Conservation and architecture of housekeeping genes in the model marine diatom *Thalassiosira pseudonana*

Zhengke Li[1,2], Yong Zhang[2,3], Wei Li[4], Andrew J. Irwin[5] and Zoe V. Finkel[2]

[1]School of Food and Biological Engineering, Shaanxi University of Science and Technology, Weiyang University Park, Xi'an, Shaanxi 710021, China; [2]Department of Oceanography, Dalhousie University, 1355 Oxford St, Halifax, NS B3H 4R2, Canada; [3]College of Environmental Science and Engineering, Fujian Key Laboratory of Pollution Control and Resource Recycling, Fujian Normal University, No. 8 Shangsan Road, Fuzhou, Fujian 350007, China; [4]College of Life and Environmental Sciences, Huangshan University, 39 Xihai Road, Huangshan, Anhui 245041, China; [5]Department of Mathematics & Statistics, Dalhousie University, 1355 Oxford St, Halifax, NS B3H 4R2, Canada

Author for correspondence:
Zhengke Li
Email: zkli@dal.ca

Received: *2 December 2021*
Accepted: *6 February 2022*

New Phytologist (2022)
doi: 10.1111/nph.18039

Key words: diatom, environmental stress, housekeeping genes, internal reference genes, *Thalassiosira pseudonana*.

Summary

- Housekeeping genes (HKGs) are constitutively expressed with low variation across tissues/conditions. They are thought to be highly conserved and fundamental to cellular maintenance, with distinctive genomic features.
- Here, we identify 1505 HKGs in the unicellular marine diatom *Thalassiosira pseudonana* based on an RNA-seq analysis of 232 samples taken under 12 experimental conditions over 0–72 h. We identify promising internal reference genes (IRGs) for *T. pseudonana* from the most stably expressed HKGs.
- A comparative analysis indicates < 18% of HKGs in *T. pseudonana* have orthologs in other eukaryotes, including other diatom species. Contrary to work on human tissues, *T. pseudonana* HKGs are longer than non-HKGs, due to elongated introns. More ancient HKGs tend to be shorter than more recent HKGs, and expression levels of HKGs decrease more rapidly with gene length relative to non-HKGs.
- Our results indicate that HKGs are highly variable across the tree of life and thus unlikely to be universally fundamental for cellular maintenance. We hypothesize that the distinct genomic features of HKGs of *T. pseudonana* may be a consequence of selection pressures associated with high expression and low variance across conditions.

Introduction

Gene expression can vary greatly with environmental conditions and, in multicellular organisms, across tissues. Substantial interest has been paid to a small proportion of genes with nearly constant expression levels across tissues, developmental or cell cycle stages, and conditions, which are commonly referred to as housekeeping genes (HKGs) (Butte *et al.*, 2001; Watson, 2004; Eisenberg & Levanon, 2013). Hundreds to a few thousand HKGs have been identified in humans and other model organisms such as *Arabidopsis* and *Zea mays* L. (Warrington *et al.*, 2000; Hsiao *et al.*, 2001; Eisenberg & Levanon, 2003, 2013; Lee *et al.*, 2007; Zhu *et al.*, 2008; She *et al.*, 2009; Chang *et al.*, 2011; Lin *et al.*, 2014; Cheng *et al.*, 2017). Transcriptomic studies of HKGs in humans and higher plants indicate that HKGs are often involved in gene expression, biogenesis of nucleotides and amino acids, and intracellular transport (Eisenberg & Levanon, 2013; Cheng *et al.*, 2017). More than 40% of HKGs in *Arabidopsis* have human orthologs, suggesting HKGs might be highly conserved across the tree of life (Cheng *et al.*, 2017). A subset of HKGs identified from these model organisms with especially low variance in expression such as actins (ACT), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), tubulins (TUB), histone, elongation factor 1-α (EFG1-α) and rRNA, have been widely used as internal reference genes (IRGs), or control genes for the internal calibration of gene expression in a range of model and nonmodel organisms (Thellin *et al.*, 1999; Nicot *et al.*, 2005).

In humans, there is evidence that HKGs have several distinguishing characteristics relative to tissue-specific genes, including higher average relative expression and a more compact structure including shorter introns, untranslated regions and coding sequence (CDS), and a slower rate of evolution (Eisenberg & Levanon, 2003; Vinogradov, 2004; Zhang & Li, 2004). Transcription consumes time and resources, so a more compact gene structure may be evolutionarily selected in HKGs with high levels of expression; this is referred to as the selection for economy hypothesis (Castillo-Davis *et al.*, 2002). Other studies, in rice, *Arabidopsis* (Ren *et al.*, 2006), yeast (Vinogradov, 2001) and humans (Zhu *et al.*, 2008), have found that highly expressed genes are less compact. These contradictory findings may be due to differences in sequencing technology used, environmental conditions examined, tissues or organism considered, and the bioinformatic criteria used to identify HKGs (Zhang *et al.*, 2015). Further work on HKGs across a wider array of organisms and conditions is needed to assess whether HKGs and IRGs and their genomic characteristics are well conserved across the tree of life and consistent with the selection for economy hypothesis.

It has been hypothesized that HKGs provide insight into the basic mechanisms underlying cellular maintenance (Eisenberg & Levanon, 2003, 2013; Vinogradov, 2004; Zhang & Li, 2004; Zhu *et al.*, 2008; Lv *et al.*, 2015). HKGs in protists have received relatively little attention in the literature compared with those in model organisms such as humans and plants. Marine diatoms are unicellular photosynthetic eukaryotes that are responsible for *c.* 20% of global primary production (Falkowski *et al.*, 1998; Field *et al.*, 1998), and thus play a critical role in ocean biogeochemistry (Tréguer *et al.*, 2018). In a pioneering study, Alexander *et al.* (2012) identified 179 potential IRGs in the model diatom *Thalassiosira pseudonana* grown under replete and phosphorus-, iron- and phosphorus-and-iron co-limited conditions. They sampled at one time point selected when growth rate deviated from the replete control conditions using Tag-Seq. They found that several genes traditionally used as IRGs in other model organisms, and applied in phytoplankton studies, were significantly variable in *T. pseudonana*, indicating that HKGs may not be well conserved across diverse eukaryotic lineages. Taking advantage of advances in sequencing technology and the increased availability of genome and proteomic sequences of diverse diatoms (Armbrust *et al.*, 2004; Bowler *et al.*, 2008; Basu *et al.*, 2017; Mock *et al.*, 2017), here we determine whether: HKGs in *T. pseudonana* are more highly expressed, or compressed, or evolved slower than non-HKGs; there is a high level of conservation in HKGs across eukaryotes; and IRGs identified by Alexander *et al.* (2012) are stable under a wide range of stressors including reduced and elevated light, reduced and elevated temperature, reduced pH, nutrient (N, P, Fe and Si) starvation, a reactive oxygen stress and a control sampled over a time course, plus a filtration and centrifugation treatment, using RNA-seq.

Materials and Methods

Strain, culture conditions and experimental treatments

The marine diatom *Thalassiosira pseudonana* ((Hustedt) Hasle et Heimdal, CCMP 1335) was obtained from the Bigelow National Center for Marine Algae and Microbiota and maintained axenically in 250-ml polycarbonate bottles in modified nutrient-replete ESAW medium (Berges *et al.*, 2001), at 20°C, pH 8.1 and a continuous photon flux density of 100 μmol m^{-2} s^{-1} (we refer to this as the control). *Thalassiosira pseudonana* was exposed to the control conditions and 10 experimental treatments: four nutrient starvation treatments (N-, P-, Si- and Fe-free media), a low- and a high-temperature treatment (14 and 26°C), a low and a high irradiance treatment (10 and 800 μmol m^{-2} s^{-1}), a low-pH treatment (pH 7.8) and a reactive oxygen species (ROS) stress treatment that was created by exogenous addition of 0.165 mM H_2O_2. Cultures were filtered, centrifuged and rinsed before the initiation of the treatments (room temperature, three rounds at 3000 ***g***, 3 min and total time ≈ 0.5 h). We sampled the control bottles both before and immediately after centrifugation at time 0, and refer to the sampling after centrifugation as the filtration and centrifugation treatment. Due to time and space constraints, we conducted the experiment in three batches, and therefore, there were three control runs and filtration and centrifugation treatment runs. Each control and treatment was conducted in four biological replicate bottles. The control was sampled at 0, 6, 24 and 72 h. The 10 environmental treatments were sampled at 2, 6, 24 and 72 h after the initiation of the treatment. In all the experimental set-ups were included a set of control samples (five time points, run in three batches) and environmental treatments (10 environmental treatments × four time points, and the filtration and centrifugation samples run in three batches, each with one time point) and four replicates for a total of 232 samples. Additional details on the experimental set-up are provided in Supporting Information Table S1 and Fig. S1 and our study that summarizes the photophysiological responses to these treatments (Li *et al.*, 2021).

Sampling, RNA extraction and sequencing

Cultures were sampled at a final cell concentration of *c.* 5 × 10^5 cells ml^{-1}, regardless of the treatment, and filtered on polycarbonate (PC) Millipore membranes (pore size: 0.8 μm; diameter: 25 mm) under gentle vacuum pressure (<18 kPa or 5 in Hg) and low light. Samples were immediately frozen in liquid nitrogen and stored at −80°C until further analysis. Total RNA was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) and RNeasy Plus Mini Kit (Qiagen, Valencia, CA, USA) following the manufacturer's instructions. In brief, samples were placed in 1 ml of TRIzol reagent with lysing matrix D (MP Biomedicals, Santa Ana, CA, USA) and homogenized using a FastPrep 24 Machine for three cycles of 30 s each at a speed setting of 8.0 m s^{-1}, and incubated with ice between cycles (MP Biomedicals), followed by a standard phenol–chloroform extraction. The aqueous phase was transferred to a gDNA eliminator column (RNeasy Plus Mini Kit, Hilden, Germany) to remove the gDNA. RNeasy Plus Mini Kit and Qiagen's RNase-free DNase Set (an on-column treatment) were used for further purification and gDNA removal. RNA quality and content were measured by a NanoDrop ND-1000 Spectrophotometer, and the concentration of the samples was then adjusted to 100 ng μl^{-1} for sequencing. All RNA samples had an A260/A280 ≥ 2.1 and A260/A230 ≥ 2.4. All samples passed the RNA integrity analysis determined using an Agilent 2100 Bioanalyzer, using the RNA 6000 Nano Kit (Agilent, Palo Alto, CA, USA). Sequencing was performed by Illumina NovaSeq 6000 System at the Genome Quebec Innovation Centre. An Illumina TruSeq Stranded mRNA library preparation method was used for paired-end 100-bp reads. All raw read files (detailed descriptions of these files are provided in Dataset S1) are available through NCBI SRA (BioProject: PRJNA734969).

Bioinformatic analyses

The quality of the raw reads was assessed by FastQC v.0.11.8 (Andrews, 2010) and FastQ Screen v.0.13.0 (Wingett & Andrews, 2018) and summarized using MultiQC (Ewels *et al.*, 2016). Trimming of the raw reads was performed to remove low-quality bases and adapter sequences with TRIMMOMATIC v.0.38 (Bolger *et al.*, 2014). Trimmed reads were mapped and aligned to the genome of *T. pseudonana* CCMP 1335 (assembly ASM14940v2, https://www.ncbi.nlm.nih.gov/genome/54) to generate the SAM files using HISAT2 (v.2.1.0) (Kim *et al.*, 2015). The SAM files were then sorted and converted into binary BAM files by SAMTOOLS (v.1.7) (Li *et al.*, 2009). The BAM files were used to calculate the gene expression transcripts per million (TPM) using STRINGTIE v.2.1.2 (Pertea *et al.*, 2015). The gene count matrix was generated from the STRINGTIE gtf results using a PYTHON script (prepDE.py) provided with STRINGTIE. The gene count matrix was used to assess the log fold change ($\log_2$FC) and the SD of log fold change (logfcsd) using DESEQ2 v.1.24.0 (Love *et al.*, 2014). We compared expressions at $t = 2$, 6, 24 and 72 h for each environmental treatment with the control at the same time, and expression in the centrifugation and filtration treatment at $t \approx 0.5$ h with the control at time $t = 0$ h. Functional annotation of the HKGs in *T. pseudonana* was performed using gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis using the R package GOSTATS (Falcon & Gentleman, 2007) with a *P*-value cut-off of 0.01, and R package KEGGPROFILE (Zhao *et al.*, 2012) with a *P*adj cut-off of 0.01, respectively. We identified the potential localization of HKGs in *T. pseudonana* based on subcellular protein targeting using HECTAR v.1.3 (https://webtools.sb-roscoff.fr/). In addition, we downloaded chromosome and subcellular location data for all *T. pseudonana* proteins from UNIPROT (https://www.uniprot.org/) and analysed the localization of the HKGs using R v.4.0.2 (R Development Core Team, 2020) and GGPLOT2 R package (Wickham, 2016).

Defining HKGs and relative expression breadth

To identify HKGs in *T. pseudonana*, we adopted criteria similar to those used in humans (Eisenberg & Levanon, 2013): (1) nonzero expression in all samples examined: TPM > 0 in all 232 samples; (2) low variance in gene expression: coefficient of variation (CV) of TPM across all samples < 200%; (3) no exceptionally high expression compared with the mean: $|\log_2(\text{TPM}/(\text{mean TPM}))| < 2$; (4) and an additional constraint of low magnitude and low variability in differential expression for all treatments: $|\log_2\text{FC}| < 1$ and SD of $\log_2\text{FC} < 0.5$. We checked the expressional stability of HKG gene expression in independent differential gene expression data for *T. pseudonana* CCMP 1335 from 56 treatments archived at DiatomPortal (http://networks.systemsbiology.net/diatom-portal/, data downloaded on 22 September 2021). We analysed and compared the expressional stability (using $|\log_2\text{FC}|$ and SD of $\log_2$FC) of HKGs and non-HKGs. To investigate the expressional stability of orthologs of HKGs of *T. pseudonana* in *Phaeodactylum tricornutum* CCAP 1055/1, we identified *Phaeodactylum* orthologs of *Thalassiosira* genes using the INPARANOID v.4.1 program and database (http://inparanoid.sbc.su.se/) (O'Brien *et al.*, 2005). We found 1505 *Thalassiosira* HKGs have 941 *Phaeodactylum* orthologs. We downloaded differential expressional data for *Phaeodactylum* from Diatomicsbase (https://www.diatomicsbase.bio.ens.psl.eu/, data downloaded on 27 September 2021). We removed differential expression data of knockdown cell lines and experiments with only one or two biological replicates. The compiled dataset included 39 conditions from 14 studies. We then compared the expressional stability (using $|\log_2\text{FC}|$ and SD of $\log_2$FC) between these 941 *Phaeodactylum* orthologs of 1505 *Thalassiosira* HKGs (ptHKGs) and the remaining *Phaeodactylum* genes (non-ptHKGs).

The expression breadth in organisms with tissues is traditionally defined as the number (or proportion) of tissues under which a gene is expressed (Zhu *et al.*, 2008). We propose that an analogous concept for unicellular organisms is the proportion of experimental conditions under which a gene is expressed (we refer to this as the expression breadth, EB). In *T. pseudonana*, the vast majority of the detected genes (11 013, 94.4%) have a median TPM > 0 under all treatments and controls. A TPM threshold of 10 yields 5611 (48.1%) genes under the treatments and controls examined. We define a relative expression breadth (REB) for *T. pseudonana* as the percentage of examined experimental treatments under which a gene is expressed with median TPM ≥ 10.

Conservation of HKGs across diverse lineages

To assess the evolutionary conservation of the HKGs of *T. pseudonana*, we identified and analysed the distribution of orthologs of the HKGs across 37 species representing diverse clades from all kingdoms. To identify *Thalassiosira*-specific HKGs in the 37 species analysed, we downloaded protein sequences of *Minidiscus variabilis* CCMP 495 from JGI (https://genome.jgi.doe.gov/) and 36 other organisms from NCBI (http://www.ncbi.nlm.nih.gov/genome). We then identified orthologs between *T. pseudonana* CCMP 1335 and *M. variabilis* CCMP 495 and *Cyclotella cryptica* CCMP 332 using protein sequences and the INPARANOID v.4.1 program (O'Brien *et al.*, 2005). A phylogenetic tree of 37 species was constructed from a database of their proteomes (http://www.ncbi.nlm.nih.gov/genome or https://genome.jgi.doe.gov/) using CVTREE3 with a k-mer size of eight (Zuo & Hao, 2015). We identified orthologous gene pairs between *T. pseudonana* and the 36 other species using INPARANOID v.8.0 (O'Brien *et al.*, 2005). We calculated the number of these pairs found in each of 36 organisms and the ratio of these counts to the total number of genes in *T. pseudonana* with orthologs in any of these 36 organisms. We blasted the HKGs only identified in *T. pseudonana* against Tara Oceans data as a check to validate whether the genes had been observed independently by the Basic Local Alignment Search Tool (BLAST) program (Altschul *et al.* 1990) (https://www.genoscope.cns.fr/tara/), including the Tara Oceans eukaryotic genome (the 'SMAGs') nucleotides and peptides, metagenomic transcriptome (MGT) nucleic sequences, and the Tara Oceans eukaryote gene catalog (the 'MATOU' unigene sequences).

In a separate analysis of genes independently identified as HKGs, we compared the 1505 HKGs we identified in *T. pseudonana*, and HKGs identified in *Homo sapiens* (3804 HKGs identified across 16 human tissues) (Eisenberg & Levanon, 2013) and *Arabidopsis thaliana* (692 HKGs identified across 11 plant tissues) (Cheng *et al.*, 2017) to identify how many HKGs were shared across these model organisms using R v.4.0.2 (R Development Core Team, 2020) and UpsetR v.1.4.0 (Conway *et al.*, 2017). We looked for orthologs of these genes among the HKGs identified in *T. pseudonana*. Note the criteria for identifying HKGs in *H. sapiens* and *A. thaliana* differed somewhat from our criteria for *T. pseudonana*: HKGs in *T. pseudonana*, *H. sapiens* and *A. thaliana* were defined using different conditions and tissues, and the threshold for expression variation was different for *A. thaliana* compared with that for *T. pseudonana* and *H. sapiens*.

The evolution of HKGs in diatoms

To assess the rate of evolutionary change in HKGs vs non-HKGs in *T. pseudonana*, we computed the ratio of the number of nonsynonymous to synonymous (dN : dS) differences in all identified orthologs between *T. pseudonana* and six other diatoms and the brown macroalga *Ectocarpus siliculosus*. We analysed the genome sequences and gff files of *T. pseudonana* CCMP 1335, *M. variabilis* CCMP 495, *C. cryptica* CCMP 332, *Thalassiosira oceanica* CCMP 1005, *P. tricornutum* CCAP 1055/1, *Fragilariopsis cylindrus* CCMP 1102, *Pseudo-nitzschia multistriata* B856 and *Ectocarpus siliculosus* Ec 32 (CCAP 1310/04) obtained from NCBI (http://www.ncbi.nlm.nih.gov/genome or https://genome.jgi.doe.gov/). The CDS and protein fasta sequences were obtained from the genome sequences using the gff files and the Gffread function in Cufflinks (Trapnell *et al.*, 2010). Orthologous gene pairs between *T. pseudonana* and the seven other species were identified from the protein fasta sequences using InParanoid v.8.0 (O'Brien *et al.*, 2005). One-to-one orthologous gene pairs were identified using an InParanoid score of 1 and selecting the first gene of another diatom when there were multiple matches to a *T. pseudonana* gene. The CDS of the orthologous gene pairs was used to calculate dN : dS using the YN00 program in the phylogenetic analysis of maximum likelihood (Paml) package (Yang, 2007) and the paml-yn00-run-pipeline (Zhang *et al.*, 2020). Multiple sequence alignment and Paml format conversion were performed by ParaAT (Zhang *et al.*, 2012) and Mafft (Nakamura *et al.*, 2018) as part of the paml-yn00-run-pipeline. Outlier genes showing abnormally high $dS \geq 10$ were not considered in our analysis.

Nucleotide substitution rate, gene length and expression level of HKGs compared across diatoms, eukaryotes and prokaryotes

We examined the relationship between the evolutionary origin of HKGs in *T. pseudonana* and the nucleotide substitution rate, gene structure and expression level. We clustered the 37 organisms described above into four phyletic groups: *T. pseudonana*, other diatoms (not *T. pseudonana*, six species), nondiatom eukaryotes (24 organisms) and prokaryotes (bacteria and archaea, six organisms). We classified the 1505 HKGs of *T. pseudonana* into five groups according to their evolutionary history as revealed by the set of orthologs identified as: T, unique to *T. pseudonana*; D, present in at least one species of diatom, but not in any nondiatom eukaryote or prokaryote; E, present in at least one diatom and at least one nondiatom eukaryote, but not in any prokaryote; P, present in all groups; and a final group for all remaining HKGs with ambiguous evolutionary history. The evolutionary rate dN : dS, gene length, CDS length and expression level (TPM) were calculated and analysed using these groups.

Evaluating IRGs in *T. pseudonana*

To identify internal reference genes (IRGs) in *T. pseudonana*, we adopted criteria similar to those used in studies on humans (Eisenberg & Levanon, 2013) and in a prior study on *T. pseudonana* (Alexander *et al.*, 2012). We selected HKGs (as described above) and altered our constraints to $TPM \geq 10$, CV of $TPM < 25\%$, $|\log_2(TPM/(\text{mean TPM}))| < 1$ (2-fold) and $|\log_2 FC| < 0.4$. To these genes, we added six IRGs previously identified in *T. pseudonana* using different methods and criteria (Alexander *et al.*, 2012). We evaluated an additional 10 commonly used IRGs identified in other eukaryotes and their orthologs from UniProt (Consortium, 2019) from a literature search (Dataset S2). The ten commonly used IRGs are as follows: ACT, GAPDH, TUB, calmodulin, histone, EFG1-α, ribosomal protein (RP), cyclin-dependent kinase (CDK), TATA box-binding protein (TBP) and probable adenine phosphoribosyltransferase (APT).

Statistical analyses

All experiments were performed using four independent biological replicates. A Mann–Whitney *U*-test was used to test whether HKGs and non-HKGs are associated with significant differences in several gene characteristics of interest. A linear regression analysis was used to test the relationship between the log gene length and log median TPM. A Spearman rank correlation analysis was used to test the monotonic trend in gene length and dN : dS with REB in HKGs and non-HKGs. Statistical analyses were all carried out using R v.4.0.2 (R Development Core Team, 2020).

Results

We identify 1505 HKGs in *T. pseudonana* ($TPM > 0$ in all samples, $CV(TPM) < 200\%$, $|\log_2(TPM/(\text{mean TPM}))| < 2$, $|\log_2 FC| < 1$ and $SD(\log_2 FC) < 0.5$). We provide a list of these genes and their annotation in Dataset S3.

The HKGs we identified in our experiment exhibit lower variation in absolute $\log_2$FC values and smaller SD in $\log_2$FC relative to non-HKGs in *T. pseudonana* data archived in the DiatomPortal dataset (Fig. S2a,b). We found most of the HKGs (*c.* 90%, 1353 of 1505) have $\log_2 FC < 1$ under 48 or more of the 56 treatments in the DiatomPortal dataset (Fig. S2c). We also found 676 HKGs have $|\log_2 FC| < 1$ across 55 or 56 treatments (Fig. S2c) and lower $\log_2$FC variation than the remaining *T. pseudonana* genes (Fig. S2d,e). The 1505 HKGs of *Thalassiosira* have 941 *P. tricornutum* orthologs (ptHKGs). The ptHKGs have smaller SD in $\log_2$FC relative to non-ptHKGs (Fig. S3a,b).

Gene expression and genomic features of the HKGs

The median TPM values of HKGs are significantly higher than those of the non-HKGs (Table 1, Mann–Whitney *U*-test, $P < 0.05$), and median TPM increases with expression breadth (% REB) across all genes (Fig. 1a). The selection of HKGs includes a threshold on TPM, but this is only one criterion, and the threshold is quite low. HKGs in *T. pseudonana* have longer gene lengths, CDS lengths per gene, exon lengths, total exon length per gene and total intron length per gene, greater number of introns per gene and exons per gene, but lower absolute $\log_2$FC values than those of non-HKGs (Table 1, Mann–Whitney *U*-test, $P < 0.05$), and no difference in intron length between HKGs and non-HKGs. These results indicate HKGs in *T. pseudonana* are more highly expressed and less compact than non-HKGs; however, both the HKGs and non-HKGs with shorter gene lengths (Fig. 1b, slope = –0.28, $P < 0.001$ for HKGs, and slope = -0.14, $P < 0.001$ for non-HKGs) and CDS lengths (Fig. S4a, slope = -0.36, $P < 0.001$ for HKGs, and slope = -0.29, $P < 0.001$ for non-HKGs) tend to have higher TPM values than longer HKGs or non-HKGs. A majority of HKGs (82%, 1232 of 1505) are expressed in all treatments (REB = 100%; Fig. S5a).

The HKGs we identified have a lower dN : dS than non-HKGs when comparing *T. pseudonana* to centric diatom species (*M. variabilis*, *C. cryptica* and *T. oceanica*) (Table 1, Mann–Whitney *U*-test, $P < 0.05$). This difference is not significant when the comparison is broadened to three other pennate diatom species (*P. tricornutum*, *F. cylindrus* and *P. multiseries*) and *E. siliculosus* (Table 1). Log dN : dS is negatively correlated with % REB, regardless of the diatom species used in the analysis (Fig. S4b; Table S2, rank correlation -0.16 to -0.29, $P < 0.001$).

Table 1 Gene statistics summarized for nonhousekeeping genes (non-HKGs) and housekeeping genes (HKGs) in *Thalassiosira pseudonana*.

Parameters	Non-HKGs (n = 10 168)	HKGs (n = 1505)
Expression level (TPM)	27.5 (90.4)	34.1 (51.8)*
Differential expression ($\log_2$FC)	0.273 (0.528)	0.151 (0.216)*
Gene length (bp)	1368 (1729)	1477 (1896)*
CDS length per gene (bp)	1158 (1478)	1278 (1632)*
Exon length (bp)	392 (610)	412 (660)*
Total exon length per gene (bp)	1224 (1542)	1341 (1690)*
Intron length (bp)	84 (97)	82 (107)
Total intron length per gene (bp)	90 (186)	99 (206)*
Intron number per gene	1 (1.53)	1 (1.56)*
Exon number per gene	2 (2.53)	2 (2.56)*
dN : dS (*Thalassiosira pseudonana–Minidiscus variabilis*) (n = 7647)	0.075 (0.088)	0.067 (0.076)*
dN : dS (*Thalassiosira pseudonana–Cyclotella cryptica*) (n = 5586)	0.089 (0.109)	0.083 (0.103)*
dN : dS (*Thalassiosira pseudonana–Thalassiosira oceanica*) (n = 5724)	0.095 (0.108)	0.086 (0.099)*
dN : dS (*Thalassiosira pseudonana–Phaeodactylum tricornutum*) (n = 5191)	0.119 (0.132)	0.119 (0.130)
dN : dS (*Thalassiosira pseudonana–Fragilariopsis cylindrus*) (n = 5184)	0.118 (0.126)	0.116 (0.124)
dN : dS (*Thalassiosira pseudonana–Pseudo-nitzschia multiseries*) (n = 4319)	0.126 (0.134)	0.128 (0.138)
dN : dS (*Thalassiosira pseudonana–Ectocarpus siliculosus*) (n = 2931)	0.149 (0.157)	0.149 (0.158)
Annotated genes by KO terms	2231 (21.9%)	467 (31.0%)
Annotated genes by GO terms	6396 (62.9%)	1074 (71.4%)

TPM, transcripts per million; bp, base pair; CDS, coding sequence; dN : dS, nonsynonymous-to-synonymous ratio; KO, Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology; and GO, gene ontology.
Values reported are medians and means (in parentheses), except for annotation data, which are counts and percentages relative to total gene count. Significant differences between means for non-HKGs and HKGs are noted with an asterisk (Mann–Whitney *U*-test, $P < 0.05$).

Conservation of HKGs

To test the conservation of HKGs in *T. pseudonana*, we searched for orthologs of our HKGs in 36 species from across the tree of life (Fig. 2a–c). We also compared HKGs identified in this study with those independently identified across 16 human tissues and 11 *Arabidopsis* tissues (Fig. 2d). We found most of the HKGs we identified in *T. pseudonana* (1454 of 1505, 96.6%) have orthologs in at least one of 36 species examined (Fig. 2a). Over half of these HKGs (779 of 1454, 52.0%) have orthologs observed in no more than six other species. Fifty-one *T. pseudonana* HKGs are found only in *T. pseudonana*, but most of them (44 of 51) can be found in Tara Oceans field data (Dataset S3); 39 of the 51 HKGs were found to have no orthologous annotations in the InParanoid, eggNOG, KO or GO databases (Dataset S3). In our analysis, 91 of the *T. pseudonana* HKGs were found only in the diatom species examined (Fig. 2b). In general, species with more recent common ancestors with *T. pseudonana* have more orthologs of the HKGs found in *T. pseudonana* (Fig. 2c). Within bacteria, archaea, eukaryotes, diatoms, prokaryotes (bacteria + archaea) and all 36 species examined, we detected 25, 20, nine, 242, three and zero common orthologs of the HKGs of *T. pseudonana* (Fig. S5b). The list of these genes and their annotations are provided in Dataset S3. The proportion of shared orthologous HKGs relative to all shared orthologs between *T. pseudonana* varied between 14% and 18% across diverse eukaryotes and between 9% and 10% across diverse prokaryotes (Fig. 2c).

We compared identified HKGs in *T. pseudonana* with HKGs identified in previous RNA-seq studies conducted on *H. sapiens* and *A. thaliana* (Fig. 2d). *Homo sapiens* and *A. thaliana* have 1303 (34% of 3804 *H. sapiens* HKGs) and 207 (30% of 609 *A. thaliana* HKGs) HKGs, respectively, that are orthologous to HKGs in *T. pseudonana*. Most HKGs in *T. pseudonana* (1215, 81% of 1505) were not identified as HKGs in *H. sapiens* and *A. thaliana*. Only a small set of 24 HKGs are common across these three organisms. In aggregate, these results suggest HKGs in *T. pseudonana* are mostly unique. We list the 24 HKGs shared across *T. pseudonana*, *H. sapiens* and *A. thaliana* and their annotations in Dataset S3.

The evolution, length and expression of different age groups of HKGs in *T. pseudonana*

We divided the HKGs with orthologs in other species into four groups: unique to *T. pseudonana* (T); found in other diatoms but

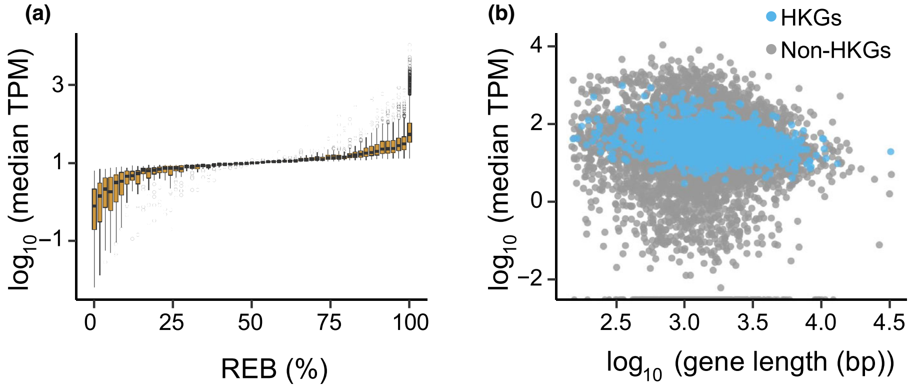

Fig. 1 Covariation between selected gene statistics (Table 1). (a) Median transcripts per million (TPM) of *Thalassiosira pseudonana* genes increases with increasing relative expression breadth (rank correlation 0.83, $P < 0.001$). We define a relative expression breadth (REB) for *T. pseudonana* as the percentage of examined experimental conditions under which a gene is expressed with median TPM ≥ 10. (b) Median TPM decreases with increasing gene length, with a more negative slope for HKGs (slope and SE –0.28 ± 0.03 for gene length, $R^2 = 0.056$, $P < 0.001$) than for nonhousekeeping genes (non-HKGs) (slope and SE −0.14 ± 0.02 for gene length, $R^2 = 0.004$, $P < 0.001$). Data points outside whiskers in boxplots are shown in black circles. bp, base pair.

not in other organisms (D+T); found in other eukaryotes and other diatoms but not in bacteria or archaea (E+D+T); and found in bacteria or archaea, other eukaryotes and other diatoms (P+E+D+T). HKGs in P+E+D+T are likely to be the most ancient and compared with HKGs in other groups had smaller dN : dS, were slightly shorter in both gene length and CDS length and were more highly expressed (Figs 2e, S6). A similar pattern holds for all *T. pseudonana* genes (HKGs and non-HKGs; Fig. S7).

Functional annotation and localization of the HKGs in *T. pseudonana*

The HKGs in *T. pseudonana* are involved in a wide array of biological functions (Figs 3a, S8). Enriched KEGG categories include transcription, repair and protein catabolism pathways, including the spliceosome (ko3040), the mRNA surveillance pathway (ko3015), basal transcription factors (ko3022), ubiquitin-mediated proteolysis (ko4120) and nucleotide excision repair (ko3420) (Fig. 3a). The enriched gene ontology (GO) categories are associated with basic biological processes including several transport processes, for example vesicle-mediated transport (GO:0016192), intra-Golgi vesicle-mediated transport (GO:0006891) and RNA splicing, via transesterification reactions (GO:0000375) (Fig. 3a). The GO-enriched molecular function categories include the structural constituent of the nuclear pore (GO:0017056), molecular adaptor activity (GO:0060090), and catalytic activity, acting on a protein (GO:0140096) (Fig. S8). GO-enriched cellular component terms include the following: Golgi apparatus (GO:0005794), protein-containing complex (GO:0032991) and the nuclear envelope (GO:0005635) (Fig. S8). All enrichment results are provided in Dataset S4.

The HKGs in *T. pseudonana* fell into six classes based on predicted protein localization using HECTAR (Fig. 3b), including chloroplast, no signal peptide or anchor, signal peptide, other localization, mitochondrion and signal anchor. The category ‘other localization’ represented most (78.9%, 1188) of the localized HKGs; however, ‘signal anchor’ had the highest ratio of HKG localization, 19.5%, relative to the gene localization of all *T. pseudonana* genes. We found that only 199 (of our 1505) HKGs had UniProt subcellular location annotations, and most of these genes had annotations of ‘membrane’ (70) and ‘nucleus’ (59) (Fig. 3c). Three UniProt subcellular location annotations (‘Golgi apparatus membrane; Membrane’, ‘Nucleus, nuclear pore complex’, and ‘Mitochondrion inner membrane’) are most highly associated with HKGs. Moreover, we found that chromosomes 4, 7, 5, 2 and 6 have both relatively higher numbers and ratios of HKGs (Fig. 3d). These results suggest HKGs appear to be mainly localized to the mitochondria and Golgi membranes and the nuclear pore complex, but we were unable to localize the majority of HKGs in *T. pseudonana* (Fig. 3b–d).

Expression stability of commonly used IRGs

Very few of the commonly used IRGs (see Dataset S3 for a summary table of IRGs used in the literature) are stably expressed in all samples and under all treatments in our experiment. We find only seven genes (18.9%) of the 37 previously identified IRGs have a CV of TPM < 25%, TPM > 50 and $|\log_2 FC| < 1$ for all treatments in our *T. pseudonana* experiment (Fig. 4). These genes include all previously identified actin genes, except ACT1, as well as TUB4, and two ubiquitin ligases (THAPSDRAFT_40148 and THAPSDRAFT_22432). Among the 30 remaining commonly used IRGs, half (15) have CV of TPM > 50% and many of them have a TPM > 50 and $|\log_2 FC| > 3$. These genes include all GAPDH genes, TUB2, CAM (THAPSDRAFT_19631 and THAPSDRAFT_263535), H4 (THAPS_35229, THAPS_35279 and THAPSDRAFT_37357), two EF1-a (THAPSDRAFT_bd1861 and THAPSDRAFT_267957) and two ubiquitin ligases (THAPSDRAFT_21152 and THAPS_23335). These 30 genes would qualify as excellent IRGs under certain conditions (see Figs S9, S10). For example, GAPDH exhibits stable expression under high-temperature and low-pH treatments. TUB genes (except TUB4) exhibit stable expression

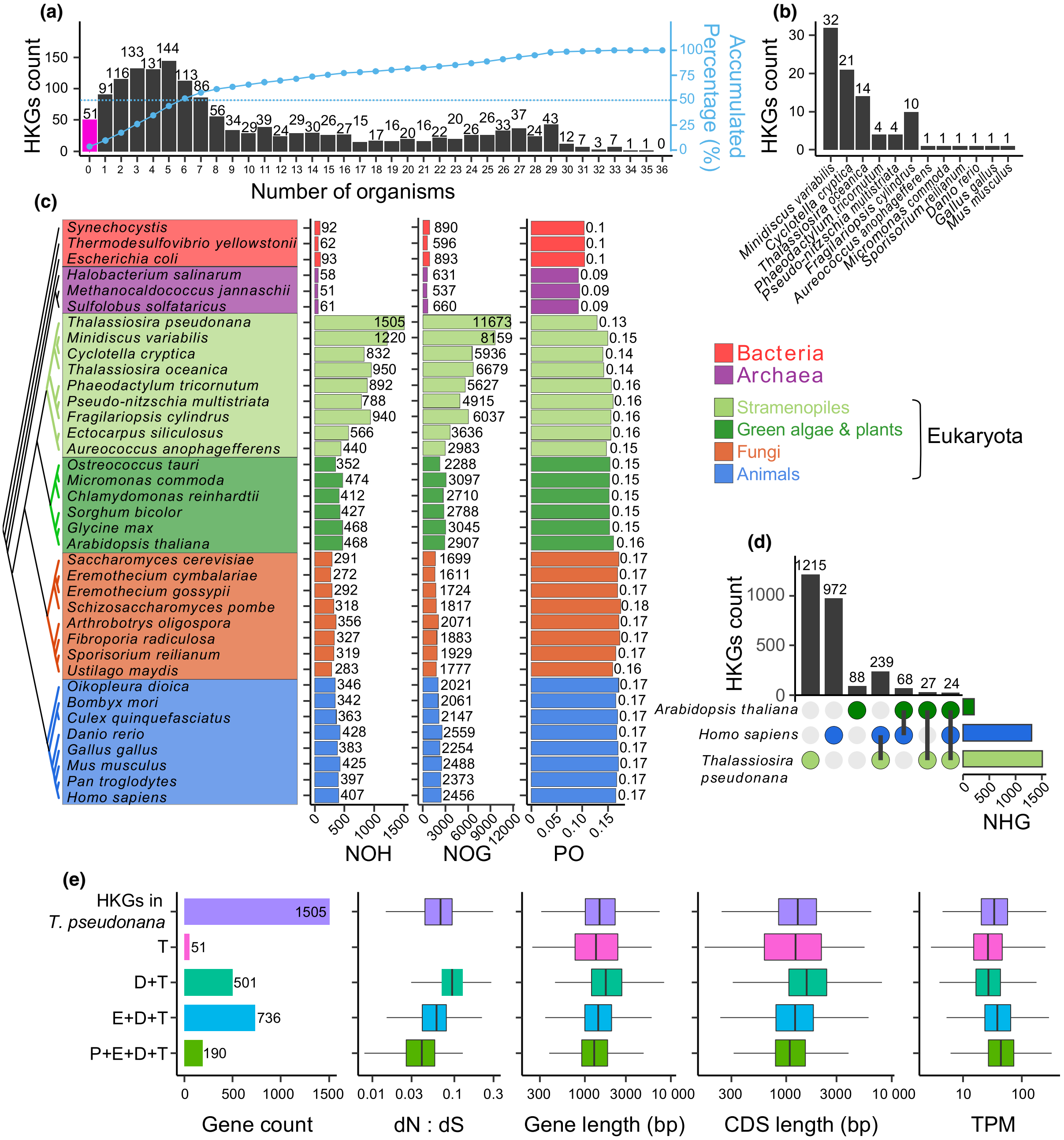

Fig. 2 Conservation and function of housekeeping genes (HKGs) in *Thalassiosira pseudonana*. (a) The number of organisms that shared different numbers of orthologs (left black *y*-axis) and accumulated percentage (right blue *y*-axis, the percentage between the sum of the number of orthologs shared in different numbers of organisms and 1505 HKGs) with the HKGs of *T. pseudonana*. The pink bar represents 51 HKGs that do not have orthologs with any other organisms. The blue horizontal dotted line at an accumulated percentage of 50% is a guide to aid identification of the number of organisms. (b) The number of HKGs of *T. pseudonana* found in one other organism. (c) HKGs in *T. pseudonana* with orthologs across the 36 organisms examined using the InParanoid program and database (http://inparanoid.sbc.su.se/). The number of orthologs in each organism, which are HKGs in *T. pseudonana* (NOH, left bar chart), the total number of orthologs in each organism matching any gene in *T. pseudonana* (NOG, middle bar chart), and the proportion of those orthologs relative to the number of orthologs in each organism matching any gene in *T. pseudonana* (PO, right bar chart). (d) The number of HKGs (NHG) in *H. sapiens*, *A. thaliana* and *T. pseudonana* identified in independent analyses, which were orthologous to any gene in *T. pseudonana* (right bar chart). These orthologs were HKGs in one, two or three of these organisms as shown in the upper bar chart and UpSet plot. (e) The relationship between the evolutionary origin of HKGs and the number of HKGs (gene count), dN : dS (calculated using orthologous gene pairs between *T. pseudonana* and *M. variabilis*; dN : dS calculated using orthologous gene pairs between *T. pseudonana* and other six species can be found in Supporting Information Fig. S4), gene length, CDS length and expression level (TPM). The 1505 HKGs are categorized into four groups according to evolutionary origin: (1) unique to *T. pseudonana* (T); (2) found in other diatoms but not in other organisms (D+T); (3) found in other eukaryotes and other diatoms but not in bacteria or archaea (E+D+T); and (4) found in bacteria or archaea, other eukaryotes and other diatoms (P+E+D+T). An additional 27 HKGs do not belong to any of the four groups (e.g. E+T or P+T). Data points outside whiskers in boxplots are not shown for better visualization. bp, base pair; dN : dS, nonsynonymous-to-synonymous ratio.

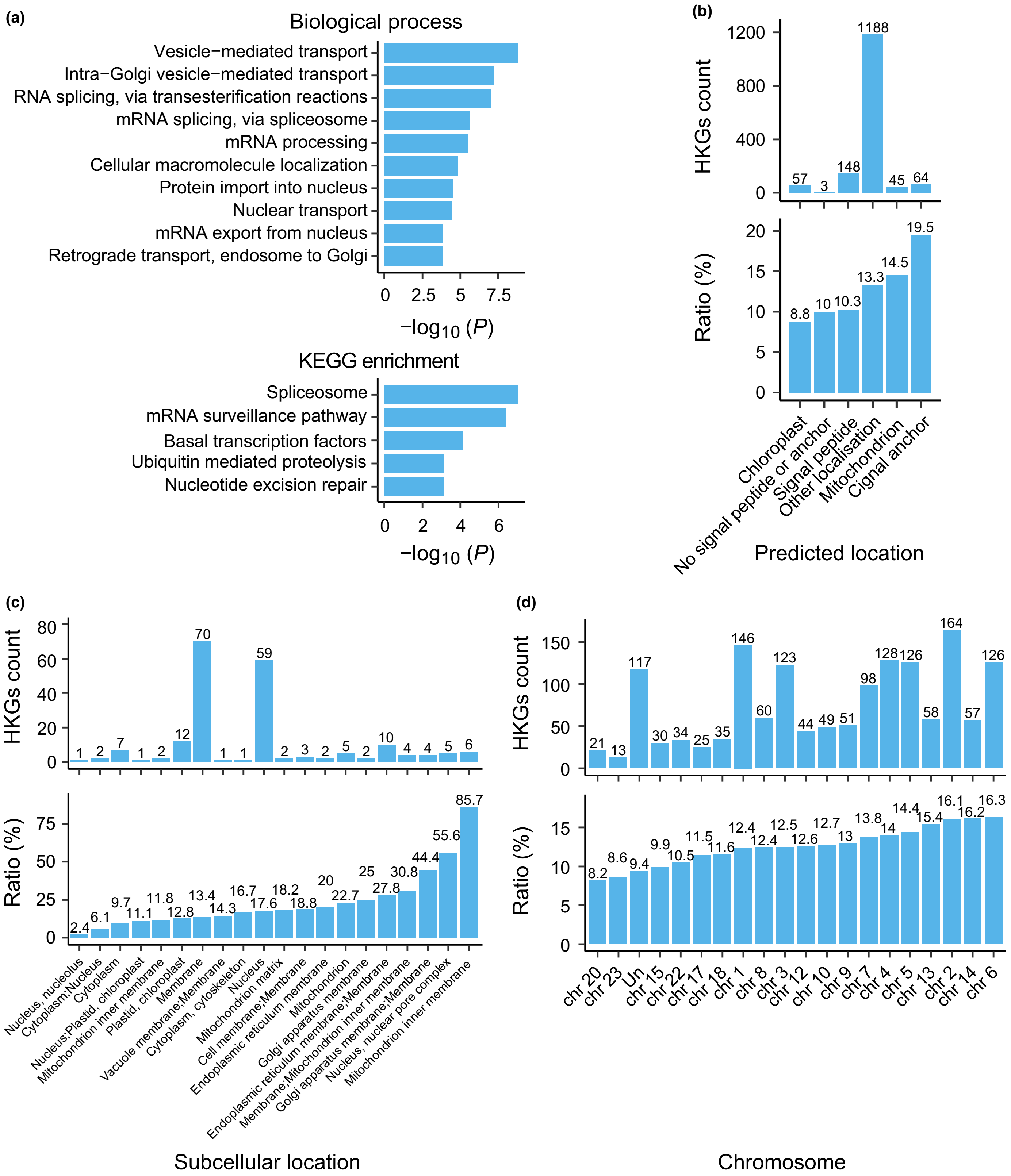

Fig. 3 Enrichment and localization of housekeeping genes (HKGs) in *Thalassiosira pseudonana*. (a) Gene ontology (GO) biological processes and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched in HKGs. Only the top-10-ranked GO terms (by *P*-value) are shown; the full result of the enrichment analysis is provided in Supporting Information Dataset S4. (b) HECTAR protein localization of HKGs (upper, counts; lower, ratio of HKG localizations to all *T. pseudonana* gene localizations). (c) UniProt subcellular localization of HKGs (upper, counts; lower, ratio of HKG localizations to all *T. pseudonana* gene localizations). (d) Chromosome localization of HKGs (upper, counts; lower, ratio of HKG localizations to all *T. pseudonana* gene localizations).

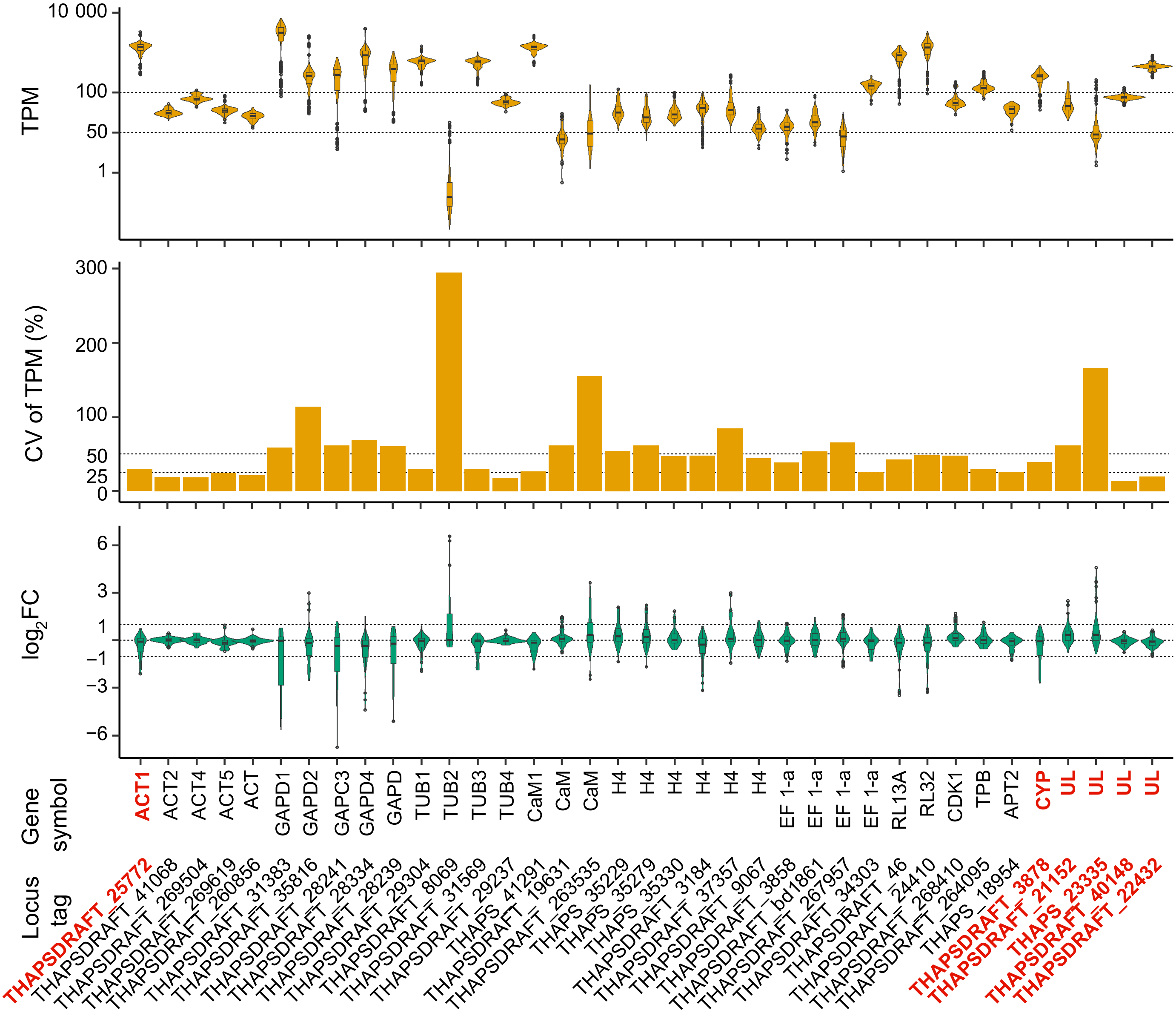

Fig. 4 Expression level (transcripts per million (TPM), log scale, box–violin plot), the coefficient of variation of TPM and differential expression (log_2FC, box–violin plot) in *Thalassiosira pseudonana* for 37 commonly used internal reference genes across all conditions examined. Horizontal dotted lines are guides to aid comparison across genes and identify the TPM or log_2FC threshold for each HKG. Gene symbols are actin (ACT), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), tubulin (TUB), calmodulin (CaM), histone H4 (H4), elongation factor 1α (EF1-α), ribosomal large subunit (RL), cyclin-dependent kinase (CDK), TATA box-binding protein (TBP), probable adenine phosphoribosyltransferase 2 (APT2), cyclophilin (CYP) and ubiquitin ligase (UL). Genes selected from Alexander *et al.* (2012) are in labelled in red. The violin plots show the full distribution of the data. Data points outside whiskers in boxplots are shown in black circles. Detailed annotations of these genes are provided in Supporting Information Dataset S4.

except under N-starvation, Si-starvation, low-light and ROS treatments.

Recommended internal reference genes for *T. pseudonana*

Here, we identify seven new IRGs in *T. pseudonana* with high TPM (≥ 10) and very low variation in expression and differential expression (CV of TPM < 25%, $log_2(TPM) < 1$ and $log_2FC < 0.4$) across all our samples and under all treatments (Fig. 5). Of these IRGs, five have UniProt protein annotations, including ANAPC4_WD40 domain-containing protein (THAPSDRAFT_2222), Ser/Thr kinase (THAPS_263127), coatomer subunit beta (THAPSDRAFT_26212), suppressor of actin mutation protein-like protein (THAPSDRAFT_31160) and WD40 repeat protein (THAPSDRAFT_36367). All seven IRGs have a minimum of two orthologous gene annotations (InParanoid/eggNOG/KO/GO). We provide detailed annotations for these IRGs (Dataset S3), TPM values, CV of TPM and log_2FC values under all treatments (Fig. S11). Expression levels for these IRGs range from a median TPM of 18.4–150.0 (Fig. 5). As far as we are aware, none of these seven IRGs have been previously identified as IRGs. If we relax our criteria from a

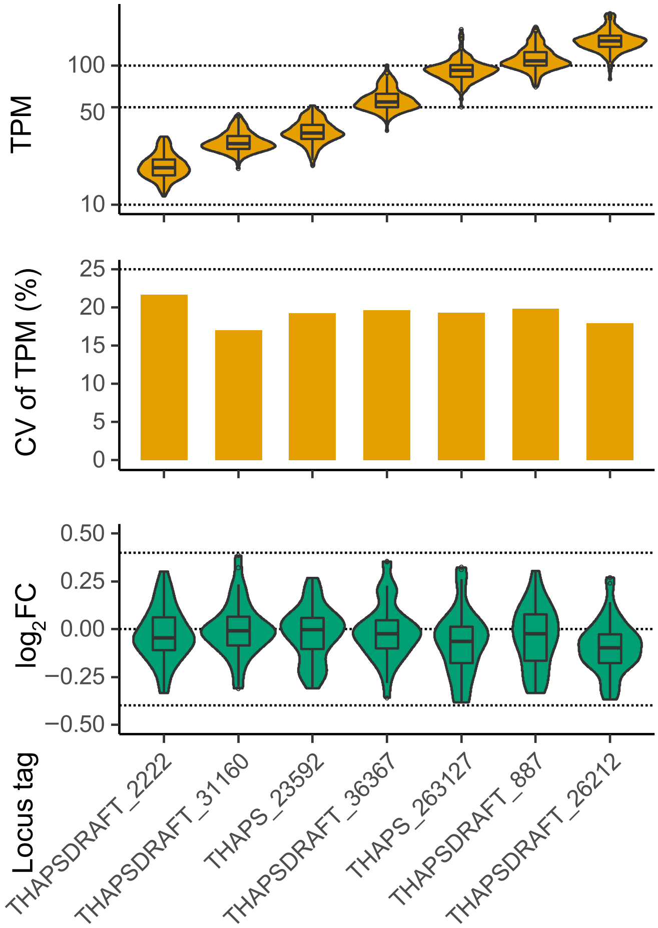

Fig. 5 Expression of the newly identified internal reference genes (IRGs) for *Thalassiosira pseudonana* showing the expression level (TPM, transcripts per million, log scale, box–violin plot), coefficient of variation (CV) of TPM and differential expression ($\log_2$FC, box–violin plot) across all conditions examined. Horizontal dotted lines are guides to aid comparison across genes and identify the TPM or $\log_2$FC threshold for each IRG. The violin plots show the full distribution of the data. Data points outside whiskers in boxplots are shown in black circles. Annotation for each gene by locus tag is provided in Supporting Information Dataset S4.

requirement that $\log_2\text{FC} < 0.4$ to $\log_2\text{FC} < 0.5$, we obtain a list of 19 IRGs. Annotation, TPMs and $\log_2$FC values under all treatments for these 19 IRGs are provided (Dataset S3; Fig. S12). This expanded list of IRGs includes HKGs and IRGs (such as THAPSDRAFT_41068 (ACT2) and THAPSDRAFT_269504 (ACT4)) that have been previously identified in *T. pseudonana* and a number of other model organisms. Note we find that the expression of IRGs can be distinctly more stable under particular experimental conditions than under the full suite of treatments (Figs S9, S10). For example, the ACT genes (except ACT1) are very stable under the P- and Fe-starvation treatments.

The seven IRGs we identified also exhibit low variation in $\log_2$FC values (between −1 and 1) under 54 or more conditions in the *T. pseudonana* data in the DiatomPortal dataset (Fig. S2f) and have five *Phaeodactylum* orthologs with stable differential expression under 39 conditions in the Diatomicsbase data (Fig. S3c).

Discussion

Much effort has been made to identify HKGs from human and higher plant tissues, and to investigate their expression and functional, structural and evolutionary features, in an attempt to better understand the mechanisms underlying basic cellular maintenance (Eisenberg & Levanon, 2003, 2013; Vinogradov, 2004; Zhang & Li, 2004; Zhu *et al.*, 2008; Lv *et al.*, 2015). Here, we identify 1505 HKGs in the model marine unicellular eukaryotic diatom *T. pseudonana* using a deep-sequencing RNA-seq approach on 232 samples taken under 12 environmental conditions taken over 0–72 h, control conditions over 0–72 h, and a filtration and centrifugation treatment taken at ≈ 0.5 h (Table S1; Fig. S1). Analyses of independent experiments conducted by several other investigators on *T. pseudonana* archived in DiatomPortal provide additional support that the HKGs identified in our experimental study exhibit lower variability in expression over a wide range of experimental conditions than non-HKGs (Fig. S2). Furthermore, we found that orthologs of the *T. pseudonana* HKGs in *P. tricornutum* also exhibit high expressional stability relative to other genes (Fig. S3).

We evaluated the expression, several structural features, evolution and their level of conservation of the 1505 HKGs we identified in *T. pseudonana* across a range of prokaryotes and eukaryotes across the tree of life and found the following: (1) the HKGs of *T. pseudonana* are more highly expressed and longer than non-HKGs, but the expression level of HKGs (and non-HKGs) is negatively correlated with gene length (Table 1; Fig. 1); (2) more ancient HKGs in *T. pseudonana* appear to evolve more slowly, and are more highly expressed and shorter than the younger HKGs (Fig. 2); and (3) HKGs are not highly conserved across the tree of life (Fig. 2). Based on our analyses, we identify seven new IRGs for *T. pseudonana* (Figs 4, 5) that also exhibit low log fold change in independent experimental data on *T. pseudonana* (Fig. S2) and orthologs in *P. tricornutum* (Fig. S3).

It has been hypothesized that highly expressed genes may be subject to selection pressure favouring a reduction in gene length to reduce the energy burden of transcription (Castillo-Davis *et al.*, 2002; Eisenberg & Levanon, 2003). This idea is referred to as the selection for economy hypothesis. The evidence for this hypothesis across eukaryotes is challenging to interpret. Several studies on human and *Mus musculus* HKGs and human and *Caenorhabditis elegans* highly expressed genes provide support for the selection for economy hypothesis (Castillo-Davis *et al.*, 2002; Eisenberg & Levanon, 2003; Li *et al.*, 2007), but studies on highly expressed genes in *Saccharomyces cerevisiae* (Vinogradov, 2001), rice and *Arabidopsis* (Ren *et al.*, 2006) appear to contradict the hypothesis. These apparently contradictory results have been ascribed to differences in genome configuration or the functional role of introns across diverse eukaryotes (Vinogradov, 2001; Ren *et al.*, 2006). Here, we found that the HKGs of *T. pseudonana* are more highly expressed than non-HKGs, but gene length was not affected as expected; median gene length is larger in HKGs than in non-HKGs (Table 1). *Thalassiosira pseudonana* genes have similar CDS length to genes in humans and *Arabidopsis* (Eisenberg & Levanon, 2003; Ren *et al.*, 2006), but *T.*

pseudonana's average gene length (including introns) is less than one-tenth the length of human genes and about half the length of plant genes (Kaul *et al.*, 2000; Eisenberg & Levanon, 2003; Ren *et al.*, 2006; Zhu *et al.*, 2008). The short gene lengths in *T. pseudonana* may mitigate or limit further selection for economy on their HKGs. The selection for economy hypothesis is still supported by the fact that highly expressed genes *T. pseudonana* do tend to be shorter (Fig. 1). In aggregate, these results suggest that gene expression is an important determinant of gene length in eukaryotes.

Evolutionary history, specifically the age of genes, may influence gene length and expression levels in HKGs in *T. pseudonana.* We tested whether older HKGs are shorter than younger HKGs by partitioning HKGs into approximate age classes. Through homology with genes in other diatoms, eukaryotes and prokaryotes, we find evidence of a negative correlation between gene age and gene length, CDS length, evolutionary rate and a positive correlation with expression level (Fig. 2e). Older HKGs in *T. pseudonana* are shorter than younger HKGs and have evolved at a slower average rate, consistent with the notion that older genes may be expected to be more strongly affected by selection for economy. This economy may enable or be offset by higher expression levels. Similarly, in humans, the fly and yeast, gene age is highly correlated with gene expression, but in contrast to our findings, older genes are longer (Alba & Castresana, 2005; Wolf *et al.*, 2009; Vishnoi *et al.*, 2010). Our finding that older genes (HKGs and non-HKGs) are shorter provides additional support for the selection for economy hypothesis in *T. pseudonana.*

HKGs are expected to be evolutionarily conserved because of their assumed contribution to basic cellular maintenance (Eisenberg & Levanon, 2003), their slow evolutionary rates (Zhang & Li, 2004; Zhu *et al.*, 2008; Lv *et al.*, 2015) and the reported overlap in HKGs between humans and *Arabidopsis* (Cheng *et al.*, 2017). However, our evidence shows that HKGs in *T. pseudonana* are largely distinct from genes in other organisms (Fig. 2a). Across a range of diverse eukaryotes, the proportion of orthologs that are HKGs in *T. pseudonana* range between 14% and 18% (Fig. 2c), but the identity of these HKGs changes from organism to organism (Fig. 2a,b). Furthermore, we identified only 24 HKGs shared among *T. pseudonana*, humans and *A. thaliana* (Fig. 2d), although the HKGs in humans and *A. thaliana* were identified somewhat differently. Technologies and criteria for identifying HKGs have changed over time, with documented inconsistencies in genes identified as housekeeping, even within humans (Zhang *et al.*, 2015). Alternative annotations (such as GO biological processes) may reveal more consistency in function of HKGs, but our analysis found HKGs spread across many GO_BP sublevel terms and did not help us find many similarities in HKGs across organisms. The notion of HKGs stretches back at least half a century (Watson, 2004), but definitions and methods of identification have not been consistent across studies. Our results suggest that many of the unique gene architecture characteristics of HKGs are probably primarily attributable to being highly expressed genes with low variance across conditions.

The most stringently selected HKGs can be used as IRGs, which are valuable as internal standards, but as with the broader class of HKGs, we find the best IRGs in *T. pseudonana* are distinct from previously identified IRGs in other organisms and are not well conserved across organisms (Table S3). We found that many commonly used IRGs (23 of 37) did not even meet our criteria for HKGs in *T. pseudonana*, since their expression varied across our panel of stressors. Since HKGs have not been previously systematically identified in eukaryotic phytoplankton, IRG candidates are typically selected from distantly related organisms and their expression stability is tested by qRT-PCR (Siaut *et al.*, 2007; Guo & Ki, 2012; Bach *et al.*, 2013; Guo *et al.*, 2013; Adelfi *et al.*, 2014; Lohbeck *et al.*, 2014; Ding *et al.*, 2015; Lee *et al.*, 2015; Deng *et al.*, 2016). Borrowing IRGs from other organisms provides a restricted and biased sample for screening, so we used our comprehensive survey to identify seven genes with more stable expression than previously used IRGs and an additional panel of IRGs that can be used under restricted sets of experimental conditions. Expressional stability of several orthologs of these IRGs in *P. tricornutum* (Fig. S3) indicates that some of these genes may be useful IRGs for other diatoms.

Our results show that some HKGs, which are defined based on a statistical analysis of expression levels across tissues or treatments, include some genes that are conserved across the tree of life, but are predominantly composed of genes that are only HKGs in a small number of organisms or a limited region of the tree of life. In contrast to HKGs in humans, HKGs in *T. pseudonana* are longer than non-HKGs. Despite this result, there is still evidence for the selection for economy hypothesis since more highly expressed or more ancient HKGs are shorter than other HKGs. Many of the gene architecture characteristics of HKGs seem to be attributable to being highly expressed genes with low variance across conditions. HKGs and IRGs are thus of great interest in genetic analysis, but are unlikely to be universally fundamental to cellular maintenance.

Acknowledgements

We thank Yingyu Hu and Rosie Sheward for help with sampling. This work was supported by the National Natural Science Foundation of China (32001180 (ZL), 41806129 (YZ)), the Natural Sciences and Engineering Research Council of Canada (NSERC) – Discovery Grants Program (ZVF, AJI), Canada Research Chairs (ZVF) Program and the Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems/CBIOMES (grant nos. 549937 (ZVF) and 549935 (AJI)). The authors declare that they have no conflict of interest.

Author contributions

ZL and ZVF designed research. ZL performed the bioinformatic analysis. ZL, YZ and WL performed the experiments. ZL, AJI and ZVF wrote the manuscript. All authors contributed to the corrections of the manuscript.

ORCID

Zoe V. Finkel https://orcid.org/0000-0003-4212-3917
Andrew J. Irwin https://orcid.org/0000-0001-7784-2319
Wei Li https://orcid.org/0000-0002-4127-8311
Zhengke Li https://orcid.org/0000-0001-8735-2313
Yong Zhang https://orcid.org/0000-0001-8805-1205

Data availability

The data that support the findings of this study are available in the Supporting Information of this article (Figs S1–S12; Tables S1–S3; Datasets S1–S4). The sequencing data of RNA-seq of this study are available through NCBI SRA (BioProject: PRJNA734969).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.

Adelfi MG, Borra M, Sanges R, Montresor M, Fontana A, Ferrante MI. 2014. Selection and validation of reference genes for qPCR analysis in the pennate diatoms *Pseudo-nitzschia multistriata* and *P. arenysensis*. *Journal of Experimental Marine Biology and Ecology* **451**: 74–81.

Alba MM, Castresana J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. *Molecular Biology and Evolution* **22**: 598–606.

Alexander H, Jenkins BD, Rynearson TA, Saito MA, Mercier ML, Dyhrman ST. 2012. Identifying reference genes with stable expression from high throughput sequence data. *Frontiers in Microbiology* **3**: 385.

Andrews S. 2010. *FastQC: a quality control tool for high throughput sequence data.* Cambridge, UK: Babraham Bioinformatics, Babraham Institute.

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M *et al.* 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**: 79–86.

Bach LT, Mackinder LC, Schulz KG, Wheeler G, Schroeder DC, Brownlee C, Riebesell U. 2013. Dissecting the impact of CO_2 and pH on the mechanisms of photosynthesis and calcification in the coccolithophore *Emiliania huxleyi*. *New Phytologist* **199**: 121–134.

Basu S, Patil S, Mapleson D, Russo MT, Vitale L, Fevola C, Maumus F, Casotti R, Mock T, Caccamo M *et al.* 2017. Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. *New Phytologist* **215**: 140–156.

Berges JA, Franklin DJ, Harrison PJ. 2001. Evolution of an artificial seawater medium: improvements in enriched seawater, artificial water over the last two decades. *Journal of Phycology* **37**: 1138–1145.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP *et al.* 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**: 239–244.

Butte AJ, Dzau VJ, Glueck SB. 2001. Further defining housekeeping, or "maintenance", genes focus on "a compendium of gene expression in normal human tissues". *Physiological Genomics* **7**: 95–96.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. *Nature Genetics* **31**: 415–418.

Chang CW, Cheng WC, Chen CR, Shu WY, Tsai ML, Huang CL, Hsu IC. 2011. Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS ONE* **6**: e22859.

Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. 2017. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant Journal* **89**: 789–804.

Consortium U. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47**: D506–D515.

Conway JR, Lex A, Gehlenborg N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**: 2938–2940.

Deng Y, Hu Z, Ma Z, Tang YZ. 2016. Validation of reference genes for gene expression studies in the dinoflagellate *Akashiwo sanguinea* by quantitative real-time RT-PCR. *Acta Oceanologica Sinica* **35**: 106–113.

Ding Y, Sun H, Zhang R, Yang Q, Liu Y, Zang X, Zhang X. 2015. Selection of reference gene from *Gracilaria lemaneiformis* under temperature stress. *Journal of Applied Phycology* **27**: 1365–1372.

Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. *Trends in Genetics* **19**: 362–365.

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends in Genetics* **29**: 569–574.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**: 3047–3048.

Falcon S, Gentleman R. 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**: 257–258.

Falkowski PG, Barber RT, Smetacek V. 1998. Biogeochemical controls and feedbacks on ocean primary production. *Science* **281**: 200–206.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**: 237–240.

Guo R, Ki JS. 2012. Evaluation and validation of internal control genes for studying gene expression in the dinoflagellate *Prorocentrum minimum* using real-time PCR. *European Journal of Protistology* **48**: 199–206.

Guo R, Lee MA, Ki JS. 2013. Normalization genes for mRNA expression in the marine diatom *Ditylum brightwellii* following exposure to thermal and toxic chemical stresses. *Journal of Applied Phycology* **25**: 1101–1109.

Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P *et al.* 2001. A compendium of gene expression in normal human tissues. *Physiological Genomics* **7**: 97–104.

Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin X. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**: 357–360.

Lee M-A, Guo R, Ebenezer V, Ki J-S. 2015. Evaluation and selection of reference genes for ecotoxicogenomic study of the green alga *Closterium ehrenbergii* using quantitative real-time PCR. *Ecotoxicology* **24**: 863–872.

Lee SR, Jo MJ, Lee JE, Koh SS, Kim SY. 2007. Identification of novel universal housekeeping genes by statistical analysis of microarray data. *BMB Reports* **40**: 226–231.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li SW, Feng L, Niu DK. 2007. Selection for the miniaturization of highly expressed genes. *Biochemical and Biophysical Research Communications* **360**: 586–592.

Li ZK, Li W, Zhang Y, Hu Y, Sheward R, Irwin AJ, Finkel ZV. 2021. Dynamic photophysiological stress response of a model diatom to ten environmental stresses. *Journal of Phycology* **57**: 484–495.

Lin F, Jiang L, Liu Y, Lv Y, Dai H, Zhao H. 2014. Genome-wide identification of housekeeping genes in maize. *Plant Molecular Biology* **86**: 543–554.

Lohbeck KT, Riebesell U, Reusch TB. 2014. Gene expression changes in the coccolithophore *Emiliania huxleyi* after 500 generations of selection to ocean acidification. *Proceedings of the Royal Society B: Biological Sciences* **281**: 20140003.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**: 550.

Lv W, Zheng J, Luan M, Shi M, Zhu H, Zhang M, Lv H, Shang Z, Duan L, Zhang R *et al.* 2015. Comparing the evolutionary conservation between human essential genes, human orthologs of mouse essential genes and human housekeeping genes. *Briefings in Bioinformatics* **16**: 922–931.

Mock T, Otillar RP, Strauss J, McMullan M, Paajanen P, Schmutz J, Salamov A, Sanges R, Toseland A, Ward BJ *et al.* 2017. Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. *Nature* 541: 536–540.

Nakamura T, Yamada KD, Tomii K, Katoh K. 2018. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* 34: 2490–2492.

Nicot N, Hausman JF, Hoffmann L, Evers D. 2005. Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress. *Journal of Experimental Botany* 56: 2907–2914.

O'Brien KP, Remm M, Sonnhammer EL. 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research* 33: D476–D480.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* 33: 290–295.

R Development Core Team. 2020. *R: a language and environment for statistical computing, v. 4.0.3*. [WWW document] URL www.r-project.org [accessed 10 October 2020].

Ren XY, Vorst O, Fiers MW, Stiekema WJ, Nap JP. 2006. In plants, highly expressed genes are the least compact. *Trends in Genetics* 22: 528–532.

She X, Rohl CA, Castle JC, Kulkarni AV, Johnson JM, Chen R. 2009. Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC Genomics* 10: 269.

Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia A, Falciatore A, Bowler C. 2007. Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. *Gene* 406: 23–35.

Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A, Heinen E. 1999. Housekeeping genes as internal standards: use and limits. *Journal of Biotechnology* 75: 291–295.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* 28: 511–515.

Tréguer P, Bowler C, Moriceau B, Dutkiewicz S, Gehlen M, Aumont O, Bittner L, Dugdale R, Finkel Z, Iudicone D *et al.* 2018. Influence of diatom diversity on the ocean biological carbon pump. *Nature Geoscience* 11: 27–37.

Vinogradov AE. 2001. Intron length and codon usage. *Journal of Molecular Evolution* 52: 2–5.

Vinogradov AE. 2004. Compactness of human housekeeping genes: selection for economy or genomic design? *Trends in Genetics* 20: 248–253.

Vishnoi A, Kryazhimskiy S, Bazykin GA, Hannenhalli S, Plotkin JB. 2010. Young proteins experience more variable selection pressures than old proteins. *Genome Research* 20: 1574–1581.

Warrington JA, Nair A, Mahadevappa M, Tsyganskaya M. 2000. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiological Genomics* 2: 143–147.

Watson JD. 2004. *Molecular biology of the gene, 5th edn*. Noida, IN, USA: Pearson Education India.

Wickham H. 2016. *ggplot2: elegant graphics for data analysis*. R package v.3.3.2. [WWW document] URL. https://cran.r-project.org/web/packages/ggplot2/index.html [accessed 17 June 2020].

Wingett SW, Andrews S. 2018. FastQ screen: a tool for multi-genome mapping and quality control. *F1000Research* 7: 1338.

Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences, USA* 106: 7273–7280.

Yang Z. 2007. PAML4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586–1591.

Zhang L, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular Biology and Evolution* 21: 236–239.

Zhang Y, Li D, Sun B. 2015. Do housekeeping genes exist? *PLoS ONE* 10: e0123691.

Zhang Y, Li Q, Xu L, Qiao X, Liu C, Zhang S. 2020. Comparative analysis of the P-type ATPase gene family in seven *Rosaceae* species and an expression analysis in pear (*Pyrus bretschneideri* Rehd.). *Genomics* 112: 2550–2563.

Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, Dai L. 2012. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochemical and Biophysical Research Communications* 419: 779–781.

Zhao S, Guo Y, Shyr Y. 2012. *KEGGprofile: An annotation and visualization package for multi-types and multi-groups expression data in KEGG pathway*. R package v.1.32.0. [WWW document] URL https://bioconductor.org/packages/KEGGprofile/ [accessed 21 October 2020].

Zhu J, He F, Hu S, Yu J. 2008. On the nature of human housekeeping genes. *Trends in Genetics* 24: 481–484.

Zuo G, Hao B. 2015. CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. *Genomics, Proteomics & Bioinformatics* 13: 321–331.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 A schematic diagram of the experimental set-up.

Fig. S2 Differential expression analysis of *Thalassiosira pseudonana* housekeeping genes and internal reference genes extracted from *T. pseudonana* data in DiatomPortal (56 treatments analysed).

Fig. S3 Differential expression analysis of orthologs of *Thalassiosira pseudonana* housekeeping genes and internal reference genes in *Phaeodactylum tricornutum* archived in Diatomicsbase (39 treatments analysed).

Fig. S4 Covariation between selected gene statistics (Table 1).

Fig. S5 The number of housekeeping genes with different relative expression breadth (REB, %) and with orthologs across different groups of organisms from Fig. 2(c).

Fig. S6 The relationship between the evolutionary origin of housekeeping genes and their evolution rate (non-synonymous to synonymous (dN : dS)).

Fig. S7 The relationship between the evolutionary origin of all 11 673 genes of *Thalassiosira pseudonana* and their evolution rate (non-synonymous to synonymous (dN : dS)), gene length, coding sequence length and expression level transcripts per million (TPM).

Fig. S8 Gene ontology cellular component and molecular function analysis of the housekeeping genes in *Thalassiosira pseudonana.*

Fig. S9 Expression level TPM of 37 previous commonly used internal reference genes in *Thalassiosira pseudonana* under different conditions.

Fig. S10 Differential expression ($\log_2$FC) of 37 previous commonly used internal reference genes in *Thalassiosira pseudonana* under different conditions.

Fig. S11 Expression level TPM and differential expression ($\log_2$FC) of the seven internal reference genes in *Thalassiosira pseudonana* under different conditions.

Fig. S12 Expression level TPM, coefficient of variation (CV) of TPM (%), and differential expression ($\log_2$FC) of the 19 genes under the identification of IRGs in *Thalassiosira pseudonana* under all conditions.

Table S1 Experimental set-up for each treatment.

Table S2 The rank correlation ρ between non-synonymous to synonymous (dN : dS) values and TPM, $\log_2$FC and relative expression breadth using Spearman correlation analysis.

Table S3 The distribution of the seven new internal reference genes with orthologs in 36 other organisms from Fig. 2(c).

Dataset S1 Description of raw read files.

Dataset S2 Summary of published literature identifying internal reference genes in phytoplankton, protists and some macroalgae.

Dataset S3 Annotations of housekeeping genes and internal reference genes from this work and selected literature.

Dataset S4 Kyoto encyclopedia of genes and genomes and gene ontology enrichment analysis of housekeeping genes.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.