Gray whale transcriptome reveals longevity adaptations associated with DNA repair, autophagy and ubiquitination

Toren D[1,2], Kulaga A[2,3], Jethva M[1], Rubin E[1], Snezhkina AV[4], Kudryavtseva AV[4], Nowicki D[5], Tacutu R[2,6], Moskalev AA[4,7,8]*, Fraifeld VE[1]*

[1] The Shraga Segal Department of Microbiology, Immunology and Genetics, Center for Multidisciplinary Research on Aging, Ben-Gurion University of the Negev, POB 653, Beer Sheva, 8410501, Israel
[2] Computational Biology of Aging Group, Institute of Biochemistry, Romanian Academy, Bucharest, 060031, Romania
[3] Humenhance OÜ, Tallinn, Estonia
[4] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991, Russian Federation
[5] Institute of MMS of NASU, Center for Cybernetics, Kiev, Ukraine
[6] Chronos Biosystems SRL, Bucharest, 060117, Romania
[7] Institute of Biology of Komi Science Center of Ural Branch of RAS, Syktyvkar, 167982, Russian Federation
[8] Moscow Institute of Physics and Technology, Dolgoprudny, 141701, Russian Federation

***Corresponding authors**

Abstract

One important question in aging research is how differences in genomics and transcriptomics determine maximum lifespan in various species. Despite recent progress, much is still unclear on the topic, partly due to the lack of samples in non-model organisms and due to challenges in direct comparisons of transcriptomes from different species. The novel ranking-based method that we employ here is used to analyze gene expression in the gray whale and compare its *de novo* assembled transcriptome with that of other long- and short-lived mammals. Gray whales are among the top 1% longest-lived mammals. Despite the extreme environment, or maybe due to a remarkable adaptation to its habitat (intermittent hypoxia, Arctic water and high pressure), gray whales reach at least the age of 77 years. In this work, we show that long-lived mammals share common gene expression patterns between themselves, including high expression of DNA maintenance and repair, autophagy, ubiquitination, apoptosis, and immune responses. Additionally, the level of expression for gray whale orthologs of pro- and anti-longevity genes found in model organisms is in support of their alleged role and direction in lifespan determination. Remarkably, among highly expressed pro-longevity genes many are stress-related, reflecting an adaptation to extreme environmental conditions. The conducted analysis suggests that the gray whale potentially possesses high resistance to cancer and stress, at least in part ensuring its longevity. This new transcriptome assembly also provides important resources to support the efforts of maintaining the endangered population of gray whales.

Keywords: longevity, gray whale, transcriptomics, comparative analysis, DNA repair.

INTRODUCTION

Long-lived animals represent a unique model for investigating the evolution of longevity. The recently conducted transcriptome analysis of the longest-lived mammal, the bowhead whale (*Balaena mysticetus*), showed that transcriptional patterns could, up to a certain extent, explain its extraordinary longevity and resistance to cancer and other age-related diseases [1, 2]. The gray whale (*Eschrichtius robustus*) is also among top 1% longest-lived mammals. It ranks 8th out of 1012 mammalian species with known maximum lifespan [3]. Despite the extreme environment, or maybe due to a remarkable adaptation to its habitat (intermittent hypoxia, cold Arctic water and high pressure), gray whales reach at least the age of 77 years, according to currently available data [3].

The *Eschrichtius robustus* is the only member of the *Eschrichtiidae* family from the order Cetacea [4]. It is considered a "living fossil" because of its short, coarse baleen plates and lack of a dorsal fin [5]. Gray whales were almost extinct in the middle of the 20th century and despite being protected by law (limited hunting for food being permitted only for the indigenous population of Chukotka) they are still considered to be an endangered species [6]. As such, having the opportunity to get insights from its transcriptome is of great importance to aging research.

Here, we comprehensively analyzed the transcriptome of the gray whale in two tissues (liver and kidney), focusing on the possible links between longevity and the expression of individual genes or group of genes. For this purpose, we also compared the gray whale transcriptome [7] with that of two other whales, bowhead whale (*Balaena mysticetus*) and minke whale (*Balaenoptera acutorostrata*), and with the transcriptomes of several mammalian species of different longevities.

RESULTS AND DISCUSSION

The gray whale transcriptome assembly and annotation

Kidney and liver mRNA from a young-adult female whale were sequenced using the MiSeq System Illumina sequencer. *De novo* transcriptomes were assembled, yielding in total 114,233 contigs. Orthologous protein-coding genes encoded in the transcriptome were identified using the Sprot [8] algorithm and BLASTing [9] against UniProt [10] sequences. In total, we identified 12,072 protein-coding genes with orthologs in other mammalian species (of these 11,456 are expressed in liver and 8,363 in kidney), with a high overlap between tissues (64% of the total). In addition to the identified genes, a large number of the unidentified contigs also fit some of the required criteria for being coding sequences (certain length and expression level). For the full list of unknown genes, please see Dataset S1.

Enrichment of top-most expressed genes

To understand which gene subsets are most expressed in both tissues and/or may be representative for the most active processes in an organism, we next conducted a Gene Ontology (GO) enrichment analysis of the top 100 genes with the highest contig count compared to all genes (background). Please see Dataset S2 for the list of all contigs and Dataset S3 for the enrichment results. Several interesting features were observed for both liver and kidney. First, we found a high presence of GO categories related to extracellular matrix (ECM), cell-cell/cell-ECM interactions and exosomes. Interestingly, this is in line with our previous analysis, which highlighted the potential importance of these categories (specifically, focal adhesion and adherens junction proteins) in lifespan determination and in linking the human longevity and age-related diseases [11]. Second, the top-most expressed genes were significantly enriched in categories related to immuno-inflammatory responses, organ regeneration and regulation of cell proliferation. This could be especially important for extremely large animals, which needed to develop tumor suppressor mechanisms in order to compensate for having more somatic cells, and hence a higher risk to develop cancer [12]. Finally, as expected, based on observations in other species examined so far [13], many top-expressed genes in the gray whale fall into the mitochondria-associated categories. The full list of enriched categories is presented in Dataset S3 for liver and kidney, respectively.

Unknown genes

Aside from the 12,072 protein-coding genes identified in the gray whale transcriptome, 35% of all assembled contigs remained unidentified. Overall, this percentage of unannotated sequences is not unexpectedly large in the case of *de novo* transcriptome assembly [1, 2]. Generally, unidentified sequences might be the result of (i) mapping errors or (ii) still uncharacterized sequences. To further investigate unannotated sequences, while excluding false positives due to mapping, we next selected only those contigs with a high-count number (hence highly expressed) and with a sequence length comparable to that of a common mRNA size (threshold of >200 bp). In total, almost 600 unannotated sequences fit the above criteria. A big part of the unknown genes without annotations have a very high expression level in the gray whale transcriptome. For example, among the first 1000 top-expressed genes, 92 are unannotated, suggesting that their functional significance should be further investigated (Dataset S1).

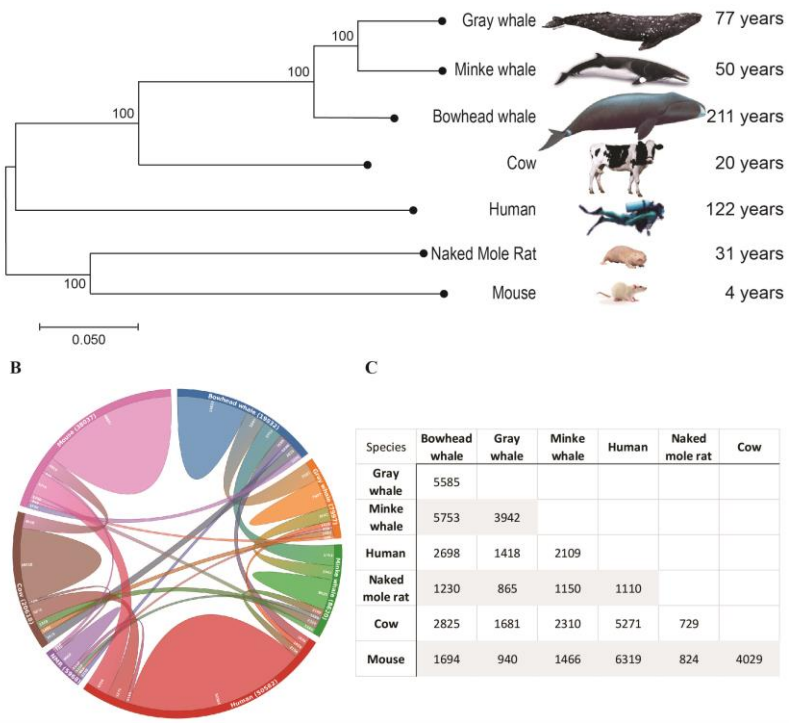
To look closer at these unannotated sequences, we chose the 20 top-expressed transcripts and manually BLASTed them [9] against the NCBI RefSeq nucleotide database [14], searching for similar sequences/domains/motifs (Table S1). Interestingly, all these transcripts were also found highly expressed in other transcriptomes, including those of the bowhead whale and minke whale (data not shown). Swiss-

Prot [10], Pfam [15] and the MEME Suite [16] tools did not reveal any significant domain or motif matches in the top-expressed unannotated sequences. Several sequences of predicted/hypothesized proteins or transcripts with high similarity were however found by BLAST in other species. Among them are genes hypothetically related to ribosomal RNA, metabolic processes (Zinc fingers), ERK signaling and cytoskeleton remodeling (ACTB), detoxification of reactive oxygen species, ubiquitination, proliferation, differentiation, and carcinogenesis (PRDX3, CALR), immune response (histocompatibility complex class II and extended class II sub-region, NDRG1, PTMA, GPS2), mitochondria, ATP metabolism and iron binding (PAH, CYP1A1), senescence and autophagy in cancer (CREG1, CAAX box protein), lipoprotein and cholesterol metabolic process (APOA2), regulation of cell differentiation, serpin family (SERPINF2). Interestingly, many of above-mentioned genes are involved in aging associated pathways [17-19].

A broader question is whether some of these unannotated genes are essential for basic/fundamental biological processes. In this regard, Hutchison et al. have recently synthesized a functionally viable artificial genome containing only 473 essential genes. Of them, almost a third (149 genes) had no known biological functions [20]. The fact that unknown genes still exist even in an extremely reduced genome supports our suggestion that many unannotated genes found in the gray whale transcriptome should be further investigation as they might be involved in fundamental processes.

Novel approach: Comparative transcriptome analysis of gray whale vs other mammals

To identify expression patterns that are species-specific or shared across species, we then undertook a comparative analysis of the transcriptomes of the gray whale and of six other mammalian species, including two other whales (bowhead and minke whales) and four terrestrial mammals (naked mole rat, mouse, cow, and humans) (Fig. 1). These species were selected because (i) the bowhead and minke whales are genetically close to the gray whale, and bowhead whale is the longest-lived mammal (according to existing records); (ii) the naked mole rat arguably presents characteristics of “negligible senescence” and has an impressive lifespan of almost 8 times longer than rodents with comparable body size; (iii) humans are the second longest-lived mammal (according to current records) and are considered among species with exceptional longevity; (iii) the mouse is a good reference for a short-lived organism; and (iv) the cow is a species with a maximum lifespan that is close to the average within the mammalian class [3] and the most evolutionarily related to whales among modern land mammals [21].



Species	Bowhead whale	Gray whale	Minke whale	Human	Naked mole rat	Cow
Gray whale	5585					
Minke whale	5753	3942				
Human	2698	1418	2109			
Naked mole rat	1230	865	1150	1110		
Cow	2825	1681	2310	5271	729	
Mouse	1694	940	1466	6319	824	4029

Fig. 1. Relationship of the gray whale to other mammalian species. (A) Phylogenetic tree of species selected for analysis. The tree with the highest log-likelihood is shown. The percentage of trees in which the associated taxa clustered together (bootstrap values) is shown next to the branches. **(B)** Overlap of Uniref90 protein clusters between the examined species. Uniref90 clusters (which contain proteins with 90% sequence similarity) were predicted from open reading frames (ORFs) extracted from the coding transcripts. The Gray whale transcriptome was assembled in the current study (please see methods), while transcriptomes for other species were taken from previous studies, publicly available at the NCBI database. Presented in brackets is the number of Uniref90 entries predicted from each transcriptome. **(C)** Overlap of Uniref90 protein clusters between the examined species in table format.

Unfortunately, expression levels for individual genes from transcriptomes of different species cannot be compared directly. Additionally, both absolute and normalized expression values could be affected by technical issues. The main culprits are generally the heterogeneity in sequencing methods and/or sequencing equipment, variation in the sample preparation protocols, potential errors in alignments and transcriptome assemblies [22]. To minimize these biases, in our analysis: (i) comparison of species was performed at the level of groups of functionally-linked genes (e.g., GO categories-- biological processes) instead of comparing individual genes directly; (ii) within each species, rankings of normalized transcript counts were used (transcripts per million, TPM) instead of absolute contig counts. For this purpose, the protein-coding transcripts were grouped by GOs and their normalized counts were computed within each GO term. The counts were then ranked, from 1 (the GO with highest composite-expression) to 9779 (the GO with lowest composite-expression) in the gray whale's transcriptome. Rank values for all species examined are presented in Dataset S4.

To validate the ranking-based approach we compared the intra-species variance in the ranks of a given GO term to the inter-species variance. As expected, the intra-species variance was much lower than the inter-species variance. Phylogenetically closely related species (in our case, gray, bowhead and minke whales) had more similar ranks of a given GO term than more distant species (specifically, whales, on the one hand, and mice and cows, on the other hand). The Spearman correlation analysis indicated a higher similarity between GO transcription level ranks in the same species or closely related whale species (Fig. S1). The values of Spearman's rank correlation coefficients were 0.97 between the two mice experiments, 0.98 between cow experiments, and ~0.85 for the whales' category, with a high statistical significance ($p < E-25$) for both liver and kidney tissues. As expected, the correlation coefficients between the ranks of the gray whale and those of other mammalian species (naked mole rat, humans, mouse, cow) were much lower (~0.6, $p < E-25$). These results provide an indirect (however, rather strong) evidence for the validity of GO rank values for comparison of transcriptomes.

Since whales are among the longest-lived mammals, we specifically analyzed the shared processes that are highly expressed in long-lived species, while being lowly expressed in shorter-lived mammals. Our analysis of ranks showed that all longed-lived mammals (whales, humans, naked mole rats) correlate positively between them and negatively with short-lived species (cows and mice) in the following GO categories: DNA maintenance and repair (Fig. 2 A1, A2), autophagy (Fig. 2 B1, B2), ubiquitination (Fig. 2 C1, C2), and to some degree apoptosis (GO categories: "Positive regulation of intrinsic apoptotic signaling pathway in response to DNA damage", "Endonucleolytic cleavage", "Positive regulation of endodeoxyribonuclease activity", "Positive regulation of cysteine-type endopeptidase activity involved in execution phase of apoptosis") (Dataset S4). Also, a high similarity between long-lived species was found in the expression of immune response-related GO terms ("Positive regulation of interleukin-2 production", "Positive regulation of T cell receptor signaling pathway", "Positive regulation of activated T cell proliferation"). This is in line with previous findings, with a high expression of genes associated with immune response being recently reported for the bowhead whale transcriptome [1, 2].

Notably, high expression ranks for DNA maintenance and repair, autophagy, ubiquitination (Fig. 2A-C), apoptosis, positive regulation of T cell receptor signaling pathway and activated T cell proliferation (Dataset S4) were also observed for the naked mole rat, a rodent species with exceptional longevity and a remarkable resistance to cancer [23]. In contrast, in the mouse and cow transcriptome, the expression ranks of the aforementioned categories were much lower, whereas in humans, the ranks of most GO terms were closer to whales and naked mole rats (Fig. 2A-C; Dataset S4).

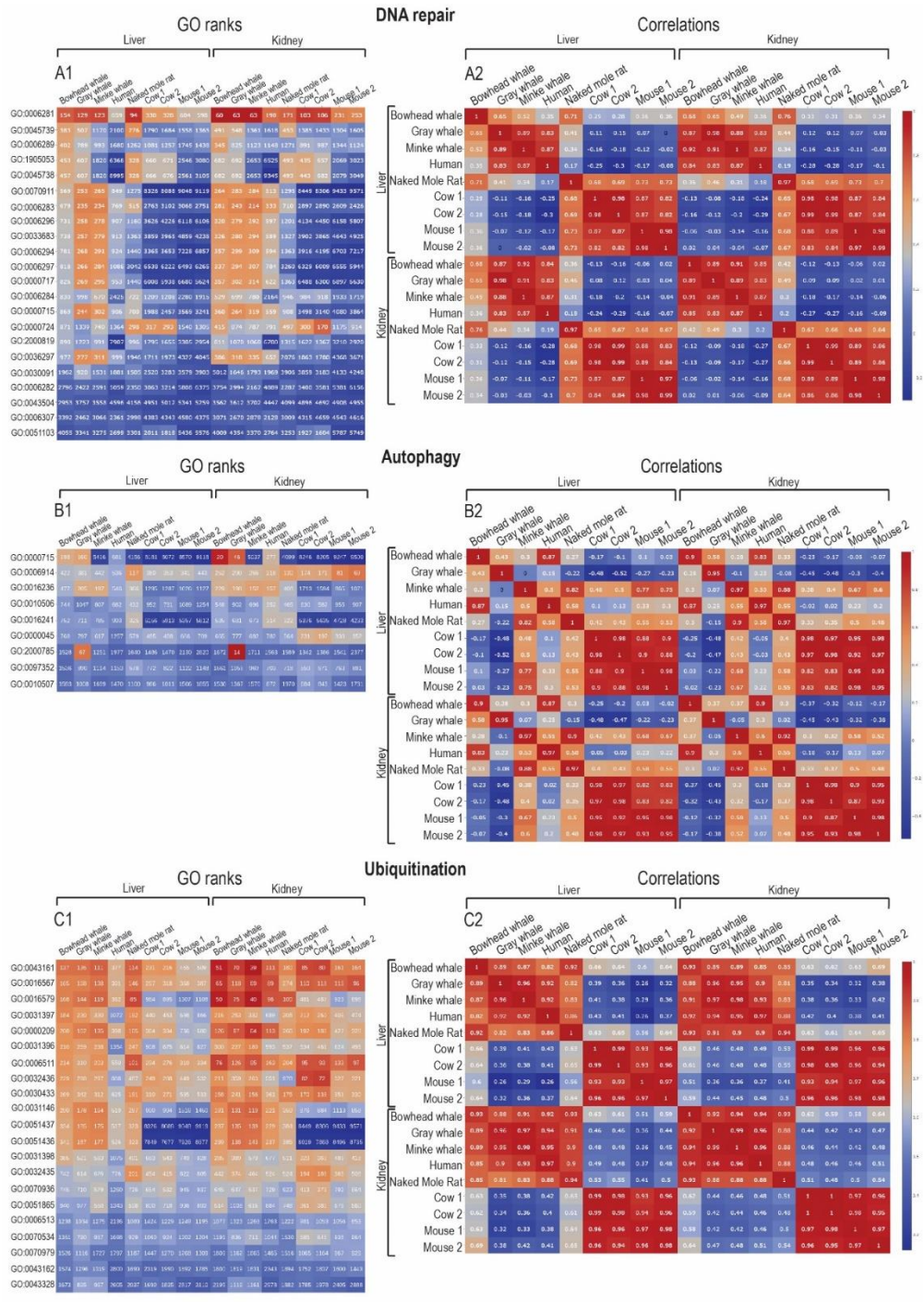


Fig. 2. Heatmap of cross-species transcriptome comparative analysis for GO terms. Presented in the figure are the following categories: **A. DNA repair; B. Autophagy; C. Ubiquitination.** Both liver and kidney tissues are included for all the seven species compared. The same rank ranges and normalization (transcripts per million reads, TPMs) are used for all species. **A1, B1, C1:** Ranks of the top 1,000 expressed GO categories. Ranks range between 1 (top expressed GO term) and 6,260 (least expressed GO term). In the color scheme gradients of red indicate highest expression, gray -- middle expression, and blue -- lowest

expression. A list of names for the GO terms shown in the figure is available in Dataset S5. **A2, B2, C2:** Correlations of GO ranks between each two species. The Spearman's coefficient is shown in the cells of the heatmap. Color gradient indicates the correlation level, from red (highest correlation) to blue (lowest correlation).

The above-mentioned processes are well known to be involved in anti-cancer mechanisms, for preventing cell transformation or elimination of potentially cancerogenous or cancer cells [12, 24], and in a larger sense in determining mammalian longevity [1, 25-27]. This is highly relevant, because while the bowhead whale is known for its resistance to cancer [12, 19], for the gray whale, such data is still absent. In light of our findings on the similarity in the expression of anti-cancer processes, it would probably be reasonable to expect that the gray whale might also be resistant to cancer. Additional to the anti-cancer similarities, correlation analysis of expression ranks revealed some interesting patterns for certain GO groups when comparing long-lived with short-lived species. For instance, expression ranks of DNA repair processes show a positive correlation within the groups of long-lived (whales, humans, naked mole rats; $p < 5E\text{-}4$) and short-lived (cows and mice, 2 transcriptomes each; $p < 5E\text{-}4$) species, respectively (Fig. 2 A2). Complementarily, there is a negative correlation for expression ranks between long- and short-lived species ($p < 5E\text{-}4$). A similar result can also be observed for autophagy, ubiquitination (Fig. 2 A2, B2, C2) and other processes like apoptosis and immune response (Dataset S4). Altogether, considering the above high correlations of longevity-associated processes in the gray whale and other long-lived species, the currently accepted record for the gray whale maximum lifespan of 77 years might be an underestimate, stemming from limited available information.

Apart from a high similarity of overall transcription patterns with bowhead and minke whales, the gray whale exhibits a number of particularly high expression ranks for certain gene categories, including GOs relevant to several longevity-related processes. Among them are GO categories like ATP synthesis coupled proton transport, cilia-related processes, regulation of autophagosome assembly, immune responses (the latter two being much higher even than in the bowhead whale), regulation of Wnt signaling pathway, the sensory processes related to the inner ear, cardiac muscle cell differentiation, and neural precursor cell proliferation (see Dataset S4).

Analysis of longevity-associated genes (LAGs) in the gray whale's transcriptome

With regards to longevity determination, hundreds of genes have been identified to have an impact, when genetically manipulated, on the lifespan of model organisms like yeast, worm, fruit fly and mouse. These genes have been previously defined as longevity-associated genes (LAGs) and include two categories: a)

LAGs which promote longevity, and b) LAGs which reduce lifespan and promote an accelerated aging phenotype (a phenotype that along with shorter lifespan includes features of premature aging) [27, 28].

Although intuitive, the hypothesis that in long-lived species, pro-longevity genes should be expressed at a higher level than anti-longevity genes, has not been fully examined. Here, we tested this assumption by analyzing the level of LAG expression in the gray whale transcriptome. Nucleotide sequences for all LAGs in the GenAge database [3] were extracted from RefSeq database [14] and BLASTed against the gray whale transcriptome. In total, we used Vsearch [29] to look for 1,422 known LAGs discovered in mouse, fruit fly, and roundworm, of which 601 are pro-longevity and 821 anti-longevity genes. After manual curation and the removal of redundant genes, our analysis shows that 113 (19%) of the transcripts with detectable expression in the gray whale correspond to known pro-longevity genes and only 77 (9%) to anti-longevity genes (Dataset S5).

A noteworthy finding is the lower than expected number of LAG homologues found both for anti- and pro-longevity genes in the gray whale, particularly in view of the assumed high evolutionary conservation of LAGs [27]. This however has a technical explanation, namely the fact that the gray whale did not have a reference genome until now, and that the transcriptome analysis was performed *de novo*, whereas for model organisms, there are already numerous well annotated transcriptomes and genomes. Thus, an incomplete mapping of LAGs is to be expected due to this reason.

In our analysis, in all transcriptomes examined, we found more pro- than anti-longevity genes being expressed. This is despite the fact that, based on the GenAge database [3] (currently the most comprehensive repository in terms of LAGs) more anti-longevity genes than pro-longevity genes were reported in the literature for each model organism. Several explanations for this finding can be suggested: 1) some anti-longevity genes might not be in the genome entirely because they have been evolutionarily less favored in long-lived species; 2) some anti-longevity genes might have a very small expression that is silenced/less detectable in young adults (which is also the case of our gray whale samples) but they might be expressed later in life, a phenomenon known as antagonistic pleiotropy; 3) some anti-longevity genes might be expressed through the entire lifespan, however at a constantly low level; and finally, 4) some anti-longevity genes might be expressed only on a conditional basis, for example in stress responses. For the full list of expressed pro- and anti-longevity genes please see Dataset S5.

Two additional points should be stressed in this context: 1) first, as we previously showed, despite high evolutionary distances, the lifespan effects obtained when manipulating orthologous LAGs in different model organisms leads mostly to concordant results [27], and 2) second, the experiments resulting in lifespan extension, and even more so, those where overexpressing a LAG results in lifespan extension, are more

definite in terms of evaluating the impact of a given gene on longevity than those resulting in lifespan reduction [27]. This means that focusing on LAGs from overexpression studies will be more definite and unambiguous. Then, if a pro-longevity LAG was found by overexpression, it is intuitive to expect that this LAG is also highly expressed in longed-lived mammals such as gray whales. Complementarily, when the lifespan-extending effect for a given anti-longevity LAG was found by knock-out or knock-down experiments, its expression in adult gray whales should most likely be at a relatively low level or undetectable. The same could be expected for the overexpression experiments of anti-longevity genes which result in lifespan reduction. The following results of our analysis support these suggestions. Indeed, the vast majority of LAGs found through overexpression experiments, whose orthologs were also identified in the gray whale transcriptome, were pro-longevity (n = 30) and only three were anti-longevity (Table 1). Remarkably, the normalized expression level (TPM) of these 30 pro-longevity genes in the gray whale transcriptome was several-fold higher than the average expression of anti-longevity genes (Fig. 3). Even compared to the average gene expression of the whole gray whale transcriptome, pro-longevity genes displayed 2.6-fold and 4.9-fold higher expression in liver and kidney tissues, respectively ($p < 1.37E-6$; $p < 8.42E-6$). Similar results were obtained when comparing the median values of expression (5.8-fold and 6.5-fold increase in liver and kidney, respectively; $p < 4.2E-4$; 3.77E-5) (Figure 3). In contrast to pro-longevity genes, the three anti-longevity genes showed lower than average expression (68% in liver and 34% in kidney; $p < 5E-4$) compared to the whole transcriptome.

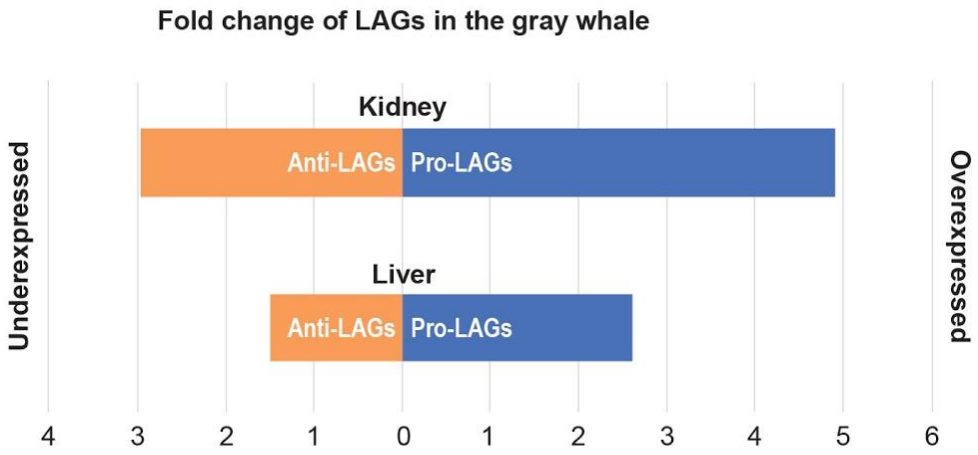


Fig. 3. Fold-change in expression of longevity-associated genes (LAGs) in the *de novo* transcriptome of the gray whale. For this analysis, LAGs were considered only from overexpression experiments in model organisms (*C. elegans*, *D. melanogaster*, *M. musculus*). Displayed are the average fold changes for pro- and anti-longevity genes expressed in the gray whale transcriptome.

As described above, in the gray whale transcriptome, pathways related to DNA repair, autophagy and ubiquitination appear to have an increased activity. In regard to this, it was previously suggested that upregulation of stress response genes could result in pro-longevity effects [18]. This is further supported by our previous study showing that overexpression of stress-related LAGs results in most cases in lifespan extension [27]. Remarkably, in the gray whale transcriptome, a great portion of the pro-longevity LAGs found through overexpression experiments (19 out of 30) are stress-related genes, which display expression values higher than average or median. For example, the heat shock 70kDa protein 1A gene is highly expressed in the gray whale (~25-fold increase, both in liver and kidney), and approximately 118 times higher ($p < 5E-4$) than in a short-lived species (mouse). Interestingly, the expression of this gene is also high in the naked mole rat (8 fold-increase vs average expression of the whole naked mole rat transcriptome). In contrast, the expression of insulin-like growth factor 1 (IGF-1), an anti-longevity gene, is significantly lower in the gray whale (as well as in other whale species) than in the mouse (18.7 and 10.6-fold decrease in liver and kidney, respectively) (Dataset S5).

If knock-out or knock-down of a given LAG results in lifespan extension, it could be expected that its expression in an adult gray whale would be at a relatively low level or undetectable. In line with this expectation, 57 out of the 72 genes (79%) from this group had lower than average expression in liver and 39 out of the 72 genes (54%) had lower than average expression in kidney (Dataset S5). This trend was more obvious for liver (3.8-fold decrease in gene expression) than for kidney (1.2-fold decrease). Complementarily, as shown in Table 1, the gray whale orthologous genes of pro-longevity LAGs, whose overexpression led to an increase in the lifespan of model organisms, displayed a markedly higher expression than average in the gray whale transcriptome as well. At the same time, gray whale orthologous genes of pro-longevity LAGs, whose downregulation results in reduced lifespan in model organisms, were found to be expressed at a low level or to be unexpressed at all (Dataset S5). Our approach opens a new avenue for a wide comparative analysis of LAG expression in long- and short-lived species, which could be an important topic for future investigation. The analysis of the gray whale transcriptome and comparison with other mammalian species suggest that the gray whale potentially possesses high resistance to cancer and stress, in part ensuring its longevity. This new transcriptome assembly and its analysis also provide important resources to support the efforts for maintaining gray whale population.

Table 1. Longevity-associated genes (LAGs) from overexpression experiments found in the *de novo* transcriptome of the gray whale. All transcripts from the gray whale transcriptome were aligned to known nematode, fly and mouse LAGs. For comparison, average TPM for the whole transcriptome is 8.8 for both liver and kidney. Median TPM for the whole transcriptome is 1.2 and 1.7 for liver and kidney, respectively.

Orthologs from	LAG type	Number of genes	Liver Average (TPM)	Liver Median (TPM)	Kidney Average (TPM)	Kidney Median (TPM)
Mice	Pro	13	30	9	67	11
	Anti	1	13.6	13.6	0.4	0.4
Flies	Pro	12	24	6	33	12
	Anti	1	2.8	2.8	2.7	2.7
Worms	Pro	5	4	2	5	6
	Anti	1	0.7	0.7	5.7	5.7
Total	Pro	30	23	7	43	11
	Anti	3	6	3	3	3

Methods

DNA and RNA sampling in tissues

Tissue samples from kidney (N = 2) and liver (N = 2) were acquired from one adult Gray whale (*Eschrichtius robustus*) female on 31 May 2013, at the seashore Lorino, Chukotka Autonomous Okrug, during the 2013 native Eskimo subsistence harvests, by the indigenous population of Chukotka Autonomous Okrug (at the Mechigmen bay of the Bering Sea, Lorino). The Eskimo have permission to hunt gray whales for food and during one of the hunts, tissue biopsies were taken. No animals were killed specifically for the current study. Sample collection and preparation was previously described in Moskalev et al., 2017[7].

Transcriptome sequencing and Assembly

Total RNA was isolated from frozen tissues using the RNeasy Mini Kit (QIAGEN, Germany) according to the manufacturer's protocol. RNA quantification was performed on the NanoDrop 1000 (NanoDrop Technologies, USA), and the RNA integrity was assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies, USA). RNA was further treated with DNase I (Thermo Fisher Scientific, USA) and purified using the RNA Clean & Concentrator-5 kit (Zymo Research, USA). The cDNA libraries were prepared using the Illumina TruSeq RNA Sample Preparation Kit v2 (LT protocol) as described in Moskalev et al., 2014 [18]. The libraries were sequenced on the Illumina MiSeq System (USA) using the MiSeq Reagent Kit v2 for 500 (2 × 250) cycles. On average, fragment size was 300bp (insertion + adapters), with the insertion size about 180bp. Trimming the samples was performed using Trimmomatic[30]. To identify orthologous protein coding transcripts, we employed a BLAST algorithm against the SwissProt database [31]. *De novo* assembly for the four samples was done with the Trinity software stack [32] and yielded 114,233 contigs. The assembly quality was checked with BUSCO [33].

Gene Expression Analysis, Filtering, and Expression

To identify orthologous protein coding transcripts a BLAST algorithm was run against SwissProt [31]. Summarization of the results (*de novo* + annotation) was then generated using an in-house built perl script. Normalization of contig counts was done by computing transcripts per million (TPMs). Genes that were identified and included into subsequent analyses had TPM values of at least 1.

Analysis of differential gene expression

Gene ontology (GO) enrichment analysis was carried out in R, using the TopGO package v2.26.0 [34] (http://bioconductor.org/packages/release/bioc/html/topGO.html). Adjusted p-values lower than 0.05 were considered significant. Only reads with one count per million in at least three samples were included in this analysis.

Comparisons of transcriptomes

To compare the gene expressions in gray whale (*Eschrichtius robustus*) across multiple species, we retrieved liver and kidney RNA-Seq data for the bowhead whale (*Balaena mysticetus*), minke whale (*Balaenoptera acutorostrata*), naked mole rat (*Heterocephalus glaber*), human (*Homo sapiens*), cow (*Bos taurus*) and mouse (*Mus musculus*). In this work, only the transcriptome of the gray whale was assembled *de novo*; for all other species previously published transcriptomes were used.

For each species, kidney and liver RNA-Seq assemblies were chosen and quantified with Salmon [35]. For mouse and cow, two samples (mouse 1 and 2, cow 1 and 2) were taken in order to investigate variation between samples of the same species. To avoid technical errors - only coding transcripts were taken, and the protein sequences translated from them were aligned. To establish a single point of reference - transcripts were mapped to Uniref90 protein clusters [36]. For the protein homology search we used DIAMOND[37]. The used e-value cutoff was 1E-03. To convert results into ranks - the GO processes from Uniref90 clusters annotations were aggregated by GO biological processes. The protein-coding transcripts, from all transcriptomes, are grouped by GOs and their composite TPMs are computed within each GO term. Terms are ranked based on the composite expression from 1 to 9,779 (number of GO terms in the gray whale's transcriptome).

LAGs analysis

The comparison with known mouse, fly and worm LAGs was done using data from GenAge database (build 19, 24/06/2017)[3]. Yeast, bacteria and fungus LAGs were excluded from the analysis, resulting in a total of 1168 genes being included into the analysis. Of these, 486 had pro-longevity annotations and 682 had anti-longevity annotations (note: several LAGs are annotated as both pro- and anti-longevity depending on the

performed genetic interventions). Nucleotide sequences for these LAGs were exported and aligned using the Vsearch [29] (github.com/torognes/vsearch) sequence alignment tools (parameters are provided inside wdl workflows in the github repository: github.com/antonkulaga/gray-whale-expressions). The results were then manually cleaned to remove duplicates.

Statistics

For comparison between different transcriptomes the Spearman rank correlation was used. The Spearman correlation of two samples is defined as a regular (Pearson's) correlation between ranks of values in each sample. To compute the test for significance (*p*-value) using *t* is distributed approximately as Student's t-distribution with n − 2 degrees of freedom under the null hypothesis [38]. A justification for this result relies on a permutation argument[39].

For LAGs analysis we used a hypothesis that expression of each LAG is described by Poisson stream of reads. It obeys Poisson distribution with intensity parameter l computed using known TPM values as well as the total number of reads. Cumulative expression is a sum of large number of Poisson random variables, so it can be approximated using normal distribution according to Central limit theorem. LAGs expression values have distribution like ratio of two Gaussian (normal) variables. We used (numerically computed) CDF of this distribution to get the *p*-values [40].

URLs

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession NTJE000000000. Data available at www.ncbi.nlm.nih.gov/nuccore/NTJE000000000 and www.ncbi.nlm.nih.gov/Traces/wgs/?val=NTJE01.

Code availability: github.com/antonkulaga/gray-whale-expressions

Accession numbers

Bowhead whale: liver tissue SRR1685415, kidney tissue SRR1685390.

Minke whale: liver tissue SRR919296, kidney tissue SRR919295.

Naked mole rat: liver tissue SRR2123747, kidney tissue SRR2124226.

Human: liver tissue GSM1698568, kidney tissue GSM1698570.

Mouse: liver tissue SM1400574, kidney tissue GSM219518.

Cow: liver tissue GSM1020724, kidney tissue GSM1020723.

Acknowledgments

We are very grateful for the help received from Dr. Vered Chalifa-Caspi and Dr. Michal Gordon, from the Bioinformatics Core Facility in the Ben-Gurion University of the Negev. We would also like to thank Dr. Sorel Cahan for his useful comments on the manuscript.

Contributions

This collaborative study was carried out by the research groups of VEF, AAM and RT. All the samples used in the project have been provided by AAM's group.

Bioinformatics data processing and analyses of genetic variation data were carried out by DT, AK, MJ. DN performed the statistical analysis. Library construction, sequencing and genome assembly for the draft reference genome were carried out by AVS, AVK and DT. VEF coordinated the research and together with AAM and RT provided supervision for the project. ER provided additional insights on the bioinformatics methodology. VEF, AAM, DT and RT conceptually designed the experiments and the analyses. DT, RT and VEF wrote and edited the manuscript. All authors have read and approved the final manuscript.

Funding

This work was supported by the Russian Science Foundation grant N 14–50-00060 (to Dr. Moskalev), by the Competitiveness Operational Programme 2014-2020, POC-A.1-A.1.1.4-E-2015 awarded to Dr. Tacutu, and by the Dr. Amir Abramovich Research Fund (to Dr. Fraifeld).

Competing interests

The authors declare no competing financial interests.

REFERENCES

1. Keane, M. et al. Insights into the Evolution of Longevity from the Bowhead Whale Genome. *Cell Rep* **10**, 112-122 (2015).
2. Seim, I. et al. The transcriptome of the bowhead whale Balaena mysticetus reveals adaptations of the longest-lived mammal. *Aging-Us* **6**, 879-899 (2014).
3. Tacutu, R. et al. Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Res* **46**, D1083-D1090 (2018).
4. Wolman, A.A. Gray whale, Eschrichtius robustus (Lilljeborg, 1861). *Handbook of Marine Mammals, vol. 3, The Sirenians and Baleen Whales* (1985).
5. Nollman, J. The Charged Border: Where Whales and Humans Meet (1st éd.). *Henry Holt & Co* (1999).
6. Weller, D., Burdin, A., Wursig, B., Taylor, B. & Brownell Jr, R. The western gray whale: a review of past exploitation, current status and potential threats. (2002).
7. Moskalev, A.A. et al. De novo assembling and primary analysis of genome and transcriptome of gray whale Eschrichtius robustus. *Bmc Evol Biol* **17** (2017).
8. Galgonek, J., Hoksza, D. & Skopal, T. SProt: sphere-based protein structure similarity algorithm. *Proteome Sci* **9** (2011).
9. Johnson, M. et al. NCBI BLAST: a better web interface. *Nucleic Acids Res* **36**, W5-9 (2008).
10. Apweiler, R. et al. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115-D119 (2004).
11. Wolfson, M., Budovsky, A., Tacutu, R. & Fraifeld, V. The signaling hubs at the crossroad of longevity and age-related disease networks. *Int J Biochem Cell B* **41**, 516-520 (2009).

12. Seluanov, A., Gladyshev, V.N., Vijg, J. & Gorbunova, V. Mechanisms of cancer resistance in long-lived mammals. *Nat Rev Cancer* **18**, 433-441 (2018).
13. Mercer, T.R. et al. The Human Mitochondrial Transcriptome. *Cell* **146**, 645-658 (2011).
14. Haft, D.H. et al. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res* **46**, D851-D860 (2018).
15. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427-D432 (2018).
16. Bailey, T.L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-W208 (2009).
17. Ma, S. & Gladyshev, V.N. Molecular signatures of longevity: Insights from cross-species comparative studies. *Semin Cell Dev Biol* **70**, 190-203 (2017).
18. Moskalev, A., Aliper, A., Smit-McBride, Z., Buzdin, A. & Zhavoronkov, A. Genetics and epigenetics of aging and longevity. *Cell Cycle* **13**, 1063-1077 (2014).
19. Tian, X., Seluanov, A. & Gorbunova, V. Molecular Mechanisms Determining Lifespan in Short- and Long-Lived Species. *Trends Endocrin Met* **28**, 722-734 (2017).
20. Hutchison, C.A. et al. Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).
21. Nikaido, M., Rooney, A.P. & Okada, N. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: Hippopotamuses are the closest extant relatives of whales. *P Natl Acad Sci USA* **96**, 10261-10266 (1999).
22. Su, Z. et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* **32**, 903 (2014).
23. Tian, X. et al. High-molecular-mass hyaluronan mediates the cancer resistance of the naked mole rat. *Nature* **499**, 346-349 (2013).
24. Cuervo, A.M. et al. Autophagy and aging: the importance of maintaining" clean" cells. *Autophagy* **1**, 131-140 (2005).
25. Kevei, E. & Hoppe, T. Ubiquitin sets the timer: impacts on aging and longevity. *Nat Struct Mol Biol* **21**, 290-292 (2014).
26. Li, Y. & de Magalhães, J.P. Accelerated protein evolution analysis reveals genes and pathways associated with the evolution of mammalian longevity. *Age* **35**, 301-314 (2013).
27. Yanai, H., Budovsky, A., Barzilay, T., Tacutu, R. & Fraifeld, V.E. Wide-scale comparative analysis of longevity genes and interventions. *Aging Cell* **16**, 1267-1275 (2017).
28. Budovsky, A., Abramovich, A., Cohen, R., Chalifa-Caspi, V. & Fraifeld, V. Longevity network: construction and implications. *Mech Ageing Dev* **128**, 117-124 (2007).
29. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahe, F. VSEARCH: a versatile open source tool for metagenomics. *Peerj* **4** (2016).
30. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
31. Bateman, A. et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158-D169 (2017).
32. Grabherr, M.G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-U130 (2011).
33. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).
34. Alexa, A. & Rahnenfuhrer, J. topGO: Enrichment analysis for Gene Ontology. R package version 2.28. 0. *CRAN* (2016).
35. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods* **14**, 417 (2017).

36. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B. & Wu, C.H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926-932 (2015).
37. Buchfink, B., Xie, C. & Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60 (2015).
38. Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (Cambridge, 1992).
39. Kendall, M. & Stuart, A. (ISBN, 1973).
40. Hinkley, D.V. On the ratio of two correlated normal random variables. *Biometrika* **56**, 635-639 (1969).

Supplementary Information:

This PDF file includes:

Fig. S1. Heatmap representation of correlation coefficients between GO expression ranks of mammalian species.

Table S1. The 20 top-expressed unannotated genes (contigs with a high-count number and a sequence length comparable to the size of common mRNAs) in the *de novo* transcriptome of the gray whale.

Other supplementary materials for this manuscript include the following:

Datasets S1 to S3:

Dataset S1 – Full list of unknown genes from the gray whale transcriptome.

Dataset S2 – List of contigs with TPM in the gray whale transcriptome.

Dataset S3 – TopGO enrichment results.

Dataset S4 – Rank values for all species examined.

Dataset S5 – List of LAGs found in the gray whale transcriptome.