



# communications biology

## ARTICLE

<https://doi.org/10.1038/s42003-021-01946-0>

OPEN

# Diapause vs. reproductive programs: transcriptional phenotypes in a keystone copepod

Petra H. Lenz<sup>1</sup>, Vittoria Roncalli <sup>1,2✉</sup>, Matthew C. Cieslak<sup>1</sup>, Ann M. Tarrant <sup>3</sup>, Ann M. Castelfranco<sup>1</sup> & Daniel K. Hartline<sup>1</sup>

Many arthropods undergo a seasonal dormancy termed “diapause” to optimize timing of reproduction in highly seasonal environments. In the North Atlantic, the copepod *Calanus finmarchicus* completes one to three generations annually with some individuals maturing into adults, while others interrupt their development to enter diapause. It is unknown which, why and when individuals enter the diapause program. Transcriptomic data from copepods on known programs were analyzed using dimensionality reduction of gene expression and functional analyses to identify program-specific genes and biological processes. These analyses elucidated physiological differences and established protocols that distinguish between programs. Differences in gene expression were associated with maturation of individuals on the reproductive program, while those on the diapause program showed little change over time. Only two of six filters effectively separated copepods by developmental program. The first one included all genes annotated to RNA metabolism and this was confirmed using differential gene expression analysis. The second filter identified 54 differentially expressed genes that were consistently up-regulated in individuals on the diapause program in comparison with those on the reproductive program. Annotated to oogenesis, RNA metabolism and fatty acid biosynthesis, these genes are both indicators for diapause preparation and good candidates for functional studies.

<sup>1</sup>Pacific Biosciences Research Center, University of Hawai‘i at Mānoa, Honolulu, HI, USA. <sup>2</sup>Stazione Zoologica Anton Dohrn, Napoli, Italy. <sup>3</sup>Woods Hole Oceanographic Institution, Woods Hole, MA, USA. ✉email: [vittoria.roncalli@szn.it](mailto:vittoria.roncalli@szn.it)

Large copepods are at the base of the metazoan food web of high-latitude marine ecosystems that support highly productive fisheries<sup>1–3</sup>. Low recruitment of young-of-the-year fish larvae in the North Atlantic, North Pacific and the Bering Sea have been correlated with below-average abundances of these lipid-rich copepods<sup>4–8</sup>. While their population abundances do correlate negatively with temperature<sup>5,9</sup>, observed temperatures are well within known species' tolerances<sup>10</sup>, suggesting that indirect effects may be more important. Changes in ocean circulation patterns and the timing and magnitude of spring phytoplankton blooms could have major impacts on zooplankton abundances and distributions<sup>11–13</sup>. Furthermore, lipid-rich copepods have complex life histories and depend on a seasonal dormancy (diapause) to ensure the continued presence of a strong spring population in a system. Thus, poor spring recruitment due to changes in diapause could be a tipping point for a local ecosystem. However, the copepods' adaptive potential and phenotypic plasticity are unknown and require a more precise understanding of diapause and how it is controlled before environmental tipping-points can be predicted.

Our current understanding of the life cycle and ecology of lipid-rich copepods has emerged mostly from studies on *Calanus finmarchicus*, a keystone species that plays a central role in North Atlantic food webs<sup>2,4,14–16</sup>. Its annual cycle begins with generation G0 when copepods emerge from diapause as pre-adults (copepodid stage CV), molt into adults, mate, and reproduce<sup>17,18</sup>. The offspring (G1) of the G0 generation then make a critical “choice”: some individuals develop directly through six naupliar and six copepodid stages into adults (“reproductive program”) and produce another generation (G2), while others develop to the CV stage, then migrate to depth and enter diapause (“diapause program”)<sup>14,19–21</sup>. In the Gulf of Maine, *C. finmarchicus* can complete up to three such generations annually, with each generation in turn appearing to contribute to the overwintering population<sup>19,22</sup>. In contrast, those of very high latitudes, such as the Norwegian Sea, have only one (G1) generation per year; all are on the diapause program<sup>23</sup>. Copepods destined to diapause accumulate lipids that fuel the dormant period and contribute energetically to reproduction post-diapause<sup>17,24</sup>.

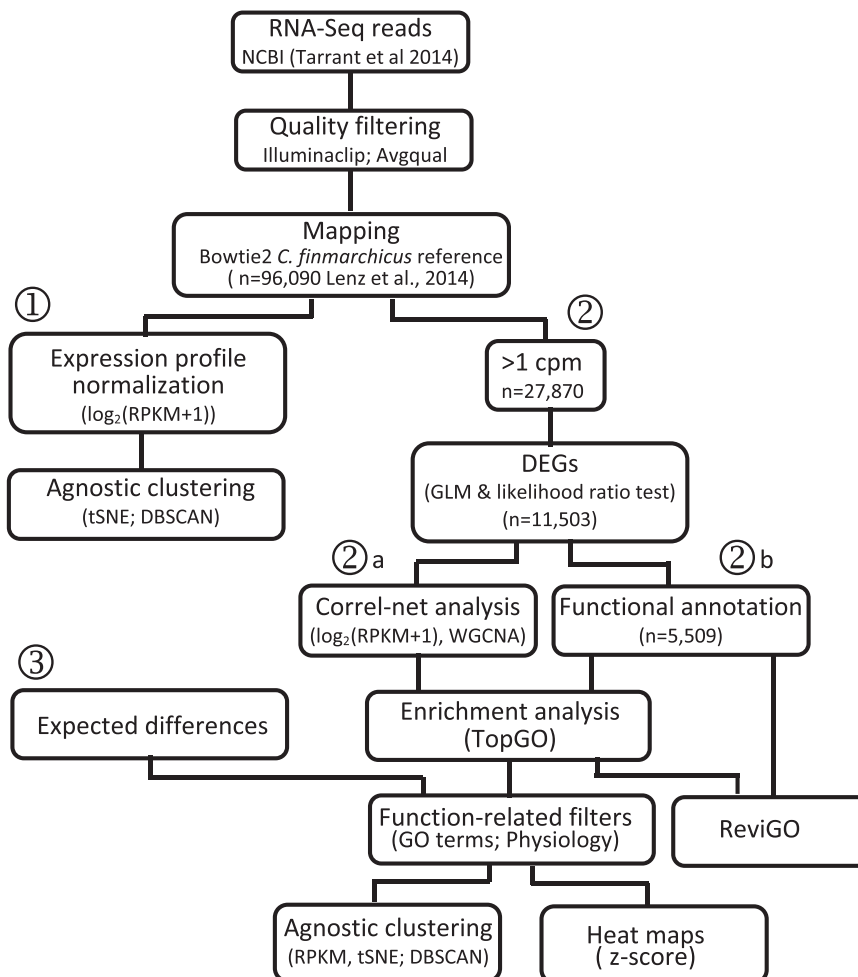
Predicted changes in phenology in response to ocean warming<sup>25</sup> raise two central questions about diapause in *C. finmarchicus*. The first one, how many, which and when copepods from each generation initiate the diapause program, is critical for assessing recruitment in the following year and for predicting the population cycle in the current year. Copepods that migrate to depth take their accumulated lipids with them and thus reduce lipid availability for upper trophic levels in surface waters. Those entering into diapause sequester biomass and lipids, effectively removing carbon at least temporarily from the upper 100 m and placing it into long-term storage for later availability to the ecosystem<sup>15,26–28</sup>. The second question relates to the basic biology and evolution of post-embryonic diapause in copepods. Developmental programs that include dormancy have evolved multiple times in arthropods<sup>29</sup>. The copepod diapause differs in that it is unlikely to be regulated by temperature and/or photoperiod<sup>14</sup>. A central question then is how does the copepod diapause compare with that of other organisms? What physiological characteristics are shared and which ones differ? While depressed metabolic rates and an arrest in development characterize diapause<sup>24,30</sup>, gene expression studies suggest that the specific molecular mechanisms that control diapause vary among taxa<sup>29,31–34</sup>. Modern transcriptomic technology permits examining a comprehensive array of genes involved in all aspects of an organism's life, and thus offers an opportunity to address both questions.

The inability to distinguish between uninterrupted (reproductive program) vs. interrupted (diapause program) life-history phenotypes is an impediment to a mechanistic understanding of how the decision to enter the diapause program changes depending on genotype, environment, and season. While major programmatic differences in physiology have been demonstrated in insects, these studies have relied on controlled experimental conditions<sup>35–37</sup>. In contrast, developmental program is difficult to assess in field-collected individuals of species with facultative diapause and unknown triggers for entering the diapause program. This includes *C. finmarchicus*<sup>23,24</sup>. However, once program-specific patterns in gene expression have been characterized, the how-when-why of diapause initiation can be investigated. A transcriptomic approach that reliably distinguishes reproductive-program from diapause-program stage CV individuals could transform *C. finmarchicus* population studies by enabling tracking of how the number (and proportion) of diapause-program CVs changes during the season.

**Analysis strategy for an existing RNA-Seq dataset.** Our goal was to determine whether the two programs could be separated by their respective gene expression (transcriptomic) phenotypes and whether this difference would lead to new insights into the physiological basis of the diapause program. The analysis was based on an RNA-Seq dataset generated by Tarrant and colleagues that included predominantly *C. finmarchicus* pre-adult copepodid stage CVs on different developmental programs<sup>23,38</sup>. These data allowed a broad-based comparison of transcriptional profiles of copepods on either the reproductive or the diapause program.

The RNA-Seq dataset comprised two-time points each, obtained from two sources of copepod: a laboratory-cultured group that was on the reproductive program, and a field-collected group from Trondheimsfjord that was on the diapause program. On close examination, the latter group was discovered to contain a limited admixture of two additional congeners, also on the diapause program, which we found had little impact on the results (see “Methods” and details in the Supplementary Note). The laboratory culture had been originally isolated from the same local fjord<sup>39</sup>. The laboratory-culture experimental groups were based on the number of days after molting into copepodid stage CV. Such history was unknown for the field copepodids. The analysis was tailored to identify gene expression differences that could be linked to the diapause program with the goal of excluding culture vs. field effects, or differences related to maturation within the molt-cycle.

To reliably separate stage CV individuals by the program without a priori knowledge, we employed three strategies to identify distinguishing gene expression patterns embedded in the data (Fig. 1). The first strategy applied a dimensionality-reduction algorithm, the t-Distributed Stochastic Neighbor Embedding technique (t-SNE) to cluster samples agnostically by similarity in gene expression patterns (Fig. 1)<sup>40,41</sup>. The second strategy focused on the identification of differentially expressed genes (DEGs) followed by downstream correlation network analysis and examination of predicted gene function (strategy 2a)<sup>32,42,43</sup>. The third approach was focused on functional analysis of expression differences and comparison with expected physiological and transcriptional differences (strategies 2b, 3)<sup>23,35,36,44–48</sup>. This targeted strategy builds computational filters to generate sets of genes based on relevant biological processes and gene ontology (GO) terms that reliably separate samples by the developmental program. The goal of the analysis was to design filters that identified processes that were independent of or minimally



**Fig. 1 Diagram of workflow.** Three different strategies used to assess the physiological ecology of copepods in the different samples are laid out using circled numbers. Initial steps included downloading of RNA-Seq data, removal of low-quality reads and sequence trimming followed by mapping against the Gulf of Maine *Calanus finmarchicus* annotated reference transcriptome (96 K transcripts) using Bowtie2. The mapped count data were normalized and log-transformed before dimensionality reduction by t-SNE and identification of clusters using DBSCAN (strategy 1). For differential gene expression analysis (DEGs), the mapped count data were entered into EdgeR for statistical testing using a generalized linear model (GLM) (strategy 2). The downstream analysis involved weighted correlation network analysis (WGCNA) on the log-transformed expression of the DEGs (sub-strategy 2a). SwissProt-based annotations for the DEGs were retrieved from the reference transcriptome and distribution of DEG function was visualized using ReviGO (sub-strategy 2b). DEGs from the GLM analysis and WGCNA modules were assessed for enriched processes (TopGO). Enrichment results in combination with expected differences in physiology were used to generate GO filters and retrieve log-transformed relative expression of all genes in the reference annotated to a specific filter for additional t-SNE and DBSCAN analyses (strategy 3). Gene expression patterns were visualized as z-scores in heatmaps.

affected by the environment (culture vs field) and/or time (early vs. late).

## Results and discussion

**Transcriptional phenotypes and analysis of differentially expressed genes.** Transcriptional phenotypic similarities among samples were assessed by applying the t-SNE algorithm to the *C. finmarchicus* expression data (see “Methods”). The t-SNE algorithm, widely used to distinguish among cell types within a single tissue, can identify homogeneous transcriptional phenotypic groups without a priori knowledge of “origin” or “treatment”<sup>49,50</sup>. It considers all transcripts simultaneously, grouping like phenotypes together, while excluding non-similar ones. The 16 samples aggregated into three separate clusters (Fig. 2). All field samples (diapause program), early (EF), and late (LF), belonged to the same transcriptional phenotype (i.e., in a single cluster), while the early (EC) and late (LC) culture samples (reproductive program) separated into distinct phenotypes.

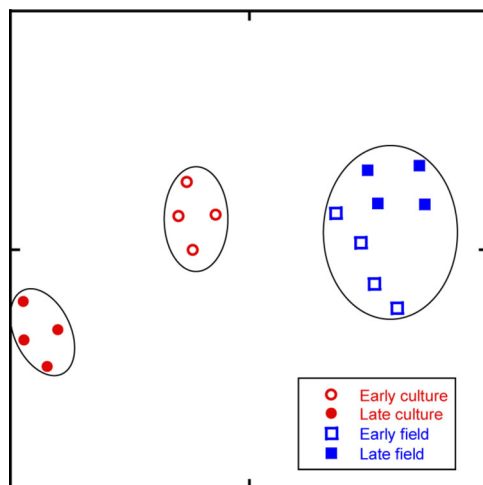
Maturation during the CV stage for individuals on the reproductive program involves large changes in expression as reported previously<sup>38</sup>. The LC individuals were approaching the final molt, while the EC individuals were only about ¼ of the way through the ~2-week molt cycle (see “Methods”). The difference between early and late culture could not be attributed to “environmental” factors, since culture conditions remained constant during the experiment, nor could they be attributed to differences in the program since none of the cultured copepods entered diapause. In contrast, the field-collected individuals clustered together, despite the two-week separation between sampling points. The samples presumably derived from the same population in the fjord and represent different stages in progress within the diapause program. Asynchrony within the field population might have partially obscured any temporal changes in expression. However, because diapause involves the developmental arrest and a lengthening of life span, the similarity in expression patterns between early and late fields may well be intrinsic to CVs on the diapause program.

**Differential gene expression and functional analysis.** A generalized linear model (GLM) identified over 11,000 DEGs among the four treatments (Table 1; strategy 2, Fig. 1). Consistent with the t-SNE results the smallest number of DEGs were found between the two sets of field samples, while the largest numbers of DEGs were between late culture CVs and those collected from the field (early and late field). The analysis also identified a large number of DEGs between early and late culture (6908), which is similar to the number reported previously for this comparison (“unique comps:” 7470) using a different reference transcriptome and short-read mapping program<sup>38</sup>.

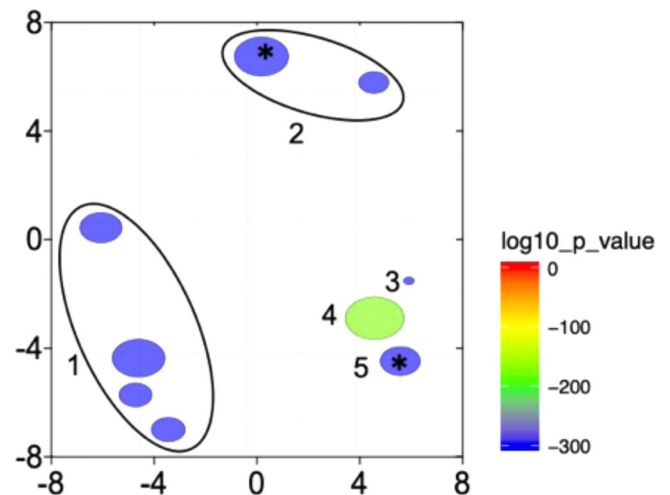
A search of the reference transcriptome for functional annotations returned nearly half of the DEGs with gene ontology (GO) terms ( $n = 5509$ ; strategy 2, Fig. 1). The ReviGO summarization grouped DEGs into nine GO terms (Fig. 3). The broad categories of ‘development’ (group 1) and ‘lipid metabolism’ (group 2) included four and two GO terms respectively. Development included GO terms associated with reproduction (e.g., ‘germ cell development’, ‘developmental process involved in reproduction’). Enrichment analysis identified two metabolic

processes as over-represented among the DEGs (‘very long-chain fatty acid metabolic process’ and ‘RNA metabolic process’). These processes might well be expected to be over-represented between individuals on reproductive vs diapause programs.

Differences between samples were further analyzed using correlation network analysis (WGCNA) to group DEGs into modules with highly correlated expression patterns (strategy 2a, Fig. 1). WGCNA identified four modules using the 11 K DEGs. The largest numbers of DEGs were assigned to two modules (blue > 3500; turquoise > 5000) (Fig. 4A). Expression patterns of these two modules differentiated between field and culture



**Fig. 2 Dimensionality reduction and cluster identification of expression data from all genes using t-SNE and DBSCAN.** Two-dimensional t-SNE plot of normalized and log-transformed expression profiles from mapped read data generated by Bowtie2 for the samples from the four groups (diapause program: EF, LF; reproductive program: EC, LC). The perplexity parameter was set to 5, and 50,000 iterations of the t-SNE algorithm were run. The DBSCAN algorithm was followed by the calculation of the Dunn index to determine the optimal grouping of points into clusters (enclosed in black ovals). Samples are coded by fill (open: early [E]; filled: late [L]) and by symbol and color (blue squares: field [F]; red circles: culture [C]).

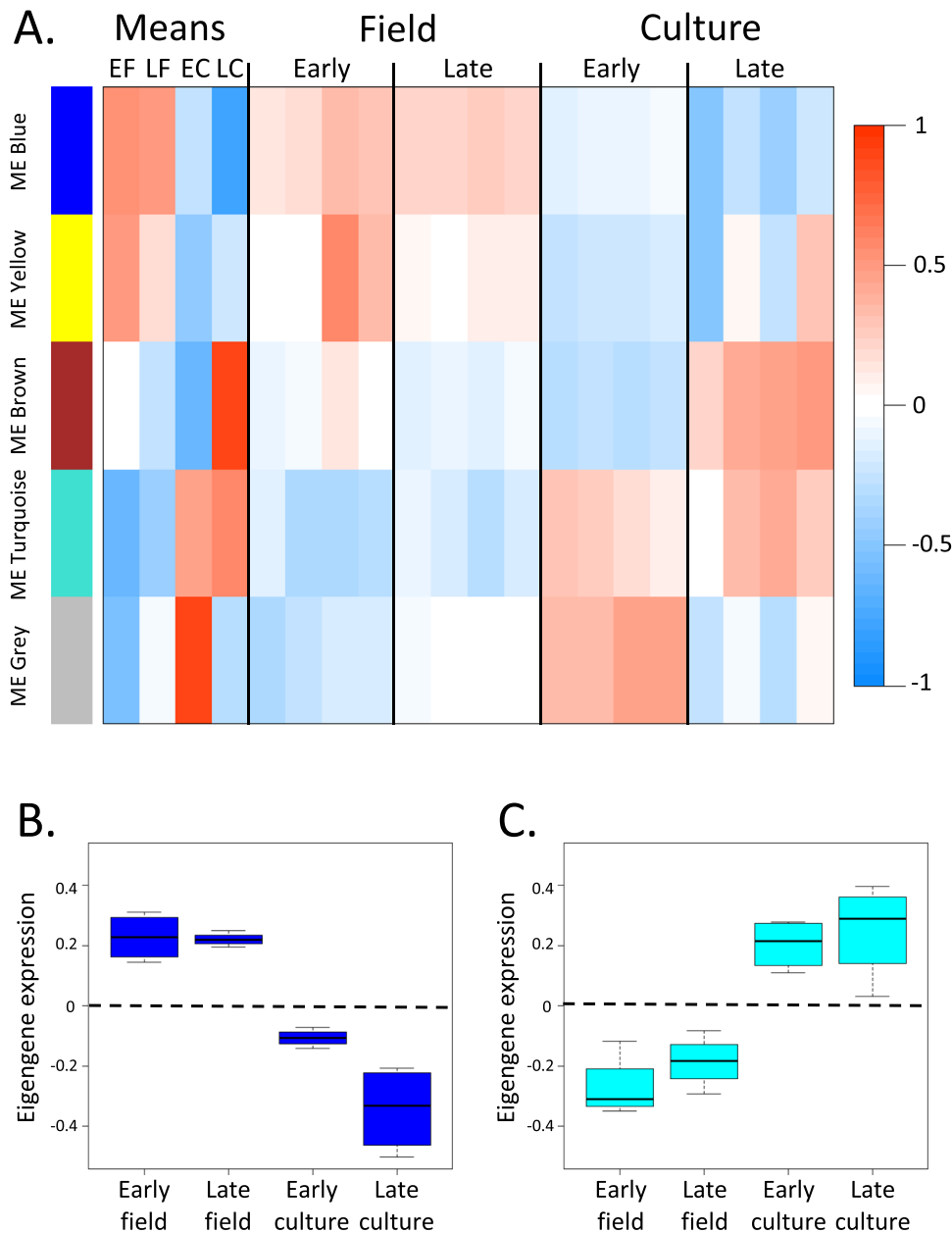


**Fig. 3 Biological processes represented among the differentially expressed genes (DEGs).** ReviGO two-dimensional representation of all GO terms represented among the total number of annotated DEGs ( $n = 5509$ ). The redundancy reduction filter was set to “small” (0.5). Each “bubble” represents a GO term; bubble color scales by the  $p$ -value determined by EdgeR (color scale, bottom right) and the bubble size indicates the frequency of the GO term in the underlying gene-ontology annotation database. Based on the Gene Ontology hierarchical organization, GO terms with the same GO parent have been circled (black line) and are indicated as a single number. GO term annotation: (1) Development/reproduction (‘developmental process involved in reproduction’ [GO:0003006], ‘reproduction’ [GO:0000003], ‘cellular developmental process’ [GO:0048869], ‘germ cell development’ [GO:0007281]). (2) Lipid metabolism (‘lipid metabolic process’ [GO:0006629], \*‘long-chain fatty acid metabolic process’ [GO:0001676]); (3) ‘Response to stress’ [GO:0006950]; (4) ‘Signal transduction’ [GO:0007165]; (5) \*‘RNA metabolic process’ [GO:0016070]. Asterisks (\*) mark GO terms that were already represented among the DEGs but were significantly enriched ([GO:0001676] and [GO:0016070]).

**Table 1 Statistical comparison of gene expression across four groups of stage CV *Calanus* that differed by source (field/diapause program vs. culture/reproductive program) and time point (early vs. late).**

Statistical test	Comparison	DEGs	Upregulated	Downregulated
Generalized linear model		11,503		
Pairwise likelihood ratio test	EF vs LF	1739	982	757
	EF vs EC	7197	3271	
	EF vs LC	10,077	6987	
	LF vs EC	7675	3715	
	LF vs LC	9939	4883	
	EC vs LC	6908	3818	

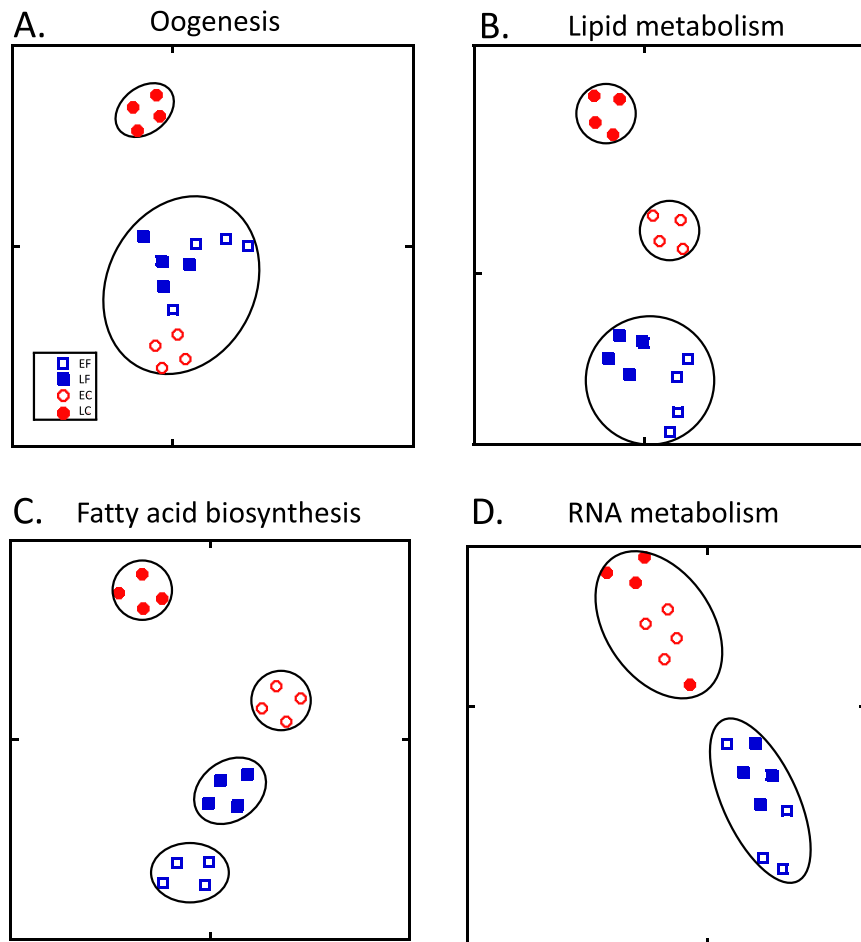
Summary of the number of differentially expressed genes (DEGs). DEGs were identified using a generalized linear model followed by downstream pairwise likelihood ratio tests (significant for  $p$ -value  $\leq 0.05$ ;  $p$ -value adjusted with the Benjamini-Hochberg procedure to control for false discovery rate [FDR]). Field-collected: EF (early field), LF (late field). Culture: EC (early culture), LC (late culture).



**Fig. 4 Correlation network analysis (WGCNA) of DEGs showing modules of genes with similar expression patterns. A** WGCNA network significance correlation matrix. Heatmap of correlation of module eigengenes with sample traits (rows correspond to modules [labeled by color] and columns to groups or individual samples). The first four columns represent the correlation of module eigengenes with each group (diapause program: early field [EF], late field [LF]; reproductive program: early culture [EC], late culture [LC]). Columns on the right (16) are the correlations of the eigengene expression for each module with the individual samples. Direction and the strength of correlation are indicated by color with blue showing negative (downregulation) and red showing positive (upregulation) (color scale on right). Number of genes in the four major modules: blue ( $n = 3827$ ), yellow ( $n = 745$ ), brown ( $n = 1133$ ), turquoise ( $n = 5689$ ). A small number of DEGs were placed into the “unassigned” gray module ( $n = 109$ ). **B** Box and whiskers plot of the blue module eigengene expression for each group ( $n = 4$ ). **C** Box and whiskers plot of the turquoise module eigengene values for each group ( $n = 4$ ). The box displays the median and interquartile range, while the whiskers give the minimum and maximum values for each group.

samples, as shown in the box and whiskers plots of module eigengene expression (Fig. 4B, C, see “Methods”). DEGs in the blue module were positively correlated with CVs on the diapause program (warm colors in field, cool colors in culture), while the DEGs in the turquoise module had the opposite expression pattern (Fig. 4A). Enrichment analysis of the GO terms represented among the DEGs identified a single over-represented process in each module: ‘glycerophospholipid biosynthetic process’ (GO:0046474) in the blue module and ‘positive regulation of RNA metabolic process’ (GO:0051254) in

the turquoise module. ‘Glycerophospholipid biosynthetic process’ is associated with the formation of glycerophospholipids, which are constituents of membranes and lipoproteins. It is a ‘child term’ of ‘lipid metabolic process’ (GO:0006629; bubble 2, Fig. 3). ‘Positive regulation of RNA metabolic process’ is a child term of ‘RNA metabolic process’ (GO:0016070), which was identified as an enriched process in the overall analysis (bubble 5, Fig. 3). In summary, this analysis identified only two key biological processes that drive gene expression differences between culture/reproduction and field/diapause programs.



**Fig. 5 t-SNE plots for subsets of transcripts filtered according to membership in different gene ontology (GO) terms and their child terms.** Circular profiles enclose clusters as determined by DBSCAN algorithm. **A** ‘Oogenesis’ filter [GO:0048477]; **B** ‘Lipid metabolic process’ filter [GO:0006629]; **C** ‘Fatty acid biosynthesis’ filter [GO:0006633]; **D** ‘RNA metabolic process’ filter [GO:016070]. Only the ‘RNA metabolic process’ filter divided the samples into separate field and culture transcriptional phenotypes. Symbol coding: field samples [F]: squares; culture samples [C]: circles; early samples [E]: open symbols; late samples [L]: closed symbols. All panels: perplexity = 5; number of iterations = 50,000 identified using DBSCAN with MinPts = 3 and the Eps value that maximized the Dunn index.

### Transcriptional analysis of expected physiological differences.

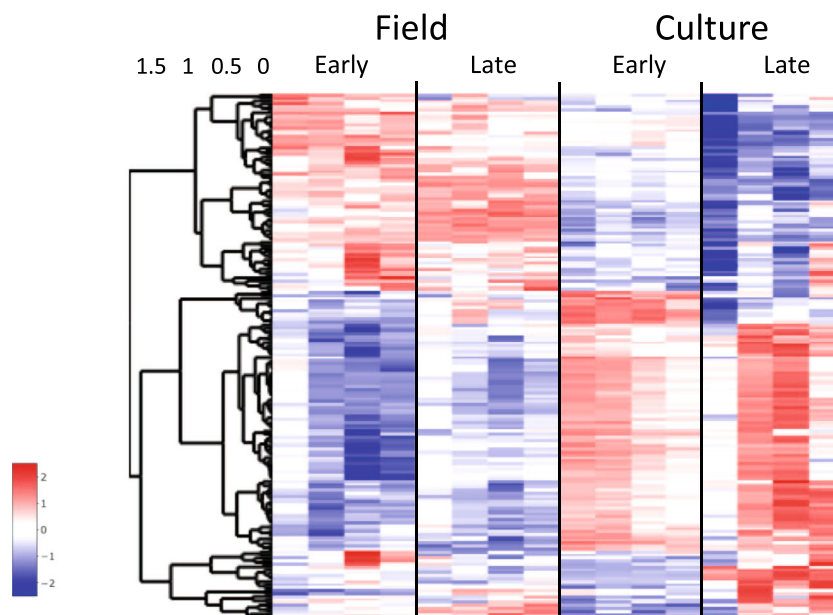
The goal of the third strategy was to analyze differences in expression by employing a priori knowledge on differences in developmental, metabolic and regulatory processes described in insects and expected in copepods<sup>23,36,45,46,51</sup>. Diapause preparation includes metabolic changes that lead to fat accumulation in arthropods and the build-up of wax ester stores in *C. finmarchicus* and other calanid copepods<sup>20,23,30,47,52–55</sup>. In contrast, maturing females require energetic resources for provisioning oocytes<sup>17,56</sup>. The differential gene expression analysis presented above broadly supports this, but has provided few details. In combination with heatmaps of the DEGs, we used gene ontology (GO) filters to establish transcriptional phenotypes associated with all genes annotated to specific processes in the reference transcriptome independent of whether they were differentially expressed (strategy 3, Figs. 1, 5).

Gonad development and early oogenesis occur during stage CV in individuals on the reproductive<sup>57</sup>, but not the diapause program, a difference that was confirmed by microscopic examination of cultured and field-collected individuals done concurrently with the transcriptomics<sup>23</sup>. In the reference transcriptome, 584 genes were annotated to oogenesis (GO:0048477, and its child terms). Dimensionality reduction by t-SNE of relative expression of these genes separated the

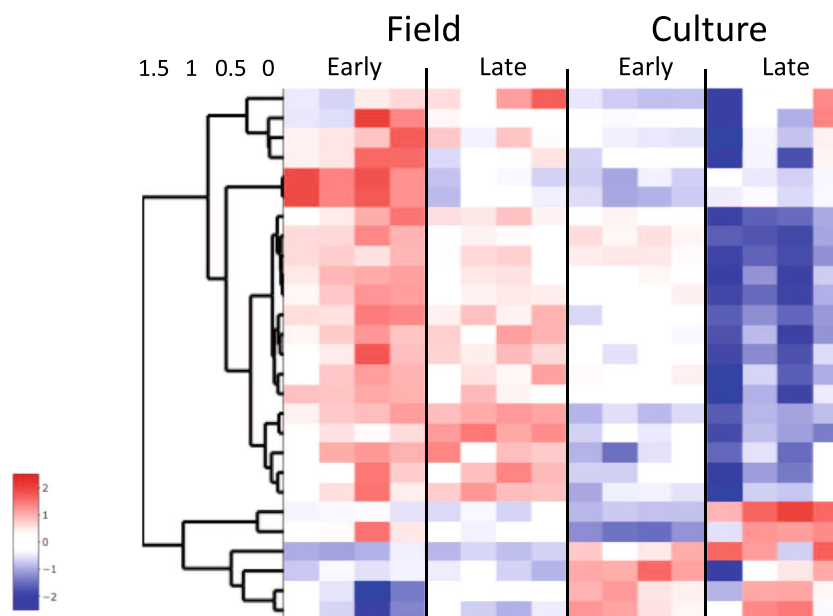
16 samples into two clusters (Fig. 5A). The late culture individuals, which were approaching maturity, aggregated into a distinct cluster. However, there was no substantial separation of the 12 remaining samples, which were widely distributed within a single cluster. A heatmap of relative expression of the 178 DEGs annotated with the oogenesis GO term is consistent with the t-SNE result: somewhat more than half of the oogenesis genes were upregulated and the rest downregulated in late culture CV samples when compared with all other samples (Fig. 6). For the remaining 12 samples, even if more variable, the expression pattern of several genes separated the field samples from the early culture samples. Thus, a general oogenesis filter may prove useful in the identification of CVs approaching the final molt, but it may be less successful in separating recently molted CVs on the reproductive program from those on the diapause program.

Genes involved in lipid metabolism are expected to be differentially expressed between reproductive- and diapause-program CV individuals given fat accumulation during diapause preparation<sup>23</sup>. Processes linked to lipid metabolism were found to be enriched among all DEGs, and a child term was enriched in the blue (diapause-correlated) module of the WGCNA analysis. To pursue this further, two lipid metabolism filters were applied to the whole transcriptome: ‘lipid metabolic process’ (GO:0006629 and its child terms) with 717 genes and ‘fatty acid





**Fig. 6 Expression heatmap showing z-scores of DEGs involved in oogenesis.** Differentially expressed genes between field/diapause program and culture/reproductive program and early and late CV copepodids annotated with the ‘oogenesis’ [GO:0048477] GO term and its child terms ( $n = 178$ ). Color-coding for each gene indicates the magnitude of expression as z-scores of each individual sample. Relative expression of each sample is given in a separate column (ordered by group) as labeled at the top. Genes (rows) were ordered by hierarchical clustering.

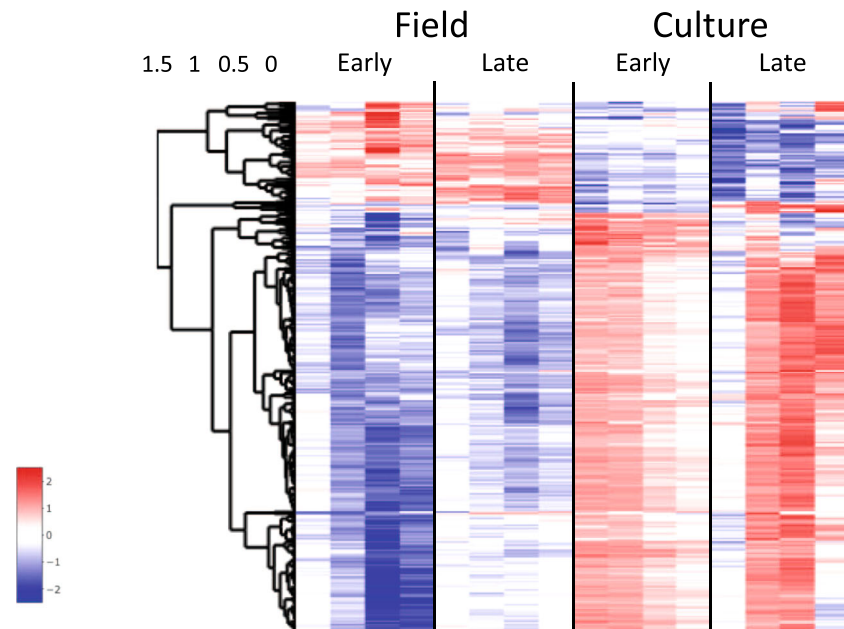


**Fig. 7 Expression heatmap showing z-scores of DEGs involved in fatty acid biosynthesis.** Differentially expressed genes between field/diapause program and culture/reproductive program and early and late CV copepodids annotated with the GO term ‘fatty acid biosynthetic process’[GO:0006633] or its child terms ( $n = 23$ ). Color-coding for each gene indicates the magnitude of expression as z-scores of each individual sample. Relative expression of each sample is given in a separate column (ordered by group) as labeled at the top. Genes (rows) were ordered by hierarchical clustering.

biosynthetic process’ (GO:0006633 and its child terms) with 70 genes. A t-SNE analysis that included all genes annotated to the first of these separated the 16 samples into three clusters (Fig. 5B) that were qualitatively similar to the t-SNE analysis that included all genes (Fig. 2). The culture samples segregated into an early and a late group suggesting that maturation during the CV stage includes regulation of lipid metabolism.

The more specific filter of ‘fatty acid biosynthesis’ separated the samples into four clusters, with the early and late field samples placed into distinct transcriptional phenotypes (Fig. 5C). Such a

pattern could be explained by the regulation of fatty acid metabolism along the CV’s progression towards diapause in the field, and/or it could reflect responses to environmental differences between the two sampling times. Thus, a limitation of a GO filter associated with lipid metabolism is that expression differences may occur in response to environmental factors such as food quantity and quality, as reported in another diapausing calanid, *Neocalanus flemingeri*<sup>40,42</sup>. Nevertheless, the 23 DEGs annotated to ‘fatty acid biosynthesis process’ (GO:0006633) show a general upregulation of genes associated with lipid synthesis in



**Fig. 8 Expression heatmap showing z-scores of DEGs involved in RNA metabolism.** Differentially expressed genes between field/diapause program and culture/reproductive program and early and late CV copepodids annotated with the GO term ‘RNA metabolic process’ [GO: 0016070] or its child terms ( $n = 335$ ). Color-coding for each gene indicates the magnitude of expression as z-scores of each individual sample. Relative expression of each sample is given in a separate column (ordered by group) as labeled at the top. Genes (rows) were ordered by hierarchical clustering.

field-collected individuals in comparison with cultured individuals (Fig. 7). However, the expression pattern was quite variable among individual samples. Consistent with diapause preparation in field CVs, we observed the upregulation of enzymes involved in wax ester biosynthesis (two *diacylglycerol O-acyltransferases 1* and two *fatty acyl-CoA reductases*), a process directly related to lipid accumulation.

Downregulation of genes involved in RNA and DNA metabolism during diapause has been demonstrated in insects, copepods, and other arthropods<sup>32,36,58</sup>. While the environmental triggers of diapause in calanid copepods remain unknown, in insects the developmental program can be pre-set by varying day length. This allowed Poelchau and colleagues to compare gene expression in non-diapause (“ND”) with diapause-bound (“D”) individuals in the mosquito *Aedes albopictus* during embryogenesis<sup>35</sup>. Downregulation of genes involved in metabolism, energy production, and protein synthesis, including a child term of ‘RNA metabolic process’ was already apparent during pre-diapause. A similar pattern was observed in *C. finmarchicus*. Genes involved in ‘RNA metabolic process’ were downregulated in field CV individuals and this process was enriched among the DEGs (Figs. 3, 4, see turquoise module).

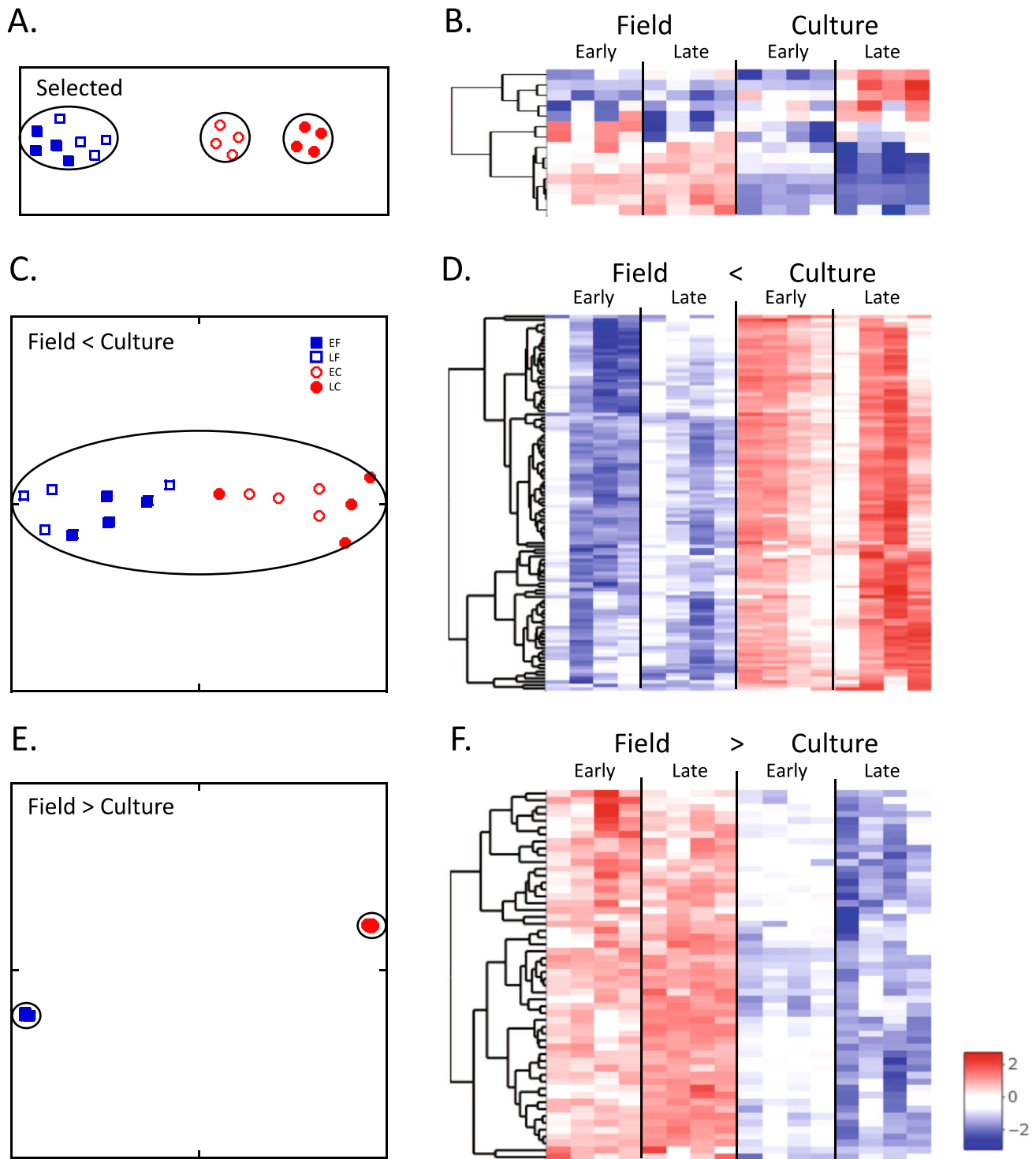
Application of a general filter for ‘RNA metabolic process’ (GO:0016070 and child terms,  $n = 1064$ ) followed by t-SNE separated the 16 samples into two clusters consisting of either culture or field samples (Fig. 5D). This filter did not show differences in gene expression associated with maturation in CVs on the reproductive program (i.e., clustering both EC and LC together). The pattern in the heatmap of 335 DEGs is consistent with the t-SNE analysis and clearly separated the field from the culture samples, in spite of individual variability among replicate samples (Fig. 8). Most of these DEGs showed low expression in diapause-bound (field) individuals and substantially higher expression in culture individuals. Among the DEGs are several genes encoding proteins involved in RNA processing such as *pre-mRNA splicing factors*, *spliceosomes*, and *mRNA decay activators* (Fig. 8). This signal was more pronounced and pervasive in *C. finmarchicus* than in the mosquito<sup>35</sup>. While it is possible that

environmental factors contributed to this separation of culture and field individuals, neither ‘RNA metabolic process’ nor any of its child terms were identified as enriched among the differentially expressed genes reported in diapause-bound *N. flemingeri* collected from locations with order of magnitude differences in food resources<sup>42</sup>.

Another approach to identify calanids on the diapause program has been to explore potential biomarker genes by selecting a set of genes a priori based on comparisons between presumably active or dormant field-collected individuals. Such comparisons exist for *C. finmarchicus* and *Calanus sinicus* with samples collected from different depths and profiling relative gene expression using a variety of molecular methods<sup>59–61</sup>. Differentially expressed genes from these studies were then cross-referenced to genes regulated just prior to diapause in insects, *Artemia* and/or *Caenorhabditis elegans*<sup>35,62–65</sup>. Using this approach, we identified 14 potential candidates for biomarker genes (Fig. 9A, B, Supplementary Table S1). These genes did not include any annotated to GO terms used in the previous filters. Based on our analysis, seven genes were differentially expressed (three *serpins* [out of 4], two *nitric oxide synthases* [out of eight], one *phosphoenolpyruvate carboxylase kinase* [out of 1], one *RAS-related protein Rab-10* [out of 1]), and relative expression of these genes differed between field and culture as shown in the heatmap (Fig. 9B). However, a t-SNE analysis of the relative expression of these 14 genes failed to separate the samples into cohesive field and culture clusters, but rather generated three clusters, similar to the pattern generated in the initial analysis that included all genes (Fig. 9A).

**Workflow to separate CVs by program using RNA-Seq of individuals.** While we focused on a set of 16 pooled RNA-Seq samples from four known treatment groups, the goal was to develop a protocol to determine which and how many CVs are on the diapause program in a natural population. Gene expression profiles generated for individual CVs collected from the environment could be assessed for the developmental program by





**Fig. 9 “Designer” filters with high-responding genes to separate field from culture transcriptional phenotypes. A, B** Filter selecting genes based on evidence for involvement in diapause preparation, as described in the text. **C–F** Transcripts were selected from DEGs annotated with the GO terms ‘oogenesis’ (GO:0048477), ‘fatty-acid biosynthesis’ (GO:0006633) and ‘RNA metabolic process’ (GO:0016070). **C–D** Filter selecting transcripts having z-scores in all culture samples higher than any in the field samples. **E–F** Filter selecting transcripts having z-scores in all field samples higher than any in the culture samples. **A, C, E** t-SNE plots, perplexity = 5; max iterations 50,000; clusters identified by DBSCAN encircled. Key in **(C)** applies to all t-SNE plots: EF early field, LF late field, EC early culture, LC late culture; **B, D, F** ordered heatmaps.

producing expression profiles for the ca. 1000 genes annotated to ‘RNA metabolic process’ in the reference transcriptomes and applying t-SNE. We hypothesize that applying t-SNE to these profiles will separate individuals into two clusters based on developmental program, which can then be confirmed using

differential gene expression. Individuals can be separated by cluster membership and tested for expected gene expression differences between the diapause and reproductive programs.

Another approach is a search for robust indicator genes. Ecological studies require testing large numbers of individuals

across time and space, which calls for protocols capable of high-throughput of samples based on RT-qPCR or nCounter (NanoString®) technologies<sup>66</sup>. These technologies need a smaller set of indicator genes with a robust signal-to-noise ratio (Fig. 9). We searched for a set of genes with consistent and large differences in expression between culture and field samples among the DEGs annotated to the three GO terms that we tested for differentiating between programs (oogenesis, GO:0048477,  $n = 178$ ; fatty-acid biosynthesis, GO:0006633  $n = 23$ ; and RNA metabolic process, GO:0016070,  $n = 335$ ). A transcript was included when all of its expression z-scores in culture samples were either above or below all of its values among the field samples (i.e., no overlap in relative expression). This selection method is clearly shown in the respective heatmaps for the two filters with all of one color for the field and the opposite color for the culture samples (Fig. 9D, F). Genes that were upregulated in culture (i.e., reproductive program) compared with field included 111 such transcripts from oogenesis ( $n = 32$ ) and RNA metabolic process ( $n = 79$ ), but none from fatty acid biosynthesis passed this filter (Fig. 9C, D). Relative expression of these genes was highly variable and did not separate the CVs by the developmental program as shown by the single cluster in the t-SNE plot (Fig. 9C).

In contrast, a filter comprising genes that were more highly expressed in the field (i.e., diapause program) than culture samples, produced two tight and distinct clusters in t-SNE (Fig. 9E). The 54 transcripts in this filter included representatives from all three GO terms: oogenesis ( $n = 19$ ), RNA metabolic process ( $n = 28$ ), and fatty acid biosynthesis ( $n = 7$ ) as shown in the heatmap (Fig. 9F; Supplementary Table S2). While it is premature to speculate on their specific functions with respect to the diapause program in *C. finmarchicus*, these genes are good candidates for further investigation. Two transcripts on this list, one *diacylglycerol O-acyltransferase 1* and one *fatty acyl-CoA reductase* are predicted to be involved in wax ester biosynthesis, while another *AMP-activated protein kinase (AMPK) gamma 1* is involved in the regulation of cellular energy metabolism.

## Conclusions

An existing RNA-Seq dataset was analyzed to develop a workflow for environmental transcriptomics that can classify pre-adult CV *C. finmarchicus* individuals by developmental program. Through a combination of statistical and functional analyses, we propose two workflows. The first relies on a global gene expression analysis (RNA-Seq) and involves applying a gene ontology filter (RNA metabolic process) followed by t-SNE clustering to separate samples into groups for statistical comparison. The second workflow employs an indicator strategy for high-throughput gene expression technologies. A designer filter identified 54 genes that were consistently upregulated in individuals on the diapause program compared with those on the reproductive program. The t-SNE analysis of the relative expression of these genes separated the samples into two distinct transcriptional phenotypes based on the developmental program. While these workflows need further testing in natural populations, they may be broadly applicable to *C. finmarchicus* and other diapausing calanid copepods. These molecular approaches can be used to assess reproductive strategies within an environmental context. Furthermore, the specific genes and pathways identified in this analysis may be good candidates for elucidating the physiological processes that differentiate the two developmental programs, including determining when the decision to diapause is made in copepods.

## Methods

**Calanus finmarchicus reference transcriptome.** The study used an existing Gulf of Maine *Calanus finmarchicus* transcriptome for mapping the short sequence reads (NCBI BioProject PRJNA236528)<sup>67</sup>. Briefly, this reference was assembled

from 100 bp short-sequence reads from six developmental stages and had been annotated against the SwissProt protein database ([www.uniprot.org](http://www.uniprot.org)). Annotation identified 28,616 transcripts with significant similarity to known proteins ( $E$ -value cutoff =  $10^{-3}$ ) and 10,334 transcripts with significant GO annotations ( $E$ -value cut-off =  $10^{-6}$ ; <http://geneontology.org/>)<sup>67–69</sup>. The reference with 96 K transcripts had no contamination from other *Calanus* species and was characterized by very low ambiguous mapping (<1% ‘mapped more than once’ by Bowtie2)<sup>68</sup>.

**RNA-Seq data description, retrieval, and pre-processing.** Short-read sequences for 16 samples were downloaded from the short-sequence read archive (SRA) in the National Center for Biotechnology Information (NCBI) database (Table S1, Supplementary Note; Illumina HiSeq2000, 50 bp, paired-end with  $\geq 30$  M spots per sample, (BioProject: PRJNA 231164)<sup>38</sup>. For each sample, RNA had been extracted from pools of stage CV individuals (5–7)<sup>38</sup>. The dataset included four replicate samples for each of the two time-points in both the laboratory-cultured population and the field-collected wild population<sup>23,38</sup>. The experimental design and number of replicates provided the necessary statistical power for this analysis, which focused on distinguishing between two developmental programs. Additional details on the experiments can be found in previous studies<sup>23,38</sup> and in the biosample descriptions in the NCBI database. Previous analysis of the data focused on characterizing transcriptional changes associated with maturation in stage CV copepodids on the reproductive program<sup>38</sup>. In a second study, differences in the developmental program were sought by analyzing pathways associated with lipid metabolism for temporal changes in gene expression of biomarkers in culture and field CVs using RT-qPCR. While differences in relative expression were noted, this analysis was not detailed enough to discriminate between “within stage maturation” and developmental program<sup>23</sup>. Neither study included an analysis of the high-throughput sequencing of the field samples, which is the central approach used in the current study.

Briefly, the laboratory-cultured samples consisted of recently molted ( $\leq 24$  h) stage CV copepodids that had been isolated and incubated separately until harvested at three (early culture, “EC”) and 10 days (late culture, “LC”) post-molt. The time points represented early and late stages in the molt cycle, which under the experimental conditions had a median duration of 13.5 days<sup>38</sup>. During the incubation, copepods were maintained on the standard culture diet<sup>38,39</sup>. Microscopic examination of other individuals from each experimental set of CVs confirmed the presence of early development of gonads at both three and 10 days post-molt. At three days post-molt, all individuals were in the pre-apolysis jaw phase, and by days 9 through 11, 45% had matured into post-apolysis jaw phases consistent with progression toward the terminal molt.

The diapause-program copepodids had been collected from the field at Trollet Station in Trondheimsfjord ( $63^{\circ}29'N$ ,  $10^{\circ}18'E$ ) with a zooplankton net towed vertically from 50 to 0 m on 28 May 2013 (early field, “EF”) and 14 days later on 10 June 2013 (late field, “LF”)<sup>23</sup>. Microscopic examination of CVs revealed that, consistent with pre-diapause, all had undifferentiated gonads and were in the pre-apolysis jaw phase<sup>23</sup>. The juvenile copepods were not sorted according to sex, and presumptive males and females were included in laboratory and field samples. Although the field samples were originally thought to contain only *C. finmarchicus* CVs, recent studies reported that *C. glacialis* and *C. helgolandicus* can co-occur with *C. finmarchicus* in the region including Trondheimsfjord<sup>70–72</sup>. The three congeners are morphologically very similar and can only be identified reliably to species using genetic tools (Choquet et al.<sup>71</sup>). We confirmed the presence of the congeners in the field samples using a molecular approach (see below, Supplementary Note).

In Trondheimsfjord, *C. glacialis* and *C. helgolandicus* are on the same diapause-bound program as is *C. finmarchicus*<sup>73</sup>, and thus are not expected to diverge greatly in their transcriptional phenotypes. Nevertheless, we examined the possibility of bias impacting the analysis due to species composition differences between field and culture samples. We concluded that there was no significant bias, and the multi-step analyses that led us to this conclusion are described in detail in the Supplementary Note.

Briefly, we assessed the species composition of each sample by using species differences in the mtCOI sequences<sup>74</sup> and quantifying reads mapping to each sequence. Significant contamination was limited to the field samples. Congener composition of most samples was below 32%, which combined with an estimated 30% cross-mapping efficiency to congeneric references<sup>75</sup>, indicated a modest 11% estimate for mean cross-mapping levels (Table S1, Supplementary Note). We then used publicly available congeneric read sets to identify the most cross-map-prone transcripts in our *C. finmarchicus* reference. About half of the transcripts susceptible to cross-mapping were among the transcripts with significant expression ( $>1$  count per million reads [cpm]). This proportion was maintained in most of the analyses we performed in our more targeted transcript selections, indicating a uniform contribution from cross-mapped sources (Table S2, Supplementary Note). However, there was some enrichment of cross-mapped transcripts, so in our last test we compared the t-SNE analyses for each transcript set with a paired set that excluded all transcripts with cross-mapped reads (Fig. S1, Supplementary Note). The effects were minimal, and duplicated the transcriptional phenotype results when the conserved transcripts with some contamination from cross-mapped reads were included in the set. Thus, the main text refers to samples

as *C. finmarchicus* samples, this being the dominant species present and the species used as the bioinformatic reference.

**Mapping of short reads and computation of relative gene expression.** After quality filtering to remove sequences with a Phred score  $\leq 20$ , short sequence reads from each sample were mapped against the *C. finmarchicus* reference transcriptome to generate gene expression profiles (Fig. 1) using Bowtie2 software (default settings; v.2.1.0)<sup>76</sup> (Table S1, Supplementary Note). After the mapping step, RPKM (reads per kilobase of transcript length per million mapped reads) were calculated to normalize relative gene expression [i.e., for transcript *i* from sample *j*,  $\text{RPKM}(i,j) = \text{reads}(i,j)/[(\text{length}(i)/1000)*(\text{mapped\_reads}(j)/1000000)]$ <sup>77</sup>. We next  $\log_2$  transformed the relative expression data after adding a pseudocount of 1 to the RPKM value for each transcript (i.e.,  $\log_2[\text{RPKM}+1]$ ) (Fig. 1). These  $\log_2$ -transformed relative expression data were used in all dimensionality-reduction analyses and to calculate z-scores for each transcript and sample. Z-scores were used in heatmaps for expression comparisons across samples.

#### Dimensionality reduction and identification of transcriptional phenotypes.

The dimensionality reduction method t-distributed Stochastic Neighbor Embedding (t-SNE) was used to visualize variation in gene expression across samples<sup>41,78</sup> (Fig. 1, strategy 1). The t-SNE algorithm reduces the high dimensional gene expression profiles to a two-dimensional representation while seeking to conserve the local relationships among the samples. We have found t-SNE to be better for identifying copepod transcriptional phenotypes than other dimensionality-reduction methods such as principal component analysis (PCA)<sup>40</sup>. We applied t-SNE as implemented in the R package Rtsne (Rtsne URL: <https://github.com/jkrijthe/Rtsne>)<sup>79</sup> to the  $\log_2$ -transformed RPKM values for either the entire set of transcripts ( $n = 96,090$ ), or for subsets of transcripts filtered for specific GO terms (see below; Fig. 1, strategy 3). After pre-testing, program parameters were set as follows: perplexity = 5, maximum number of iterations = 50,000 and the remaining parameters equal to their default values. In addition, the t-SNE algorithm was run multiple times to ensure that the output was representative (i.e., to ensure that the phenotypes so identified were robust)<sup>40</sup>. The results were plotted as a 2-D scatterplot in the t-SNE coordinates. To provide an objective method of identifying which samples formed clusters, the density-based clustering algorithm, DBSCAN (with *MinPts* = 3) was applied to the t-SNE results (coordinates of points)<sup>40,80</sup>. The clustering cut-off (*Eps* parameter) was chosen to maximize the Dunn index score<sup>81</sup>. Both the DBSCAN algorithm and the Dunn index were run in R (dbscan: <https://CRAN.R-project.org/package=dbscan>; clusterCrit: <https://CRAN.R-project.org/package=clusterCrit>)<sup>40,82,83</sup>.

#### Differential gene expression and weighted gene correlation network analysis (WGCNA).

The “mapped reads” file generated by Bowtie2 was used as the input to the BioConductor package EdgeR to identify differentially expressed genes (DEGs)<sup>84</sup> (Fig. 1, strategy 2). Prior to the statistical analysis, transcripts with very low expression levels (those failing to have at least 1 cpm in 4 of the 16 samples) were removed leaving a total of 27,870 transcripts (out of the original 96,060 in the reference). As implemented by EdgeR, libraries were normalized using the TMM method (trimmed mean of M values). The negative binomial generalized linear model (GLM) identified DEGs across samples (glmFit function) with *p*-values adjusted for false discovery rate (FDR; Benjamini–Hochberg procedure). The GLM analysis was followed by pairwise comparisons using the downstream likelihood ratio test (glmLRT) to identify significant differences in gene expression between each treatment pair (*p*-value  $\leq 0.05$ , corrected for FDR).

Patterns of differential gene expression among samples were explored using weighted gene correlation network analysis (WGCNA), a technique for finding modules of highly correlated genes across treatments<sup>85,86</sup>. Downstream analysis of modules or a representative of the gene expression profiles in each module, such as the “eigengene”, provides a network-based method for data reduction. The WGCNA analysis was performed on the  $\log_2$ -transformed ( $\log_2[\text{RPKM}+1]$ ) gene expression of all DEGs (11 K, Fig. 1, strategy 2). The analysis used an unsigned, weighted network with a soft threshold power of 14 and minimum module size (*minModuleSize*) set to 100. Modules were determined by applying the automatic block-wise module detection function of the WGCNA package. The module eigengene, defined as the first principal component of the module gene expression matrix, gives a weighted average of the module expression profiles and was used to investigate the relationship between modules and biologically interesting sample traits. Pearson correlations between module eigengenes and membership in a specific experimental group were computed. A heatmap was generated to visualize these correlations by experimental group and for the individual samples to allow comparison of expression patterns across replicates. We used boxplots to display the descriptive statistics (median, first (25%) quartile, third (75%) quartile, minimum and maximum) of module eigengene expression for each experimental group (EF, LF, EC, and LC). Annotated genes assigned into WGCNA modules were tested for enriched GO terms (see below).

**Functional analysis and filtering of genes using gene ontology.** Functional analysis of the DEGs was based on the *C. finmarchicus* transcriptome. Briefly, DEGs were cross-referenced with the annotated transcriptome and nearly half were

found to have GO term annotations. ReviGO software was used to summarize and visualize in two-dimensional space the biological processes represented among the DEGs<sup>87</sup>. The list of GO-annotated DEGs (all) and their *p*-values were summarized using a very stringent filter (similarity setting to “small” = 0.5), which substantially reduced the redundancy intrinsic to the Gene Ontology hierarchy.

Enrichment analysis was performed using TopGO software<sup>88</sup> on DEGs with GO annotations. As implemented by TopGO, a Fisher exact test with a Benjamini–Hochberg correction (*p*-values  $\leq 0.05$  [v. 2.88.0, set to the default algorithm “weight01”]) was used to compare the DEGs identified for each sample pair ( $n = 6$ ) against all transcripts with GO terms in the reference transcriptome<sup>67</sup>.

Based on the enrichment results (strategy 2, Fig. 1) and pre-determined functional hypotheses (strategy 3), several GO filters were applied to workflow strategies 2 and 3 (Fig. 1). Specifically, the AmiGO software GO Online SQL Environment (GOOSE)(October, 2019: <http://amigo2.berkeleybop.org/goose/cgi-bin/goose>) was used to search descendants of target GO terms to obtain all transcripts annotated to a specific process. For this, the LEAD SQLwiki on the AmiGO Labs prototype page, using the example called “find descendants of the node ‘nucleus’ was edited to replace ‘nucleus’ with the specific GO term to be used for the filter. The annotated reference transcriptome was then used to retrieve all transcripts within each functional category defined by specific GO terms and their child terms. In addition, GO lists were searched for DEGs, and heatmaps were generated using z-scores (see above) and the software package heatmaply in R, which clusters genes by expression similarity (heatmaply: <https://github.com/talgali/heatmaply>)<sup>89</sup>.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

#### Data availability

The RNA-Seq data analyzed in this study are available on the National Center for Biotechnology Information (NCBI) database under the BioProject PRJNA 231164. The files generated in this study with relative expression per contig (counts, RPKM,  $\log_2[\text{RPKM}+1]$  and z-scores) are available in DryAd<sup>90</sup>.

Received: 22 May 2020; Accepted: 1 March 2021;

Published online: 29 March 2021

#### References

- Record, N. R. et al. Copepod diapause and the biogeography of the marine lipidscape. *J. Biogeogr.* **45**, 2238–2251 (2018).
- Conover, R. J. & Corner, E. D. S. Respiration and nitrogen excretion by some marine zooplankton in relation to their life cycles. *J. Mar. Biol. Assoc. UK* **48**, 49–75 (1968).
- Kattner, G. et al. Perspectives on marine zooplankton lipids. *Can. J. Fish. Aquat. Sci.* **64**, 1628–1639 (2007).
- Beaugrand, G., Brander, K. M., Lindley, J. A., Souissi, S. & Reid, P. C. Plankton effect on cod recruitment in the North Sea. *Nature* **426**, 661–664 (2003).
- Coyle, K. et al. Climate change in the southeastern Bering Sea: impacts on pollock stocks and implications for the oscillating control hypothesis. *Fish. Oceanogr.* **20**, 139–156 (2011).
- Liu, H., Bi, H. & Peterson, W. T. Large-scale forcing of environmental conditions on subarctic copepods in the northern California Current system. *Prog. Oceanogr.* **134**, 404–412 (2015).
- Peterson, W. T. et al. The pelagic ecosystem in the Northern California Current off Oregon during the 2014–2016 warm anomalies within the context of the past 20 years. *J. Geophys. Res. Oceans* **122**, 7267–7290 (2017).
- Bi, H., Peterson, W. T., Lamb, J. & Casillas, E. Copepods and salmon: characterizing the spatial distribution of juvenile salmon along the Washington and Oregon coast, USA. *Fish. Oceanogr.* **20**, 125–138 (2011).
- Kirby, R. R. & Beaugrand, G. Trophic amplification of climate warming. *Proc. R. Soc. B* **276**, 4095–4103 (2009).
- Hirche, H.-J. Temperature and plankton II. Effect on respiration and swimming activity in copepods from the Greenland Sea. *Mar. Biol.* **94**, 347–356 (1987).
- Mahara, N., Pakhomov, E. A., Jackson, J. M. & Hunt, B. P. Seasonal zooplankton development in a temperate semi-enclosed basin: two years with different spring bloom timing. *J. Plankton Res.* **41**, 309–328 (2019).
- Hooff, R. C. & Peterson, W. T. Copepod biodiversity as an indicator of changes in ocean and climate conditions of the northern California current ecosystem. *Limnol. Oceanogr.* **51**, 2607–2620 (2006).
- Keister, J. E., Di Lorenzo, E., Morgan, C., Combes, V. & Peterson, W. Zooplankton species composition is linked to ocean transport in the Northern California Current. *Glob. Change Biol.* **17**, 2498–2511 (2011).



14. Johnson, C. L. et al. Characteristics of *Calanus finmarchicus* dormancy patterns in the Northwest Atlantic. *ICES J. Mar. Sci.* **65**, 339–350 (2008).
15. Ji, R. B., Edwards, M., Mackas, D. L., Runge, J. A. & Thomas, A. C. Marine plankton phenology and life history in a changing climate: current research and future directions. *J. Plankton Res.* **32**, 1355–1368 (2010).
16. Weydmann, A., Walczowski, W., Carstensen, J. & Kwaśniewski, S. Warming of Subarctic waters accelerates development of a key marine zooplankton *Calanus finmarchicus*. *Glob. Change Biol.* **24**, 172–183 (2018).
17. Niehoff, B., Madsen, S., Hansen, B. & Nielsen, T. Reproductive cycles of three dominant *Calanus* species in Disko Bay, West Greenland. *Mar. Biol.* **140**, 567–576 (2002).
18. Meise, C. J. & O'Reilly, J. E. Spatial and seasonal patterns in abundance and age-composition of *Calanus finmarchicus* in the Gulf of Maine and on Georges Bank: 1977–1987. *Deep-Sea Res. II* **43**, 1473–1501 (1996).
19. Fiksen, Ø. The adaptive timing of diapause—a search for evolutionarily robust strategies in *Calanus finmarchicus*. *ICES J. Mar. Sci.* **57**, 1825–1833 (2000).
20. Miller, C. B., Crain, J. A. & Morgan, C. A. Oil storage variability in *Calanus finmarchicus*. *ICES J. Mar. Sci.* **57**, 1786–1799 (2000).
21. Miller, C. B., Cowles, T. J., Wiebe, P. H., Copley, N. J. & Grigg, H. Phenology in *Calanus finmarchicus* - Hypotheses about control mechanisms. *Mar. Ecol. Prog. Ser.* **72**, 79–91 (1991).
22. Speirs, D. C. et al. Ocean-scale modelling of the distribution, abundance, and seasonal dynamics of the copepod *Calanus finmarchicus*. *Mar. Ecol. Prog. Ser.* **313**, 173–192 (2006).
23. Tarrant, A. M. et al. Transcriptional profiling of metabolic transitions during development and diapause preparation in the copepod *Calanus finmarchicus*. *Integr. Comp. Biol.* **56**, 1157–1169 (2016).
24. Baumgartner, M. F. & Tarrant, A. M. The physiology and ecology of diapause in marine copepods. *Annu. Rev. Mar. Sci.* **9**, 387–411 (2017).
25. Wilson, R. J., Banas, N. S., Heath, M. R. & Speirs, D. C. Projected impacts of 21st century climate change on diapause in *Calanus finmarchicus*. *Glob. Change Biol.* **22**, 3332–3340 (2016).
26. Jónasdóttir, S. H., Visser, A. W., Richardson, K. & Heath, M. R. Seasonal copepod lipid pump promotes carbon sequestration in the deep North Atlantic. *Proc. Natl Acad. Sci. USA.* **112**, 12122–12126 (2015).
27. Jónasdóttir, S. H., Wilson, R. J., Gislason, A. & Heath, M. R. Lipid content in overwintering *Calanus finmarchicus* across the Subpolar Eastern North Atlantic Ocean. *Limnol. Oceanogr.* **64**, 2029–2043 (2019).
28. Varpe, Ø. Fitness and phenology: annual routines and zooplankton adaptations to seasonal cycles. *J. Plankton Res.* **34**, 267–276 (2012).
29. Denlinger, D. L., Yocum, G. D. & Rinehart, J. P. in *Insect Endocrinology* (ed Gilbert, L. I.) 430–463 (Academic Press, 2012).
30. Hirche, H. J. Diapause in the marine copepod, *Calanus finmarchicus* - a review. *Ophelia* **44**, 129–143 (1996).
31. Häfker, N. S. et al. *Calanus finmarchicus* seasonal cycle and diapause in relation to gene expression, physiology, and endogenous clocks. *Limnol. Oceanogr.* **63**, 2815–2838 (2018).
32. Roncalli, V. et al. Physiological characterization of the emergence from diapause: a transcriptomics approach. *Sci. Rep.* **8**, 12577 (2018).
33. Roncalli, V., Cieslak, M. C., Hopcroft, R. R. & Lenz, P. H. Capital breeding in a diapausing copepod: a transcriptomics analysis. *Front. Mar. Sci.* **7**, 56 (2020).
34. MacRae, T. H. Gene expression, metabolic regulation and stress tolerance during diapause. *Cell. Mol. Life Sci.* **67**, 2405–2424 (2010).
35. Poelchau, M. F., Reynolds, J. A., Elsik, C. G., Denlinger, D. L. & Armbruster, P. A. Deep sequencing reveals complex mechanisms of diapause preparation in the invasive mosquito, *Aedes albopictus*. *Proc. R. Soc. B* **280** (2013).
36. Ragland, G. J. & Keep, E. Comparative transcriptomics support evolutionary convergence of diapause responses across Insecta. *Physiol. Entomol.* **42**, 246–256 (2017).
37. Košťál, V. Eco-physiological phases of insect diapause. *J. Insect Physiol.* **52**, 113–127 (2006).
38. Tarrant, A. M. et al. Transcriptional profiling of reproductive development, lipid storage and molting throughout the last juvenile stage of the marine copepod *Calanus finmarchicus*. *Front. Zool.* **11**, 1 (2014).
39. Jensen, L. K. et al. A multi-generation *Calanus finmarchicus* culturing system for use in long-term oil exposure experiments. *J. Exp. Mar. Biol. Ecol.* **333**, 71–78 (2006).
40. Cieslak, M. C., Castelfranco, A. M., Roncalli, V., Lenz, P. H. & Hartline, D. K. t-Distributed Stochastic Neighbor Embedding (t-SNE): a tool for eco-physiological transcriptomic analysis. *Mar. Genomics* **51**, 100723 (2020).
41. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
42. Roncalli, V., Cieslak, M. C., Germano, M., Hopcroft, R. R. & Lenz, P. H. Regional heterogeneity impacts gene expression in the sub-arctic zooplankton *Neocalanus flemingeri* in the northern Gulf of Alaska. *Commun. Biol.* **2**, 1–13 (2019).
43. Johnson, K. M., Wong, J. M., Hoshijima, U., Sugano, C. S. & Hofmann, G. E. Seasonal transcriptomes of the Antarctic pteropod *Limacina helicina antarctica*. *Mar. Env. Res.* **143**, 49–59 (2019).
44. Denlinger, D. L. Regulation of diapause. *Annu. Rev. Entomol.* **47**, 93–122 (2002).
45. Denlinger, D. L. & Armbruster, P. A. Mosquito diapause. *Annu. Rev. Entomol.* **59**, 73–93 (2014).
46. Hahn, D. A. & Denlinger, D. L. Energetics of insect diapause. *Annu. Rev. Entomol.* **56**, 103–121 (2011).
47. Sim, C. & Denlinger, D. L. Transcription profiling and regulation of fat metabolism genes in diapausing adults of the mosquito *Culex pipiens*. *Physiol. Genomics* **39**, 202–209 (2009).
48. Sim, C. & Denlinger, D. L. Insulin signaling and the regulation of insect diapause. *Front. Physiol.* **4**, 189 (2013).
49. Andrews, T. S. & Hemberg, M. Identifying cell populations with scRNASeq. *Mol. Asp. Med.* **59**, 114–122 (2018).
50. Habib, N. et al. Div-Seq: single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925–928 (2016).
51. Arrese, E. L. & Soulages, J. L. Insect fat body: energy, metabolism, and regulation. *Annu. Rev. Entomol.* **55**, 207–225 (2010).
52. Hahn, D. A. & Denlinger, D. L. Meeting the energetic demands of insect diapause: nutrient storage and utilization. *J. Insect Physiol.* **53**, 760–773 (2007).
53. Lee, R. F., Hagen, W. & Kattner, G. Lipid storage in marine zooplankton. *Mar. Ecol. Prog. Ser.* **307**, 273–306 (2006).
54. Kattner, G. & Hagen, W. Polar herbivorous copepods—different pathways in lipid biosynthesis. *ICES J. Mar. Sci.* **52**, 329–335 (1995).
55. Miller, C. B., Morgan, C. A., Prahl, F. G. & Sparrow, M. A. Storage lipids of the copepod *Calanus finmarchicus* from Georges Bank and the Gulf of Maine. *Limnol. Oceanogr.* **43**, 488–497 (1998).
56. Hirche, H. J. & Niehoff, B. Reproduction of the Arctic copepod *Calanus hyperboreus* in the Greenland Sea-field and laboratory observations. *Pol. Biol.* **16**, 209–219 (1996).
57. Niehoff, B. & Hirche, H.-J. Oogenesis and gonad maturation in the copepod *Calanus finmarchicus* and the prediction of egg production from preserved samples. *Pol. Biol.* **16**, 601–612 (1996).
58. Košťál, V., Štětina, T., Poupardin, R., Korbelová, J. & Bruce, A. W. Conceptual framework of the eco-physiological phases of insect diapause development justified by transcriptomic profiling. *Proc. Natl Acad. Sci. USA.* **114**, 8532–8537 (2017).
59. Aruda, A. M., Baumgartner, M. F., Reitzel, A. M. & Tarrant, A. M. Heat shock protein expression during stress and diapause in the marine copepod *Calanus finmarchicus*. *J. Insect Physiol.* **57**, 665–675 (2011).
60. Unal, E., Bucklin, A., Lenz, P. H. & Towle, D. W. Gene expression of the marine copepod *Calanus finmarchicus*: responses to small-scale environmental variation in the Gulf of Maine (NW Atlantic Ocean). *J. Exp. Mar. Biol. Ecol.* **446**, 76–85 (2013).
61. Ning, J., Wang, M. X., Li, C. L. & Sun, S. Transcriptome sequencing and *de novo* analysis of the copepod *Calanus sinicus* using 454 GS FLX. *PLoS ONE* **8**, e63741 (2013).
62. Zhang, Q., Lu, Y.-X. & Xu, W.-H. Proteomic and metabolomic profiles of larval hemolymph associated with diapause in the cotton bollworm, *Helicoverpa armigera*. *BMC Genomics* **14**, 751 (2013).
63. Hansen, M. et al. A role for autophagy in the extension of lifespan by dietary restriction in *C. elegans*. *PLoS Genet.* **4**, e24 (2008).
64. Qiu, Z. & MacRae, T. H. ArHsp21, a developmentally regulated small heat-shock protein synthesized in diapausing embryos of *Artemia franciscana*. *Biochem. J.* **411**, 605–611 (2008).
65. Lu, M.-X. et al. Diapause, signal and molecular characteristics of overwintering *Chilo suppressalis* (Insecta: Lepidoptera: Pyralidae). *Sci. Rep.* **3**, 1–9 (2013).
66. Forreryd, A., Johansson, H., Albrekt, A.-S. & Lindstedt, M. Evaluation of high throughput gene expression platforms using a genomic biomarker signature for prediction of skin sensitization. *BMC Genomics* **15**, 379 (2014).
67. Lenz, P. H. et al. *De novo* assembly of a transcriptome for *Calanus finmarchicus* (Crustacea, Copepoda)—the dominant zooplankton of the North Atlantic Ocean. *PLoS ONE* **9**, e88589 (2014).
68. Roncalli, V., Cieslak, M. C. & Lenz, P. H. Transcriptomic responses of the calanoid copepod *Calanus finmarchicus* to the saxitoxin producing dinoflagellate *Alexandrium fundyense*. *Sci. Rep.* **6**, 25708 (2016).
69. Roncalli, V., Cieslak, M. C. & Lenz, P. H. Data from: Transcriptomic responses of the calanoid copepod *Calanus finmarchicus* to the saxitoxin producing dinoflagellate *Alexandrium fundyense*. *Dryad, Dataset* (2016).
70. Choquet, M. et al. Genetics redraws pelagic biogeography of *Calanus*. *Biol. Lett.* **13**, 20170588 (2017).
71. Choquet, M. et al. Can morphology reliably distinguish between the copepods *Calanus finmarchicus* and *C. glacialis*, or is DNA the only way? *Limnol. Oceanogr.: Methods* **16**, 237–252 (2018).

72. Skottene, E. et al. A crude awakening: effects of crude oil on lipid metabolism in calanoid copepods terminating diapause. *Biol. Bull.* **237**, 90–110 (2019).
73. Melle, W. & Skjoldal, H. R. Reproduction and development of *Calanus finmarchicus*, *C. glacialis* and *C. hyperboreus* in the Barents Sea. *Mar. Ecol. Prog. Ser.* **169**, 211–228 (1998).
74. Weydmann, A. et al. Mitochondrial genomes of the key zooplankton copepods Arctic *Calanus glacialis* and North Atlantic *Calanus finmarchicus* with the longest crustacean non-coding regions. *Sci. Rep.* **7**, 1–11 (2017).
75. Lenz, P. H., Lieberman, B., Cieslak, M. C., Roncalli, V. & Hartline, D. K. Transcriptomics and metatranscriptomics in zooplankton: wave of the future? *J. Plankton Res.* **43**, 3–9 (2021).
76. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* <https://doi.org/10.1186/Gb-2009-10-3-R25> (2009).
77. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621 (2008).
78. van der Maaten, L. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
79. Krijthe, J. H. Rtsne: t-Distributed Stochastic Neighbor Embedding using a Barnes-Hut implementation, version 0.13. (2015).
80. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. Second International Conference on Knowledge Discovery and Data Mining (KDD-96)* **96**, 226–231 (1996).
81. Dunn, J. C. Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* **4**, 95–104 (1974).
82. Hahsler, M. & Piekenbrock, M. Dbscan: density based clustering of applications with noise (DBSCAN) and related algorithms. *R. package version 1*, 1–3 (2018).
83. Desgraupes, B. ClusterCrit: Clustering Indices. R package version 1.2.8. (2018).
84. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
85. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* **9**, 559 (2008).
86. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol.* **4**, 17 (2005).
87. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).
88. Alexa, A. & Rahnenfuhrer, J. topGO: enrichment analysis for gene ontology. *R. package version 2*, 2010 (2010).
89. Galili, T., O'Callaghan, A., Sidi, J. & Sievert, C. heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics* **34**, 1600–1602 (2018).
90. Lenz, P. H. et al. Diapause vs. reproductive programs: transcriptional phenotypes in *Calanus finmarchicus*. *Dryad, Dataset*, <https://doi.org/10.5061/dryad.12jm63xw7> (2021).

## Acknowledgements

We would like to thank Mark F. Baumgartner, Bjørn Henrik Hansen, Dag Altin, Trond Nordtug, and Anders J. Olsen for the original study and for making the RNA-Seq data publicly available. We greatly appreciate the administrative and secretarial support provided by Lynn Hata and technical assistance from Brandon Lieberman. This work was supported by National Science Foundation Grants (NSF) OCE-1459235 and OCE-1756767 to P.H.L., D.K.H. and AE Christie and OPP-1746087 to A.M.T. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the National Science Foundation. This is SOEST contribution number 11290.

## Author contributions

P.H.L., V.R., and D.K.H. conceived the study; M.C.C., D.K.H., A.M.C., A.M.T., V.R., and P.H.L. analyzed the data and evaluated the conclusions; P.H.L., V.R., and D.K.H. wrote the first draft of the manuscript; V.R., D.K.H., A.M.C., M.C.C., A.M.T., and P.H.L. revised the manuscript. All authors approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s42003-021-01946-0>.

**Correspondence** and requests for materials should be addressed to V.R.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021