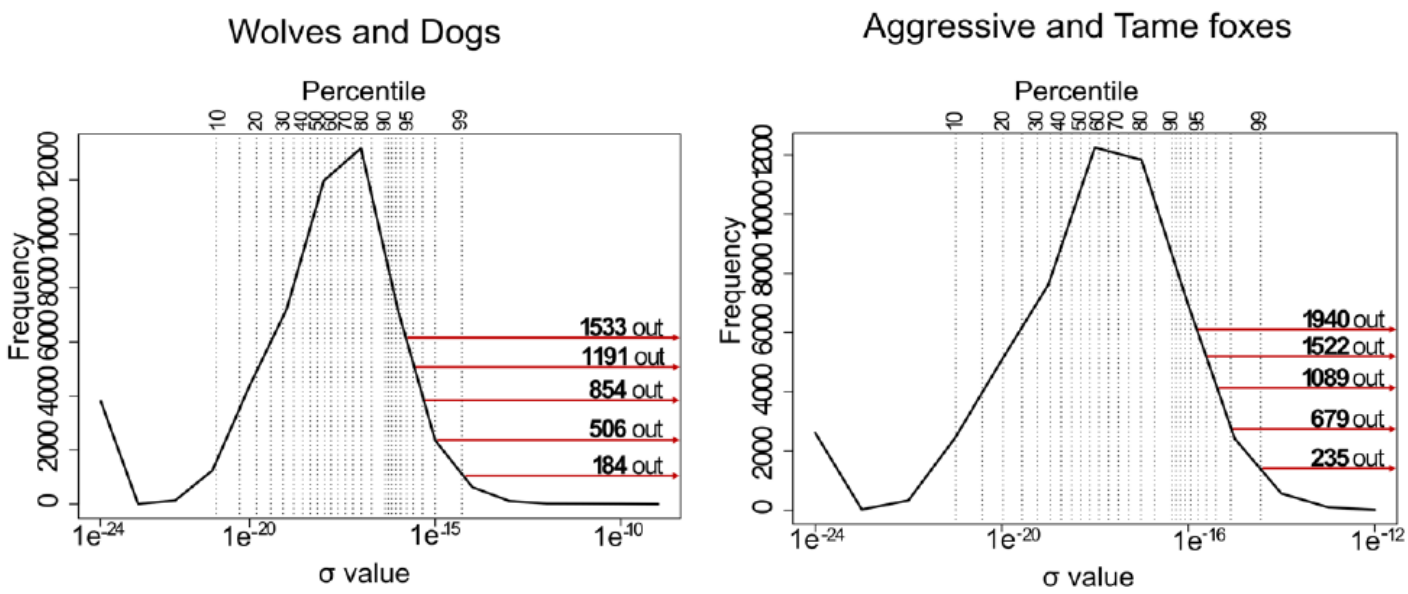

Figure 2

Histograms of the distribution of the variance (σ) values (x-axis) and their absolute frequency (y-axis) present in each contrast: wolves vs dogs and aggressive vs tame foxes. Dashed lines indicate the position of the percentiles (thresholds). Percentiles between 99th and 70th were explored for filtering and red arrows indicate the number of transcripts removed in the first five filtering steps.

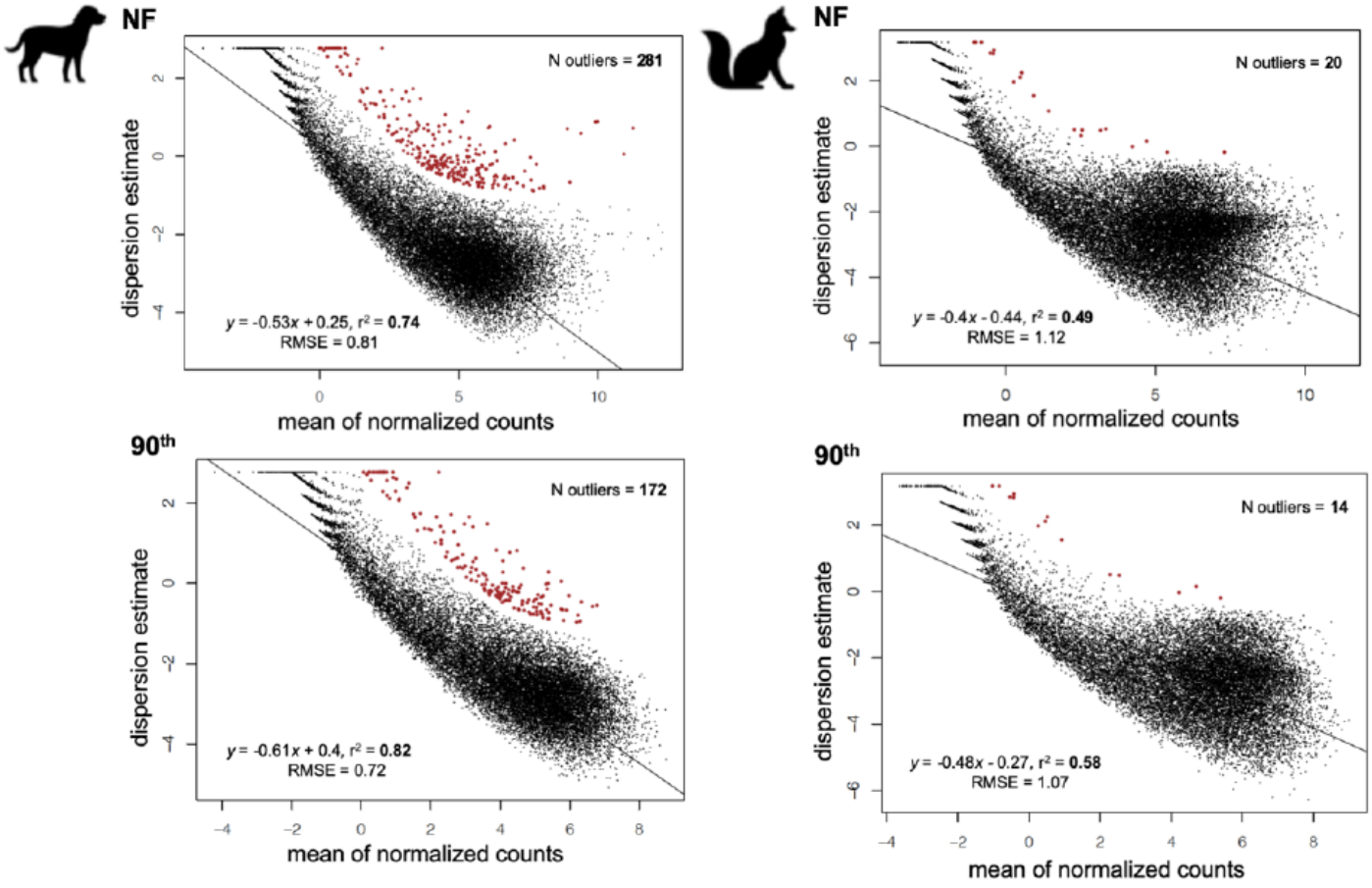

Figure 3

Plots of final dispersion estimates over the average expression length for wolves vs dogs (left) and aggressive vs tame foxes (right) contrasts calculated using DESeq2 for the non-filtered (NF) and 10% filtered (90th) datasets. Each black dot represents one transcript and red dots represent outliers. Values for the equation of the regression line, r-square and root mean square error (RMSE) are presented at the bottom of each graph. Both x and y-axis were transformed into a logarithm scale.

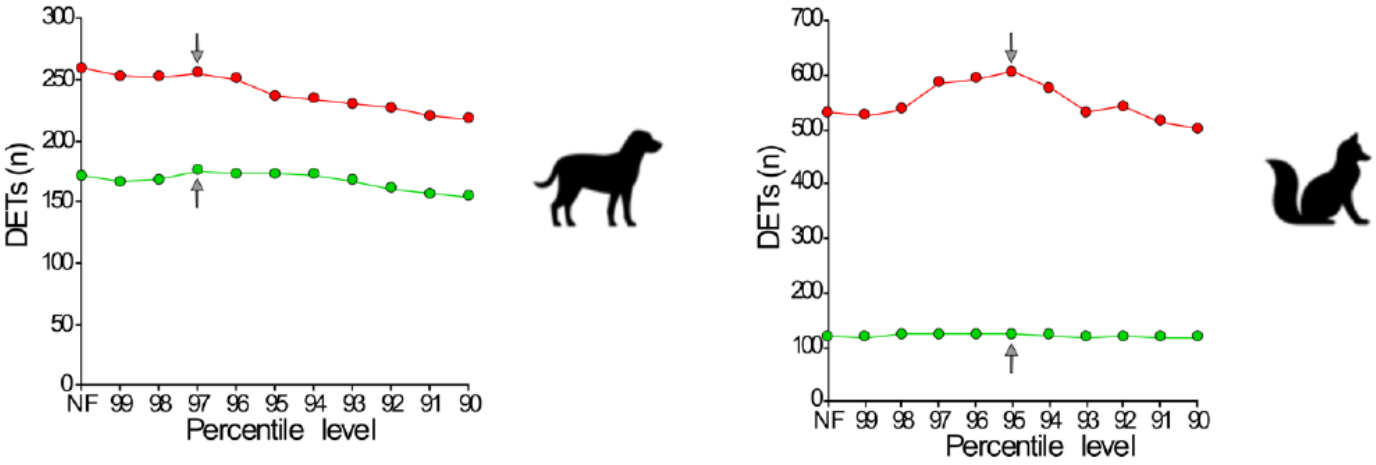

Figure 4

The number of differentially expressed transcripts (DETs) (y-axis) identified using non-filtered and filtered datasets based on the first 10 percentiles of the variance distribution (x-axis) in dogs (left) and in tame foxes (right). Over and under expressed genes are represented by red and green dots, respectively, and gray arrows represent the selected threshold for each contrast.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.pdf