TVScript: A Tool for Exploring the Effects of Intra-Condition Variation on the Detection of Differentially Expressed Genes

Diana Lobo
CIBIO/InBIO, Universidade do Porto
Raquel Godinho
CIBIO/InBIO, Universidade do Porto
John Archer (john.archer@cibio.up.pt)
CIBIO/InBIO, Universidade do Porto

Research Article

Keywords: RNA-seq technology, TVScript, intra-condition variation, transcriptome evolution

Posted Date: April 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-468506/v1

Abstract

The evolution of RNA-Seq technologies yielded datasets that are of immense scientific value. Commonly, such data is generated within differential expression studies, where datasets derived from individual samples are grouped into conditions, and gene expression patterns quantified. The number of archived datasets is increasing and revisiting many at an inter-study level provides an in-depth view into transcriptome evolution. The biggest hurdle is in dealing with variation of read counts at an individual transcript level between common conditions. We present a tool, TVScript, that quantifies intra-condition variation, and subsequently, removes reference-based transcripts that are associated with high levels of this. TVScript is demonstrated at inter and intra-study levels, using data from brain samples of dogs, wolves and foxes (aggressive and tame), where a marked improvement in the distribution of the gene-wise dispersion estimates, the metric utilized by the majority of differential expression tools, lowered the number of outliers detected. We provide support for seven candidate genes with potential for being involved with selection for tameness, and that appear to play a crucial role in canine domestication. We also identify several genes previously identified as being differentially expressed, but that possessed high intra-condition variation, weakening their relevance. TVScript is available at: https://sourceforge.net/projects/tvscript/.

Introduction

Developments in RNA-seq technology [1] have revolutionized transcriptomics studies by allowing for a rapid hi-resolution view of gene expression under varying conditions, compartments, and/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 set. Expression profiles are then compared between conditions to identify differentially expressed genes [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 can 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 [6,7], which is amplified at an inter-study level, as there is little control over the sample environments or experimental setups.

Differential expression tools generally compute a p-value for each gene that is based on the overall distribution of normalized read counts and that reflects the possibility 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 within RNA-seq data including, EDASeq [8], RUV2 [9], sva [3], and PEER [10]. However, when these methods are compared, highly variable results are observed [11] and no consensus exists on the best approach to apply. In this study, we

present TVScript, a tool for the identification and removal of transcripts associated with high levels of intra-condition read count variation. Our assumption is that, within a single condition, if a given transcript possesses large amounts of variation across normalized read count values, then an accurate expression pattern for that transcript cannot be determined relative to the condition itself, regardless of what statistical correction is applied. For such transcripts, this variation suggests that the condition-associated datasets are discordant with each other and thus, comparisons to other conditions in order to identify differentially expressed genes can only yield spurious results and, at a minimum, they should be highlighted in downstream analysis, if not completely removed. The metric TVScript uses to identify a given transcript displaying intra-condition variation, is the comparison of the variation present within the pairwise differences of normalized read counts for that transcript across datasets associated with the condition, to that obtained for each transcript of the reference within each of the two conditions being compared. Additionally, within a single condition, the range of these values is an intuitive indication of the compatibility of the data for usage within a differential expression analysis.

We demonstrate the usefulness of TVScript at an inter-study level using multiple published RNA-seq datasets from brain samples of dogs and wolves, and at an intra-study level to brain samples of foxes from domestication experiments. Domestic dogs present marked behaviour differences from wolves, their wild ancestors, due to the evolution of unique social cognitive capabilities [12–14], whilst a lineage of tame red foxes has been recently discovered to have originated in fur farms in Canada [15], which resulted from deliberated selection against fear and aggression over several generations of cross-breeding [16–18]. We used TVScript, in conjunction with DESeq2, to explore patterns of gene expression between i) wolves and dogs and ii) aggressive and tame foxes and highlight genes associated with behavioural traits involved in both domestication events. DESeq2 has been previously demonstrated to be consistent in identifying differentially expressed genes by estimating gene-wise dispersions and shrinking these estimates to generate more accurate estimations of dispersions to model the counts [19]. The results of our analysis provides support for seven candidate genes with potential for being involved with selection for tameness, and that appear to play a crucial role in canine domestication. Additionally, we identify several genes that were previously identified as being differentially expressed, but that possessed high intra-condition variation, thus highlighting the need to discuss such variation when presenting the results of differential expression experiments. The software, along with usage details and sample data, is available at: https://sourceforge.net/projects/tvscript/.

Materials And Methods

RNA-seq datasets, mapping and conditions

Available RNA-seq datasets from the brain of dogs, wolves, and foxes were downloaded from the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EMBL-EBI) (see Supplementary Table S1 online). A total of 44 samples belonging to five studies [14,17,20–22] were used. Within supplementary table S1 all available details for all samples, including the relative location of

the tissue, age, and sex of animals, replicate information, and sequencing details are accounted for. The sample codes will be referred to throughout the rest of this manuscript. Reads from each of the 44 samples were mapped to the dog reference transcriptome [22], containing 26,107 annotated transcripts (Ensembl CanFam3.1, release 92) using Bowtie2.3.4.1 [23]. 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 [24]. These count files were then 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.

Software

The steps that TVScript implements to identify transcripts associated high levels of intra-condition variation are: (i) Each input dataset, containing count values representing a particular sample, is allocated to either condition A or B, as indicated within the configuration file; (ii) Counts are normalized by dividing them by the length of the reference transcript they are associated with and by the sum of all counts within the file for a given sample; (iii) For each reference transcript (t), the absolute pairwise differences between normalized read counts across all samples within condition A are calculated; (iv) The corresponding variance (σ_{At}) within these pairwise distances is then also calculated; (v) Steps (iii) and (iv) are repeated for condition B to obtain values for each σ_{Bt}; (vi) Variance scores from each condition are placed in ascending order to associate them with corresponding percentiles; (vii) Reference transcripts are removed based on the variance associated with these percentiles, provided as user input within the configuration file. Note: The user can obtain these variance values by pre-running software on the input datasets with the − 1 parameter described in the readme file. The software will then output the full intra-condition pairwise distance variance distribution and the associated percentile for each value; (viii) Raw read counts associated with the remaining transcripts are then outputted into separate files that correspond to each input dataset. The names of output files are specified in the configuration file. These modified count files are the subsequent input files for the differential expression tool DESeq2.

Filtering transcripts and differential gene expression

For each contrast, TVScript was run using variance thresholds ranging from the 70th to the 90th percentiles (in steps of five), and from the 91st to the 99th (in steps of one). Steps of one were used in the latter in order to explore this range containing the minority transcripts associated with the highest levels of intra-condition variation in more detail. For each level, only read counts associated with transcripts that passed the filter level specified on the configuration file were maintained. These were outputted into individual modified read count files, one corresponding to each of the original inputs. DESeq2 was then used to perform differential gene expression analysis using these count files as input [2]. The gene-wise dispersions estimations calculated by DESeq2 are inversely related to the mean since lower mean counts are affected by variation to a higher degree. For each contrast involving the non-filtered and each level of

filtered data, estimates of dispersion were calculated and used in linear regression analysis in relation to the mean of normalized count values using the R statistical programming package [25]. The number of differentially expressed transcripts between conditions of each contrast (wolves vs. dogs and aggressive vs. tame foxes) was assessed prior to and post-filtering increments using a p-adj < 0.05 threshold for significance. A principal component analysis (PCA) was performed, using the *plotPCA* function from DESeq2 with non-filtered normalized count values, to visualize the overall effects of experimental covariates.

Gene annotation and gene family analysis

Independently for each contrast, differentially expressed transcripts obtained using the non-filtered and filtered datasets were annotated to the correspondent gene ID, using the R package BioMart [26] 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 both. Under expressed genes were treated in the same manner. For a given contrast, a gene family was only maintained if all the associated genes agreed in relation to their direction of differential expression.

Results

Mapping

Mapping of the 44 samples against the dog reference transcriptome revealed a success of 60% and 58% for dogs and wolves, respectively (see Supplementary Figure S1 online), as expected due to their recent divergence (~ 27 kya) [27]. Similar portions of reads failing to map (~ 40%) have been previously reported for dog brain samples [20] and are 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 [28]. For the fox datasets, an average of 50% of reads mapped to the dog reference transcriptome (see Supplementary Figure S1 online). The lower percentage of mapping for foxes was expected due to an increased genetic divergence to dogs (~ 10 mya) [29] together with the other aforementioned factors.

Software

TVScript has been written in the Java programming language and runs on all operating systems with installed Java Runtime Environment 8.0 or higher. The input for the software is a set of count files that are each associated with a specific sample. Within each count file, counts represent the number of reads that map to each transcript of the reference set. Along with these count files, a configuration file, describing how samples should be allocated into one of two conditions, is required. The output is a corresponding set of count files, but with the counts associated with the most variable transcripts

removed across all, and that can be directly used by differential expression analysis tools such as DESeq2 [2] with no further modification. The software, along with source code usage details and sample data, is available at: https://sourceforge.net/projects/tvscript/.

Intra-condition variation

Across all transcripts prior to filtering with TVScript, the mean intra-condition variation observed between wolves and dogs was not significantly different (Wilcoxon-test, p-value < 0.198, Fig. 1a and b), while between aggressive and tame foxes a significant difference was observed (Wilcoxon-test, p-value < $2.2e^{-16}$, Fig. 1c). In the latter, tame fox samples exhibited a higher number of transcripts associated with increased variability, likely due to five samples seen to differentiate from the remaining in the PCA plot of normalized count values (axis PC1 explained 80% of the variance; Fig. 1d). Using TVScript and based on the combined variance distribution for wolves and dogs, and separately, for aggressive and tame fox samples, transcripts were removed from the reference according to a series of threshold values (Fig. 2a and b, and Supplementary Table S2 online). 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

Prior to filtering, differential expression analysis yielded 430 differentially expressed genes between wolves and dogs. Of those, 259 were over expressed in dogs while 171 were under expressed (Table 1). Between aggressive and tame foxes, 651 differentially expressed genes were observed, of which, 532 and 119 were over and under expressed, respectively. Post filtering, within the first ten steps of size one from the 99th to the 90th percentiles, the number of differentially expressed genes identified, peaks at the 97th (n = 430; over = 255, under = 175) and the 95th percentiles (n = 730; over = 607, under = 123) in dogs and tame foxes (Fig. 3), respectively. These peaks suggest that, in the case of our datasets, 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 differentially expressed genes. These filtered datasets were selected as inputs for the gene annotation step. Following annotation, 49 over expressed genes in dogs (6 genes) and tame foxes (43 genes) within the non-filtered datasets (Supplementary Tables S3 and S4 online) were absent from the filtered datasets at 3% and 5% thresholds (Supplementary Tables S5 and S6 online). Similarly, seven under expressed genes present within the non-filtered data were absent from each of the under expressed sets in dogs and tame foxes within the filtered data.

Table 1
The number of the total differentially expressed transcripts (DETs) and the correspondent over (OE) and under (UE) expressed transcripts obtained for each contrast, wolves vs. dogs and aggressive vs. tame foxes, prior to (NF) and post-filtering steps, using DESeq2 ($p < 0.05$).

	Wolves vs Dogs			**Aggressive vs Tame Foxes**		
Percentile	DETs (n)	OE (n)	UE (n)	DETs (n)	OE (n)	UE (n)
NF	430	259	171	651	532	119
99	419	253	166	644	526	118
98	420	252	168	660	537	123
97	430	255	175	710	586	124
98	423	250	173	717	594	123
95	409	236	173	730	607	123
94	406	234	172	699	577	122
93	397	230	167	648	531	117
92	388	227	161	663	543	120
91	377	220	157	632	515	117
90	372	218	154	618	501	117
85	348	205	143	585	469	116
80	325	196	129	485	378	107
75	281	171	110	429	327	102
70	279	174	105	394	295	99

The regression analysis between the dispersion estimates over the mean of normalized counts, from the non-filtered data, revealed a better fitting for the contrast wolves vs. dogs (Fig. 4a), 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 shrunk towards the fit curve, has decreased (Supplementary Figure S2 and Table S7 online). Removal of the top 10% of 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) (Fig. 4c). For the fox contrast, the linear regression did not fit well ($r^2 = 0.49$), due to an elevated number of transcripts being dispersed (Fig. 4b). In this case, the shrinkage was more extensive (Supplementary Figure S2 online). Nevertheless, a similar increase in the

r^2, and decrease in the RMSE and the number of outliers, was observed after removing the 10% of transcripts associated with intra-condition variation (Fig. 4d, also Supplementary Figure S2 and Table S7 online).

Genes and gene families

Between the filtered data at 3% and 5% thresholds, 21 gene families, containing 50 genes, were observed to be simultaneously over expressed within dogs and tame foxes (Table 2). 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 3). 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 S8 online), however, the RGR gene family, which included a shared gene between dogs and tame foxes, was lost.

Table 2

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 3
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 growing number of published RNA-seq datasets, and the associated number of conclusions involving biological systems, at times depending on one or few transcripts identified as being differentially expressed, the quantification of variation present between replicates at an individual transcript level is sometimes overlooked. Given the importance of the results postulated around such individual transcripts and the growing ability to base highly informative studies around archived transcriptomics datasets, it is imperative that care is taken in understanding such variation to a greater degree than the black-box approach that many of today's user-friendly software tools provide. In aid of this, we developed easy-to-use software to quantify intra-condition variation at an individual transcript level and, if desired, to use the output to remove reference-based transcripts associated with highly variable read counts. By applying our method, we demonstrated the effects of reducing the level of noise that standard differential expression tools are required to accommodate when determining differential expression patterns, in relation to inter and intra-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 differentially expressed genes, where the correlation between the mean of normalized counts and dispersion estimates per gene improved when removing transcripts displaying the highest levels of intra-condition variation (Fig. 3 and Supplementary Table 7 online). By removing such transcripts from the reference, prior to usage within differential expression software, gene-wise estimates of dispersion became more accurate [2]. Such noise at an inter, and to a lesser extent intra, study level arises from a range of characterized and uncharacterized sources including: i) biological differences between samples such as age, sex, diet, and health; ii) in silica error involving assembly tools producing poorly understood chimeras within the reference transcriptome [30]; iii) ambiguities in read mapping to such references [31]; iv) normalization of count data derived from such mapped reads [32]; and v) *in vitro* error during library preparation protocols [33,34]. Each of these can affect read counts across samples that are considered within the same condition of interest. Thus, when individual transcripts are hypothesized to be part of an informative result, the variation within the data that surrounds them ought to be thoroughly explored.

Samples from fox intra-study datasets contained a higher number of transcripts associated with high amounts of intra-condition variation. This high discordance between read counts from samples allocated to the same condition has resulted in a more spread distribution of dispersion estimates. When the detection of differentially expressed genes is dependent on how far a certain point is from a curve of best fit, having points that are highly spread will affect the detection of outliers and increase the number of false positives, likely explaining the low number of outliers detected among fox data (Table 1). Following the removal of the 3% and 5% of transcripts associated with the highest levels of variation between wolves vs. dogs and aggressive vs. tame foxes, respectively, the number of differentially expressed genes had the largest increase (Fig. 3). 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 (Supplementary Tables 3 to 5 online), demonstrating how dependent the final list of differentially expressed genes is on

the accuracy of gene-wise dispersion estimates. Prior to filtering, when results of the differential expression analysis in our intra-study case example were compared with those of the original publication [17], 92% of the genes previously identified as being differentially expressed were in agreement. Post filtering however 7% of these were no longer differentially expressed, highlighting the need for careful understanding of variance within transcriptomic datasets. Our method has been effective in improving 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 most rely on shared information across genes for dispersion estimation e.g. edgeR [35], BBSeq [36], DSS [37], baySeq [38] and ShrinkBayes [39].

At the 3% and 5% cutoffs, amongst the 50 over expressed genes identified, across the 21 shared gene families, seven genes were shared between dogs and tame foxes (Table 2). Of these seven 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 [16,40,41], whole-genome sequencing [42–44], and RNA data both using microarrays and RNA-seq [12,17,45–47]. Up until recently, almost no gene overlap had been observed between gene expression profiles involving pairs of domesticated and wild animals [14]. However, a newly published paper performing population genomic and brain transcriptional comparisons in seven bird and mammal domesticated species has revealed a strong convergent pattern in genes implicated in neurotransmission and neuroplasticity [48]. These functions are compatible with those found in our analysis. The shared gene ITGA7 belongs to a gene family that is known to play an essential role in the control of neuronal connectivity [49] and the inflammatory response [50]. Other genes from this family, for example, ITGA8, have been previously observed to be over expressed in tame foxes [46], and here we also observed its over expression in dogs providing further evidence of the family's role in tameness. Similar functions are associated with the shared genes CHRNA5 [51,52] and TRIB2 [53] from the cholinergic and tribbles family, respectively. Additionally, we found a shared gene involved in sensing local environmental stimuli, the MYO7A, whose mutation results in loss of hearing and vision [54]. Amongst the three gene families identified as under expressed (Table 3), we found the shared gene STMND1, which deficiency in the amygdala of mice was connected to a deficiency in innate and learned fear [55], a behaviour that speculatively could also have an important role in domestication.

In summary, we have presented a software to aid in the quantification, and if desired the removal, of intra-condition variation from RNA-seq count data and demonstrated its usage in improving the distribution of gene-wise dispersion estimates in an attempt to reduce the number of false positives in differential gene expression analysis. We would like our approach to highlight that studies mixing RNA-seq data from different sources should firstly, ensure that such datasets are suitable for performing differential expression on in relation to intra-condition variation at an individual transcript level, and secondly, characterize and discuss the variation present within the data prior to drawing conclusion on individual genes observed to be differentially expressed. Finally, in our case study of wolves, dogs and foxes, we

have provided further support for candidate genes involved with selection for tameness, and that appear to play a crucial role in the canine domestication.

Declarations

Acknowledgements

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 Funds through FCT under the references POCI-01-0145-FEDER-029115 and PTDC/BIA-EVL/29115/2017.

Competing interests

The author(s) declare no competing interests.

Data availability

The datasets analyzed during the current study were already publicly available. Detailed information about all data sources used in this study is described in Table S1.

Author Contributions

DL, RG and JA designed the study; JA and DL conceived and designed methodology. JA and DL implemented the software. DL compiled datasets and performed analysis; DL, RG and JA wrote the manuscript. All authors gave final approval for publication.

References

1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
2. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550(2014).
3. Hansen, K. D., Wu, Z., Irizarry, R. A. & Leek, J. T. Sequencing technology does not eliminate biological variability. *Nat. Biotechnol.* **29**, 572–573 (2011).
4. Liu, Y., Zhou, J. & White, K. P. RNA-seq differential expression studies: More sequence or more replication? *Bioinformatics.* **30**, 301–304 (2014).
5. Rau, A., Marot, G. & Jaffrézic, F. Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics 15:91(2014).
6. Xu, Z. & Asakawa, S. Physiological RNA dynamics in RNA-Seq analysis. *Brief. Bioinform.* **20**, 1725–1733 (2019).
7. McIntyre, L. M. *et al.* RNA-seq: Technical variability and sampling. BMC Genomics 12:293(2011).

8. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics.* https://doi.org/10.1186/1471-2105-12-480 (2011).
9. Gagnon-Bartsch, J. A. & Speed, T. P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics.* **13**, 539–552 (2012).
10. 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. 6(5): e100(2010).
11. Li, S. *et al.* Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* **32**, 888–895 (2014).
12. Li, Y. *et al.* Domestication of the dog from the Wolf was promoted by enhanced excitatory synaptic plasticity: A hypothesis. *Genome Biol. Evol.* **6**, 3115–3121 (2014).
13. Li, Y. *et al.* Artificial selection on brain-expressed genes during the domestication of dog. *Mol. Biol. Evol.* **30**, 1867–1876 (2013).
14. Albert, F. W. *et al.* A Comparison of Brain Gene Expression Levels in Domesticated and Wild Animals. PLoS Genet. 8:e1002962(2012).
15. Lord, K. A., Larson, G., Coppinger, R. P. & Karlsson, E. K. The History of Farm Foxes Undermines the Animal Domestication Syndrome. *Trends Ecol. Evol.* https://doi.org/10.1016/j.tree.2019.10.011 (2019).
16. Kukekova, A. *et al.* Mapping loci for fox domestication: Deconstruction/Reconstruction of a behavioral phenotype. *Behav. Genet.* **41**, 593–606 (2011).
17. Wang, X. *et al.* Genomic responses to selection for tame/aggressive behaviors in the silver fox (Vulpes vulpes). Proc. Natl. Acad. Sci. 115, 10398–10403(2018).
18. Hekman, J. *et al.* Anterior Pituitary Transcriptome Suggests Differences in ACTH Release in Tame and Aggressive Foxes. G3; Genes|Genomes|Genetics 8, 859–873(2018).
19. Costa-Silva, J., Domingues, D. & Lopes, F. M. RNA-Seq differential expression analysis: An extended review and a software tool. PLoS One 12:e019015(2017).
20. Roy, M. *et al.* Analysis of the canine brain transcriptome with an emphasis on the hypothalamus and cerebral cortex. *Mamm. Genome.* **24**, 484–499 (2013).
21. Fushan, A. A. *et al.* Gene expression defines natural changes in mammalian lifespan. *Aging Cell.* **14**, 352–365 (2015).
22. Hoeppner, M. P. *et al.* An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. PLoS One 9(3):91172(2014).
23. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* https://doi.org/10.1038/nmeth.1923 (2012).
24. Bushnell & Brian BBMap: A Fast, Accurate, Splice-Aware Aligner. in Conference: 9th Annual Genomics of Energy {\&} Environment Meeting(2014). doi:10.1186/1471-2105-13-238.

25. (2020), R, R. C. T. A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* https://www.r-project.org/ (2020).
26. Durinck, S. *et al.* BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics.* **21**, 3439–3440 (2005).
27. 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.* **25**, 1515–1519 (2015).
28. Archer, J., Whiteley, G., Casewell, N. R., Harrison, R. A. & Wagstaff, S. C. VTBuilder: A tool for the assembly of multi isoform transcriptomes.*BMC Bioinformatics***15**, (2014).
29. Wayne, R. K. *et al.* Molecular Systematics of the Canidae. *Syst. Biol.* **46**, 622–653 (1997).
30. Hsieh, P. H., Oyang, Y. J. & Chen, C. Y. Effect of de novo transcriptome assembly on transcript quantification.*Sci. Rep.***9**, (2019).
31. Reinert, K., Langmead, B., Weese, D. & Evers, D. J. Alignment of Next-Generation Sequencing Reads. *Annu. Rev. Genomics Hum. Genet.* **16**, 133–151 (2015).
32. Evans, C., Hardin, J. & Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.* **19**, 776–792 (2018).
33. Brodin, J. *et al.* PCR-Induced Transitions Are the Major Source of Error in Cleaned Ultra-Deep Pyrosequencing Data.*PLoS One***8**, (2013).
34. Ma, X. *et al.* Analysis of error profiles in deep next-generation sequencing data.*Genome Biol.***20**, (2019).
35. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* **26**, 139–140 (2010).
36. Zhou, Y. H., Xia, K. & Wright, F. A. A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics.* https://doi.org/10.1093/bioinformatics/btr449 (2011).
37. Wu, H., Wang, C. & Wu, Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics.* **14**, 232–243 (2013).
38. Hardcastle, T. J., Kelly, K. A. & BaySeq Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11:422(2010).
39. Van De Wiel, M. A. *et al.* Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics.* **14**, 113–128 (2013).
40. 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.* **12**, 330–337 (2013).
41. Albert, F. W. *et al.* Genetic architecture of tameness in a rat model of animal domestication. *Genetics.* **182**, 541–554 (2009).
42. Carneiro, M. *et al.* Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science (80-.).* **345**, 1074–1079 (2014).

43. Freedman, A. H. *et al.* Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs. PLoS Genet. 12:e100585(2016).
44. Kukekova, A. *et al.* The red fox genome assembly identifies genomic regions associated with tame and aggressive behaviors. *Nat. Ecol. Evol.* **2**, 1479–1491 (2018).
45. Saetre, P. *et al.* From wild wolf to domestic dog: Gene expression changes in the brain. *Mol. Brain Res.* **126**, 198–206 (2004).
46. Kukekova, A. *et al.* Sequence comparison of prefrontal cortical brain transcriptome from a tame and an aggressive silver fox (Vulpes vulpes). BMC Genomics 12:482(2011).
47. Heyne, H. O. *et al.* Genetic influences on brain gene expression in rats selected for tameness and aggression. *Genetics.* **198**, 1277–1290 (2014).
48. Hou, Y. *et al.* Genome-wide analysis reveals molecular convergence underlying domestication in 7 bird and mammals.*BMC Genomics***21**, (2020).
49. Lilja, J. & Ivaska, J. Integrin activity in neuronal connectivity. J. Cell Sci. 131:jcs212(2018).
50. González-Amaro, R. & Sánchez-Madrid, F. Cell adhesion molecules: selectins and integrins. *Crit. Rev. Immunol.* **19**, 389–429 (1999).
51. Winterer, G. *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.* **153**, 1448–1458 (2010).
52. Zhang, H., Kranzler, H. R., Poling, J., Gruen, J. R. & Gelernter, J. Cognitive flexibility is associated with KIBRA variant and modulated by recent tobacco use. *Neuropsychopharmacology.* **34**, 2508–2516 (2009).
53. Eyers, P. A., Keeshan, K. & Kannan, N. Tribbles in the 21st Century: The Evolving Roles of Tribbles Pseudokinases in Biology and Disease. *Trends Cell Biol.* **27**, 284–298 (2017).
54. Miller, K. A. *et al.* Inner Ear Morphology Is Perturbed in Two Novel Mouse Models of Recessive Deafness. PLoS One 7(12):e512(2012).
55. Martel, G., Nishi, A. & Shumyatsky, G. P. Stathmin reveals dissociable roles of the basolateral amygdala in parental and social behaviors. Proc. Natl. Acad. Sci. U. S. A. 105, 14620–14625(2008).

Figures

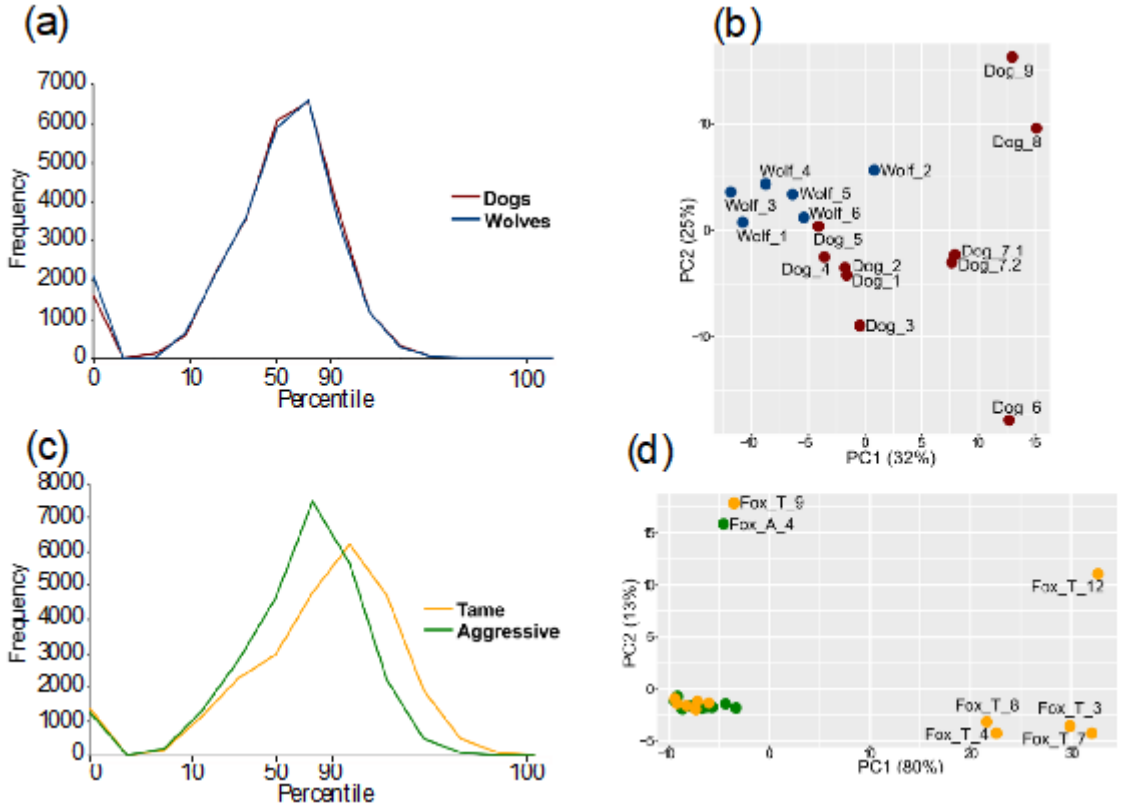

Figure 1

Percentile range of intra-condition variation scores (x-axis) observed prior to filtering for wolves and dogs (a), as well for aggressive and tame foxes (c). Corresponding PCA plots based on normalized and non-filtered data of the individual datasets comparing wolves and dogs (b) and aggressive and tame foxes (d; only samples displayed in a distant cluster are labeled).

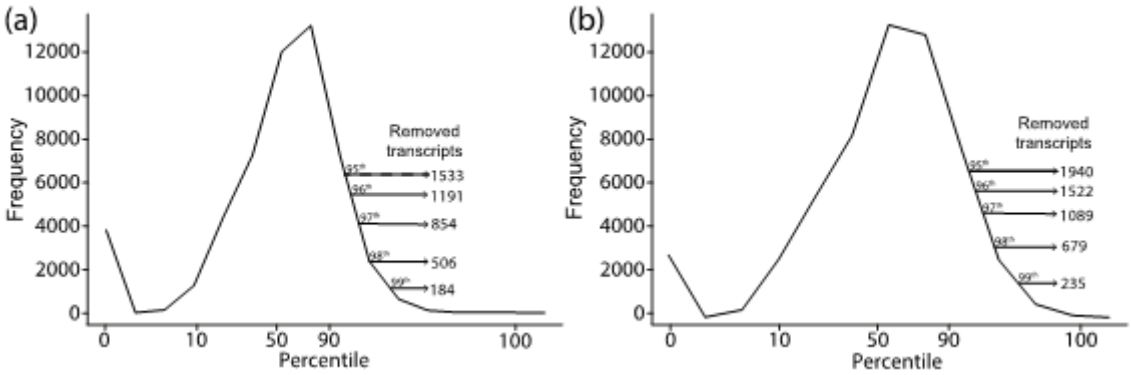

Figure 2

Percentile range of combined intra-condition variance scores (x-axis) present in each contrast: wolves vs. dogs and aggressive vs. tame foxes depicting the number of transcripts above values from the 95th the 99th percentiles. Arrows indicate the number of transcripts removed for the associated percentiles, from the 95th to the 99th.

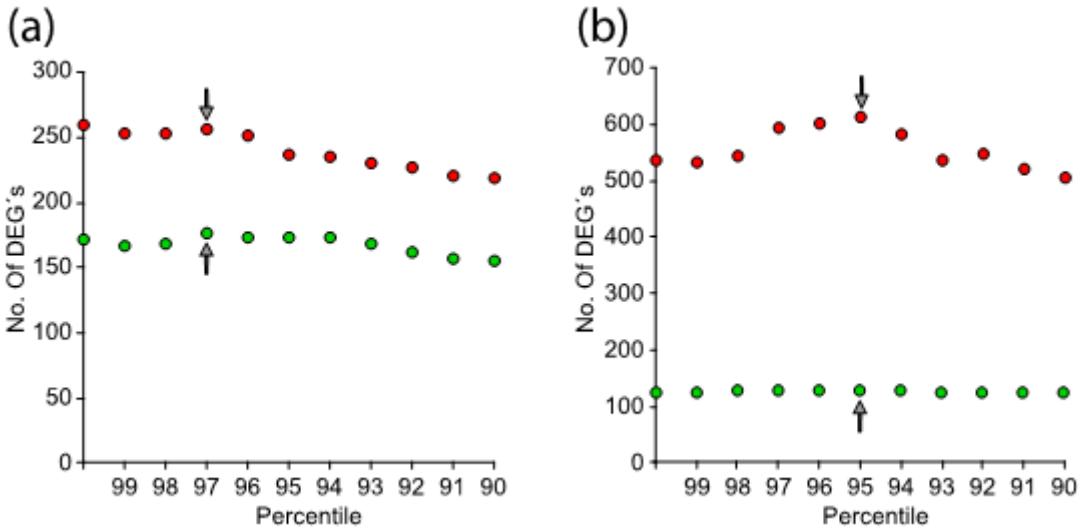

Figure 3

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

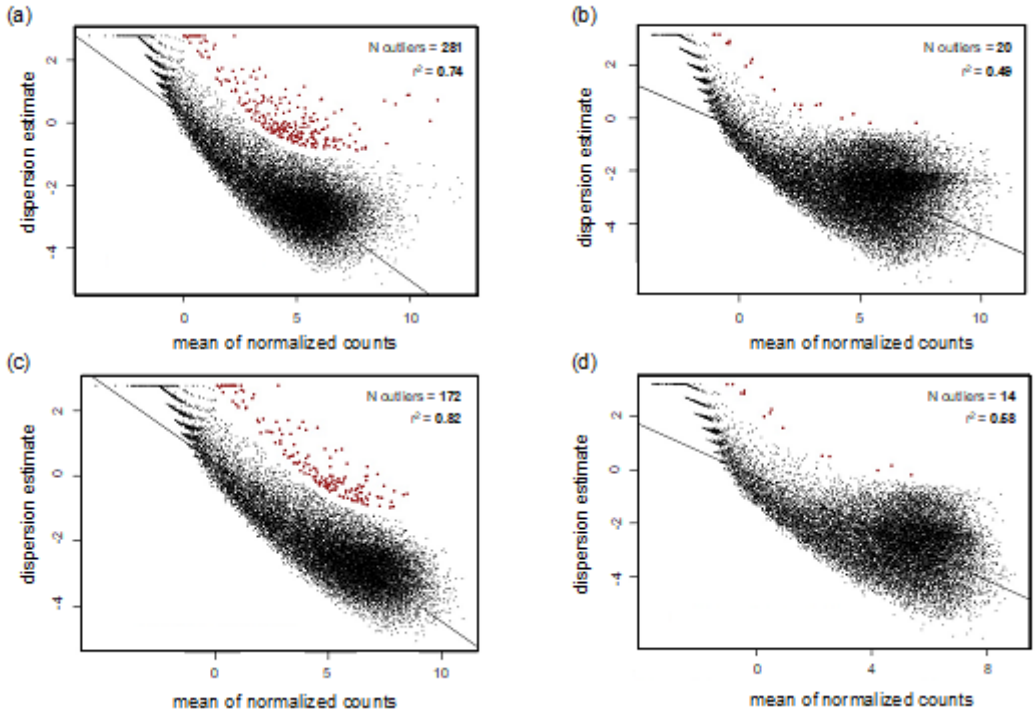

Figure 4

Plots of final dispersion estimates for wolves vs. dogs (a and c) and aggressive vs. tame foxes (b and d) calculated using DESeq2 for the non-filtered (NF, a and b) and 10% filtered (90th, c and d) datasets. Each black dot represents one transcript and red dots represent outliers. Values for the number of outliers and

r-square are presented at the top of each graph. Both x and y-axis were transformed into a logarithm scale.

Supplementary Files

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

- supplementaryfiguresandtablesloboetal.docx