Contents lists available at ScienceDirect

Data in brief

Data Article

Reference quantitative transcriptome dataset for adult *Caenorhabditis elegans*

Allison Piovesan, Francesca Antonaros, Pierluigi Strippoli, Lorenza Vitale, Maria Chiara Pelleri[*], Maria Caracausi

Department of Experimental, Diagnostic and Specialty Medicine, (DIMES), Unit of Histology, Embryology and Applied Biology, University of Bologna, Via Belmeloro 8, 40126 Bologna, BO, Italy

ARTICLE INFO

ABSTRACT

Caenorhabditis elegans is a nematode widely used in biology and genomics as a model organism. We provide an integrated, quantitative reference map for the transcriptome of whole, wild type Bristol N2 strain *C. elegans* worms. The map has been obtained by meta-analysis of 110 gene expression profiles available in Gene Expression Omnibus (GEO) repository and integrated using the computational biology tool Transcriptome Mapper (TRAM). Following probe assignment to the relative locus and intra- and inter-sample normalization (in particular using the scaled quantile method), a mean, consensus reference value is provided for 45,932 transcripts, along with standard deviation. Expression values are all mapped in the context of genomic coordinates. The map provides easy access to relationships among expression values of different genes in this standard condition, highlights genomic segments with relatively high over-/under-expression and may serve as a reference to test for gene expression variation for both individual genes and the whole transcriptome in specific biological conditions (e.g. mutated strains or differently grown worms).

* Corresponding author.
E-mail addresses: allison.piovesan2@unibo.it (A. Piovesan), francesca.antonaros2@unibo.it (F. Antonaros), pierluigi.strippoli@unibo.it (P. Strippoli), lorenza.vitale@unibo.it (L. Vitale), mariachiara.pelleri2@unibo.it (M.C. Pelleri), maria.caracausi2@unibo.it (M. Caracausi).

Specifications table

Subject area	*Biology*
More specific subject area	*Genomics, bioinformatics*
Type of data	*Table*
How data was acquired	*Microarray data repository: Gene Expression Omnibus (GEO) provided by the National Center for Biotechnology Information (NCBI) at https://www.ncbi.nlm.nih.gov/geo/; elaboration: Transcriptome mapper (TRAM) software, version 1.3*
Data format	*Excel Table (.xlsx) - Data analyzed by TRAM software and exported as spreadsheet*
Experimental factors	*Database search, dataset selection, TRAM (Transcriptome Mapper) analysis*
Experimental features	*Meta-analysis of wild type, Bristol N2 strain Caenorhabditis elegans adult worms*
Data source location	*Data sources are listed in the Supplementary Table 1; Meta-analysis results have been obtained in Bologna, Italy, DIMES Department at University of Bologna; TRAM software set up for Caenorhabditis elegans with the results obtained in this analysis has been released*
Data accessibility	*Data sources are available at https://www.ncbi.nlm.nih.gov/geo/; analyzed data are with this article*
Related research article	*L. Lenzi, F. Facchin, F. Piva, M. Giulietti, M.C. Pelleri, F. Frabetti, L. Vitale, R. Casadei, S. Canaider, S. Bortoluzzi et al, TRAM (Transcriptome Mapper): database-driven creation and analysis of transcriptome maps from multiple sources, BMC Genomics. 12 (2011) 121*

Value of the data

- Reference table for a quantitative gene expression value for each of the 45,932 Caenorhabditis elegans transcripts, offering the possibility for immediate establishment of quantitative relative ratio of expression for every pair of desired genes as well as analysis of global patterns of expression with any tool of gene expression profile elaboration.
- Benchmark to identify variation in individual gene expression value following comparison with gene profiles derived by worms in different biological conditions, e.g. different developmental stages, different feeding conditions or treatments, strains with knockdown of specific genes or with any type of genetic difference.
- Possibility to select genes with the desired features of the expression values (high/low, with high or low standard deviation from the mean among a large number of individuals, usefulness as a reference gene in gene expression studies).
- Possibility to select genomic segments with high/low expression values (mean of expression values of the genes contained in the segment), thus also identifying genomic open chromatin domains.
- The quantitative reference values of the enzyme mRNAs might be used in metabolic network models for the validation of hypotheses about the relationships among mRNA levels, corresponding enzymatic proteins and the quantities of their substrates or products obtained by metabolome experiments.

1. Data

1.1. Caenorhabditis elegans *transcriptome map*

Caenorhabditis elegans is a nematode widely used in biology and genomics as a model organism [1,2]. We provide an integrated, quantitative reference map for the transcriptome of whole, wild type Bristol N2 strain *Caenorhabditis elegans* worms. The map has been obtained by meta-analysis of 110 gene expression profiles available in Gene Expression Omnibus (GEO) repository (https://www.ncbi.nlm.nih.gov/geo/) and integrated using the computational biology tool Transcriptome Mapper (TRAM) [3]. Gene expression profiles were derived from expression microarray experiments and fulfilled the described exclusion and inclusion criteria (Materials and Methods section).

Sample identifiers (GEO accession numbers) and main sample features are listed in Supplementary Table 1. Following probe assignment to the relative locus and intra- and inter-sample normalization (in particular using the scaled quantile method), a mean, consensus reference value is provided for 45,932 transcripts, along with standard deviation (Supplementary Table 2). Expression values are all mapped in the context of genomic coordinates. The over-/under-expressed genomic segments shown in Table 1 were selected using the "Map" mode graphical representation. Detailed results are also released within the TRAM software available at: http://apollo11.isto.unibo.it/software/.

1.2. Reference gene search

In the *C. elegans* transcriptome map, the search for reference genes with the described criteria (Materials and Methods section) retrieved 3 loci (Table 2). The *rpl4* locus, encoding 60S ribosomal

protein L4, shows the most favorable combination of high level expression, high number of samples and low standard deviation.

2. Experimental design, materials and methods

2.1. Database search and selection

Caenorhabditis elegans is a nematode widely used in biology and genomics as a model organism [1,2].

A search in GEO gene expression data repository for any available samples listing gene expression values for whole wild type, Bristol N2 strain *C. elegans* worms was conducted in November 2018 querying for: "*Caenorhabditis elegans*"[Organism] AND "Expression profiling by array"[Filter] AND adult. 250 datasets were found, and 50 randomly selected datasets (the first 50 presented by default order) were further studied to identify any individual, pertinent gene expression values list. The criteria for inclusion were: RNA extracted from whole Bristol N2 wild type adult (or young adult) worms at any age (day 2 - day 15); hermaphrodite/male sex. Criteria for exclusion were: larval stage, worms treated with empty or not empty vectors, worms not fed with living *E. coli* or fasting, exposition to DMSO (dimethyl sulfoxide) as vehicle control, grown at 25 °C when 20 °C condition was available.

Although RNA sequencing (RNA-Seq), the other high-throughput method used to assess gene expression, is considered to be more sensitive and to have a broader dynamic range than RNA microarrays [4], the latter remains an accurate tool for measuring the levels of gene expression [5], also offering some specific advantages over RNA-Seq [6], and thus continuing to provide useful data-mining resources.

2.2. TRAM analysis

TRAM software [3] allows the importation of gene expression data from any source (expression microarray, RNA-Seq or proteomic platforms). It performs the integration of all data related to the same biological source by decoding probe set identifiers to gene symbols via UniGene data parsing [7], normalizes data from multiple platforms using intra-sample and inter-sample normalization (scaled quantile normalization) [8], and creates a graphical representation of gene expression profiles along the chromosomes also determining the statistical significance of differential expression of chromosomal segments in comparison with the other segments in the biological condition studied. When two conditions A and B are compared, it is able to calculate differential expression of each segment between them. The statistical method used by TRAM to this aim is hypergeometric distribution, a recognized algorithm able to test the probability 'p' that colocalization of over-/under-expressed genes within the same chromosomal segment may be due to chance [3].

We used an updated version of TRAM (TRAM 1.3) [8], including enhanced resolution of gene identifiers through an updated NCBI Gene database, updated platform annotation files and UniGene data parsing. TRAM set up for *C. elegans* was performed in November, 2018 following the software user guide. The gene expression profiles fulfilling the criteria for exclusion and inclusion were imported as Pool A. Pool B is available for comparisons with a different biological condition.

The value for each locus is defined as the mean value of all available values for that locus. The genome wide gene expression median value was used in order to determine percentiles of expression for each gene.

Using the "Map" mode graphical representation we searched for over-/under-expressed genome segments which have a window size of 20,000 bp (base pairs) and a shift of 10,000 bp. These values were chosen according to the ratio between human and *C. elegans* mean gene length (as determined by searching the recent GeneBase database available for humans [9,10] and running an analogous NCBI Gene *C. elegans* data import in GeneBase for worms - data unshown). The expression value for each genomic segment is the mean of the expression values of the loci included in that segment. Loci for which mean value was derived from less than five biological samples were not considered. A segment is defined as over-/under-expressed by descriptive statistics if it has an expression value within the highest and the lowest 2.5th percentile among all genomic segments and contains at least three genes which have an expression value within the highest and the lowest 2.5th percentile (default

Table 1

The genomic segments significantly over-/under-expressed in the _C. elegans_ transcriptome map. Over-expressed genes are in bold, under-expressed genes are with an asterisk and in bold. "+" or "-" signs indicate a value above or below the genome median, respectively. In order to simplify, segments with over-/under-expressed gene content fully included in a segment listed here are not shown.

#	Chromosome	Segment Start	Segment End	Expression Value	q-value	Genes in the segment
Over-expressed segments						
1	chrIV	11,330,001	11,350,000	2308.35	0.00003906	F54E12.2+ **his-55 his-56 his-58 his-57** klp-12+
2	chrIII	10,970,001	10,990,000	2,113.98	0.00002970	dhc-4- **col-92 col-93 col-94**
3	chrV	11,070,001	11,090,000	1,762.53	0.00002970	**act-3 act-2 act-1** Y42A5A.1-
4	chrIV	11,390,001	11,410,000	1,588.97	0.00010384	tag-89- dsl-6- **his-66** cyp-31A2+ **his-63 his-64**
5	chrIV	7,470,001	7,490,000	1,482.27	0.00000041	**plk-3** F55G1.6 + F55G1.9+ rod-1+ **his-61 his-62 his-60 his-59**
6	chrIV	11,320,001	11,340,000	1,223.15	0.00002952	B0035.18 + B0035.6+ his-47+ **his-48** his-46+ his-45 + Cel.6357- F54E12.2+ **his-55 his-56 his-58**
7	chrI	2,060,001	2,080,000	1,035.73	0.00014806	Y37E3.1+ rpb-10+ moag-4+ arl-13- **rla-1 Y37E3.8 phb-1**
8	chrV	8,880,001	8,900,000	1,029.50	0.00000060	K06C4.1- **his-28 his-27 his-22** his-20+ his-19+ **his-17 his-18** frpr-13-
9	chrMT	1	20,000	975.75	0.00000001	**ND1** ATP6+ **CYTB COX3 ND4 COX1 COX2** ND5+
10	chrIII	7,170,001	7,190,000	917.83	0.00043732	acs-4+ srb-8- srb-7- **rps-14 rpl-36** F37C12.3+ epg-4+ F37C12.1 + F37C12.14- F37C12.10+ **rps-21**
11	chrI	10,550,001	10,570,000	899.64	0.00033432	F25H2.4+ **ndk-1** F25H2.6 + F25H2.7- ubc-25 + F25H2.15+ pas-5+ **rla-0 tct-1** F25H2.14+
12	chrX	7,300,001	7,320,000	772.99	0.00019922	sur-5+ **his-38** T08A9.6- **spp-3** spp-2+ spp-6- spp-4+ **spp-5**
13	chrII	13,810,001	13,830,000	755.25	0.00019922	ZK131.11+ **his-16 his-13** his-12+ his-11+ **his-10** his-9+ his-42+
14	chrV	2,310,001	2,330,000	722.00	0.00005651	Y19D10B.1- Y19D10B.6- **pud-1.2 pud-2.2 pud-3**
15	chrII	8,560,001	8,580,000	634.59	0.00025710	**iff-2** stc-1+ F54C9.3+ col-38+ **rpl-5** bcs-1+ F54C9.7- **puf-5** F54C9.9+
16	chrIV	5,050,001	5,070,000	618.41	0.00014806	**msp-55** C09B9.7- **msp-57 msp-53** R13H9.5 + R13H9.6+ rmd-6+
17	chrIV	8,330,001	8,350,000	611.61	0.00002612	his-29+ **his-30** lys-10- his-31+ **his-32 his-34** lgc-6- lgc-5- **F17E9.5** F17E9.4+
18	chrV	8,530,001	8,550,000	604.67	0.00025710	otpl-5- **his-8 his-7** his-6+ **his-5** his-39+ **otpl-4*** asns-1- stdh-4-
19	chrIII	9,740,001	9,760,000	528.29	0.00001188	T05G5.1+ **iff-1 cdk-1** T05G5.4- T05G5.5+ **ech-6** vps-53+ **rmd-1**
Under-expressed segments						
1	chrIV	5,870,001	5,890,000	5.03	0.00045937	**srv-17* srv-18* srv-19*** srv-20- srv-21- srv-22- **srv-23*** H04M03.11- glf-1-
2	chrI	12,380,001	12,400,000	4.89	0.00338533	gly-16- **T15D6.5*** nhr-77- **glct-3*** T15D6.8- T15D6.9- **T15D6.10*** T15D6.11- T15D6.12-
3	chrV	3,060,001	3,080,000	4.79	0.00257570	srt-15- srt-16- **srh-185*** str-40* C50H11.13- **srt-10*** srt-9- srt-5-
4	chrII	3,690,001	3,710,000	4.73	0.00448442	**srx-101* srx-100* srx-102*** srx-104- srx-105- srx-106- srx-107- srx-108- srx-109- srx-110-
5	chrIV	5,860,001	5,880,000	4.57	0.00198431	spe-27- **srv-17* srv-18* srv-19*** srv-20- srv-21- srv-22-
6	chrV	15,300,001	15,320,000	4.56	0.00257570	**srx-49* srx-48*** T26H8.5- **srz-10*** irld-62- srt-22- nhr-246- ZK1037.13-
7	chrIV	9,480,001	9,500,000	4.35	0.00018020	**cng-3*** gadr-2- **sru-2*** sru-1- sru-6- **sru-3* sru-4***
8	chrV	16,670,001	16,690,000	4.27	0.00338533	str-61- F14F8.8- **srz-2*** srz-1- srz-103- srw-44- **srw-36* srw-43*** srz-102-
9	chrI	12,690,001	12,710,000	4.21	0.00001683	**sra-17*** F28C12.6- **sra-18* sra-19*** sra-20- **sra-21* sra-22*** sra-23- **sra-24*** T06G6.11- T06G6.3-
10	chrI	13,100,001	13,120,000	3.94	0.00134398	

Table 1 (continued)

#	Chromosome	Segment Start	Segment End	Expression Value	q-value	Genes in the segment
						Y26D4A.21- C17H1.2- **pals-2*** pals-12- **pals-4*** **C17H1.1***
11	chrV	2,930,001	2,950,000	3.82	0.00257570	C31B8.16- C31B8.1- srh-247- **srw-141*** srw-143- srw-137- **srh-87*** **srw-128***
12	chrV	9,820,001	9,840,000	3.81	0.00134398	**sru-32*** **str-193*** str-2- sru-40- **sru-38*** srsx-21-
13	chrIV	14,140,001	14,160,000	3.57	0.00134398	srz-31- **srz-30*** oac-37- **oac-38*** H12I19.8- **R05A10.8***
14	chrV	2,740,001	2,760,000	3.54	0.00002042	**srg-61*** srg-62- **srx-25*** **srx-24*** **srx-26*** srx-27- **srx-28*** srg-58-
15	chrV	16,680,001	16,700,000	3.35	0.00001183	srw-44- **srw-36*** **srw-43*** srz-102- **srw-42*** **srz-101*** **srw-41***
16	chrV	16,460,001	16,480,000	3.23	0.00018020	srh-142- T08G3.7- sru-44- **srh-138*** **srw-35*** **srx-126*** **T08G3.11***
17	chrV	2,940,001	2,960,000	3.21	0.00030016	srw-143- srw-137- **srh-87*** **srw-128*** srw-122- **srh-248*** **srw-142*** srw-144-
18	chrV	2,950,001	2,970,000	3.15	0.00198431	srw-122- **srh-248*** **srw-142*** srw-144- srw-138- srh-88- **srw-116***
19	chrV	6,800,001	6,820,000	3.11	0.00134398	**dmsr-10*** **dmsr-11*** dmsr-12- T15B7.10- **dmsr-14*** dmsr-13-
20	chrIV	9,280,001	9,300,000	3.01	0.00040976	nhr-267- **nhr-264*** **F49C12.1*** **F49C12.2***
21	chrI	13,110,001	13,130,000	2.73	0.00023681	**pals-4*** **C17H1.1*** **pals-11***

parameters) among all genes. The statistical significance of the over-/under-expression of the over-/under-expressed genome segments, respectively, is then assessed by statistical tests based on hypergeometric distribution, a recognized algorithm able to test the probability "p" that colocalization of three over-/under-expressed genes within the same chromosomal segment may be due to chance and corrected for possible multiple comparisons causing false discovery rate (FDR) due to the high number of segments in a genome. A segment was considered to be statistically significantly over-/under-expressed for q < 0.05 [3,8].

Apart from gene expression analyses, these data might be used in metabolic network models [11] for the validation of hypotheses about the relationships among mRNA levels, corresponding enzymatic proteins and the quantities of their substrates or products obtained by metabolome experiments [12], also using genomic location data [13]. In addition these data might be used to calculate the recently described transcriptomic GC content [14], i.e. the guanine-cytosine percentage calculated in the mRNA amount actually expressed in a tissue or cell type, to search for variation of this parameter among different biological conditions.

Sample expression values equal to or lower than "0" ($\leq$0) will be thresholded by TRAM [15,16] to 95% of the minimum positive value present in that sample, in order to obtain meaningful numbers when dividing "Sample Pool A" values by "Sample Pool B" values. Assuming that in these cases an expression level is too low to be detected under the used experimental conditions, this transformation still allows a ratio between values in Pool A and values in Pool B to be obtained, which is useful to highlight differential gene expression.

Table 2
List of the best predicted reference genes from the whole adult *C. elegans* quantitative transcriptome map. Chr = chromosome; SD = standard deviation.

Gene name	Chr	Expression Value	Sample Number	SD as % of Expression	Description
rpl4	chrI	2603.77	102	19.85	60S ribosomal protein L4
riok-3	chrIII	165.73	61	18.99	Serine/threonine-protein kinase RIO3
Y48G1C.1	chrI	149.21	55	19.88	hypothetical protein

2.3. Reference gene search

The ideal reference, or control, gene for the study of gene expression in a given organism should be expressed at a medium-high level for easy detection and at a constant/stable level throughout different samples also undergoing different treatments [17]. A search of reference genes best suitable for the study of whole adult *C. elegans* was performed in the transcriptome map created as described above using the following parameters in combination: expression value > 100 in order to select genes expressed above the mean value (that is posed equal to 100), therefore at an appreciable level; number of samples $\geq$ of half the total number of samples of the map in order to select commonly expressed genes ($\geq$55); standard deviation (SD), expressed as a percentage of the mean value, $\leq$20 in order to identify genes with a very low expression variation among different samples [17].

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2019.104152.

References

[1] C. elegans, Sequencing Consortium, Genome sequence of the nematode C. elegans: a platform for investigating biology, Science 282 (1998) 2012–2018.
[2] J. Apfeld, S. Alper, What can we learn about human disease from the nematode C. elegans? Methods Mol. Biol. 1706 (2018) 53–75. https://doi.org/10.1007/978-1-4939-7471-9_4.
[3] L. Lenzi, F. Facchin, F. Piva, M. Giulietti, M.C. Pelleri, F. Frabetti, L. Vitale, R. Casadei, S. Canaider, S. Bortoluzzi, et al., TRAM (Transcriptome Mapper): database-driven creation and analysis of transcriptome maps from multiple sources, BMC Genomics 12 (2011) 121. https://doi.org/10.1186/1471-2164-12-121.
[4] S. Zhao, W.P. Fung-Leung, A. Bittner, K. Ngo, X. Liu, Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells,, PLoS One 9 (2014) e78644, https://doi.org/10.1371/journal.pone.0078644.
[5] Y. Zhang, O.S. Akintola, K.J. Liu, B. Sun, Membrane gene ontology bias in sequencing and microarray obtained by housekeeping-gene analysis, Gene 575 (2016) 559–566. https://doi.org/10.1016/j.gene.2015.09.041.
[6] J.A. Timmons, P.J. Atherton, O. Larsson, S. Sood, I.O. Blokhin, R.J. Brogan, C.H. Volmar, A.R. Josse, C. Slentz, C. Wahlestedt, et al., A coding and non-coding transcriptomic perspective on the genomics of human metabolic disease, Nucleic Acids Res. 46 (2018) 7772–7792. https://doi.org/10.1093/nar/gky570.
[7] L. Lenzi, F. Frabetti, F. Facchin, R. Casadei, L. Vitale, S. Canaider, P. Carinci, M. Zannotti, P. Strippoli, UniGene Tabulator: a full parser for the UniGene format, Bioinformatics 22 (2006) 2570–2571. https://doi.org/10.1093/bioinformatics/btl425.
[8] M.C. Pelleri, C. Cattani, L. Vitale, F. Antonaros, P. Strippoli, C. Locatelli, G. Cocchi, A. Piovesan, M. Caracausi, Integrated quantitative transcriptome maps of human trisomy 21 tissues and cells, Front. Genet. 9 (2018) 125. https://doi.org/10.3389/fgene.2018.00125.
[9] A. Piovesan, M. Caracausi, F. Antonaros, M.C. Pelleri, L. Vitale, GeneBase 1.1: a Tool to Summarize Data from NCBI Gene Datasets and its Application to an Update of Human Gene Statistics, Database, Oxford), 2016, 2016, https://doi.org/10.1093/database/baw153.
[10] A. Piovesan, M. Caracausi, M. Ricci, P. Strippoli, L. Vitale, M.C. Pelleri, Identification of minimal eukaryotic introns through GeneBase, a user-friendly tool for parsing the NCBI Gene databank, DNA Res. 22 (2015) 495–503. https://doi.org/10.1093/dnares/dsv028.
[11] J. Schellenberger, R. Que, R.M. Fleming, I. Thiele, J.D. Orth, A.M. Feist, D.C. Zielinski, A. Bordbar, N.E. Lewis, S. Rahmanian, et al., Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0, Nat. Protoc. 6 (2011) 1290–1307. https://doi.org/10.1038/nprot.2011.308.
[12] L.M. Johnson, L.M. Chandler, S.K. Davies, C.F. Baer, Network architecture and mutational sensitivity of the C. elegans metabolome, Front Mol. Biosci. 5 (2018) 69. https://doi.org/10.3389/fmolb.2018.00069.

[13] M. Caracausi, V. Ghini, C. Locatelli, M. Mericio, A. Piovesan, F. Antonaros, M.C. Pelleri, L. Vitale, R.A. Vacca, F. Bedetti, et al., Plasma and urinary metabolomic profiles of Down syndrome correlate with alteration of mitochondrial metabolism, Sci. Rep. 8 (2018) 2977. https://doi.org/10.1038/s41598-018-20834-y.

[14] A. Piovesan, M.C. Pelleri, F. Antonaros, P. Strippoli, M. Caracausi, L. Vitale, On the length, weight and GC content of the human genome, BMC Res. Notes 12 (2019) 106. https://doi.org/10.1186/s13104-019-4137-z.

[15] M. Caracausi, A. Piovesan, L. Vitale, M.C. Pelleri, Integrated transcriptome map highlights structural and functional aspects of the normal human heart, J. Cell. Physiol. 232 (2017) 759–770. https://doi.org/10.1002/jcp.25471.

[16] L. Vitale, A. Piovesan, F. Antonaros, P. Strippoli, M.C. Pelleri, M. Caracausi, A molecular view of the normal human thyroid structure and function reconstructed from its reference transcriptome map, BMC Genomics 18 (2017) 739. https://doi.org/10.1186/s12864-017-4049-z.

[17] M. Caracausi, A. Piovesan, F. Antonaros, P. Strippoli, L. Vitale, M.C. Pelleri, Systematic identification of human housekeeping genes possibly useful as references in gene expression studies, Mol. Med. Rep. 16 (2017) 2397–2410. https://doi.org/10.3892/mmr.2017.6944.