RNA-Seq data analysis for Planarian with tensor decomposition-based unsupervised feature extraction

Makoto Kashima[1], Nobuyoshi Kumagai[2], Hiromi Hirata[1], and Y-h. Taguchi[3,*]

[1]Department of Biology, Aoyama Gakuin Univeristy, Tokyo 150-8366, Japan
[2]Department of Life Science, Gakushuin University, Tokyo 171-8588, Japan
[3]Department of Physics, Chuo University, Tokyo 112-8551, Japan
[*]tag@granular.com

ABSTRACT

RNA-Seq data analysis of non-model organisms is often difficult because of the lack of a well-annotated genome. In model organisms, after short reads are mapped to the genome, it is possible to focus on the analysis of regions well-annotated regions. However, in non-model organisms, contigs can be generated by *de novo* assembling. This can result in a large number of transcripts, making it difficult to easily remove redundancy. A large number of transcripts can also lead to difficulty in the recognition of differentially expressed transcripts (DETs) between more than two experimental conditions, because *P*-values must be corrected by considering multiple comparison corrections whose effect is enhanced as the number of transcripts increases. Heavily corrected *P*-values often fail to take sufficiently small *P*-values as significant. In this study, we applied a recently proposed tensor decomposition (TD)-based unsupervised feature extraction (FE) to the RNA-seq data obtained for a non-model organism, Planarian; we successfully obtained a limited number of transcripts whose expression was altered between normal and defective samples as well as during time development. TD-based unsupervised FE is expected to be an effective tool that can identify a limited number of DETs, even when a poorly annotated genome is available.

Introduction

Identification of differentially expressed transcripts (DETs)[1] between more than two distinct experimental conditions is the starting point of RNA-seq data analysis, as DETs are expected to be related to the experimental conditions considered, for example, diseases. Once DETs are successfully identified, it is possible to study the biological properties that are enriched in a set. Thus, unless DETs can be successfully identified, it is very difficult to make use of RNA-seq data to identify the biological processes that take place during experiments using RNA-seq datasets.

DET identification in model organisms has been well-established. Once short reads are successfully mapped to the genome, we can concentrate those mapped to well-annotated regions; for example, protein-coding genes. Then, the number of reads can be summed up within the considered regions, for example, the gene body, then it is possible to estimate the amount of expression of individual genes. In contrast, DET identification of non-model organisms is not straightforward. Because of the lack of a well-annotated genome, no short reads can be mapped to it. Instead of the genome, it is possible to assemble the contigs by assembling *de novo* and short reads can then be mapped toward these contigs. However, there is one problem with this strategy; it is never possible to be confident that the parts of the assembled contigs are unique. Some contigs may overlap with some parts of another contig. This results in so-called multiple mapping, which drastically reduces the accuracy of the estimated amount of transcript. In addition, the number of contigs can often be larger than the true number of transcripts because of the above-mentioned redundancy of contigs. As there are no ways to estimate which part of the contigs is redundant, the number of contigs can often be in millions. This causes serious problems because *P*-values are computed by statistical tests that check how significant the observed distinct expression of a transcript is under the null hypothesis, and the hypothesis assumes that the equivalence of experimental conditions must be corrected by considering multiple comparison corrections. As the number of multiple comparisons is equivalent to the number of contigs, too large contigs (e.g. millions) can incorrectly identify the difference between more than two experimental conditions as non-significant.

To address this problem, we applied tensor decomposition (TD)-based unsupervised feature extraction (FE)[2] to the RNA-seq data of Planarian, a non-model organism without a well-annotated genome. Although the number of contigs generated by *de novo* assembly is as many as 2.8×10^5, which is much larger than the expected number of true genes of Planarian, we could successfully select a limited number of contigs whose expression was altered during time development as well as was distinct between normal and defective samples. This suggests the usefulness of TD-based unsupervised FE when it is applied to

RNA-seq data analysis of non-model organisms, from which too many redundant contigs are often obtained.

Results

First, we attempted to identify which $u_{\ell_1 j}$ is associated with the desired property. Figure 1 shows u_{2j} that represents distinction between the RNAi treatments. The RNAi of *Djhp1-1*, *Djima1-1*, *DjpiwiB*, and *DjpiwiC* is distinct from that of *gfp* and *DjpiwiA*. While *gfp* is the control treatment and KD of *piwiA* is known not to affect (inhibit) the regeneration of planarians, RNAi of *Djhp1-1*, *Djima1-1*, *DjpiwiB*, and *DjpiwiC* causes regenerative defects [3–5]. Thus, u_{2j} coincides with the distinction between normal and defective samples.

Next, we attempted to identify which $u_{\ell_2 t}$ coincides with time development. Figure 2 shows u_{2t} that exhibits the desired time development. Several days after RNAi treatment, RNAi appeared to have an effect, which then gradually decreased. Amputation was performed seven days after RNAi treatment, which corresponds to the last date when u_{2t} takes the larger negative values (the sign of u_{2t} does not have any meaning because only the product of $G(\ell_1\ell_2\ell_3\ell_4)u_{\ell_1 j}u_{\ell_2 t}u_{\ell_3 k}u_{\ell_4 i}$ matters, not individual terms in the product (see eq.(1)). Thus, this time development seemed to coincide with the experimental procedures. The reason why u_{2t} takes almost zero might be because it requires several days until RNAi starts to affect the transcriptome.

We also noticed that u_{1k} has almost constant values independent of the biological replicate, k (Fig. 3); this suggests that the expression of genes associated with u_{1k} is likely to be common among the six biological replicates.

We then determine which ℓ_4 had the largest $|G(2,2,1,\ell_4)|$. Table 1 shows $G(2,2,1,\ell_4)$. $\ell_4 = 1,3,4$ had a larger $|G|$ value. It is not possible to employ all three ℓ_4s to compute P_i because $|G(1,1,1,1)|$ is much larger than $G(2,2,1,1)$ (not shown here). This means that u_{1i} is more coincident with $\ell_1 = \ell_2 = \ell_3 = 1$, which we are not interested in. Thus, we decided to employ $\ell_4 = 3,4$ to attribute P-values to the ith contig (that is, $\Omega_{\ell_4} = \{3,4\}$ in Eq.(2)). P_is are corrected as described in the Methods section, and contigs with adjusted P_i less than 0.01 were selected.

After applying TD-based unsupervised FE to RNA-seq data obtained from Planarian as described in the Methods section, we obtained 155 contigs as those whose expression was altered during time development as well as distinct between normal and defective samples. To confirm whether we could successfully identify genes with altered expression, we applied statistical tests that validated significant dependence upon time as well as significant distinction between normal and defective samples. First, we applied the t test to determine whether gene expression was distinct between $t \leq 8$ and $t \geq 9$ (Table 2). There were at least non-zero genes expressed differently between $t \leq 8$ and $t \geq 9$.

We also checked whether genes were expressed differently on individual days between normal and defective samples (Table 3). In this case, more than half of the genes were differentially expressed between normal and defective samples. Thus, our analysis was successful.

Discussion

We employed unpopular TD-based unsupervised FE to identify DET despite the existence of many other conventional methods. This was to show that the problem is too difficult to resolve using other conventional methods. Therefore, we applied DESeq2[6] to the present dataset. As a result, we identified as few as 10 and 5 contigs for those expressed differently between normal and defective samples, respectively, and those expressed dependent upon time; these numbers are very low when compared with the number of contigs identified by TD-based unsupervised FE, 155. DESeq2 failed because there were as many as 278167 contigs, whereas the number of samples was as low as $6 \times 15 \times 6 = 540$; its ratio was as large as 5×10^2. It is a difficult problem to tackle using standard conventional methods designed for a much lower ratio of the number of features to the number of samples.

On the other hand, biological evaluation of the 155 contigs is difficult. Although we tried to perform a Basic Local Alignment Search Tool (BLAST) match between 155 contigs and a well-annotated mouse transcript library, we could not find any significant enriched GO terms within contig annotation based upon BLAST search toward the mouse genome. Instead, we performed a BLAST search of all organisms, individually, of 155 contigs (see Supplementary Materials). We found that almost half of the 155 contigs had a common match with two known long planarian transcripts; AK388828.1 and AK389113.1, and that these are likely to be alternatively spliced transcripts of the transcript. TRINITY_DN1947 (Fig. 4). Although we are unsure why Trinity[7] failed to merge these redundant contigs into one, this suggests that TD-based unsupervised FE can detect transcripts that share the same expression patterns. If individual contigs are alternative spliced short transcripts of a long transcript, which is likely to be TRINITY_DN1947, there are similar expression patterns.

In conclusion, although we were unable to assess the biological significance of the 155 contigs obtained, we believe that the methodological advantages of TD-based unsupervised FE were successfully demonstrated. First, TD-based unsupervised FE could identify more contigs that are expressed distinctly between normal and defective samples, as well as being expressed in a time-dependent manner (days after RNAi). In addition, half of the identified 155 transcripts were likely alternative spliced transcripts of a longer transcript. This supports the ability of TD-based unsupervised FE to select contigs that share similar expression patterns, since alternative spliced transcripts from longer transcripts are likely to share the same expression pattern.

As a result, TD-based unsupervised FE is expected to be an effective tool to be applied to DET detection using the redundant contigs obtained by applying the *de novo* assembly of short reads from RNA-seq applied to non-model organisms that lack a well-annotated genome.

Methods

Maintenance of planarian

A clonal strain of planarian *D. japonica*, a sexualizing super planarian (2n = 16), was used in this study[8]. The planarians were fed chicken liver once or twice a week. The planarians that were approximately 5 mm in length were starved for at least one week prior to the following experiments.

Double-stranded RNA (dsRNA) synthesis

dsRNA was synthesised as described by Rouhana et al.[9]. We prepared templates flanked by the T7 promoter for dsRNA synthesis by using polymerase chain reaction (PCR) for each EST (*DjpiwiA* [*Dj_aH_000_03609HH*], *DjpiwiB* [*Dj_aH_221_M14*], *DjpiwiC* [*Dj_aH_000_05977HH*], *Djhp1-1* [*Dj_aH_000_01636HH*], *Djima1-1* [*Dj_aH_313_F03*]). The primers for the PCR reaction were as follows (5' to 3'): Zap Linker + T7: GATCACTAATACGACTCACTATAGGGGAATTCGGCACGAGG M13 Rev: GTTTTCCCAGTCACGACGTTGTAA.

Feeding RNA interference

RNA interference was conducted as described previously[9]. dsRNA-containing food consisting of 25 μL of chicken liver solution, 6 μL of 2% agarose, and 6.5 μL of 2.0 μg/μL dsRNA was synthesised in vitro. We fed the planarians twice at intervals of two days. Control animals were fed food containing eGFP dsRNA. Six individual planarians were euthanized every day from day 1 to 16 after the second feeding.

Total RNA purification

Total RNA was extracted from each individual planarian using the "Direct-TRI" method[10]. Briefly, a planarian was lysed using TRI Reagent-LS (Molecular Research Center, Cincinnati, OH, USA). Then, the lysate was purified using an AcroPrep Advance 96-well long tip filter plate for nucleic acid binding (Pall, Port Washington, NY, USA). RNA was eluted with 30 µL nuclease-free water.

RNA-Seq library preparation and sequencing

3' mRNA-Seq was conducted using the Lasy-Seq ver. 1.1 Protocol (https://sites.google.com/view/lasy-seq/)[11,12]. Briefly, 100 ng of total RNA was reverse transcribed using an RT primer with an index and SuperScript IV reverse transcriptase (Thermo Fisher Scientific, Waltham, MA, USA). All RT mixtures of the samples were pooled and purified using an equal volume of AMpure XP beads (Beckman Coulter, Brea, CA, USA) according to the manufacturer's instructions. Second-strand synthesis was conducted on the pooled samples using RNaseH (5 U/μL, Enzymatics, Beverly, MA, USA), and DNA polymerase I (10 U/μL, Enzymatics, Beverly, MA, USA). To avoid carryover of large amounts of rRNAs, the mixture was subjected to RNase treatment using RNase T1 (Thermo Fisher Scientific, Waltham, MA, USA). Then, purification was conducted with a 0.8× volume of AMpure XP beads. Fragmentation, end-repair, and A-tailing were conducted using a 5× WGS fragmentation mix (Enzymatics, Beverly, MA, USA). The adapter for Lasy-Seq was ligated using 5× Ligation Mix (Enzymatics, Beverly, MA, USA), and the adapter-ligated DNA was purified twice with a 0.8× volume of AMpure XP beads. After the optimisation of PCR cycles for library amplification by qPCR using Evagreen, 20× in water (Biotium, Fremont, CA, USA) and the QuantStudio5 Real-Time PCR System (Applied Biosystems, Waltham, MA, USA), the library was amplified using KAPA HiFi HotStart ReadyMix (KAPA BIOSYSTEMS, Wilmington, MA, USA) on the ProFlex PCR System (Applied Biosystems, Waltham, MA, USA). The amplified library was purified using an equal volume of AMpure XP beads. One microliter of the library was then used for electrophoresis using a Bioanalyzer 2100 with the Agilent High Sensitivity DNA kit (Agilent Technologies, Santa Clara, CA, USA) to check for quality. Sequencing of 150 bp paired-end reads was performed using HiSeq X Ten (Illumina, San Diego, CA, USA).

Mapping and gene quantification

Read 1 reads were processed with fastp (version 0.21.0)[13] using the following parameters: trim_poly_x -w 20 –adapter_sequence=AGATCGGA –adapter_sequence_r2= AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-l 31. The trimmed reads were then mapped to a *D. japonica* reference sequence of GJEZ00000000 in the TSA repository using BWA mem (version 0.7.17-r1188)[14] with the default parameters. The read count for each gene was calculated with salmon using -l IU, which specifies the library type (version v0.12.0)[15]. All quantification results and sample information of the RNA-seq analysis were deposited as GSE174855 in the GEO repository.

TD-based unsupervised FE

The number of reads mapped to individual contigs was formatted as a tensor, $x_{ijtk} \in \mathbb{R}^{278167\times 6\times 15\times 6}$, which represents the expression of the ith contig of the kth biological replicates at the tth day after the jth RNAi was performed. The 14th dataset was excluded because it included serious outliers. Thus, $t \leq 13$ corresponds to t days after the treatment, whereas $t \geq 15$ corresponds to $t+1$ days after the treatment. x_{ijtk} is normalised such that $\sum_i x_{ijtk} = 0$ and $\sum_i x_{ijtk}^2 = 278167$. Then, we applied higher order singular value decomposition (HOSVD)[2] and got

$$x_{ijtk} = \sum_{\ell_1=1}^{6}\sum_{\ell_2=1}^{15}\sum_{\ell_3=1}^{6}\sum_{\ell_4=1}^{278167} G(\ell_1\ell_2\ell_3\ell_4)u_{\ell_1 j}u_{\ell_2 t}u_{\ell_3 k}u_{\ell_4 i} \tag{1}$$

where $G \in \mathbb{R}^{6\times 15\times 6\times 278167}$ is the core tensor that represents the weight of the product $u_{\ell_1 j}u_{\ell_2 t}u_{\ell_3 k}u_{\ell_4 i}$, $u_{\ell_1 j} \in \mathbb{R}^{6\times 6}$. $u_{\ell_2 t} \in \mathbb{R}^{15\times 15}$ and $u_{\ell_3 k} \in \mathbb{R}^{6\times 6}$, and $u_{\ell_4 i} \in \mathbb{R}^{278167\times 278167}$ are singular value matrices that are orthogonal matrices.

First, we needed to identify $u_{\ell_1 j}$, $u_{\ell_2 t}$, and $u_{\ell_3 k}$ that satisfy

- $u_{\ell_1 j}$ that exhibits distinction between normal and defective samples.
- $u_{\ell_2 k}$ with time (days) dependence.
- $u_{\ell_3 k}$ independent of individual biological replicates, k, i.e. with constant values.

After identifying ℓ_1, ℓ_2, and ℓ_3 that satisfy the above requirements, we sought a set of ℓ_4s, Ω_{ℓ_4}, associated with the larger $|G(\ell_1\ell_2\ell_3\ell_4)|$s given ℓ_1, ℓ_2, and ℓ_3, because $u_{\ell_4 i}$s are expected to have larger absolute values for the genes whose expression is associated with the above requirement. Then, we attributed P-values to the ith contig by assuming that $u_{\ell_4 i}$ obeys a multiple Gaussian distribution (null hypothesis)

$$P_i = P_{\chi^2}\left[> \sum_{\ell_4 \in \Omega_{\ell_4}} \left(\frac{u_{\ell_4 i}}{\sigma_{\ell_4}} \right)^2 \right], \tag{2}$$

where $P_{\chi^2}[> x]$ is the cumulative probability distribution of the χ^2 distribution when the argument takes larger values than x. The obtained P_is are corrected by the BH criterion[2], and contigs associated with an adjusted P_i less than 0.01 were selected.

DESeq2

DESeq2[6] (version 1.30.0) was applied to either comparison between two groups of RNAi treatments, that is, *gfp*, *DjpiwiA* and others (*Djhp1-1*, *Djima1-1*, *DjpiwiB*, *DjpiwiC*) or that between two time periods, that is, $t = 1, 2, \cdots, 8$, and others ($t = 9, 10, 11, 12, 13, 15, 16$). We selected genes associated with adjusted P-values less than 0.01, which is the same as that used for TD-based unsupervised FE. All other parameters were taken as the default values.

References

1. Taguchi, Y.-H. Comparative transcriptomics analysis. In Ranganathan, S., Gribskov, M., Nakai, K. & Schönbach, C. (eds.) *Encyclopedia of Bioinformatics and Computational Biology*, 814–818, DOI: https://doi.org/10.1016/B978-0-12-809633-8.20163-5 (Academic Press, Oxford, 2019).
2. Taguchi, Y.-H. *Unsupervised Feature Extraction Applied to Bioinformatics* (Springer International Publishing, 2020).
3. Shibata, N. *et al.* Inheritance of a nuclear PIWI from pluripotent stem cells by somatic descendants ensures differentiation by silencing transposons in planarian. *Dev. Cell* **37**, 226–237, DOI: 10.1016/j.devcel.2016.04.009 (2016).
4. Hubert, A. *et al.* A functional genomics screen identifies an Importin-α homolog as a regulator of stem cell function and tissue patterning during planarian regeneration. *BMC Genomics* **16**, 1–18, DOI: 10.1186/s12864-015-1979-1 (2015).
5. Zeng, A. *et al.* Heterochromatin protein 1 promotes self-renewal and triggers regenerative proliferation in adult stem cells. *J. Cell Biol.* **201**, 409–425, DOI: 10.1083/jcb.201207172 (2013).
6. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, DOI: 10.1186/s13059-014-0550-8 (2014).
7. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512, DOI: 10.1038/nprot.2013.084 (2013).

8. Shibata, N. *et al.* Comprehensive gene expression analyses in pluripotent stem cells of a planarian, Dugesia japonica. *Int. J. Dev. Biol.* **56**, 93–102, DOI: 10.1387/ijdb.113434ns (2012).
9. Rouhana, L. *et al.* RNA interference by feeding in vitro-synthesized double-stranded RNA to planarians: Methodology and dynamics. *Dev. Dyn.* **242**, 718–730, DOI: 10.1002/dvdy.23950 (2013).
10. Ujibe, K., Nishimura, K., Kashima, M. & Hirata, H. Direct-TRI: High-throughput RNA-extracting method for all stages of zebrafish development. *bio-protocol* **in press**.
11. Kamitani, M., Kashima, M., Tezuka, A. & Nagano, A. J. Lasy-Seq: a high-throughput library preparation method for RNA-Seq and its application in the analysis of plant responses to fluctuating temperatures. *Sci. Reports* **9**, 7091, DOI: 10.1038/s41598-019-43600-0 (2019).
12. Kashima, M., Kamitani, M., Nomura, Y., Hirata, H. & Nagano, A. J. DeLTa-Seq: direct-lysate targeted RNA-Seq from crude tissue lysate. (8/15 words) Running Title: Development of direct-lysate targeted RNA-Seq method Corresponding Author. *bioRxiv* 2020.09.15.299180, DOI: 10.1101/2020.09.15.299180 (2020).
13. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, DOI: 10.1093/bioinformatics/bty560 (2018).
14. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, DOI: 10.1093/bioinformatics/btp324 (2009). 1303.3997.
15. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419, DOI: 10.1038/nmeth.4197 (2017). 1505.02710.

Acknowledgements

This work was supported by KAKENHI [grant numbers 19H05270, 20H04848, and 20K12067] to Y.T. and KAKENHI [grant numbers 19K16149] to M.K.

Competing interests

The author(s) declare no competing interests.

Author contributions statement

M.K. and N.K. conceived and conducted the experiments, and all authors analysed the results and reviewed the manuscript.

Supplementary materials

BLAST search results toward 155 contigs.

Table 1. $G(2,2,1,\ell_4)$

ℓ_1	$G(2,2,1,\ell_4)$
1	227.87614
2	-8.36258
3	-113.10063
4	250.20079
5	-43.37827
6	-64.02292
7	37.49047
8	12.84754
9	31.20391
10	-21.70284

Table 2. The number of contigs identified by the t test to be distinct between $t \leq 8$ and $t \geq 9$.

RNAi	netative	positive
gfp	89	66
Djhp1-1	96	59
Djima1-1	139	16
DjpiwiA	89	66
DjpiwiB	102	53
DjpiwiC	151	4

Table 3. The number of contigs identified by the t test to be distinct between normal and defective samples.

days	negative	positive
1	76	79
2	118	37
3	41	114
4	69	86
5	78	77
6	90	65
7	48	107
8	73	82
9	125	30
10	93	62
11	65	90
12	46	109
13	58	97
15	95	60
16	50	105

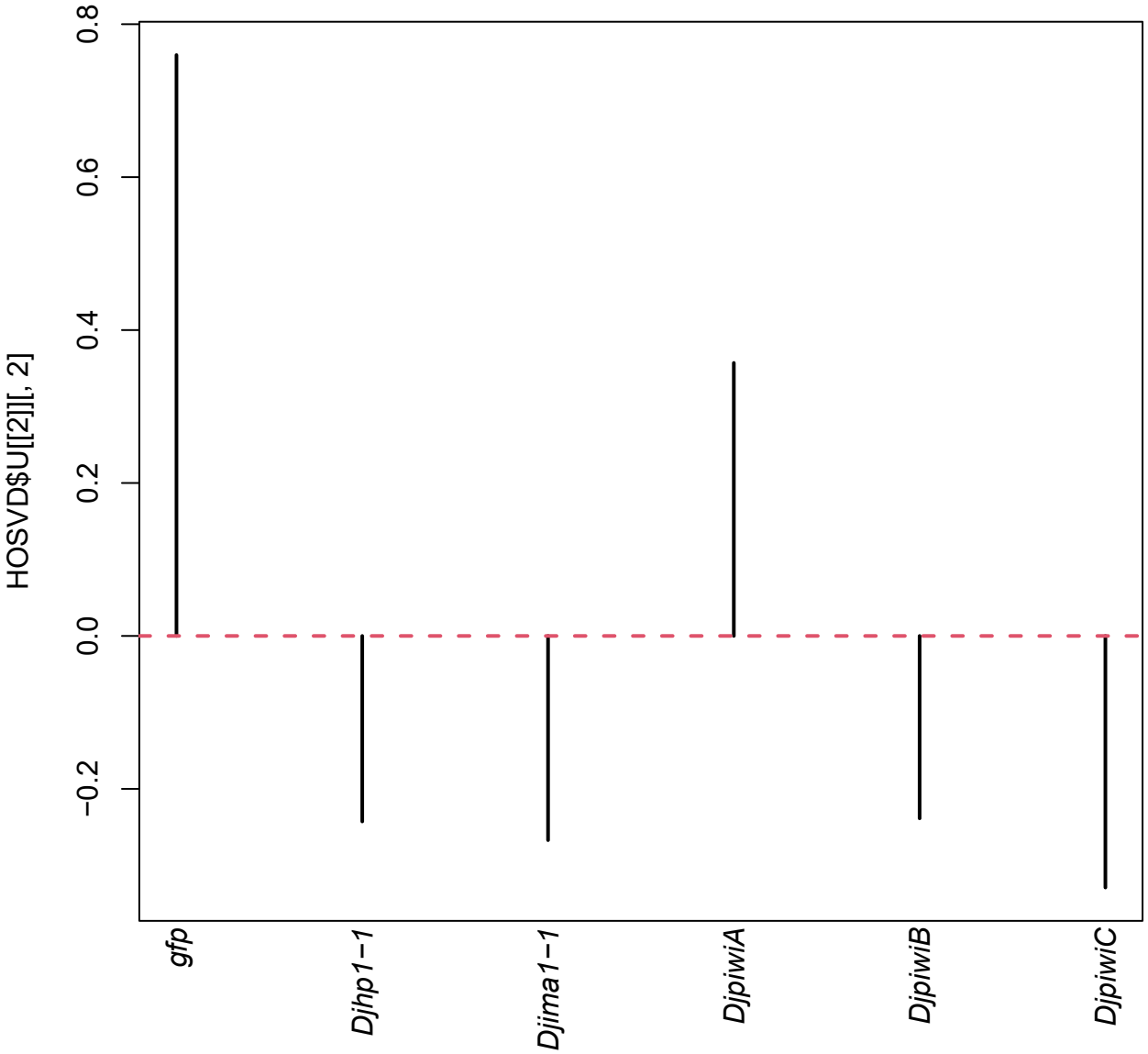

Figure 1. u_{2j} that represent 2nd singular value vectors attributed to RNAi experiments

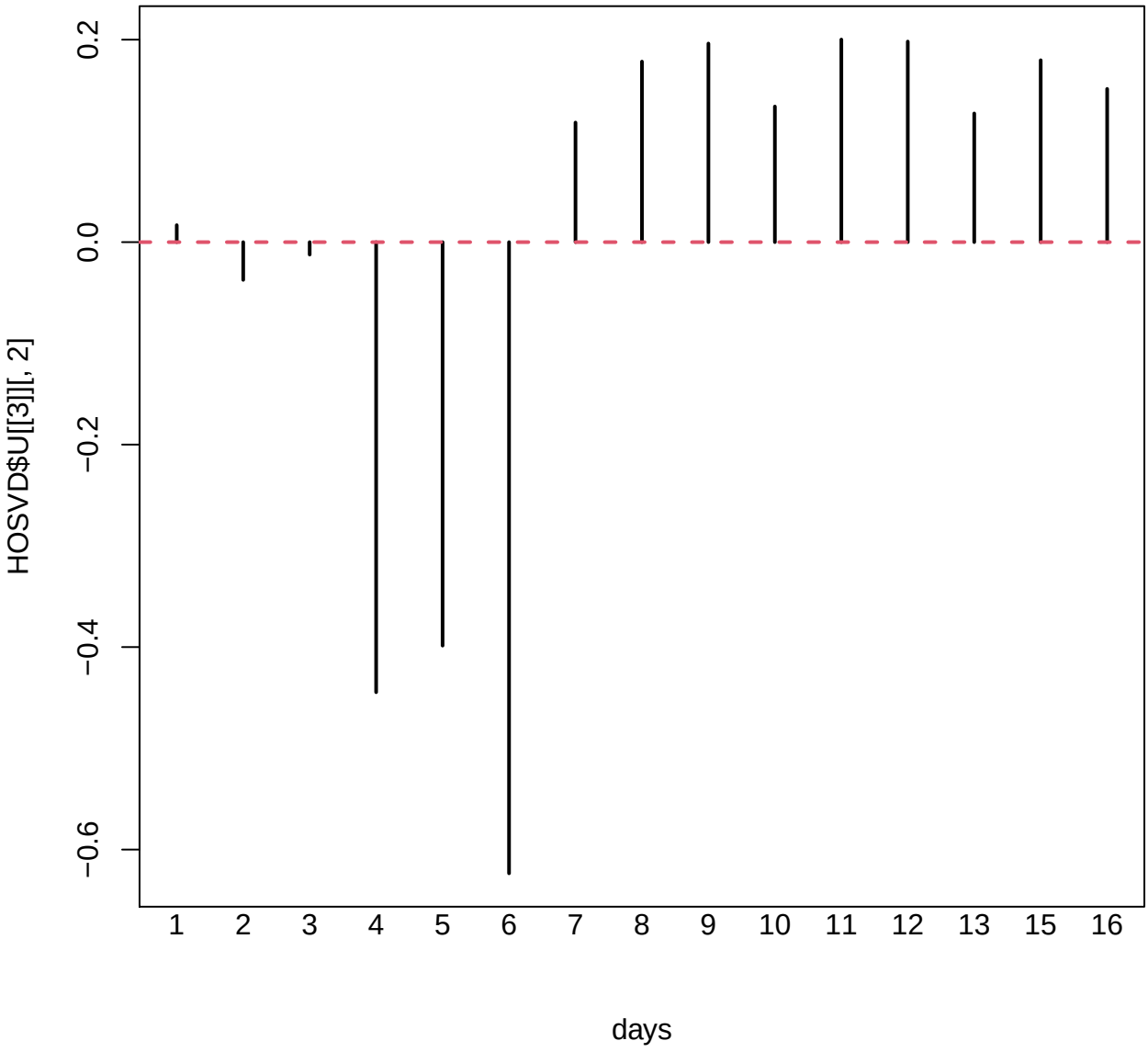

Figure 2. u_{2t} that represent 2nd singular value vectors attributed to days after the treatments

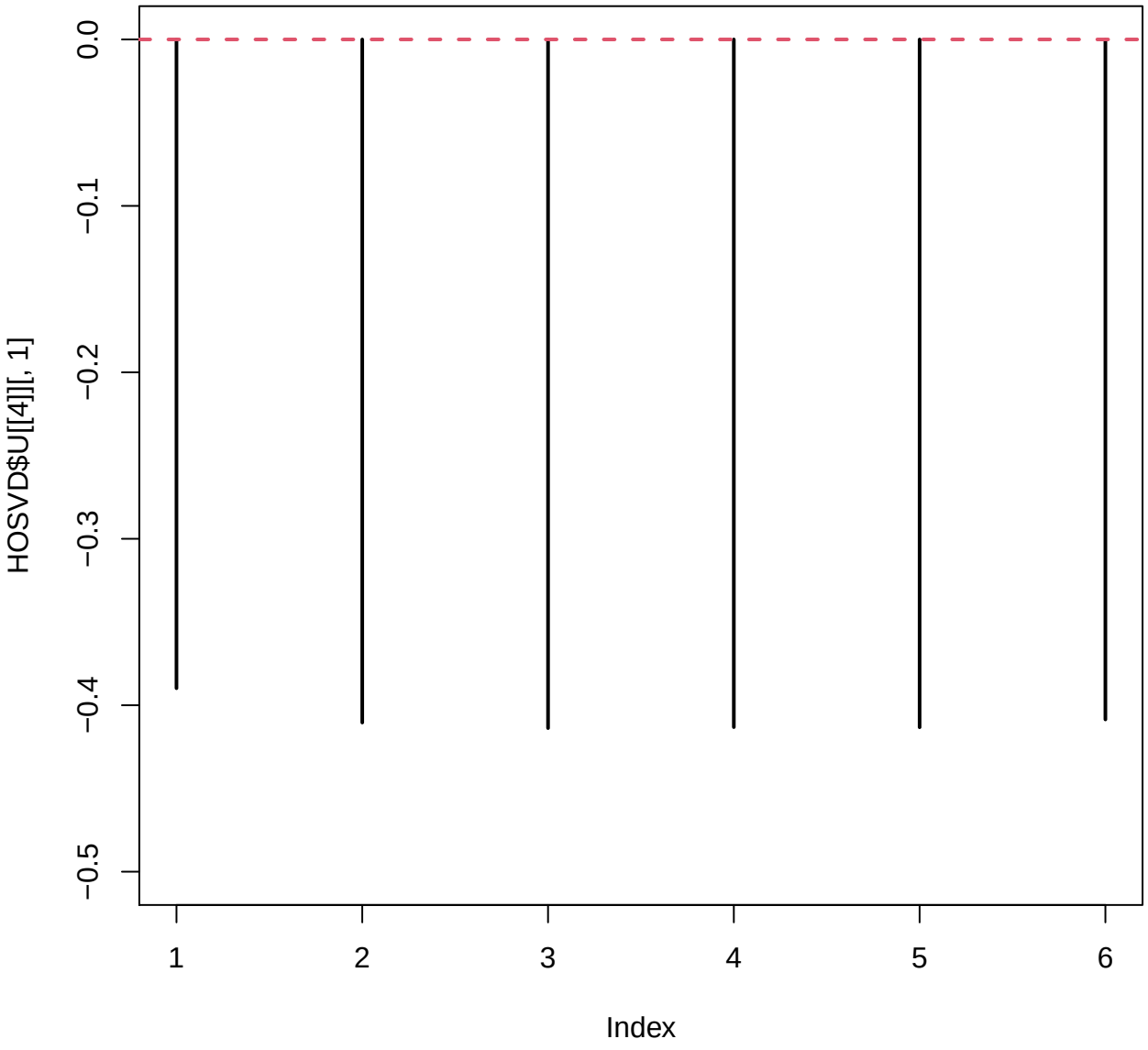

Figure 3. u_{1k} that represent 1st singular value vectors attributed to biological replicates

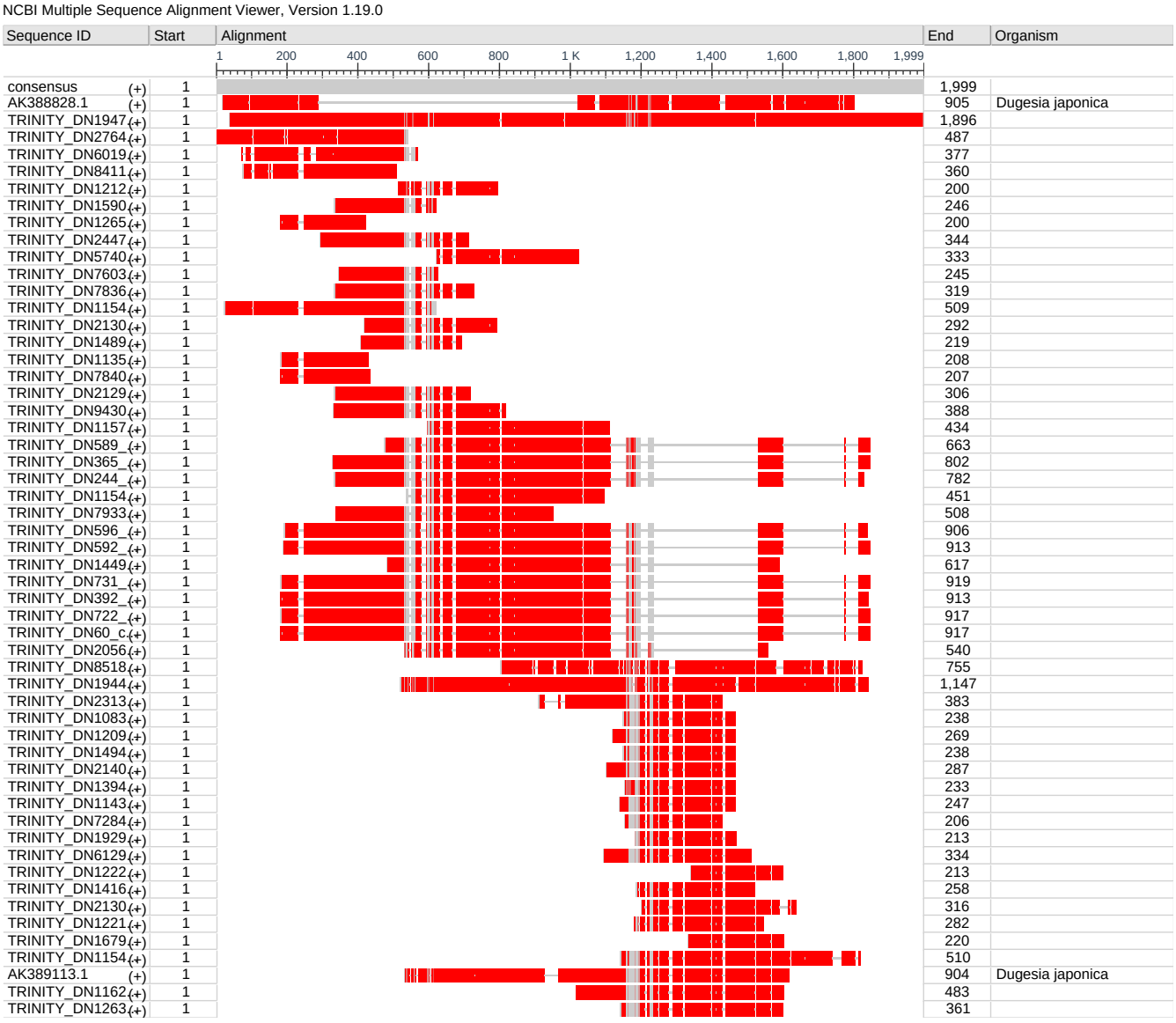

Figure 4. Multiple alignment of contigs identified by using BLAST search to be aligned to known planarian transcripts, AK388828.1 or AK389113.1.