Marine Genomics 51 (xxxx) xxxx

Contents lists available at ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen

Method paper

# t-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological transcriptomic analysis

Matthew C. Cieslak[a], Ann M. Castelfranco[a], Vittoria Roncalli[a,b], Petra H. Lenz[a,*], Daniel K. Hartline[a]

[a] Pacific Biosciences Research Center, University of Hawai'i at Mānoa, 1993 East-West Rd., Honolulu, HI 96822, USA
[b] Department of Genetics, Microbiology and Statistics, Facultat de Biologia, IRBio, Universitat de Barcelona, Av. Diagonal 643, 08028 Barcelona, Spain

ARTICLE INFO

*Keywords:*
Omics
RNA-Seq
Bioinformatics
Zooplankton
Copepod

ABSTRACT

High-throughput RNA sequencing (RNA-Seq) has transformed the ecophysiological assessment of individual plankton species and communities. However, the technology generates complex data consisting of millions of short-read sequences that can be difficult to analyze and interpret. New bioinformatics workflows are needed to guide experimentation, environmental sampling, and to develop and test hypotheses. One complexity-reducing tool that has been used successfully in other fields is "t-distributed Stochastic Neighbor Embedding" (t-SNE). Its application to transcriptomic data from marine pelagic and benthic systems has yet to be explored. The present study demonstrates an application for evaluating RNA-Seq data using previously published, conventionally analyzed studies on the copepods *Calanus finmarchicus* and *Neocalanus flemingeri*. In one application, gene expression profiles were compared among different developmental stages. In another, they were compared among experimental conditions. In a third, they were compared among environmental samples from different locations. The profile categories identified by t-SNE were validated by reference to published results using differential gene expression and Gene Ontology (GO) analyses. The analyses demonstrate how individual samples can be evaluated for differences in global gene expression, as well as differences in expression related to specific biological processes, such as lipid metabolism and responses to stress. As RNA-Seq data from plankton species and communities become more common, t-SNE analysis should provide a powerful tool for determining trends and classifying samples into groups with similar transcriptional physiology, independent of collection site or time.

## 1. Introduction

Key to assessing the health of an ecosystem is to know the physiological states of its inhabitants. Physiology governs progressions through life cycles, adaptiveness to environment including ability to cope with stressors, and reproductive success. But the "physiological state" of an organism is not easily determined because it comprises a multitude of interacting chemical and physical processes governed by gene expression and gene regulatory networks operating simultaneously in multiple organs and tissues of the body. In addressing the problem at the gene-expression level, much of ecology, including biological oceanography, currently depends on identifying the condition of pre-selected biological processes using single-gene (or small gene-set) biomarkers for expression profiling of transcriptional physiological state. For example, gene expression studies on key zooplankton species such as *Calanus finmarchicus*, have typically focused on relative expression of target genes as biomarkers of physiological responsiveness (e.g. stress markers: Voznesensky et al., 2004; Tarrant et al., 2008; Hansen et al., 2008, 2010; Aruda et al., 2011; Roncalli et al., 2016c). More recently a broader approach has been taken: two *de novo* reference transcriptomes have been assembled for this species using high-throughput Illumina sequencing (RNA-Seq) of short RNA fragments (Lenz et al., 2014; Tarrant et al., 2014). These references have been used to investigate global differences in gene expression as a function both of development and experimental treatment (Lenz et al., 2014; Tarrant et al., 2014, 2016; Roncalli et al., 2016a, 2016b, 2017). In these studies, data analysis was focused on comparing pairs of samples, and downstream analysis examined differentially expressed genes (DEGs) between treatments. In some studies, the data were then used to identify a small number of genes to serve as biomarkers, thus reenlisting a targeted-gene approach (e.g., Tarrant et al., 2014). Because RNA-Seq produces millions of short sequences ("reads") from RNA fragments that generate gene expression profiles for an entire small planktonic organism, these high dimensionality datasets can be difficult to interpret and compare across samples.

* Corresponding author at: Pacific Biosciences Research Center, University of Hawai'i at Mānoa, 1993 East-West Rd., Honolulu, HI 96822, USA.
*E-mail address:* petra@hawaii.edu (P.H. Lenz).

https://doi.org/10.1016/j.margen.2019.100723
Received 5 September 2019; Received in revised form 20 October 2019; Accepted 1 November 2019

**Table 1**
Summary of publicly accessible RNA-Seq data used in the method development.

| Bioproject | Species | Experiment | Developmental stage | Samples |
|---|---|---|---|---|
| PRJNA236528 | *Calanus finmarchicus* | Development | Multiple | 6 |
| PRJNA312028 | *Calanus finmarchicus* | Toxic alga | Adult females | 18 |
| PRJNA328961 | *Calanus finmarchicus* | Development | Multiple | 16 |
| PRJNA356331 | *Calanus finmarchicus* | Toxic alga | Late nauplii | 6 |
| PRJNA496596 | *Neocalanus flemingeri* | Ecological survey | Stage CV | 18 |

The field of cell biology has gone beyond the search for biomarkers within transcriptomes, succeeding in using most of the information available in transcriptome-wide profiles analyzed "agnostically" without *a priori* assumptions about how samples might be categorized into groups for statistical comparisons (reviewed by Andrews and Hemberg, 2018). Such approaches produce a more complete, unbiased and nuanced determination of physiological state and/or cell type (e.g., Sebé-Pedrós et al., 2018). In such approaches, profiles of the "*N*" genes an organism expresses are represented mathematically by an *N*-dimensional state-vector. Each axis in *N*-space corresponds to one of the genes, with its coordinate value corresponding to the gene's expression level. How similar in transcriptomic profile two organisms are may then be determined by assessing how "close" their two transcriptional state-vectors are in *N*-space. A collection of organisms (or cells or treatment responses) with similar physiological states generates a cluster of state-vectors in *N*-space. The challenge then becomes how to recognize and interpret these clusters of *N*-vectors. Here we show the application and robustness of a technique termed "t-distributed Stochastic Neighbor Embedding," or "t-SNE" (van der Maaten and Hinton, 2008). This state-of-the-art technique is being used increasingly for dimensionality-reduction of large datasets. It has not, to our knowledge, been applied to high-dimensional biological data in oceanography or marine biology. We demonstrate its use for guiding the investigation and interpretation of transcriptional data by showing its effectiveness when applied to previously-published well-analyzed field-collections and experiments on the calanoid copepods *Calanus finmarchicus* and *Neocalanus flemingeri*.

## 2. t-SNE

Dimensionality reduction is a necessary step in the extraction of the most significant features from a complex set of expression profiles from different samples, treatments or origins that involve thousands of simultaneously-sampled genes. This consists of mapping the high-dimensional state-vectors onto a low-dimensional space (typically a plane) without losing critical information on the relatedness of the component samples. Some standard methods for this in use in biological oceanography, among other fields, include principal component analysis (PCA; e.g. Hotelling, 1933; Ma and Dai, 2011) and multi-dimensional scaling (MDS; e.g., Torgerson, 1952; Tenenbaum et al., 2000; Tzeng et al., 2008; Groenen and Borg, 2014). Another related tool is hierarchical clustering (e.g., with the Bray-Curtis similarity index), which is used for grouping samples according to similarity (Batta-Lona et al., 2017). Like other dimensionality-reduction methods, t-SNE generates a 2-dimensional (or 3D) visualization of sample interrelations that allows close similarities between samples to be identified by the relative location of mapped points. The methods all strive to retain in their mapping the proximity of similar samples while placing dissimilar samples at greater distances. However, because of t-SNE's nonlinearity and its ability to control the trade-off between local and global relationships among points, it usually produces more visually-compelling clusters when compared with the other methods (see e.g. Taskesen and Reinders, 2016). t-SNE can be readily applied to transcriptomic as well as other large high-dimensionality datasets (e.g. Shekhar et al., 2016; Wu et al., 2017; Lima and Rheuban, 2018). At the outset, however, we reiterate a caution frequently voiced in the literature, that as powerful as the technique is, it is important not to over-interpret the plots it generates. We will address some of these issues below.

In this paper, we present no new research data, but rather we re-analyze previously published peer-reviewed results to demonstrate and validate the application of the t-SNE algorithm to these data. We test and evaluate t-SNE over a range of parameters, using transcriptomic data from multiple sources, different sequencing depths, and with the application of filters. We propose guidelines for optimizing the method to a particular dataset obtained from field samples or lab experiments on individuals or small numbers of zooplankton, and we demonstrate how the algorithm can lead to novel and robust interpretations of the exemplar datasets.

## 3. Materials and methods

### 3.1. Dataset description

RNA-Seq data available through Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI) were used in the t-SNE analysis (Table 1; Supplementary Table S1). These datasets are for two species of high-latitude copepods: *Calanus finmarchicus* and *Neocalanus flemingeri*, and detailed descriptions of the data and analyses have been published previously (Christie et al., 2014, 2016; Lenz et al., 2014; Roncalli et al., 2016a, 2016b, 2017, 2019).

The *Calanus finmarchicus* dataset included sequencing data from different developmental stages and for individuals incubated for two and five days under different experimental conditions (control, low dose and high dose of the toxic dinoflagellate *Alexandrium fundyense*) (Table 1). The data from the developmental stages (2011, 2012) were from either wild-caught (copepodid stages CIV, CV and adult females) from the Gulf of Maine or reared in the laboratory from eggs produced by wild-caught females. All *C. finmarchicus* samples contained multiple individuals ranging from three (adult females, CVs) to 500 (embryos).

The *Neocalanus flemingeri* dataset consisted of pre-adults (copepodid CV) that were wild-caught over a 6-day period (May 5–10, 2015) at six stations along the Seward Line and in Prince William Sound, Alaska (Roncalli et al., 2019). RNA-Seq was performed on individual copepods that had been collected using a vertical tow from 100 m to the surface with a CalVET net and preserved in RNAlater RNA Stabilization Reagent (QIAGEN) within two hours of collection. The robustness of t-SNE performance under different levels of sample transcriptional heterogeneity was tested by comparing runs on *N. flemingeri* samples, made up of single individuals, with runs on *C. finmarchicus*, comprising multiple individuals.

### 3.2. Workflow for t-SNE application to RNA-Seq datasets

Fig. 1 diagrams the t-SNE workflow that we applied to the datasets. The Illumina RNA-Seq quality-filtered reads (Table 1) were mapped using Bowtie2 (v. 2.1.0) (Langmead et al., 2009) against a specific reference transcriptome. The *C. finmarchicus* reference contained 96,090 transcripts and was derived from a *de novo* assembly (Trinity, v. 2012-03-17-IU_ZIH_TUNED; Haas et al., 2013) from the 2011 data set (Lenz et al., 2014; Roncalli et al., 2016a, 2016b). The *N flemingeri* reference contained 51,743 transcripts and was assembled by Trinity (2.0.6) from a single individual (GAK1-S83R1) (Roncalli et al., 2019). The resulting

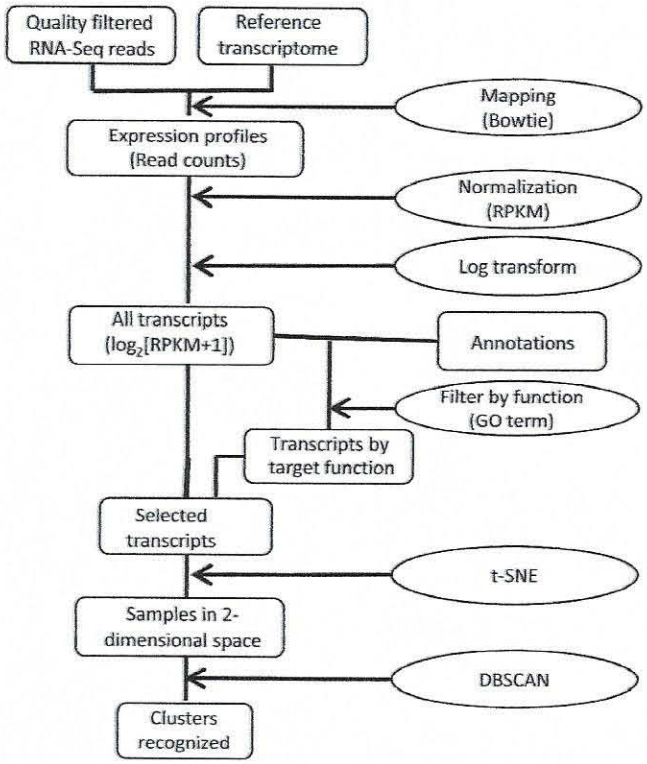


**Fig. 1. Diagram of workflow.** Quality-filtered reads from each sample are mapped against the corresponding species' reference transcriptome (see text for accession information) to produce an expression profile for each sample. The raw counts of reads mapping to each transcript are then normalized to give the profile in reads per kilobase per million mapped reads (RPKM) and log transformed after adding a pseudocount of 1. Optionally, profiles consisting of transcripts successfully annotated and assigned GO terms can be filtered according to a selected GO term. Profiles are then processed by the R package *Rtsne* and plotted as a 2-D scatter plot, one point per sample. Optionally, clusters of sample points can be recognized by application of the R package *dbscan*.

expression profiles (read-counts) were normalized across samples by the length of each transcript, being computed as the number of reads per kilobase of transcript length per million mapped reads (RPKM; Mortazavi et al., 2008; see also Supplementary Methods S1). After adding a pseudocount of 1 to the RPKM value for each transcript, the multi-dimensional dataset of relative expression for thousands of genes for each sample was log-transformed to bring the data closer to a normal distribution (reduce heteroscedasticity). The entire dataset, or a subset filtered for specific Gene Ontology (GO) terms for functional studies using annotated references (see below) was then processed by the R package *Rtsne* (see next). The results were plotted and, in some cases, subjected to further cluster analysis using DBSCAN (see below). An example of the R code used is given in Supplementary methods S2.

### 3.3. t-SNE analysis and theory

Dimensionality reduction methods aim to represent a high-dimensional data set $X = \{x_1, x_2, \ldots, x_N\}$, here consisting of the relative expression of several thousands of transcripts, by a set $Y$ of vectors $y_i$ in two or three dimensions that preserves much of the structure of the original data set and can be displayed as a scatterplot. A nonlinear, dimensionality reduction method, t-SNE seeks to minimize the divergence between the probability distribution, $P$, of pairwise similarities of the data in the high-dimensional space and the probability distribution, $Q$, of pairwise similarities of the corresponding low-dimensional points (van der Maaten and Hinton, 2008). The similarity between two high-dimensional data points, $x_i$ and $x_j$ is based on their Euclidean distance. Joint probabilities $p_{ij}$ that measure the pairwise similarity between the high-dimensional points are defined by symmetrizing the conditional probability that $x_i$ would have $x_j$ as its neighbor, if neighbors were determined in proportion to their probability density under a Gaussian distribution centered at $x_i$. The variance of the Gaussian, $\sigma_i^2$, is set such that the perplexity (defined as $2^{H(Pi)}$, where $H(P_i)$ is the Shannon entropy of the probability distribution $P_i$) of the conditional probability distribution equals a parameter value set by the user (default = 30). Hence, the higher the density of points surrounding $x_i$, the smaller the value of $\sigma_i^2$; that is, the value of the perplexity parameter sets the number of effective neighbors of $x_i$. Perplexity adjusts the trade-off between local and global inclusiveness, with low values favoring local geometries and high values more global ones. The pairwise similarities between corresponding low-dimensional points are computed using a normalized Student's t-distribution with 1 degree of freedom (the "t" in "t-SNE"), which has heavier tails than the Gaussian. The heavy tails of the t-distribution allow moderate distances between points in the high-dimensional space to be modeled by much larger distances in the low dimensional space, which in turn allows small distances to be better represented. The low-dimensional points ($y_i$) are determined by finding points, which minimize the Kulback-Leibler (KL) divergence between the joint probability distributions $P$ and $Q$. This cost function emphasizes modeling high values of $p_{ij}$ by high values of $q_{ij}$, that is, similar objects in the high dimensional space by nearby points in the low dimensional space. Minimization is done using a gradient descent method with an adaptive learning rate scheme and a random initial solution $Y = \{y_1, \ldots, y_N\}$ drawn from a normal distribution. The number of gradient descent iterations can be set (default = 1000) and it must be set sufficiently high to stabilize the pattern of the resulting low dimensional points. We used a variant of t-SNE that utilizes a Barnes-Hut algorithm (Barnes and Hut, 1986) to approximate the t-SNE gradient and is implemented in R (Rtsne URL: https://github.com/jkrijthe/Rtsne;Krijthe, 2015). The Barnes-Hut variant of t-SNE substantially speeds up the algorithm and allows t-SNE to be applied to much larger datasets that would be computationally intractable with the original t-SNE algorithm (van der Maaten, 2014). A consequence of using this algorithm is the requirement that the perplexity parameter must be less than or equal to $N/3$, where $N$ is the number of samples in the original dataset. In addition, the algorithm begins by running PCA to reduce the dimensions of the original data to 30. The resulting 30-dimensional representation is then reduced to two dimensions by t-SNE. This speeds up the computation with little change in the resulting 2-dimensional dataset $Y$. The t-SNE algorithm can be run without the initial PCA step. A consequence of the stochasticity of the t-SNE algorithm is that different runs (realizations) give different results. These different realizations vary primarily in the orientation of the scatterplot within the plane and in the numeric values of the coordinate axes, but not in the grouping of the points within the scatterplots. Thus we follow the common practice of omitting labels from the coordinate axes. A standard technique for handling the stochasticity of t-SNE is to run the algorithm with the same parameter values multiple times and select the solution that gives the smallest KL divergence (i.e. the smallest value of the cost function). However, we found that the variability of the KL divergence by iteration changed from run to run, so the run that gave the smallest KL divergence could be the run with the greatest variability. Hence, instead we selected a t-SNE run that captures the consensus of multiple runs.

### 3.4. Cluster recognition

While a t-SNE plot provides a good visualization of the arrangement of the data points based on similarity, it isn't always clear how to divide these points into subsets or "clusters" such that the points within a cluster are more closely related to each other than those assigned to other clusters. We approached this problem by using a density-based clustering algorithm,

DBSCAN (Ester et al., 1996). The idea underlying this approach is that for point $z$ to belong to a cluster, a neighborhood of $z$ of a given radius, $\varepsilon$ (or "*Eps*"), must contain at least a minimum number of points, that is, the density in the neighborhood has to exceed some threshold. The DBSCAN algorithm has the following distinctive properties: it doesn't require that the number of clusters be specified in advance; it allows clusters to have arbitrary and elongated shapes; it doesn't force all points to belong to a cluster (points that are not in a cluster are called "noise" points) and it requires only minimal knowledge of the input data to determine the parameter values. Two parameters need to be set for the algorithm: 1) *Eps*, the radius of the neighborhood of an interior point of a cluster, and 2) *MinPts*, the minimum number of points that must be contained in an *Eps*-neighborhood of an interior point. Together these parameters define the density threshold for a cluster. DBSCAN provides auxiliary subroutines that help in determining these input parameters (dbscan: https://CRAN.R-project.org/package=dbscan;Hahsler and Piekenbrock, 2018). To choose a value of the parameter *Eps*for a given *MinPts* = $k$, the distances from a point to its $k$ nearest neighbors are computed for all $N$ points, the resulting $k*N$ distances are then ranked from smallest to largest, and plotted as a function of that rank, to produce a $k$-nearest neighbor (kNN) distance plot. The *Eps* parameter can then be chosen to equal the distance that corresponds to the last "knee" (where the slope of the curve changes abruptly) in the plot. This choice for the *Eps* parameter works well when the data form dense clusters of points in a background of sparsely scattered "noise" points. Ester et al. (1996) indicate that for 2-dimensional data, improvement by setting $k > 4$ is minimal, while increasing computational cost. They suggest letting *MinPts* = 4 and using the distances of the 4 nearest neighbors of a point to determine *Eps*. Note, however, that since numerical distances in t-SNE coordinates are meaningless, it is the relative value of *Eps* that is of interest. We computed the Dunn index to compare the quality of the assignment of points to clusters for different values of *Eps* (Dunn, 1974). The Dunn index is defined as the ratio of the minimal distance between points of different clusters to the maximum distance between points within a cluster. Hence, increasing the separation between different clusters or decreasing the diameters of the clusters increase the Dunn index. An optimal value of *Eps* is one that results in the partition of points into clusters with the largest Dunn index. We used the R package clusterCrit to compute the Dunn index (clusterCrit: https://CRAN.R-project.org/package=clusterCrit;Desgraupes, 2018).

### 3.5. Parameter selection

Initial testing involved varying perplexity values and the number of iterations in the Rtsne program in order to select the result that best represented the consensus. This qualitative approach was justified since t-SNE is a visualization tool, not one intended for in-depth quantitative analysis. When used with this in mind, it is a very powerful adjunct for guiding more rigorous down-stream investigations. The effect of changing these two parameters on a single set of samples is shown in Supplementary Figs. S1 and S2. The optimal perplexity setting – one that optimizes the clustering of data points – depends on the number of samples included in the analysis. Rtsne will not accept perplexity above 1/3 of the number of samples. Within broad limits (values $< n/3$), it changes primarily the density of the clusters but not their integrity. The parameter controlling the number of iterations must be set sufficiently high that the arrangement of points in the display is adequately stable with changes in that number (see Supplementary Fig. S2). For the datasets presented here, we found that patterns were well stabilized by 50,000 iterations. However, the optimal number of iterations may differ depending on the nature of the data. There is also an interaction between perplexity and the number of iterations, so exploring a range for both parameters for an untested dataset is advisable. Finally, to allow reproducibility of runs, the seed for R's random number generator was always set to 42. Changing this seed, while keeping other parameters fixed generates plots that are similar in general layout, but differ in detail owing to the stochastic nature of the t-SNE algorithm. The effect of such stochastic differences between re-runs is shown in Supplementary Fig. S3. Unless otherwise stated, other parameters were set to the default values in Rtsne.

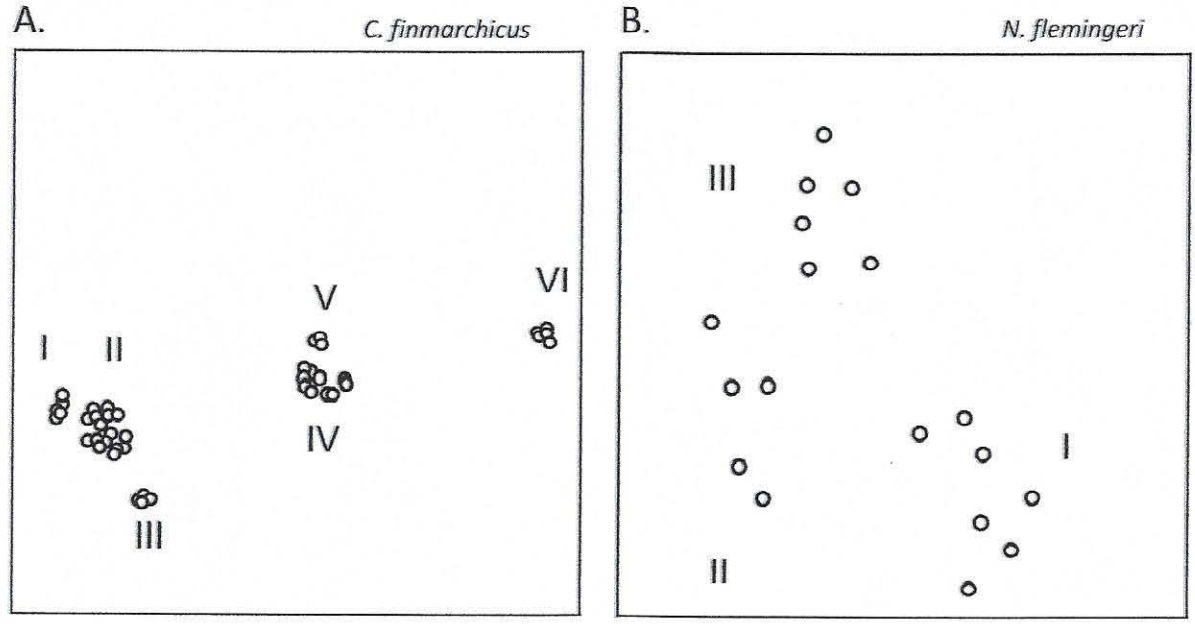


**Fig. 2.** t-SNE plots for samples from two species of high-latitude copepods. Unfiltered RPKM data from Bowtie mapping to species-specific reference transcriptomes were log-transformed before applying the algorithm (Fig. 1). The perplexity parameter was set to 5, and 50,000 iterations of the algorithm were run. A. *Calanus finmarchicus* (developmental stages plus *Alexandrium* feeding experiment, data from Lenz et al., 2014; Roncalli et al., 2016a, 2016b, 2017). 46 samples in all clustered agnostically into 6 groupings (labels I-VI), determined visually. B. *Neocalanus flemingeri* (field collections, data from Roncalli et al., 2019), 18 samples clustered agnostically into 3 groupings (I - III). See Supplementary Tables S2 and S3 for sources.

### 3.6. Selective analysis

The t-SNE algorithm, as well as other clustering algorithms can be applied agnostically, without regard to either sample or transcript identity (annotation). How sample points assort into clusters, if at all, allows preliminary determination by visual inspection of the multiplicity, sizes and variability of distinct transcriptional profiles potentially present in a high-dimensional dataset. However, its use can be extended to subsets of samples or transcripts, selected by criteria of particular interest, to determine how such "filtering" affects the clustering (i.e. insight into transcriptional similarities). We present two such applications. First, we show the effects of identifying samples within a plot according to source and then separating out a subset of interest and reapplying the algorithm, in order to eliminate the influence of samples not relevant to the specific analysis. Second, we demonstrate how transcriptional differences among samples can be related to function by restricting the algorithm to annotated transcripts included in a "Gene Ontology" (GO) term of interest (Ashburner et al., 2000). For the latter, we constructed GO-term filtered input files by finding all of the descendent terms from a higher-level term of interest, then extracting the RPKM values for all transcripts annotated with these terms. We provide exemplar scripts for such a process in Supplementary methods S3.

## 4. Results

### 4.1. Agnostic t-SNE

Fig. 2 presents examples of how the t-SNE algorithm maps transcriptional states of samples treated as an agnostic set (i.e., omitting all source information). Fig. 2A shows the plot for the 46 *Calanus finmarchicus* samples (lab-cultured, wild-caught and experimentally-manipulated) while Fig. 2B shows the plot for the 18 field samples of *Neocalanus flemingeri*. The clustering of sample points for both species is *prima facie* evidence that subsets of samples share similar transcriptional profiles and that these differ from the profiles characterizing the other clusters. In the examples shown, clusters determined visually number at least 6, for the *Calanus* set (designated I through VI in the figure) and three (I - III) for *Neocalanus*. These unbiased initial plots identify multiple distinct transcriptional profiles, proxies for physiological state, represented in the two sets of samples, without indicating what specific gene-expression differences underlie these states. Clustering was robust, occurring over a 7-fold range in the number of mapped reads per sample, and equally apparent in samples consisting of single as well as multiple individuals (see Supplementary Tables S2–S3).

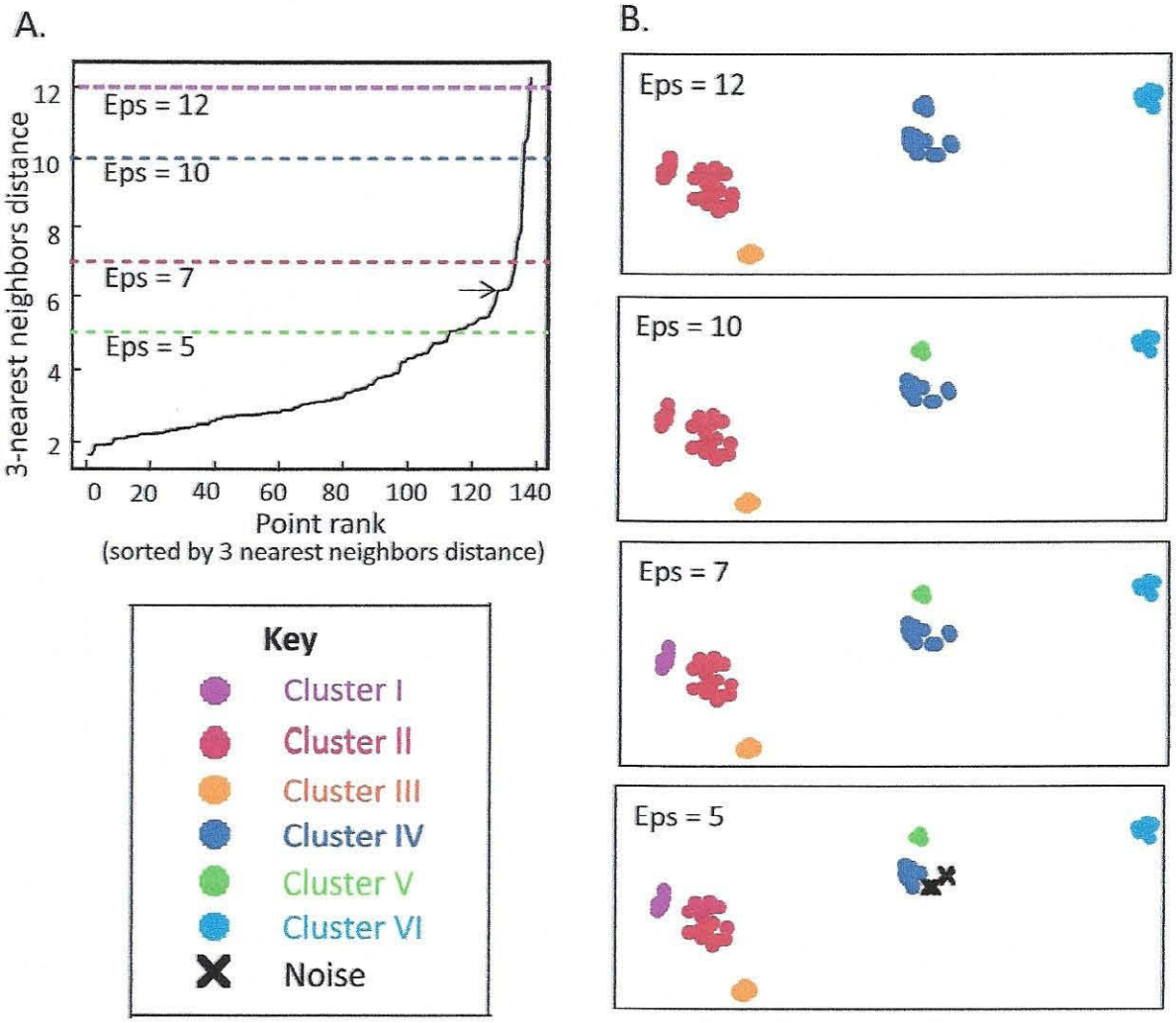


**Fig. 3. DBSCAN assignment of clusters as a function of the neighborhood-radius parameter chosen (*Eps*).** A. For each of 46 points on the t-SNE plot (Fig. 2A, *C. finmarchicus*), distances to the three nearest neighbors were tallied, then the collection of distances (3 × 46 in all) ranked and the distances plotted as a function of rank. Arrow indicates the "knee" in the plot as the best value for $5 < Eps < 7$. B. Clusters resulting from application of DBSCAN algorithm for 4 values of the *Eps* parameter and *MinPts* = 3. Values of *Eps* parameters are in terms of the t-SNE coordinates of the points.

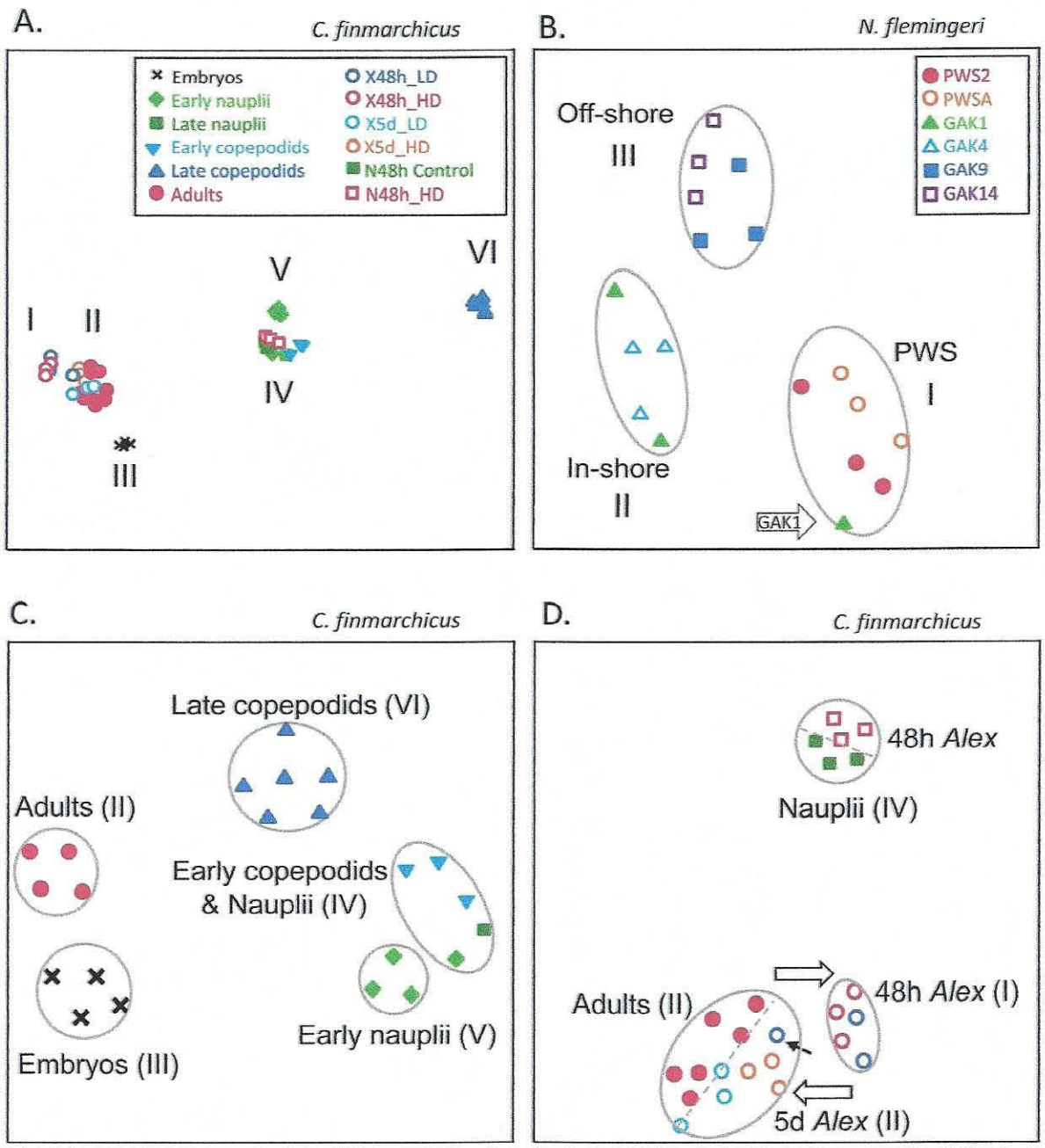


**Fig. 4.** t-SNE plots from Fig. 2, color and symbol-coded according to sample origin (see keys in the top panels). Grey ovals enclose clusters identified by DBSCAN. A. All *C. finmarchicus* samples taken together (see Supplementary Table S2; also Fig. S1 showing examples with different perplexity settings). B. All *N. flemingeri* field samples (see Supplementary Table S3). Open arrow: outlier from GAK1 included in the PWS cluster. C. and D. Subsets of *C. finmarchicus* samples with t-SNE run separately. C. Samples from different developmental stages, as indicated. Stages adjacent in developmental sequence tend to cluster together or map to adjacent clusters. DBSCAN recognized 4 or 5 clusters, depending on *Eps* (the latter case shown). D. Samples from experiments feeding adult females and late nauplii on the toxic dinoflagellate *Alexandrium fundyense*. Adult female samples came from 2-day (48 h Alex) and 5-day (5d Alex) incubations in either a low dose (LD) or a high dose (100%: HD) of *A. fundyense* (open circles); naupliar samples came from 2 days in a high dose (open squares) (controls in corresponding solid symbols). Treatment-dependent segregation is evident within both clusters II and IV (broken lines inside ovals). Open arrows show the transition from cluster II to cluster I for the 2-day *Alexandrium* treatment, and the return to cluster II by the 5th day of treatment. A single 2-day low-dose sample (solid arrow) fell into cluster II rather than I. Parameters: perplexity 5 for panels A. and B.; 7 for C. and D.; number of iterations for panels A. - C. 50,000; for panel D. 10,000.

### 4.2. Cluster identification

Since clustering is used as evidence of a shared physiological transcriptional state, it becomes important to objectively determine what constitutes a cluster. The density-based clustering algorithm DBSCAN described in Section 3.4 (Cluster recognition) provides a formal method for this. Since the data included replicates consisting of three samples, we set the DBSCAN parameter for the minimum number of points in a cluster, *MinPts*, equal to 3. Even had there been more replicates, the results of Ester et al. (1996) would suggest not using a value above 4. Using *MinPts* = 3, we then generated the sorted 3-nearest neighbor (3-NN) distance plot for the *C. finmarchicus* data set as an example (Fig. 3A). The slope of the curve increased abruptly for a distance greater than 6 (arbitrary units); a knee in the resulting curve (arrow)

indicates a good initial choice for the *Eps* parameter. The clusters resulting from four possible choices for *Eps* 5, 7, 10 and 12 are shown in Fig. 3B. As *Eps* is reduced, more clusters result, ranging from 4 to 6 in the panels shown. Both *Eps* = 5 and *Eps* = 7 resulted in 6 clusters, although for the smaller value of *Eps* four of the points were determined to be too isolated and were designated as noise points. The values of the Dunn index for the four *Eps* values were 0.30, 0.59, 0.37 and 0.65, respectively. Since a larger Dunn index value indicates a better separation of the data into clusters, this suggests that identifying either 4 or 6 distinct clusters in the t-SNE plot is an improvement over 5 clusters (*Eps* = 10). A similar approach was used in the cluster analysis reported below.

### 4.3. Restriction by source

The evidence of groups of samples having similar transcriptional

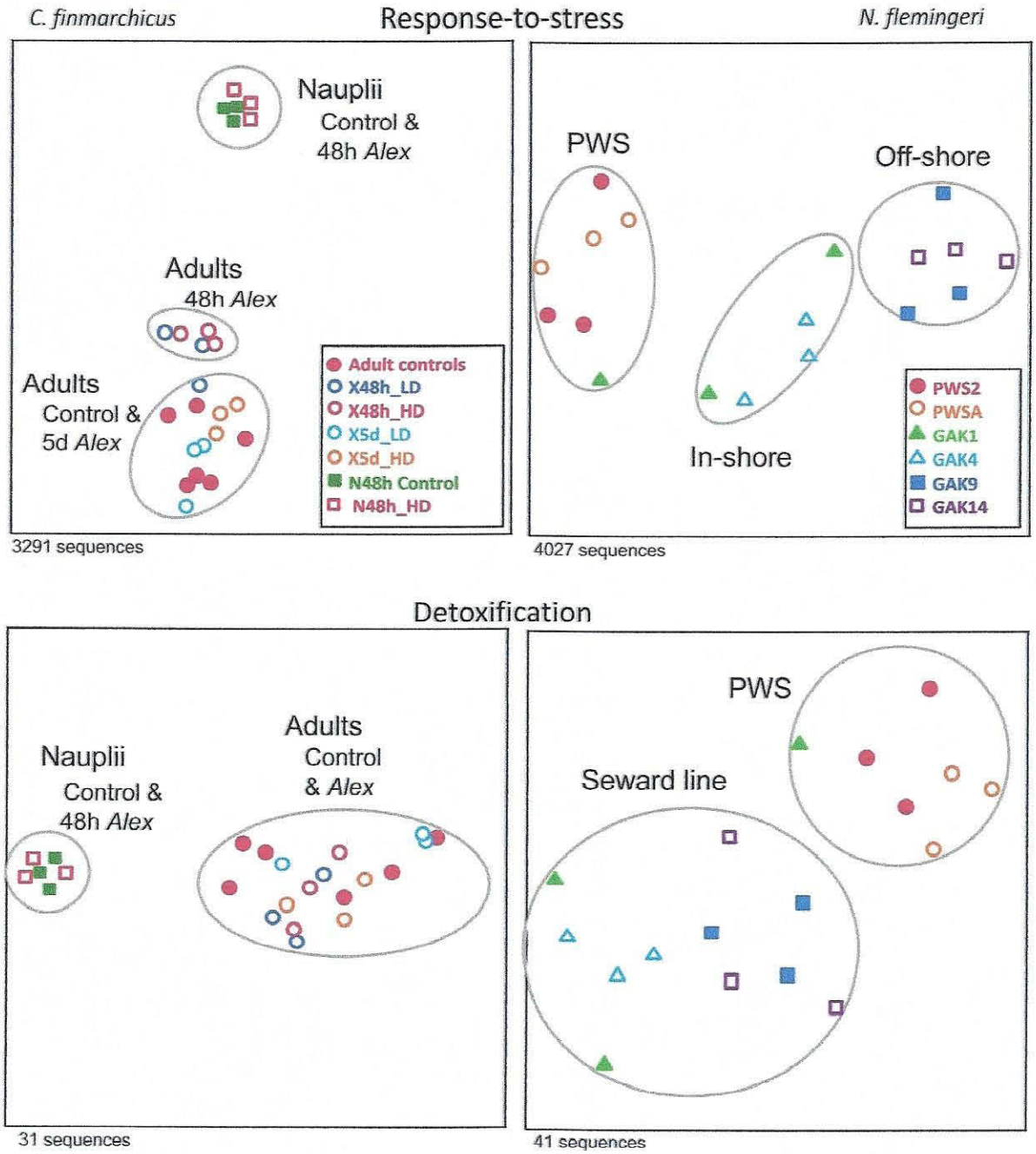


**Fig. 5. t-SNE plots of GO-term filtered transcripts for response to stress [GO:0006950] and detoxification [GO:0098754].** Panels on the left are from the *C. finmarchicus Alexandrium* exposure experiment of Roncalli et al. (2016b, 2017). Panels on the right are from the *N. flemingeri* spatially-distributed field collections (Roncalli et al., 2019). The DBSCAN algorithm was followed by calculation of the Dunn index to determine the optimal grouping of points for each plot (enclosed in grey ovals). The response to stress filter made no changes in clusters (cf. Fig. 4D), while the detoxification filter combined control and experimental *C. finmarchicus* adult samples and in-shore and off-shore *N. flemingeri* samples (cf. Fig. 4B; designated "Seward line"). Perplexity = 7 for *C. finmarchicus*; = 5 for *N. flemingeri*; iterations = 50,000 for both. The number of sequences in the reference transcriptome for each filter is given at the lower left of each panel.

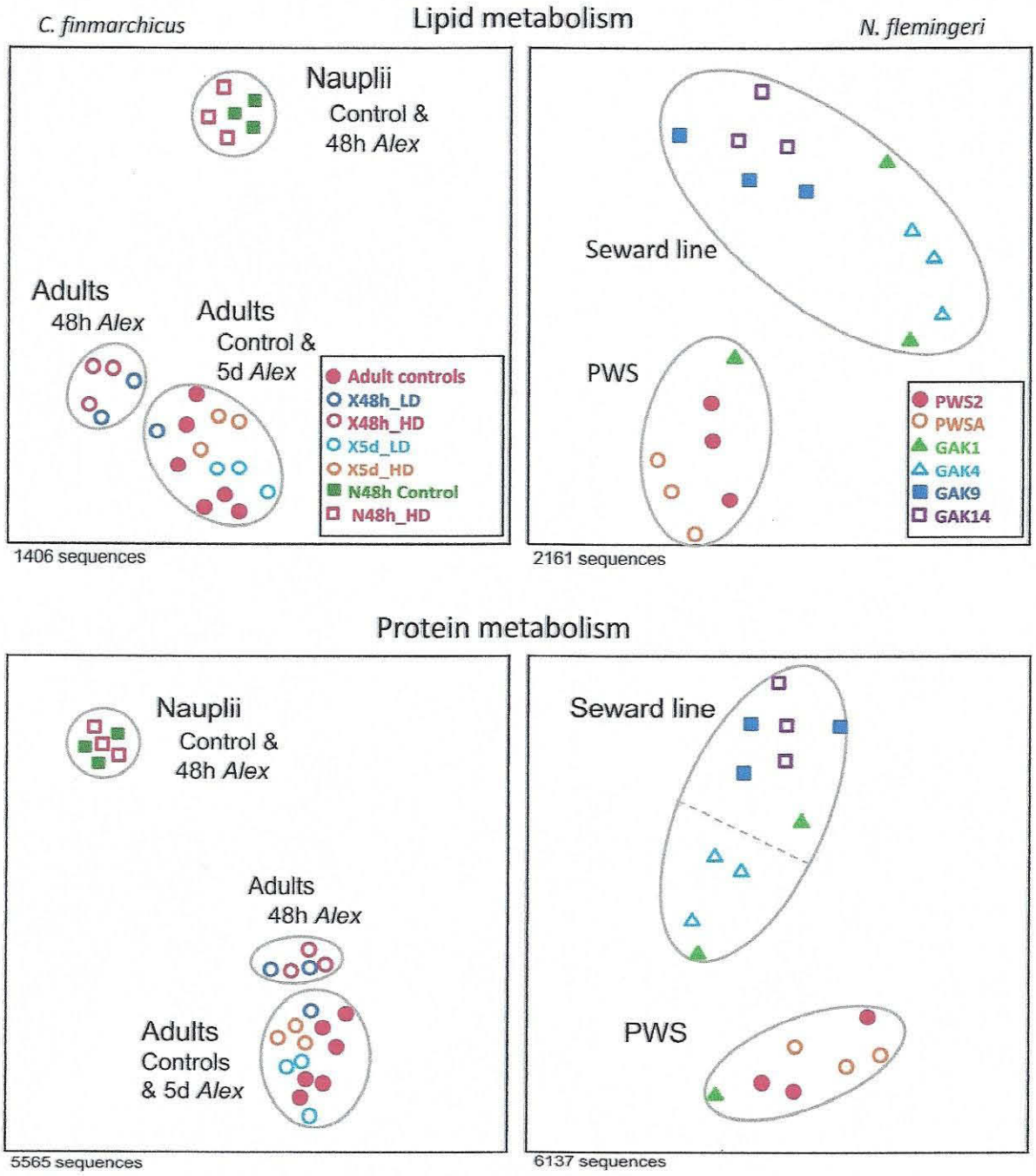


**Fig. 6. t-SNE plots of GO-term filtered transcripts for lipid metabolic process [GO:0006629] and protein metabolic process [GO:0019538].** Panels on the left are from the *C. finmarchicus Alexandrium* exposure experiment of Roncalli et al. (2016b, 2017); Panels on the right are from the *N. flemingeri* spatially-distributed field collections (Roncalli et al., 2019). The DBSCAN algorithm was followed by calculation of the Dunn index to determine the optimal grouping of points for each plot (enclosed in grey ovals). Neither filter changed the *C. finmarchicus* cluster assignments; both filters combined the in-shore and off-shore *N. flemingeri* samples. A 12% reduction in *Eps* for the *N. flemingeri* metabolic process filter resulted in 3 clusters, as indicated by the broken grey line. Perplexity = 7 for *C. finmarchicus*; = 5 for *N. flemingeri*; iterations = 50,000 for both. The number of sequences in the reference transcriptome for each filter is given at the lower left of each panel.

profiles within a diverse collection, and from each of two different species, when processed agnostically by the t-SNE algorithm, leads to the question of what underlying factors characterize the clusters. One feature is immediately apparent in our exemplar datasets when the samples are identified according to source, as shown in Figs. 4A and B for the *C. finmarchicus* and *N. flemingeri* samples, respectively. Here color and symbol-coding according to source has been added to the plots of Fig. 2A and B. Replicate and biologically similar samples map mostly to the same cluster, which is an indication that individuals from one source tend to share transcriptional physiological states. For the *C. finmarchicus* samples of Fig. 4A, Cluster I contained all but one of the adults from Day 2 of an experimental treatment (blue and red open circles); Cluster II contained the control adults (solid red circles) and those from a later time point in the experiment (light blue and orange open circles); Cluster III contained the embryos (black "x"s); and Cluster V the early nauplii (green diamonds); different developmental

stages, late nauplii (solid green and open red squares) and early copepodids (inverted light blue triangles) accounted for all but one of the samples in Cluster IV, and late-stage copepodids Cluster VI (upright blue triangles). The "noise" group identified with *Eps* = 5 (Fig. 3B) turns out to be a composite of one late-stage nauplius and three early-stage copepodids. Only a few samples from a given source are split between clusters (2 out of 46).

### 4.4. Transcriptional heterogeneity in field samples

The field samples of *N. flemingeri* are less diverse than the samples from *C. finmarchicus*, deriving exclusively from one developmental stage (CV). Despite their stage homogeneity, the sample points separated into three clusters. This was substantiated by application of the DBSCAN clustering algorithm followed by computing the Dunn index to determine the value of the *Eps* parameter that gave the optimal grouping of points, as indicated by the grey circles in Fig. 4B. Almost all samples from a given station mapped into the same cluster. The two stations in Prince William Sound (PWS) generated Cluster I; two stations from the shelf region of the Gulf of Alaska (GAK1 and GAK4) combined to produce Cluster II; and two stations from farther offshore (GAK9 and GAK14) together formed Cluster III. This implies a minimum of three distinct transcriptional profiles characterizing the individuals of the three regions. The one exception was an individual from the GAK1 station (arrow), which was more similar to the PWS individuals than to the others from the nearshore shelf region. Comparisons with other dimensionality-reduction methods are shown in Supplementary Figs. S4, S5 and S6. Neither PCA nor MDS (MDS-Sammon, MDS-Kruskal) clustered the samples as well as t-SNE.

### 4.5. Transcriptional heterogeneity in development

Restricting the number of samples prior to t-SNE can produce more refined plots. As an example, the most prominent pattern in the clustering of the combined *C finmarchicus* samples in Fig. 4A is, perhaps not surprisingly, the developmental stage of the copepod. An animal's physiology changes as it develops from embryo to adult, and this is reflected in the source-specific vector clusters. By excluding the experimental treatments from the original dataset, differences among the developmental stages stand out better, as shown in Fig. 4C. Cluster I (an experimental group) disappears, but the remaining ones, as recognized by DBSCAN, remain. The occurrence of successive developmental stages in roughly neighboring positions in the plot is particularly noteworthy, as is the similarity (cluster proximity) between adult females and embryos. The effect of varying the input parameters on this dataset is shown in Supplementary Figs. S1–S3.

### 4.6. Transcriptional heterogeneity from an experimental protocol

The second clustering seen in the original t-SNE output for *C. finmarchicus* related to the experimental animals fed on a diet of the toxic dinoflagellate *Alexandrium fundyense*. In Fig. 4A, in addition to Cluster I, certain regions of both Clusters II and IV include a mixture of experimental-treatment points (open symbols) and controls (solid symbols). A rerun of t-SNE on just the experimental samples and their controls gives somewhat clearer segregation within each cluster, as shown in Fig. 4D. As in the original plot, the naupliar samples form a separate cluster, with controls segregated in one portion of the cluster and toxic-alga treated in the other. The most pronounced clustering for the adult females, as in the original plot, is the separation of the 48-hour treatment from both the controls and the 5-day treatment. The latter treatment clustered with the controls, but the cluster had internal structure with little overlap between controls and treatment. A single "outlier" in the 48-hour treatment (arrow) also clustered with the control/5-day group.

### 4.7. Functional filtering

The analysis just described, while showing a strong correlation of transcriptomic profile with sample source provides no information on which particular genes are responsible for the similarities. The transcriptional state vector represented by a given point is derived from all expressed genes in the transcriptome. To extract more functional insight, the data supplied to the t-SNE algorithm can be pre-filtered to contain expression data restricted to pre-selected groups of genes. Both the appearance of new clusters and the disappearance of original ones can be informative. Such filters can be constructed in multiple ways. We will illustrate one that uses Gene Ontology terms. If separation into two or more clusters is observed when only genes from a selected GO term are included, it flags the corresponding function as a candidate for contributing to the transcriptional difference between samples in those clusters. Four cases are shown in Figs. 5 and 6 when different GO-term filters are applied to the datasets from the *C. finmarchicus* experiment on *Alexandrium*-exposure (left-side panels; Roncalli et al., 2016b, 2017) and the same filters to the field samples of *N. flemingeri* (right-side panels; Roncalli et al., 2019). In each of the *C. finmarchicus/Alexandrium* panels, there are, not surprisingly, at least two clusters, one for the adults and one for nauplii as in the developmental stage collection (left panels). The split into clusters by the experimental treatment that occurred in the unfiltered plot of the *C. finmarchicus* adults (Fig. 4D) was retained under three of the filters: "response to stress" [GO:0006950], "lipid metabolic process" [GO:0006629] and "protein metabolic process" [GO:0019538] (Figs. 5 and 6 left panels). This is evidence that genes that annotated into these biological processes were involved in the response to the toxic alga for *C. finmarchicus*. A similar retention of the spatially-distinct clustering occurred for *N. flemingeri* under the response to stress filter (Fig. 5 upper right panel). On the other hand, the filter using the GO term "detoxification" [GO:0098754] (Fig. 5, lower panels) eliminated the split in both the adult *C. finmarchicus* experimental samples and that between the two shelf areas in the *N. flemingeri* samples. This last condensation was also seen for the lipid metabolic process and the protein metabolic process filters applied to *N. flemingeri* samples, thus deemphasizing the likely contributions of genes in those functional categories to the separation between inner and outer shelf stations.

## 5. Discussion

### 5.1. Comparison with published results

We have argued that the separation of sample points into several distinct clusters by the t-SNE algorithm is good evidence both of transcriptional similarity within clusters and differences among clusters. In what follows, we validate this by comparing the t-SNE clustering results with the differential gene expression analyses reported in the original publications.

#### 5.1.1. Developmental transcriptional profile

The strongest "signal" in the t-SNE analysis for *C. finmarchicus* was the separation of the developmental stages into stage-specific clusters (Fig. 4C), which implies distinct transcriptional profiles for the different stages. The cause undoubtedly lies in the fact that the developmental trajectory of any organism involves complex temporal sequences of gene expression patterns. The strength and inter-annual consistency of the clustering would be expected since the stage-to-stage developmental progression is independent of many environmental factors. This is consistent with mapping results in Lenz et al. (2014), who found that large numbers of reference transcripts failed to be expressed in any one specific developmental stage (*ca.* 30% for each stage). Furthermore, a targeted analysis of the genes involved in lipid synthesis (a key pathway in pre-diapause copepods), showed stage-specific expression of genes such as *diacylglycerol o-acetyl transferase 1*. Such discrepancies in gene

expression would contribute strongly to the separation of points by the t-SNE algorithm.

The t-SNE plots had a tendency to place successive developmental stages in neighboring clusters, suggesting a measure of shared expression profiles. Several nauplii and early copepodids were even clustered together, despite having different body forms (Fig. 4A, C). A difference in cluster association of early nauplii between 2011 ($n = 1$; solid green square in cluster IV) and 2012 ($n = 3$; solid green diamonds in clusters IV and V) may be related to differences in stage bias (% of NI vs. % of NII) between the two years. Expression differences between these samples from the two years were also found in a target gene expression analysis focused on transcripts encoding enzymes in the amine biosynthetic pathways (Christie et al., 2014). The t-SNE result further validates these findings.

More surprisingly, the embryo and the adult female clusters were consistently closer to each other than to the other developmental stages (Fig. 4C), suggesting a transcriptional similarity between them. This proximity was not as readily anticipated from previous and more limited gene expression studies (e.g., Christie et al., 2014; Lenz et al., 2014). However, in retrospect, Lenz et al. (2014) noted that while the distribution of GO terms for the silent genes was for the most part equally represented among the different developmental stages, adults and embryos shared a greater representation in three GO terms than did the other stages (BP: localization [GO:0051179]; MF: catalytic activity [GO:0003824]; and CC: membrane proteins [GO:0016020]).

#### 5.1.2. Response of transcriptional profiles to toxic challenge

We illustrated the application of the t-SNE tool to experimental treatments that alter transcriptional profiles using data from Roncalli et al. (2016b) on adult female *Calanus finmarchicus* fed on a diet of the toxic dinoflagellate *Alexandrium fundyense* (Fig. 4D). The experimental samples clustered separately from the controls for the acute phase (48-hour) and rejoined the control cluster (albeit still segregated) after 5-days of treatment. Thus, the most prominent difference in gene expression occurred in the short term, with a return over time to transcriptional states closer to (but not intermixed with) the control group. A single "outlier" in the 48-hour treatment rather close to the 5-day (and control) samples is of some interest (Fig. 4D arrow), and was missed in the original analysis. Roncalli et al. (2016b) found that a relatively large fraction of genes expressed by the adult females (4–5%) were differentially expressed with respect to controls in the 48-hour treatment. This is consistent with the separate clustering in the t-SNE plots (Fig. 4D). That fraction was smaller for both the 5-day treatment of the adult females and the 48-hour treatment of late nauplii (2–3%), which would help explain the lack of separate clustering from controls in those groups.

The broad t-SNE clustering thus is consistent with the in-depth functional analysis of Roncalli et al. (2016b). Gene expression differences included a large number of differentially-expressed genes in the 48-hour samples that were associated with the cellular stress response. After 5-days' exposure, the response was characterized by a lower number of differentially-expressed genes, which is consistent with a return to cellular homeostasis (Kültz, 2005). These results also explain the persistence of t-SNE clusters under a stress-response filter (Fig. 5, upper left panel). Furthermore, lack of clustering when the "detoxification" filter is applied (Fig. 5, lower left panel) is consistent with Roncalli et al.'s (2016b, 2016c, 2017) conclusion that regulation of detoxification genes was not a significant response to the toxic diet.

#### 5.1.3. Transcriptional profiles of spatially-distributed Neocalanus flemingeri

Transcriptomic technology has the potential to revolutionize biological oceanography provided that data can be analyzed and interpreted. For example, RNA-Seq of field-collected phytoplankton led to new hypotheses on niche separation between two diatom species, which were then tested experimentally (e.g. Alexander et al., 2015a). Transcriptional profiling of *N. flemingeri* pre-adults (copepodid stage CV) demonstrated how regional heterogeneity in resource availability is affecting the physiology of a copepod during preparation for diapause (dormancy) (Roncalli et al., 2019). The t-SNE plots from six field stations in the Gulf of Alaska and a bordering embayment illustrate the application of this technology in *N. flemingeri*. Three clusters comprising two stations each suggest regional differences in transcriptional physiology within the same developmental stage of a genetically mixed population of *N. flemingeri*. It is significant that at this global level the samples fell into only three profile categories despite originating from six stations spread over 300 km of distance.

The large and consistent separation between PWS and the two offshore GAK stations described by Roncalli et al. (2019) is underscored in all t-SNE plots. The t-SNE plots using multiple filters also highlight the complexity of the gene expression patterns, as the relationship of the inshore GAK stations (GAK1 and GAK4) changes depending on GO term filter. Individuals from Prince William Sound showed up-regulation of transcripts involved in lipid biosynthesis, while in the Seward-line individuals (mostly GAK4-GAK14) up-regulation was found for genes involved in lipid catabolism and protein degradation. The occurrence of separate PWS and Seward-line t-SNE clusters using the "lipid metabolic process" GO filter was consistent with this. In addition, Roncalli et al. (2019) found differential expression of genes involved in response to stress, glutathione metabolism and protein metabolism (e.g. digestion, ubiquitination). Application of the detoxification filter, a process involved in response to stress, in the t-SNE analysis separated individuals from PWS and the shelf regions into two clusters consistent with results obtained by functional analysis. Overall, the comparison demonstrates the application of t-SNE algorithm as a powerful tool for discrimination of transcriptional differences.

The filtering approach offers possibilities for initial functional insights. The development of "designer" filters specific for known transcriptional response-patterns might be a fruitful direction for future refinements of this approach, especially as more data on ecophysiological responsiveness become available. The t-SNE approach, with or without the application of functional filters, provides a rapid assessment of transcriptional similarity among samples. Clusters become a basis for identifying "experimental groups" that can then be further analyzed to characterize transcriptional patterns and identify environmental correlates that contribute to observed differences. Thus t-SNE is an analysis tool that can focus the effort involved in down-stream analysis of differentially expressed genes and their function within a broader ecological context.

### 5.2. Novel insights

So far, we have used published results as confirmatory evidence for the validity of t-SNE clustering in identifying similar gene-expression profiles. However, the t-SNE patterns also identified several anomalies that contribute new insights to the published data. An initial assessment of the data using the t-SNE tool might have led to additional or modified downstream analyses.

In the *Alexandrium* experiment, the 48-hr time point showed differences in transcriptional response between the low-dose (LD) and high-dose (HD) treatments (Roncalli et al., 2016b). However, the t-SNE shows two LD replicates clustering with the three HD replicates, while the third LD sample clustered with the controls/5-day samples. It suggests that the odd replicate was either farther along in its response to the toxic dinoflagellate, or was delayed in responding to it.

The analysis by Roncalli et al. (2019), which categorized each station into a separate "group", used a Generalized Linear Model (GLM) to identify patterns of increased nutritional stress between Prince William Sound and out along the Seward Line (GAK1 to GAK14). The t-SNE analysis in combination with DBSCAN underscores how application of this tool could have strengthened the statistical analysis by reducing the number of "groups" to three, while increasing biological replication. In contrast, PCA and MDS did not generate clear clusters, in particular with respect to the

inner shelf stations (GAK1 and GAK4), which were characterized by greater individual variability in gene expression (Roncalli et al., 2019).

### 5.3. Caveats

As we have shown, the t-SNE algorithm has proven to be quite effective in quickly identifying clusters of similar transcriptomic profiles in samples both from lab experiments and field collections. However, as noted above, there are limitations to the proper application of the tool. A list of limitations and cautionary notes have been summarized in a web publication by M. Wattenberg and F. Viégas ("How to use t-SNE effectively", Distill, 2016. http://doi.org/10.23915/distill.00002). Several of these are relevant to the current application:

1. While the distance between points in the 2D plots represents a measure of the distance between the points in *N*-space, there are trade-offs that require caution in interpretation. Thus the compactness of a cluster (its standard deviation) is not significant because the algorithm performs a density-equalization operation (Wattenberg & Viégas, *loc. cit.*). Originally dense clusters of points tend to be expanded and dispersed clusters contracted by the algorithm. The distances between clusters gives a sense of global geometry, but it requires optimizing the setting of the "perplexity" parameter to develop. Three clusters that are unevenly spaced in the original data may become more evenly spaced by the algorithm. Structuring in the *N*-dimensional original data space can become distorted thereby. While the clusters are robust, the arrangement of clusters on the 2D plane is not necessarily informative and can switch around dramatically in different runs and with different parameters (as for example with the filters in Figs. 5 and 6).

2. Running the algorithm under multiple conditions of the controlling parameters can reduce ambiguity. Owing to its stochastic implementation, it yields somewhat different results each time it is run. Examples of this are given in the supplementary material (Fig. S3). Although the clusters developed are usually consistent, this is not assured. Also, many iterations are required before the algorithm "settles" on a consistent stable pattern (Fig. S2). Thus multiple runs are advisable, as well as checking pattern stability with different iteration lengths. Wattenberg & Viégas (*loc. cit.*) point out that a "pinched" shape in the pattern of points may indicate an insufficient number of iterations.

3. The "perplexity" parameter, as mentioned in Materials and Methods, governs the trade-off between local and global relationships among the vectors. For example, the replicates of a particular sample should be closely associated and might thus represent the minimum perplexity that needs testing (this is why we settled on a perplexity of 5 or 7 for our examples). A perplexity setting greater than the number of points tends to yield a single cluster, which is in a broad sense correct, but hardly helpful. A very low perplexity places each point in its own cluster, which in a sense is correct also (every individual is unique), but not useful. Beware of clustering artifacts from this source. The "right" perplexity depends on the desired level of global vs. local resolution. It is best resolved through multiple tests. An added complication arises in datasets containing clusters with large differences in size among them. Optimal perplexity may differ among such clusters (Wattenberg & Viégas *loc. cit.*).

### 5.4. Applications of t-SNE in marine biology and oceanography

Metagenomic, metagenetic and metatranscriptomic approaches have been applied to bacteria, phytoplankton and zooplankton communities (Martínez et al., 2013; Alexander et al., 2015a; Sommer et al., 2017). Transcriptomics of individual species and communities (metatranscriptomics) is being developed as a tool to investigate physiological responses to experimental manipulations within an ecological context (Marchetti et al., 2012; Alexander et al., 2015a, 2015b). Thus, research programs on plankton populations that are focused on periodic assessments of diversity and abundance can be expected to incorporate routine high-throughput sequencing of DNA and RNA of either target species or whole communities, which will require reduction of high-dimensional data. The t-SNE algorithm is a powerful tool for the initial parsing of big datasets prior to downstream bioinformatic and functional analyses.

## Authors' contribution

DKH, AMC and PHL conceived the study; MCC, DKH and AMC implemented and tested the t-SNE and DBSCAN applications, MCC, DKH, AMC, VR and PHL analyzed the data and evaluated the results and conclusions, DKH, AMC and PHL wrote the manuscript. All authors reviewed and approved the final manuscript.

## Availability of supporting data

Data analyzed here were downloaded from the National Center of Biotechnology Information (NCBI) under Bioprojects PRNJA236528, PRNJA328961, PRNJA312028, PRNJA356331, PRNJA496596. Accession numbers to the RNA-Seq short sequence read data are listed in Supplementary Table S1. The *de novo* assemblies used as reference transcriptomes were downloaded from NCBI: *Calanus finmarchicus* (GAXK00000000.1) and *Neocalanus flemingeri* (GHLB00000000.1). For the *C. finmarchicus* reference, the *de novo* assembly was reduced by including only a single isoform for each "comp" as described in Lenz et al. (2014).

## Declaration of Competing Interest

The authors declare that they have no competing interests.

## Acknowledgements

This work was supported by National Science Foundation (NSF) OCE-1459235 and OCE-1756767 to PHL, DKH and AE Christie and a North Pacific Research Board award NPRB 1709 to PHL. We greatly appreciate the administrative and secretarial support provided by Lynn Hata. This is SOEST contribution number 10841.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.margen.2019.100723.

## References

Alexander, H., Jenkins, B.D., Rynearson, T.A., Dyhrman, S.T., 2015a. Metatranscriptome analyses indicate resource partitioning between diatoms in the field. Proc. Nat. Acad. Sci. USA 112, E2182–E2190.

Alexander, H., Rouco, M., Haley, S.T., Wilson, S.T., Karl, D.M., Dyhrman, S.T., 2015b. Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. Proc. Nat. Acad. Sci. USA 112, E5972–E5979.

Andrews, T.S., Hemberg, M., 2018. Identifying cell populations with scRNASeq. Mol. Asp. Med. 59, 114–122.

Aruda, A.M., Baumgartner, M.F., Reitzel, A.M., Tarrant, A.M., 2011. Heat shock protein expression during stress and diapause in the marine copepod *Calanus finmarchicus*. J. Insect Physiol. 57, 665–675.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., 2000. Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25.

Barnes, J., Hut, P., 1986. A hierarchical O(N log N) force-calculation algorithm. Nature 324, 446.

Batta-Lona, P., Maas, A., O'Neill, R., Wiebe, P., Bucklin, A., 2017. Transcriptomic profiles of spring and summer populations of the Southern Ocean salp, *Salpa thompsoni*, in the Western Antarctic Peninsula region. Polar Biol. 40, 1261–1276.

Christie, A.E., Fontanilla, T.M., Roncalli, V., Cieslak, M.C., Lenz, P.H., 2014. Identification

and developmental expression of the enzymes responsible for dopamine, histamine, octopamine and serotonin biosynthesis in the copepod crustacean *Calanus finmarchicus*. Gen. Comp. Endocrinol. 195, 28–39.

Christie, A.E., Roncalli, V., Lenz, P.H., 2016. Diversity of insulin-like peptide signaling system proteins in *Calanus finmarchicus* (Crustacea; Copepoda) – possible contributors to seasonal pre-adult diapause. Gen. Comp. Endocrinol. 236, 150–169.

Desgraupes, B., 2018. ClusterCrit: Clustering Indices. R package version 1.2.8. https://CRAN.R-project.org/package=clusterCrit.

Dunn, J.C., 1974. Well-separated clusters and optimal fuzzy partitions. J. Cybern. 4, 95–104.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226–231.

Groenen, P., Borg, I., 2014. Past, present, and future of multidimensional scaling. Vis. Verbalization Data 10, 95–117.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, N., Regev, A., 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8, 1494–1512.

Hahsler, M., Piekenbrock, M., 2018. Dbscan: density based clustering of applications with noise (DBSCAN) and related algorithms. R package version 1.1–3. https://CRAN.R-project.org/package=dbscan.

Hansen, B.H., Altin, D., Vang, S.-H., Nordtug, T., Olsen, A.J., 2008. Effects of naphthalene on gene transcription in *Calanus finmarchicus* (Crustacea: Copepoda). Aquat. Toxicol. 86, 157–165.

Hansen, B.H., Altin, D., Booth, A., Vang, S.-H., Frenzel, M., Sørheim, K.R., Brakstad, O.G., Størseth, T.R., 2010. Molecular effects of diethanolamine exposure on *Calanus finmarchicus* (Crustacea: Copepoda). Aquat. Toxicol. 99, 212–222.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417.

Krijthe, J.H., 2015. Rtsne: t-Distributed Stochastic Neighbor Embedding using a Barnes-Hut implementation, version 0.13. https://github.com/jkrijthe/Rtsne.

Kültz, D., 2005. Molecular and evolutionary basis of the cellular stress response. Annu. Rev. Physiol. 67, 225–257.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25.

Lenz, P.H., Roncalli, V., Hassett, R.P., Wu, L.S., Cieslak, M.C., Hartline, D.K., Christie, A.E., 2014. *De novo* assembly of a transcriptome for *Calanus finmarchicus* (Crustacea, Copepoda)—the dominant zooplankter of the North Atlantic Ocean. PLoS One 9, e88589.

Lima, I.D., Rheuban, J.E., 2018. Topics and trends in NSF ocean sciences awards. Oceanography 31, 164–170.

Ma, S., Dai, Y., 2011. Principal component analysis based methods in bioinformatics studies. Brief. Bioinform. 12, 714–722.

Marchetti, A., Schruth, D.M., Durkin, C.A., Parker, M.S., Kodner, R.B., Berthiaume, C.T., Morales, R., Allen, A.E., Armbrust, E.V., 2012. Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. Proc. Nat. Acad. Sci. USA 109, E317–E325.

Martínez, A., Ventouras, L.A., Wilson, S.T., Karl, D.M., DeLong, E.F., 2013. Metatranscriptomic and functional metagenomic analysis of methylphosphonate utilization by marine bacteria. Front. Microbiol. 4, 340.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621.

Roncalli, V., Cieslak, M.C., Lenz, P.H., 2016a. Data from: Transcriptomic Responses of the Calanoid Copepod *Calanus finmarchicus* to the Saxitoxin Producing Dinoflagellate *Alexandrium fundyense*. Dryad. https://doi.org/10.5061/dryad.11978 Dataset.

Roncalli, V., Cieslak, M.C., Lenz, P.H., 2016b. Transcriptomic responses of the calanoid copepod *Calanus finmarchicus* to the saxitoxin producing dinoflagellate *Alexandrium fundyense*. Sci. Rep. 6, 25708.

Roncalli, V., Jungbluth, M.J., Lenz, P.H., 2016c. Glutathione S-transferase regulation in *Calanus finmarchicus* feeding on the toxic dinoflagellate *Alexandrium fundyense*. PLoS One 11, e0159563.

Roncalli, V., Lenz, P.H., Cieslak, M.C., Hartline, D.K., 2017. Complementary mechanisms for neurotoxin resistance in a copepod. Sci. Rep. 7, 14201.

Roncalli, V., Cieslak, M.C., Germano, M., Hopcroft, R.R., Lenz, P.H., 2019. Regional heterogeneity impacts gene expression in the sub-arctic zooplankter *Neocalanus flemingeri* in the northern Gulf of Alaska. Commun. Biol. 2, 1–13.

Sebé-Pedrós, A., Chomsky, E., Pang, K., Lara-Astiaso, D., Gaiti, F., Mukamel, Z., Amit, I., Hejnol, A., Degnan, B.M., Tanay, A., 2018. Early metazoan cell type diversity and the evolution of multicellular gene regulation. Nat. Ecol. Evol. 2, 1176.

Shekhar, K., Lapan, S.W., Whitney, I.E., Tran, N.M., Macosko, E.Z., Kowalczyk, M., Adiconis, X., Levin, J.Z., Nemesh, J., Goldman, M., 2016. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell 166, 1308–1323 e1330.

Sommer, S.A., Van Woudenberg, L., Cepeda, G., Lenz, P.H., Goetze, E., 2017. Vertical gradients in species richness and community composition across the twilight zone in the North Pacific Subtropical Gyre. Mol. Ecol. 26, 6136–6156.

Tarrant, A.M., Baumgartner, M.F., Verslycke, T., Johnson, C., 2008. Differential gene expression in diapausing and active *Calanus finmarchicus* (Copepoda). Mar. Ecol. Prog. Ser. 355, 193–207.

Tarrant, A.M., Baumgartner, M.F., Hansen, B.H., Altin, D., Nordtug, T., Olsen, A.J., 2014. Transcriptional profiling of reproductive development, lipid storage and molting throughout the last juvenile stage of the marine copepod *Calanus finmarchicus*. Front. Zool. 11, 1.

Tarrant, A.M., Baumgartner, M.F., Lysiak, N.S., Altin, D., Størseth, T.R., Hansen, B.H., 2016. Transcriptional profiling of metabolic transitions during development and diapause preparation in the copepod *Calanus finmarchicus*. Integr. Comp. Biol. 56, 1157–1169.

Taskesen, E., Reinders, M.J., 2016. 2D representation of transcriptomes by t-SNE exposes relatedness between human tissues. PLoS One 11, e0149853.

Tenenbaum, J.B., De Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323.

Torgerson, W.S., 1952. Multidimensional scaling: I. Theory and method. Psychometrika 17, 401–419.

Tzeng, J., Lu, H.H.-S., Li, W.-H., 2008. Multidimensional scaling for large genomic data sets. BMC Bioinforma. 9, 179.

van der Maaten, L., 2014. Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. 15, 3221–3245.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Voznesensky, M., Lenz, P.H., Spanings-Pierrot, C., Towle, D.W., 2004. Genomic approaches to detecting thermal stress in *Calanus finmarchicus* (Copepoda: Calanoida). J. Exp. Mar. Biol. Ecol. 311, 37–46.

Wattenberg, M., Viégas, F., Johnson, I., 2016. How to use t-SNE effectively. Distill 1, e2. http://10.23915/distill.00002 downloaded 2019-01-09.

Wu, J., Wang, J., Xiao, H., Ling, J., 2017. Visualization of high dimensional turbulence simulation data using t-SNE. In: 19th AIAA Non-Deterministic Approaches Conference, pp. 1770.