Genome architecture of an exceptionally invasive copepod crossing salinity boundaries

Zhenyong Du (✉ zdu53@wisc.edu)
University of Wisconsin-Madison https://orcid.org/0000-0002-4569-6713

Gregory Gelembiuk
University of Wisconsin-Madison https://orcid.org/0000-0001-7369-9287

Wynne Moss
University of Wisconsin-Madison https://orcid.org/0000-0002-2813-1710

Andrew Tritt
University of Wisconsin-Madison https://orcid.org/0000-0002-1617-449X

Carol Eunmi Lee (✉ carollee@wisc.edu)
University of Wisconsin-Madison https://orcid.org/0000-0001-6355-0542

Research Article

Keywords: Genome architecture, arthropod, Crustacea, invasion, osmoregulation, ionic regulation

Posted Date: June 2nd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3002580/v2

Running title: Genome architecture of a copepod

Genome architecture of an exceptionally invasive copepod crossing salinity boundaries

Zhenyong Du*
ORCID: 0000-0002-4569-6713

Gregory Gelembiuk[†]
ORCID: 0000-0001-7369-9287

Wynne Moss[¶]
ORCID: 0000-0002-2813-1710

Andrew Tritt[§]
ORCID: 0000-0002-1617-449X

Carol Eunmi Lee*
ORCID: 0000-0001-6355-0542

Department of Integrative Biology, 430 Lincoln Drive, Birge Hall, University of Wisconsin, Madison, WI 53706, U.S.A.

*Corresponding Authors:
Zhenyong Du, zdu53@wisc.edu; Carol E. Lee, carollee@wisc.edu

Current Addresses:
[†] Department of Entomology, 1630 Linden Dr., University of Wisconsin, Madison, WI 53706, USA

[¶] U.S. Geological Survey, Northern Rocky Mountain Science Center

[§] Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

*****Figures are embedded in this manuscript for ease of reviewing**

Abstract (250 word limit)

Background: Copepods are among the most abundant organisms on the planet and play critical functions in aquatic ecosystems. Among copepods, populations of the *Eurytemora affinis* species complex are numerically dominant in many coastal habitats and serve as the food source for major fisheries. Intriguingly, certain populations possess the unusual capacity to invade novel salinities on rapid time scales. Despite their ecological importance, high-quality genomic resources have been absent for calanoid copepods, limiting our ability to comprehensively dissect the genomic mechanisms underlying this highly invasive and adaptive capacity.

Results: Here, we present the first chromosome-level genome of a calanoid copepod, from the Atlantic clade (*Eurytemora carolleeae*) of the *E. affinis* species complex. This genome was assembled using high-coverage PacBio and Hi-C sequences of an inbred line, generated through 30 generations of full-sib mating. This genome consisting of 529.3 Mb (contig N50 = 4.2 Mb, scaffold N50 = 140.6 Mb) was anchored onto four chromosomes. Genome annotation predicted 20,262 protein-coding genes, of which ion transporter gene families were substantially expanded based on comparative analyses of 12 additional arthropod genomes. Also, we found genome-wide signatures of historical gene body methylation of the ion transporter genes and significant clustering of these genes on each chromosome.

Conclusions: This genome represents one of the most contiguous copepod genomes to date and among the highest quality of marine invertebrate genomes. As such, this genome provides an invaluable resource that could help yield fundamental insights into the ability of this copepod to adapt to rapid environmental transitions.

Keywords: Genome architecture, arthropod, Crustacea, invasion, osmoregulation, ionic regulation

Background

Copepods form the largest biomass of animals in the world's oceans, and arguably on the planet [1-3]. Among estuarine and coastal copepods, the planktonic calanoid copepod *Eurytemora affinis* species complex is a dominant grazer throughout the Northern Hemisphere, forming an enormous biomass in estuaries and coastal habitats, with census sizes in the billions [4-9]. As such, this copepod represents a major food source for some of the world's most important fisheries, such as herring, anchovy, salmon, and flounder [10-17].

Patterns of speciation within this species complex have been uncertain and taxonomic designations of clades within the species complex have been inconsistent. Populations and sibling species within this species complex are marked by a considerable degree of morphological stasis [18]. However, large genetic divergencies separate at least six geographically distinct clades [19, 20] with idiosyncratic patterns of reproductive isolation among the clades [19]. Subtle morphological differences have led to the naming of some populations and clades as novel species [21, 22]. However, the precise genomic architecture of members of this species complex has remained largely elusive.

This species complex has held intense ecological and evolutionary interest because of its extraordinary ability to invade a wide range of salinities over very short time scales [23, 24]. For an invertebrate, this copepod has the exceptionally rare ability to cross salinity boundaries from hypersaline to completely fresh water [20, 23, 25-29]. Within a few decades, saline populations from this species complex have invaded freshwater habitats multiple times independently on three continents through human activity [23, 30]. For instance, with the opening of the St. Lawrence Seaway, the Atlantic clade of the *E. affinis* complex (aka. *E. carolleeae* Alekseev & Souissi, 2011) [21] was introduced into the North American Great Lakes from saline estuarine populations ca. 65 years ago, starting with Lake Ontario in 1958 and reaching Lake Superior by 1972 [23, 31]. Likewise, populations of the Gulf clade of the *E. affinis* complex spread rapidly from the Gulf of Mexico into inland freshwater reservoirs and lakes throughout the Southeastern United States over a time period of ~60 years [23, 32]. Additionally, a European *E. affinis* population survived the transformation of a saltwater bay in the Netherlands into freshwater lakes (IJsselmeer and Markemeer) over a period of six years [23, 33]. Moreover, many *E. affinis* complex populations are likely to survive changing habitat salinities induced by climate change [24, 34]. These freshwater invasions by saline *E. affinis* complex populations were accompanied by the rapid evolution of freshwater tolerance, coupled with reduced high salinity tolerance, along with evolutionary changes in life history and ion regulatory function [25, 26, 35, 36]. Natural selection experiments in the laboratory have revealed that rapid freshwater adaptation could occur in only a few generations [25, 35, 37].

Investigating the genome architecture of this exceptionally invasive copepod species complex would provide fundamental insights into the genomic and evolutionary mechanisms facilitating their rapid habitat invasions [38, 39]. However, high-quality genome resources have long been absent for most copepod groups [40, 41]. Only four chromosome-level genome assemblies are available for copepods in the NCBI Genome database [42], namely for two parasitic copepods (Siphonostomatoida) and two species of the intertidal copepod *Tigriopus* (Harpacticoida). Such genomic resources are completely lacking for the copepod orders Calanoida and Cyclopoida. This deficit of genomic resources for copepods is quite striking, given their enormous ecological roles as grazers of the sea and their contribution of ~70% of the total zooplankton biomass [1, 43]. The *E. affinis* complex in particular has long served as a critically important model system for evolutionary, physiological, and ecological studies, with over 1000 studies published on this copepod system (Google Scholar).

Thus, we present the first chromosome-level reference genome for a calanoid copepod, specifically for *E. carolleeae,* the Atlantic clade of the copepod *E. affinis* species complex [19, 21, 23]. Our goal was to produce a high-quality genome, based on high coverage PacBio, Illumina, and Hi-C sequencing. To reduce the high level of heterozygosity present in the wild population [30], we generated an inbred line through 30 generations full-sib mating of a saline population from the St. Lawrence salt marsh (Baie de L'Isle Verte). As a result, we assembled a new genome that is far more contiguous than our prior genome based on the same inbred line, based only on Illumina sequencing [44]. Thus, we produced a reference genome that could be used to effectively uncover genetic mechanisms of environmental adaptation. Moreover, dissecting the genome architecture of this species complex could provide novel insights into its incredible capacity to invade novel environments.

Results

Chromosome-level genome assembly

The genome assembly we generated for *E. carolleeae* (Atlantic clade of the *E. affinis* complex) [21] had a much higher degree of completeness and contiguity than other available copepod genomes (Additional file 2: Table S1). Our genome assembly integrated sequence data from ~60.6× coverage PacBio Continuous Long Read (CLR) sequencing, ~14.2× coverage PacBio High-fidelity Circular Consensus Sequencing (HiFi CCS) and ~73.4× coverage Illumina short-read sequencing. These data generated a 536 megabase (Mb) assembly of 325 contigs, with a contig N50 of 4.2 Mb. This result was consistent with the estimated genome size of 509~540 Mb based on k-mer analyses (Additional file 1: Fig. S1). This assembly was further scaffolded based on ~85.6× coverage Hi-C data and filtered to generate a 529.3 Mb final assembly, with a scaffold N50 of 140.6 Mb. 95.6% of the assembly was anchored onto four pseudo-chromosomes (Fig. 1a). This genome was highly AT-rich, with a mean GC content 32.5% (Fig. 1a). This GC content was comparable to those of other calanoid copepods, but lower than those of harpacticoid copepods (Additional file 2: Table S1). The GC content of this genome was also lower than that of *Drosophila melanogaster* (42.0%) and lower than 128 out of 154 published genome assemblies of marine invertebrates in a recent survey [45]. The Benchmark of Universal Single-Copy Orthologs (BUSCO) analyses indicated that 93.1% (90.2% single-copy and 2.9% duplicated) complete BUSCOs (1013 in arthropod odb10 dataset) were captured in this genome.

This new genome was vastly improved relative to our prior assembly based on the same inbred line, generated from only Illumina sequencing [44]. In this new genome, the contig N50 was greatly improved (from 67.7 kilobase (kb) to 4.2 Mb) and the sequences were successfully scaffolded onto chromosomes. The contig N50 length we obtained here was greater than 33 out of 35 available genome assemblies for Copepoda in NCBI Genome database [42]. The two copepod assemblies with greater contig N50 length than ours are based on Oxford Nanopore sequencing and their samples are taken from wild outbred populations [46, 47]. The contig N50 of our genome was also longer than 151 out of 154 published genome assemblies of marine invertebrates in a recent survey [45]. Thus, this genome is one of the most contiguous copepod genomes to date and also among the highest quality of marine invertebrate genomes.

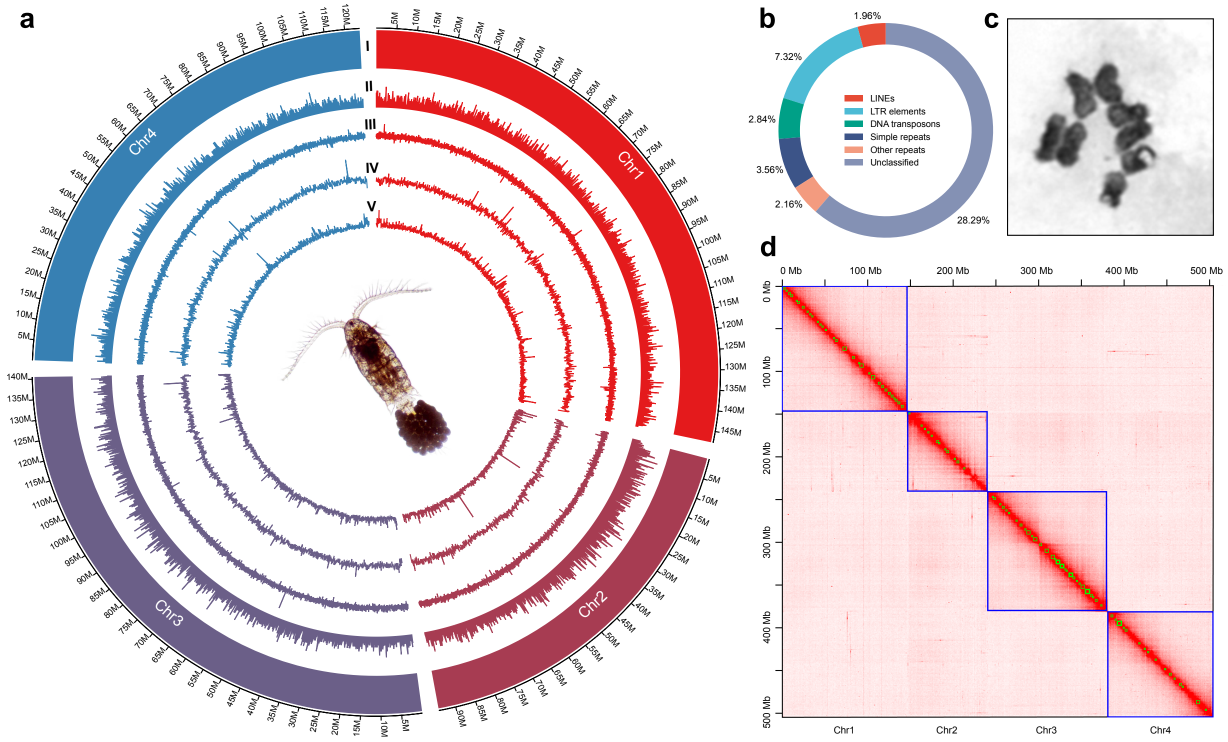

Fig. 1. Chromosome-level genome assembly of the copepod *Eurytemora carolleeae* (*E. affinis* complex, Atlantic clade). (a) Circular diagram showing the genome landscape. I. Four chromosomes on the Mb scale. II. Density of protein-coding genes. III. Distribution of GC content (Mean GC = 32.5%). IV. Distribution of repetitive sequences. V. Distribution of LTR. All distributions were calculated in 100 kb non-overlapping sliding windows. **(b)** The proportion of repetitive sequences identified in the copepod genome. The circular diagram shows their relative proportions out of the total repetitive sequences (46.12% of the genome), and the numbers labelled on the diagram represent their percentage of occupied length in the genome assembly. **(c)** Well-isolated cell that shows the karyotype of the copepod (2n = 8) at metaphase. **(d)** The Hi-C contact map of the genome generated by Juicebox.

Genome size and karyotype evolution

Among copepods, *E. carolleeae* of the *E. affinis* species complex has a small genome size and a low number of chromosomes. The genome size of *E. carolleeae* is 1C = 529.3 Mb, lower than the average size of 4.0 gigabases (Gb) for 41 calanoid copepod species and lower than the average size of 1.85 Gb for

112 copepod species from four orders, based on mostly cytological estimates and some genome sequences (Additional file 2: Table S3). For a calanoid copepod, this small genome size of *E. carolleeae* is an outlier, given that the order Calanoida exhibits larger mean genome size (Mean = 3993 Mb) than those of the other copepod orders (Mean = 315–667 Mb) (Fig. 2c). Overall, the range in genome size among copepod species is large (1C = 0.1–14.4 Gb) (Additional file 2: Tables S1 and S3) with significant differences among the four orders (Fig. 2c; Kruskal-Wallis test, H = 49.58, DF = 3, *P* = 9.8e-11).

Our *E. carolleeae* genome assembly based on Hi-C revealed only four haploid chromosomes (2n = 8) (Fig. 1d). Our karyotyping experiment confirmed the presence of four haploid chromosomes in several well isolated cells (Fig. 1c, Additional file 1: Fig. S2). This chromosome number tends to be near the low end for copepods, which varies widely among copepod species (2n = 6–42) (Figs. 2a, b; Additional file 2: Table S2) and differs significantly among the four copepod orders (Fig. 2b; Kruskal-Wallis test, H = 35.52, DF = 3, *P* = 9.5e-8). While it appears that chromosome number increased during the evolutionary history of the Calanoida, this pattern is unclear due to the unavailability of karyotype information for the most basal clade within the Calanoida and the basal clade within the Copepoda, the order Platycopioida (Fig. 2a, grey clades).

Evolutionary patterns of genomic rearrangements are difficult to discern due to lack of synteny between the genome of *E. carolleeae* and two other chromosome-level genomes from different copepod orders, namely, the tidepool copepod *Tigriopus californicus* (Harpacticoida) and the salmon louse *Lepeophtheirus salmonis* (Siphonostomatoida) (Additional file 1: Fig. S3). While the tidepool copepod and salmon louse genomes showed much greater synteny with each other than with *E. carolleeae*, a large number of chromosomal translocations between their genomes was still evident. The lack of synteny between the *E. carolleeae* and other copepod genomes indicates that major genomic rearrangements occurred during the course of copepod evolution, with far less conservation relative to vertebrates and some insects, such as butterflies and moths [48, 49].

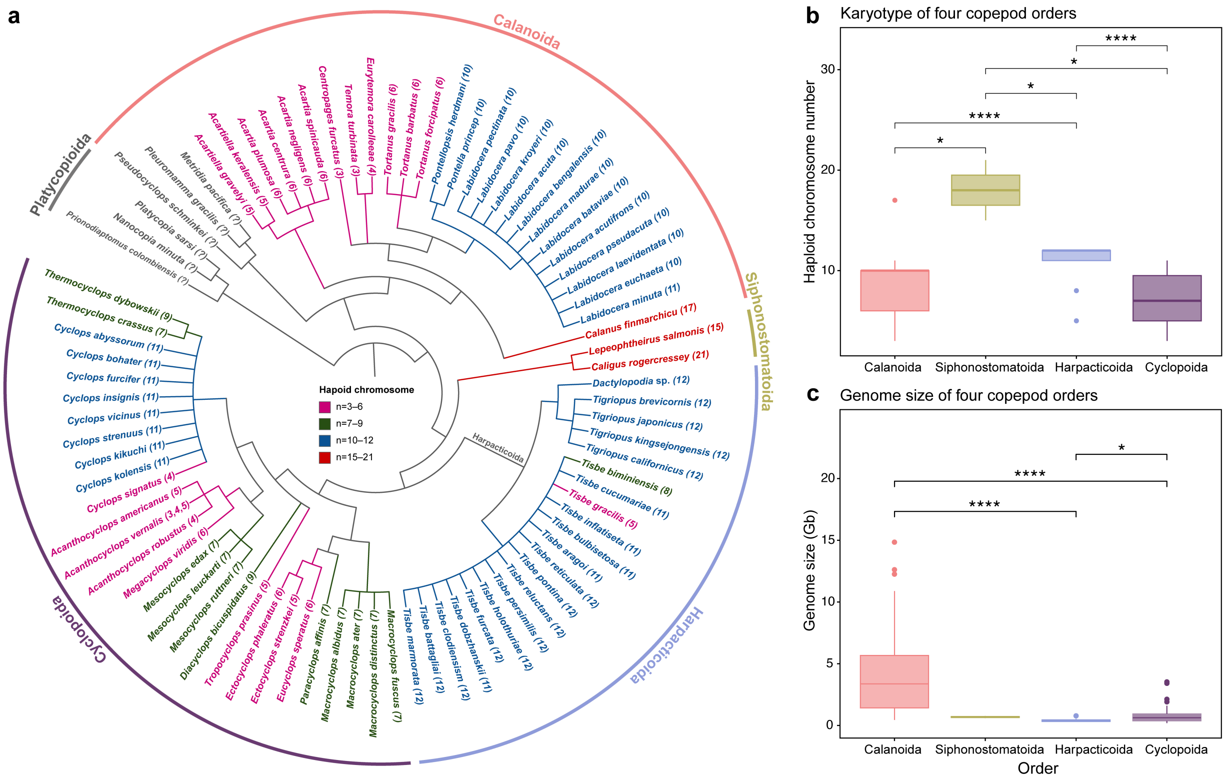

Fig. 2. Chromosome number and genome size evolution in the crustacean class Copepoda. (a) Phylogeny of copepod species from five copepod orders. The phylogenetic topology was obtained from the synthesis tree of copepods, which integrated 31 published phylogenies [50]. Chromosome numbers are shown within parentheses after the species names. Different colors of species names represent the ranges of chromosome numbers. Clades that occupy basal phylogenetic positions, but possess unknown karyotype, are shown in grey in the phylogeny. **(b)** Mean chromosome number of four copepod orders (see Additional file 2: Table S2 for details). Chromosome number differs significantly among the four orders (Kruskal-Wallis test, H = 35.52, DF = 3, *P* = 9.5e-8). **(c)** Mean genome size of four copepod orders. Calanoida mean genome size = 3993 Mb, Siphonostomatoida = 563 Mb, Harpacticoida = 315 Mb, and Cyclopoida = 667 Mb (see Additional file 2: Table S3 for details). Genome size differs significantly among the four orders (Kruskal-Wallis test, H = 49.58, DF = 3, *P* = 9.8e-11). Asterisks in (b–c) indicate the significance levels for Wilcoxon tests, where * refers to *P* < 0.05 and **** indicates *P* < 1e-4. Nonsignificant *P*-values are not shown.

Genome annotation and gene family expansions and contractions

By integrating our *de novo* repetitive sequence database with public repetitive sequence databases, we identified 46.1% of the *E. carolleeae* assembly as repetitive sequence, which comprised 244.10 Mb in length of the genome assembly (Fig. 1a, IV, V). The Long Terminal Repeat (LTR) comprised the largest percentage of the repetitive sequences (Fig. 1b, blue), other than the unclassified repetitive sequences (Fig. 1b, lavender; Additional file 2: Table S4). A total of 2426 non-coding RNA sequences were also identified and annotated in the genome, among which 1574 transfer RNA (tRNA) sequences formed the largest category (Additional file 2: Table S5). The number of non-coding RNA sequences revealed here was within the range of 386–4559 found in other copepod genomes in the NCBI Genome database [42].

A total of 20,262 protein-coding genes was predicted in the *E. carolleeae* genome, occupying 261.62 Mb in length of the genome assembly, based on abundant transcriptome data for the *E. affinis* complex, homologous proteins of other arthropods, and *ab initio* prediction (Additional file 2: Table S6). Among these genes, almost all genes (20,259) were functionally assigned based on at least one of eight functional annotation databases (Additional file 2: Table S7). This predicted number of annotated protein-coding genes is greater than those of the tidepool copepod *Tigriopus californicus* (15,500 genes) and the salmon louse *Lepeoptheirus salmonis* (13,081 genes).

The higher number of genes in our genome was not due to gene fragmentation, as our mean gene length was 12.91 kb, mean coding sequence length was 1.45 kb, and mean exon number per gene was 10.9 (Additional file 2: Table S6). In addition, this larger number of genes was not due to counting separate alleles as genes, given that we used an inbred line with heterozygosity of ~0.5% (Additional file 1: Fig. S1) and the duplicated BUSCO detected in the genome assembly was only 2.9%. To determine whether the greater gene number was caused by ancient whole genome duplication events (WGD), we examined the distribution of synonymous substitutions per site (Ks) among paralogous genes within the genome (known as Ks plot analysis) [51]. Based on the Ks plot, we found no evidence of ancient WGD in the *E. carolleeae* genome (Additional file 1: Fig. S4). Interestingly, the largest proportions of gene duplication events occurred quite recently (Ks = 0–0.04, Additional file 1: Fig. S4).

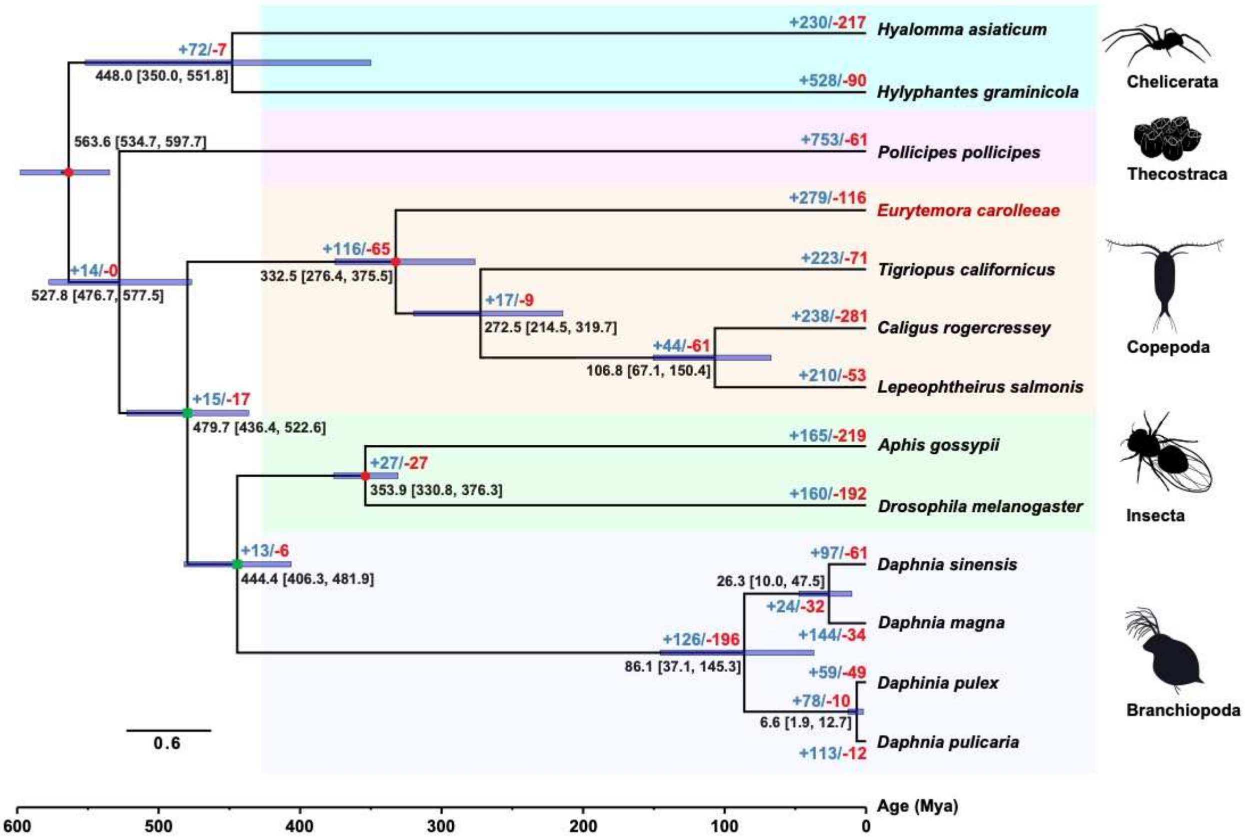

Fig. 3. Gene family expansions and contractions during the evolutionary history of the Arthropoda, with a focus on the Copepoda. Phylogenetic reconstruction of 13 high-quality arthropod genomes was performed using RAxML based on concatenated single copy ortholog genes. All nodes show bootstrap values of 100%, except for two nodes with green rectangles, which have values of 66% (left node) and 60% (right node). Red circles represent three calibrated nodes with confidence time intervals retrieved from the Timetree database and applied in MCMCTree. Mean estimated divergence times are shown at each node with brackets indicating 95% highest posterior densities. The divergence times are on a scale of millions of years ago (Mya). The numbers of expanded gene families (in blue) and contracted gene families (in red) are shown on the branch tips and next to each node.

To determine patterns of gene family gains and losses across the Arthropoda, with a focus on copepods, we conducted comparative genomic analyses using shared ortholog groups (gene families) across 12 additional arthropod species. In this comparative analysis, we included only high-quality genomes from different arthropod subphyla that were assembled with long read sequencing data to the chromosome level. A phylogeny was reconstructed using a matrix of 101 concatenated single copy ortholog genes (Additional file 2: Table S8). This phylogeny supported the topology of ((((Insecta + Branchipoda) + Copepoda) + Thecostraca) + Chelicerata); although, the relationships between Insecta, Branchiopoda, and Copepoda were not highly supported (Fig. 3, green dots at nodes). Overall, we found substantial numbers of conserved ortholog genes (4042) shared among *E. carolleeae* and three other pancrustacean species (Additional file 1: Fig. S5).

Our analysis of gene family expansions and contractions revealed a significant enrichment of ion transport-related genes in the *E. carolleeae* genome (Fig. 4, Additional file 1: Fig. S6, Additional file 2: Tables S9–S12). Compared to other arthropod genomes, we detected in this copepod genome the expansion of 279 ortholog groups (aka. gene families), corresponding to 1162 genes (Additional file 2: Table S9), and the contraction of 116 gene families, corresponding to 224 genes (Fig. 3, Additional file 2: Table S10).

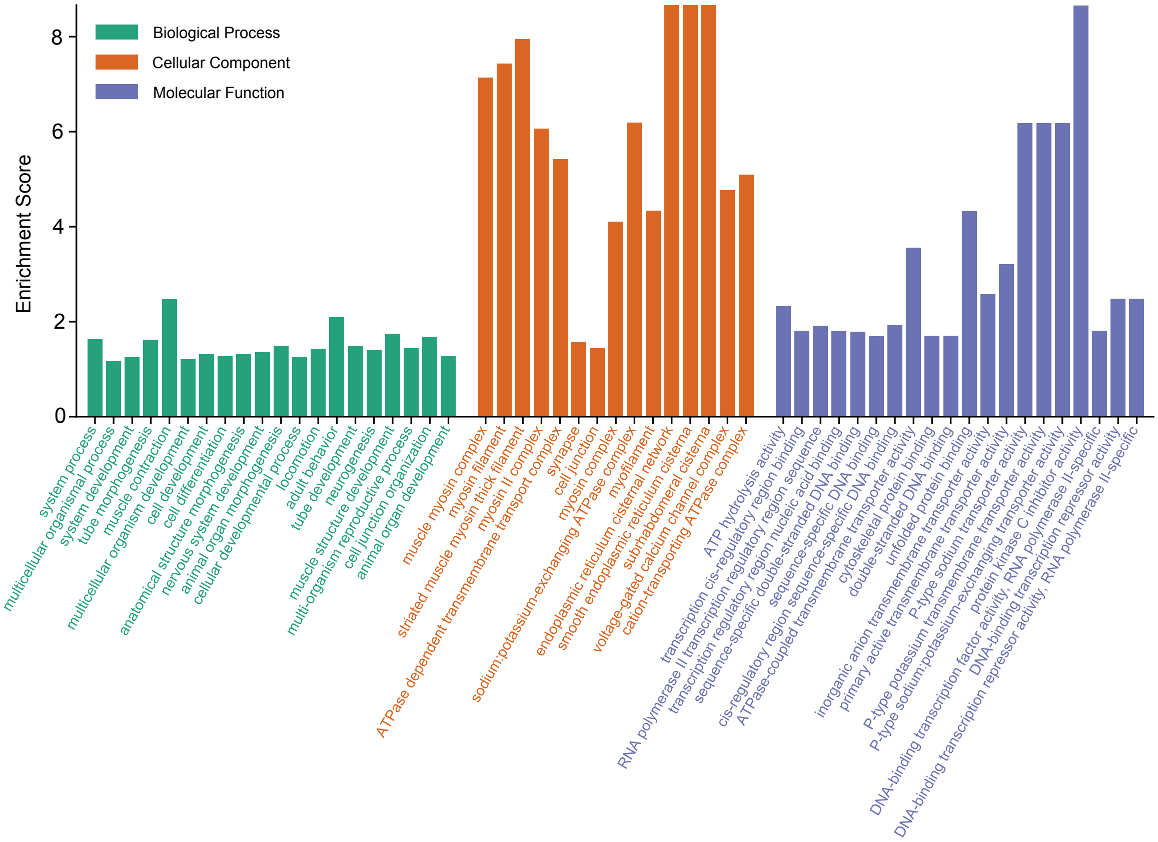

Fig. 4. Significantly enriched of gene ontology (GO) terms in the expanded set of genes in the *Eurytemora carolleeae* genome. The GO terms were sorted by *P*-value (with higher *P*-value toward the right in each category). The complete list of enriched GO terms is shown in Additional file 2: Table S11. Only the top 20 GO terms of the Biological Process and Molecular Function categories, and top 15 GO terms of Cellular Component category are shown here.

Through gene function enrichment analysis with GO and KEGG annotation, we found that 29.2% (61 out of 209) of the significantly enriched GO terms in the Molecular Function category was related to ion transport activity. Of these significant GO terms related to ion transport activity, 63.9% (39 out of 61) were related specifically to inorganic ion (cation and anion) transport activity (Fig. 4, Additional file 1: Fig. S6, Additional file 2: Tables S11 and S12). In the Cellular Component category, 7.6% (11 out of 144) of the significantly enriched GO terms were related to ion transport activity, whereas in the Biological Process category 5.6% (98 out of 1734) of the significantly enriched GO terms were related to ion transport and regulation of ion transporter activity. In the Cellular Component category, the most significantly enriched GO terms included "ATPase dependent transmembrane transport complex" (GO:0098533), "sodium: potassium-exchanging ATPase complex" (GO:0005890), "cation-transporting ATPase complex" (GO:0090533) (Fig. 4). Similarly, the most significantly enriched GO terms in the Molecular Function category included "ATPase-coupled transmembrane transporter activity" (GO:0042626), "inorganic anion transmembrane transporter activity" (GO:0015103), "primary active transmembrane transporter activity" (GO:0015399), "P-type sodium transporter activity" (GO:0008554), "P-type potassium transmembrane transporter activity" (GO:0008556), "P-type sodium: potassium-exchanging transporter activity" (GO:0005391) (Fig. 4). In the Biological Process category, significant GO terms included "regulation of sodium ion transmembrane transporter activity" (GO:2000649, GO:1902305), "regulation of sodium ion export across plasma membrane" (GO:1903276) and development related categories, such as "cell development" (GO:0048468) and "cellular developmental process" (GO:0048869). In terms of expansions of individual ion transporter gene families, such as *Na^+/H^+ antiporter* (*NHA*), *Na^+/K^+ ATPase* (*NKA*), *Ammonia transporter* (*AMT*), and *$Na^+/K^+/Cl^-$ cotransporter* (*NKCC*), the *E. carolleeae* genome has 6–8 gene paralogs, whereas *Drosophila melanogaster* typically has only two.

Genome-wide $CpG_{o/e}$ values as signatures of historical methylation of protein-coding genes

To determine genome-wide signatures of DNA methylation of our protein-coding genes, we determined the genome-wide distribution of CpG sites as indicators of DNA methylation. We calculated $CpG_{o/e}$ values, which are the ratio between the observed and expected incidence of CpG dinucleotide sites (where a cytosine [C] is followed by a guanine [G]). Most DNA methylation events occur at CpG sites and results in the production of 5-methylcytosine (5mC). Subsequently, spontaneous deamination of 5mC leads to C to T conversion [52, 53]. Thus, high levels of DNA methylation will eventually cause the depletion of CpG sites associated with genes [52, 54, 55]. Typically, genes with lower $CpG_{o/e}$ values (lower numbers of observed CpG sites than expected) might have undergone higher levels of methylation in the past. In contrast, genes with higher $CpG_{o/e}$ values might have experienced lower levels of methylation previously.

The $CpG_{o/e}$ values across all genes displayed a unimodal distribution, with a very low mean $CpG_{o/e}$ value of 0.5 in the *E. carolleeae* genome (Fig. 5a). This unimodal distribution and low mean $CpG_{o/e}$ value represents an extreme case of CpG depletion, indicating genome-wide signatures of high levels of past methylation [56]. Most of genes (19,960 out of 20,262) had a $CpG_{o/e}$ value lower than 1 (Fig. 5a). The distribution of $CpG_{o/e}$ values was not biased by the positions of genes on different chromosomes (Fig. 5b). The mean $CpG_{o/e}$ value of our genome was much lower than the unimodal distribution of *Drosophila melanogaster* (mean $CpG_{o/e}$ value around 1) [57] and its unimodal distribution differed from the bimodal distributions found in many molluscs [56] and insects [54, 57].

GO enrichment analysis for the genes with the 5% lowest and 5% highest $CpG_{o/e}$ values (1013 genes), performed to associate the occurrence of gene methylation with gene functions, revealed very different sets of gene functions in the two groups. Notably, genes with the lowest $CpG_{o/e}$ values were significantly enriched predominantly with GO terms related to ion transmembrane transport functions (Fig. 5c, Additional file 2: Table S13). Specifically, 66.7% (6 out of 9) GO terms in the Biological Process category and 60% (3 out of 5) GO terms in the Molecular Function category were related to ion transport (Fig. 5c). These GO terms included "monoatomic anion transport" (GO:0006820), "monoatomic ion transport" (GO:0006811), "inorganic cation transmembrane transport" (GO:0098662), "metal ion transmembrane transporter activity" (GO:0046873), and "salt transmembrane transporter activity" (GO:1901702). These low CpG$_{o/e}$ values for ion transporter genes suggest that these genes had extremely high levels of methylation in the past [52].

In contrast, genes with the highest CpG$_{o/e}$ values were enriched with conserved cellular functions, such as "nucleic acid binding" (GO:0003676), "RNA processing" (GO:0006396), and "RNA metabolic process" (GO:0016070) (Additional file 2: Table S14). These GO terms represent housekeeping genes that were identified as hypermethylated in previous studies [54, 57]. But, here they have relatively low levels of past methylations, so the result here seems to be opposite of what was found previously.

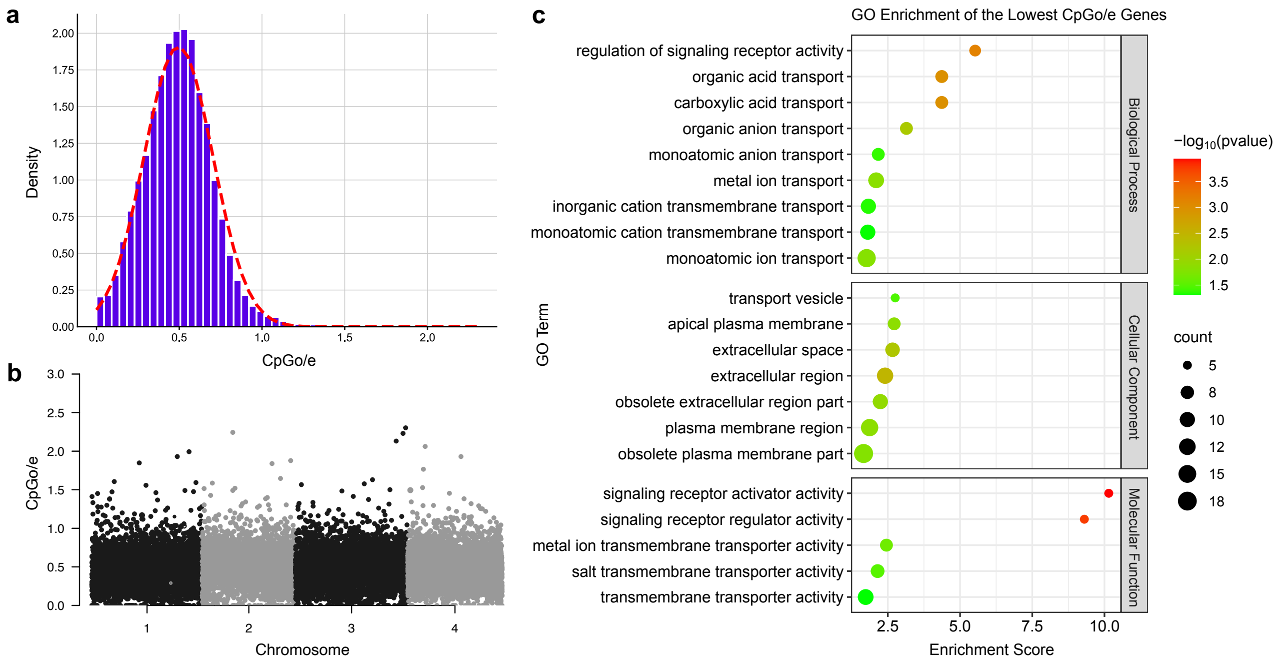

Fig. 5. Patterns of genome-wide CpG$_{o/e}$ values of gene bodies, corresponding to signatures of past gene methylation in the *E. carolleeae* genome. (a) The CpG$_{o/e}$ values of the protein-coding gene sequences display a unimodal distribution. **(b)** The distribution of CpG$_{o/e}$ values across the genome when the genes are arranged by their position on each chromosome. **(c)** GO enrichment of the 1013 genes with 5% lowest CpG$_{o/e}$ values. The significance of GO enrichment is shown by the color of the circles and the enriched gene number is indicated by the size of the circles. The ion transporter genes tend to have the lowest CpG$_{o/e}$ values, suggesting extremely high levels of methylation in the past [52].

Localization of ion transporter genes on the four chromosomes

Given that ion transport-related genes were the most enriched GO category in the *E. carolleeae* genome, we manually annotated and localized the ion transporter gene paralogs on the four chromosomes (Fig. 6a, Additional file 2: Table S15). We focused heavily on the ion transporter paralogs that are targets of natural selection during salinity transitions in *E. affinis* complex populations [30, 36, 37, 58] and likely involved in ion uptake in freshwater habitats (Figs. 6b, c). For instance, the ion transporter paralogs we mapped onto the chromosomes included the gene families *Na^+/H^+ antiporter* (*NHA*), *Na^+/K^+ ATPase* (*NKA*), *Carbonic Anhydrase* (*CA*), *Rh protein* (*Rh*), *Na^+/H^+ exchanger* (*NHE*), *$Na^+/K^+/Cl^-$ cotransporter* (*NKCC*), and *Ammonia transporter* (*AMT*) and subunits of *Vacuolar-type ATPase* (*VHA*) [58]. We found that these ion transporter gene paralogs and subunits were distributed unevenly on the different chromosomes. Specifically, 14, 14, 33, and 22 paralogs were found on Chromosomes 1 to 4, respectively (Fig. 6a). Interestingly, the highest density of ion transporters was localized on the second longest chromosome, Chromosome 3 (Chr3), which contained two-fold more paralogs than the longest chromosome (#1).

Many ion transporter paralogs of both the same and different gene families were clustered together on the chromosomes. For example, *NKCC* and *NKA-β*, *CA* and *NKA-α* on were clustered on one end of Chr3, and seven tandem *NHA* paralogs were clustered near the centromere on Chr3 (Fig. 6a). We found that the distribution of ion transporter genes on the chromosomes deviated significantly from a uniform distribution and tended to be more clustered than expected (Additional file 1: Fig. S7), both for 83 key ion transporter genes (Fig. 6a, colored vertical lines; involved in hypothesized models of ion uptake in Figs. 6b, c) (Kolmogorov-Smirnov test, Z = 4.89, *P* = 0.00) and 490 genes found with ion transporting function (Fig. 6a, vertical light blue lines) (Kolmogorov-Smirnov test, Z = 11.45, *P* = 0.00). In addition, the distributions of ion transporter genes differed significantly from those of functionally conserved housekeeping genes (Additional file 2: Tables S14 and S16) and showed a higher frequency of closely spaced genes (Additional file 1: Fig. S8), both for 83 key ion transporter genes (Additional file 1: Fig. S8a) (Chi-square goodness of fit test, $\chi^2 = 18.5$, DF = 5, *P* = 6.2e-5) and 490 genes found with ion transporting function (Additional file 1: Fig. S8b) (Chi-square goodness of fit test, $\chi^2 = 73.0$, DF = 15, P

= 1.3e-9). Notably, we found a high density of ion transporter paralogs clustered around the centromere of Chr3 (Fig. 6a, Additional file 1: Figs. S9 and S10). Although, the set of gene paralogs clustered around the centromere are not the specific ones that show coordinated gene expression or parallel evolution across multiple studies [58].

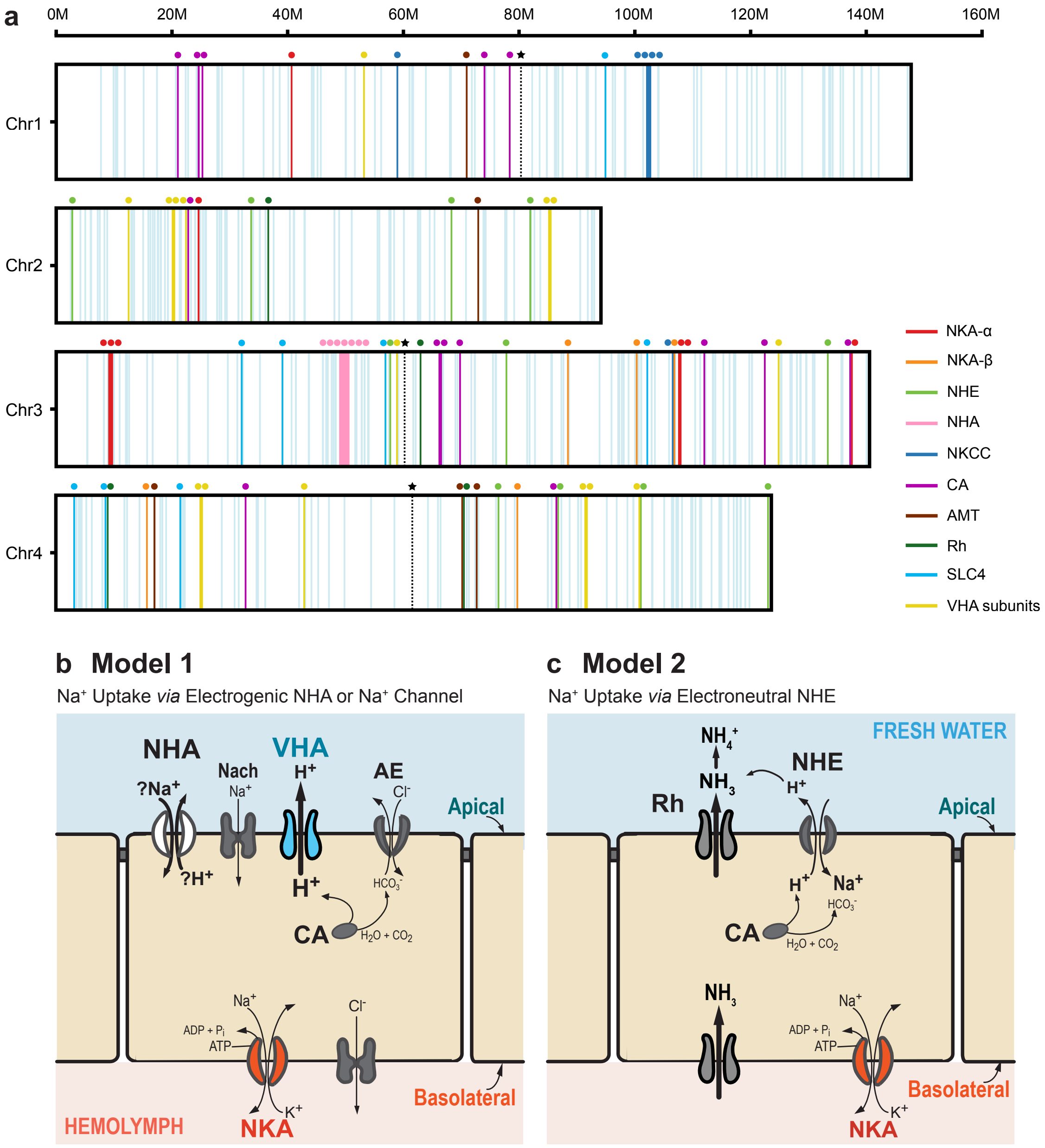

Fig. 6. Localization of ion transporter genes on *E. carolleeae* chromosomes and hypothetical models of ion uptake from fresh water. (a) Ion transporter genes mapped onto the four *E. carolleeae* chromosomes. The vertical light blue lines represent 490 genes with ion (cation and anion) transporting function based on the genome annotation. The vertical lines and circles in other colors represent 83 key genes that showed evolutionary shifts in gene expression and/or signatures of selection in prior studies

and are likely involved in hypothetical models of ion uptake. The dashed lines marked with stars indicate the positions of centromeres based on the Hi-C contact map (Fig. 1d, Additional file 1: Fig. S11). **(b, c)** Hypothetical models of ion uptake from freshwater environments. **(b)** Model 1: VHA generates an electrochemical gradient by pumping out protons, to facilitate uptake of Na^+ through an electrogenic Na^+ transporter (likely NHA). CA produces protons for VHA. **(c)** Model 2: An ammonia transporter Rh protein exports NH_3 out of the cell and this NH_3 reacts with H^+ to form NH_4^+. The deficit of extracellular H^+ concentrations cause NHE to export H^+ in exchange for Na^+. CA produces protons for NHE. These models are not comprehensive for all tissues or taxa and are not mutually exclusive.

Discussion

Features of the calanoid copepod reference genome

Copepods form the largest biomass of animals on the planet and contribute to the majority of total zooplankton biomass in aquatic habitats [1, 43]. However, despite their critical roles for ecosystem functioning and maintenance of fisheries of the planet, high-quality genomic resources had been lacking. This project generated the first chromosome-level calanoid copepod reference genome for *Eurytemora carolleeae* (Atlantic clade of the *E. affinis* species complex) [19, 21, 23], with the highest level of completeness and continuity relative to other copepod genomes [42]. Moreover, this genome ranks among the highest quality among all marine invertebrate genomes [45]. As such, this genome provides an invaluable resource for future studies of this ecologically critical group.

Fundamental features of this calanoid copepod genome are its relatively small genome size (1C = 529.3 Mb) and low chromosome number (2n = 8) (Figs. 1 and 2, Additional file 2: Tables S2 and S3) [42]. We also found extremely low synteny with genomes of other copepod species (Additional file 1: Fig. S3). The relatively small genome size of *E. carolleeae* might be a result of its large effective population size in nature [59]. The effective population size of *E. carolleeae* is approximately 10^6 in the St. Lawrence estuary, based on our previous estimates of Watterson's theta (0.0131) [30] and assuming a mutation rate of 3.46×10^{-9} based on *Drosophila melanogaster* [60].

The *E. carolleeae* genome size (1C = 529.3 Mb) is within a similar range as our previous estimate for the same population (L'Isle Verte) based on DNA cytophotometry of embryonic cells, which yielded a 2C genome size of 0.6–0.7 pg DNA/cell or 1C = 318 Mb [61]. This prior study revealed, however, that the majority of somatic cell nuclei have twice this DNA content (2C = 1.3 pg/nucleus, or 1C = 636 Mb) in the adults examined, possibly due to cells arrested at the G2 stage of the cell cycle or some degree of endopolyploidy. The occurrence of 4C nuclei has been found in other copepods [61], branchiopods [62], and in many plant species [63]. Endopolyploidy is thought to function to make more DNA available for transcription [61]. This higher DNA content of somatic cells would not have affected our genome assembly, as existing DNA would simply have been replicated. Moreover, our earlier draft genome sequence assembled from Illumina sequences [44] was based on DNA exclusively from egg sacs, namely embryonic tissue, and yielded a similar genome size of ~510 Mb (Additional file 1: Fig. S1).

In general, we found that genome size and chromosome number among copepods are not conserved but highly variable (Fig. 2, Additional file 2: Tables S2 and S3). For instance, chromosome number variation in copepods is on par with the levels of variation found in vertebrates and insects [48, 49, 64]. The high variance in chromosome number in copepods suggests an evolutionary history of chromosomal fusions and fissions [65] and associated genomic rearrangements [66]. Such genomic rearrangements might explain the low levels of synteny we found among copepod genomes (Additional file 1: Fig. S3). The relatively large genome sizes (> 1 Gb) of some copepod species, especially in the Cyclopoida (Additional file 2: Table S3), reflect only the germline genome and not the somatic genome [67-69]. Some copepods undergo chromatin diminution, which is the programmed deletion of chromatin from embryonic presomatic cells during development, resulting in a 5–75 fold reduction in somatic genome size [67, 70, 71]. There is no evidence of chromatin diminution in *E. carolleeae* [61].

Expansions of ion transporter genes in the *E. carolleeae* genome

Based on a comparative genomic analysis that included four copepods and a total of 13 arthropod species, we found substantial gene family expansion in the *E. carolleeae* genome (Fig. 3). The expanded gene families were significantly enriched with ion transporter gene categories (with 29.2% of Molecular Function, 7.6% of Cellular Component, and 5.7% of Biological Process GO terms related to ion transport). Ion transporter genes have been found repeatedly as the largest functional (GO) category under selection during salinity change in our previous evolutionary and physiological studies [20, 30, 34, 58].

The high frequency of low Ks counts (low divergence gene duplicates) in the Ks plot (Additional file 1: Fig. S4) and the occurrence of tandem ion transporter paralogs found on the chromosomes (Fig. 6) suggest that the expansions of ion transporter genes tended to occur very recently.

Low genome-wide gene body $CpG_{o/e}$ values and methylation of ion transporter genes

In the *E. carolleeae* genome, we found a genome-wide pattern of extremely low mean $CpG_{o/e}$ values of gene bodies. 98.5% of genes appeared to be CpG depleted (with $CpG_{o/e}$ values lower than 1). This CpG depletion likely contributes to the low GC content of this genome (32.5% GC). The mean $CpG_{o/e}$ value of 0.5 in the *E. carolleeae* genome was lower than those of 152 out of 154 insects and arthropods from a previous survey [62]. Based on this survey, the mean $CpG_{o/e}$ value 0.5 for *E. carolleeae* was comparable only to the low $CpG_{o/e}$ value of 0.47 for two species, the fiddler crab *Celuca pugilator* and the remipede crustacean *Xibalbanus tulumensis* [62].

Intriguingly, the ion transporter genes had the lowest $CpG_{o/e}$ values (Fig. 5), indicating complete and nearly complete depletion of CpG sites. This result suggests that the ion transporter genes have experienced extremely high levels of historical DNA methylation [52, 54]. DNA methylation of the gene body was found to be positively correlated with gene expression levels, in contrast to the suppression of gene expression by DNA methylation of gene promoter sequences [72-75]. Thus, these $CpG_{o/e}$ value ion transporter genes were likely highly expressed in the past.

Gene body methylation has been proposed to facilitate responses to environmental change and assist in acclimation by modulating gene expression [46, 76, 77]. In the *E. carolleeae* genome, the extremely low $CpG_{o/e}$ value distribution (Fig. 5a), indicating past genome-wide gene body methylation, suggests an environmental response of the low CpG genes (Fig. 5c, Additional file 2: Tables S13 and S14). The genomic signature of extremely low CpG values found in the ion transporter genes might be consistent with the critical roles these genes played during the evolutionary history of environmental fluctuations of this species complex [20, 30, 58, 78, 79] and perhaps of the genus *Eurytemora* [80].

Clustering of ion transporter genes on the four chromosomes

Previous results on the *E. affinis* complex have suggested that a set of cooperating ion transporters might undergo selection as and evolve together a unit, such that their rates of reaction would increase jointly to effectively increase rates of ion uptake [30, 34, 36, 37, 58]. In these prior studies, salinity change was accompanied by striking cases of parallel evolution, with selection acting on many of the same SNPs (single nucleotide polymorphisms) across multiple salinity gradients in wild populations and in replicate selection lines in the laboratory [30, 34, 37]. These shared targets of selection included paralogs of the ion transporters *NHA, NKA, VHA, CA, NKCC* and *Rh* [58]. Simulations of data from a laboratory evolution experiment suggest that positive epistasis among ion transporter alleles at different loci might serve as a mechanism to drive parallel selection on the same alleles in replicate selection lines [37].

We found that the ion transporter paralogs showed significant spatial clustering on the four chromosomes (Fig. 6). The distributions of these genes deviated significantly from a uniform distribution (Additional file 1: Fig. S7) and from distributions of functionally conserved genes (Additional file 1: Fig. S8). Such a clustering might facilitate the coexpression of functionally related genes or enable co-adapted alleles at different genes to be inherited together and undergo selection as a unit. The close physical linkage of beneficial alleles might be favored by selection due to reduced recombination [37, 81-83], which would break the alleles apart. Thus, such a genomic feature that maintains the clustering of beneficial alleles might serve as a contributing mechanism that facilitates rapid parallel adaptation.

However, the specific ion transporter paralogs that showed evolutionary shifts in gene expression or signatures of selection in our prior studies [30, 36, 37] were not necessarily clustered together in the genome. While many of the ion transporter paralogs under selection were localized on Chromosome #3, many others resided on the other three chromosomes (Additional file 1: Fig. S9). In particular, the ion transporter paralogs clustered near the centromere would tend to undergo low recombination and could more readily experience coordinated gene expression and/or selection as a unit of non-recombining alleles. However, the specific ion transporter paralogs that we found near any of the centromeres

(Chromosomes 1, 3, and 4; Additional file 1: Figs. S9 and S10) were not the ones that showed parallel evolution in the previous studies [58].

The significant clustering of ion transporter paralogs might be a by product of neutral processes, such as the recent expansions of ion transporter genes in the *E. carolleeae* genome (previous section; Additional file 1: Fig. S4) and the pattern of genomic rearrangements. We would need to conduct further studies to determine whether the clustering of ion transporter paralogs in the genome confers any selective benefits. While we lack evidence that the current genome-wide pattern of ion transporter gene clustering is adaptive, it is possible that the pattern of clustering could prove adaptive in other environmental contexts or in response to future environmental change.

Conclusions

The genome architecture of the calanoid copepod *E. carolleeae* appears poised to be particularly responsive to changes in habitat salinity. Characteristics of this genome, namely the substantial expansions of ion transporter genes, the extremely high signatures of past methylation of ion transporter genes, and the physical clustering of ion transporter genes might in part account for the extraordinary ability of populations of the *E. affinis* species complex to invade biogeographic boundaries into novel salinities [23, 84]. The genomic architecture described here might be relatively widespread among successful invaders crossing salinity boundaries. A large portion of the most prolific invasive species in freshwater lakes and reservoirs are immigrants from more saline waters, such as zebra mussels, quagga mussels, and many branchiopod and amphipod crustaceans [79, 84-86]. Moreover, the capacity to endure or evolve in response to salinity change is likely to become increasingly critical, as climate change is inducing drastic salinity changes throughout the globe, including rapid salinity declines in high-latitude coastal regions [87-89]. High quality genomic resources, such as the one generated by our study, will enhance our ability to gain novel insights into genomic mechanisms that enable rapid responses to environmental change and rapid invasions into novel habitats [90, 91].

Methods

Sampling and laboratory inbreeding of *E. carolleeae*

A population from the Atlantic clade (*E. carolleeae*) of the *E. affinis* species complex was originally collected in Baie de L'Isle Verte, St. Lawrence estuary, Quebec, Canada (48°00'14"N, 69°25'31"W) in October, 2008 [92]. To reduce heterozygosity of the wild population, inbred lines were generated through 30 generations (2.5 years) of full-sibling mating in the Lee laboratory of University of Wisconsin-Madison. The inbred lines were continuously reared and maintained in multiple 2L beakers in 15 Practical Salinity Unit (PSU) saline water (0.2 μm pore filtered) made with Instant Ocean, along with Primaxin (20 mg/L) to avoid bacterial infection. The copepods were fed with the marine alga *Rhodomonas salina* three time a week with water changed weekly. The inbred line VA-1 was used for this study.

Sequencing of the *E. carolleeae* genome

Approximately 3,000 adult copepods were initially collected for genome sequencing. To minimize contamination of the DNA extraction by gut contents and the microbiome, the copepods were treated with antibiotics (20 mg/L Primaxin, 0.5 mg/L Voriconazole) and D-amino acids (10 mM D-methionine, D-tryptophan, D-leucine, and 5 mM D-tyrosine) two weeks prior to DNA extraction with water changed twice a week. The copepods were treated with five additional antibiotics (20 mg/L Rifaximin, 40 mg/L Sitafloxacin, 20 mg/L Fosfomycin, 15 mg/L Metronidazole, 3 mg/L Daptomycin) for the last three days of treatment with the water changed daily. In the last 48h, the copepods were starved and fed with 90 μL/L 0.6-micron copolymer beads to remove the gut microbiome (Sigma-Aldrich, St. Louis, MO, USA).

The DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) was used for DNA extraction to obtain 48 μg of high molecular weight (HMW) genomic DNA, which was quantified by pulsed-field gel electrophoresis, Nanodrop spectrophotometry (Thermo Fisher, Wilmington, DE, USA) and Qubit 3.0 fluorometry (Thermo Fisher). The Pacific Biosciences (PacBio, Menlo Park, CA, USA) CLR library was constructed with 20 kb insert size using SMRTbell Template Prep Kit 1.0 (PacBio) following the manufacturer's protocol. The DNA library was sequenced on four PacBio Sequel SMRT Cells using the

PacBio Sequel II platform at Dovetail Genomics (Scotts Valley, CA, USA) to generate 2.6 million reads (30.3 Gb, ~60.6× coverage). To validate the assembly quality and complement the sequencing coverage, an additional 1,000 copepod individuals were collected. The CTAB-based phenol/chloroform/isoamylol DNA extraction was performed to obtain 16 μg HMW genomic DNA (Additional file 1: Online methods). A PacBio HiFi CCS library was constructed with 10–20 kb insert sizes and sequenced on a PacBio Sequel SMRT Cell 8M using the PacBio Sequel II platform at Novogene (Sacramento, CA, USA). A total of 0.59 million HiFi CCS reads (7.1 Gb, ~14.2× coverage) were generated by calling consensus from subreads generated by multiple passes of the enzyme around a circularized template. Another 0.5 μg DNA sample was used to construct a 350 bp insert size library and sequenced on the Illumina Hiseq NovoSeq 6000 platform (San Diego, CA, USA) at Novogene with 150 bp pair-end (PE) mode to generate 244.6 million reads (36.7 Gb, ~73.4× coverage).

Two Hi-C sequencing libraries were prepared following a previous protocol [93] at Dovetail Genomics. The chromatin of 500 copepods was fixed with 2% formaldehyde for cross-linking in the nucleus and extracted afterward. DNA was digested with MboI restriction endonuclease with non-ligated DNA fragments removed. The ligated DNA was sheared to ~350 bp followed by a standard Illumina library preparation protocol. The library was also sequenced on the Illumina Hiseq X Ten platform with 100 bp PE mode to generate 112 million 2×150 bp reads for library 1 and 59 million 2×150 bp reads for library 2 (for a total of 42.8 Gb, ~85.6× coverage).

Chromosome-level genome assembly of *E. carolleeae*

Genome size was estimated prior to genome assembly. Our previous Illumina genome sequencing data generated in the i5K Arthropod Genome Pilot Project [44, 94] and the newly generated Illumina sequencing data in the present study were both analyzed to estimate the genome size of *E. carolleeae*. Fastp v0.22.0 [95] was used to trim the raw sequencing reads with default parameters. Genome size was estimated based on the k-mer distribution using Jellyfish (count -m 21/25 -C -s 1G -F 2, histo -h

1,000,000). GenomeScope v2.0 [96] was used to estimate the genome size, heterozygosity, and proportion of repetitive sequence with k = 21 and 25.

The PacBio CLR data were first used solely to assemble the primary genome. The raw sequencing reads were self-corrected using NextDenovo v2.3 [97] (genome size = 500 m, seed_cutoff = 13k, read_cutoff = 1k, sort_options = -m 10g -t 2 -k 50, minimap2_options_raw = -t 8). The all-to-all alignment by minimap2 (-x ava-pb -t 8 -k17 -w17) and Nextgraph in NextDenovo (-a 1) were used to generate the primary genome assembly. NextPolish [98] was used to polish the genome assembly with both PacBio CLR reads and Illumina short reads. One round of long reads polishing and three rounds of short reads polishing (sgs_options = -max_depth 100) were performed successively to improve the assembly. To validate that the robustness of our assembly was not influenced by sequencing coverage, we combined the corrected CLR data and HiFi CCS reads and reassembled the primary genome with the same parameters using NextDenovo. The N50 statistic (defined as the sequence length of the shortest contig at 50% of the total assembly length) was used to evaluate the genome continuity of the primary assembly. The completeness of the genome assembly was assessed using BUSCO v5.2.2 at nucleotide level based on 1,013 genes in the insecta_odb10 database [99]. These two assemblies based on different datasets showed very similar quality with respect to continuity and completeness (shown in Additional file 2: Table S17). This assembly (#1) with higher contig N50 was further used in the following analyses (Additional file 2: Table S17). Purge_dups [100] was applied to remove heterozygous duplicates of the genome assembly.

For the chromosome scaffolding, Juicer [101] and 3D-DNA [102] were used to scaffold the genome assembly to the chromosome level. Juicebox v1.91 [103] was also used to manually correct the errors in scaffolding. We manually removed 11 scaffolds that were disconnected from the rest of the assembly. We identified and removed microbial sequences by searching the NT database by BLAST v2.8.1 [104].

Karyotype of the *E. carolleeae* genome

Cytogenetic analyses of the *E. carolleeae* genome was performed by the UW Cytogenetic Services in the Wisconsin State Laboratory of Hygiene (WSLH). Live copepod samples were used to isolate cells in metaphase. Cells were swollen in a hypotonic solution (0.075 M KCl) for 20 minutes at 37°C, and then fixed three times in fresh Carnoy's fixative. Cells were dropped onto slides and dried in a drying chamber. Slides were banded by GTG banding technique and scanned to find cells with well isolated chromosomes.

Genome size and chromosome number evolution across the Copepoda

To gain comparative insights into patterns of genome size and chromosome number evolution across the Copepoda, we summarized available and published data for four copepod orders. These data integrated information from both genome assemblies present in NCBI Genome database [42] and from published cytophotometric and karyological investigations (Additional file 2: Tables S2 and S3). We also retrieved records for the Copepoda from the Animal Genome Size Database [105]. The chromosome numbers were mapped onto a synthesis tree of the Copepoda that integrated 31 published phylogenies [50]. We performed statistical comparisons of the chromosome numbers and genome sizes for four copepod orders with Kruskal-Wallis and pairwise Wilcoxon tests performed in R [106].

Genome annotation of *E. carolleeae*

RepeatMasker v4.07 [107] was used to identify repetitive sequences and transposable elements in the genome based on searching in Repbase v202101 [108], Dfam v3.7 [109], a *de novo* repeat library built by RepeatModeler v1.0.8 [110], the integrated tools RECON [111], TRF v4.09 [112], and RepeatScout [113]. Long terminal repeat (LTR) searches were also performed with dependent LtrHarvest [114], CD-HIT [115], and Ltr_retriever [116] installed. We applied the MAKER v3.01 [117] pipeline to annotate protein-coding regions of our genome. Gene structure prediction was integrated using three strategies, i.e., homology-based, transcriptome-based, and *ab initio* prediction. For homology evidence, the protein sequences of *Drosophila melanogaster*, *Daphnia pulex*, *Tigriopus californicus*, *Lepeophtheirus salmonis* and *E. affinis* in NCBI Reference Sequence (RefSeq) database were fed into MAKER. For transcriptomic evidence, we used a total of 52 transcriptome data sets, including 46 of which were sequenced in our previous gene expression study under various salinity treatments [36], three of which sequenced in our previous i5K genome sequencing project [44, 94], and two of which were sequenced in the present study using samples from two other species in the *E. affinis* species complex (clades of Europe [*E. affinis* proper (Poppe, 1880)] [118] and Gulf of Mexico, Additional file 1: Online methods). These transcriptomic data sets were collected and reassembled based on our new reference genome, using HISAT v2.0.4 [119] and StringTie v2.2.1 [120]. Regarding *ab initio* gene prediction, we trained the gene predictor SNAP [121] with the gene models predicted with the above evidence. The self-trained predictor GeneMark-ES [122] was applied separately. Within MAKER, the genome was masked for repetitive regions, and protein homology and transcript sequences were aligned using BLAST. Three iterative runs of MAKER were performed, with gene predictions from each run serving as training sets for the following run. Finally, MAKER evaluated the consistency across these different forms of evidence and generated a final set of gene models.

Functional annotation of gene models was performed by BLASTP searches of the NCBI RefSeq and UniProtKB/Swiss-Prot [123] databases of invertebrates, and a separate self-established database with all gene sequences of *E. affinis* in RefSeq. GO [124], KEGG [125], COG, and eggNOG [126] databases were searched using eggNOG-mapper v2.1.9 [127]. The Pfam database in InterPro [128] was also searched by HMMER v3.2 [129].

To detect the relative ages of gene duplicates and evidence for ancient whole genome duplication (WGD), Ks frequency analysis was performed using the DupPipe pipeline [130]. All protein-coding genes were translated to identify reading frames by comparing the Genewise alignment to the best hit protein from the same homology protein sequences used in the genome annotation. Synonymous divergence (Ks) was estimated using PAML with the F3 × 4 model [131].

Transfer RNAs (tRNAs) were defined using tRNAscan-SE v2.0 [132] with default parameters. MicroRNA and small nuclear RNA were identified with BLASTN against the Rfam database v12.0 [133] and ribosomal RNA (rRNA) was identified against other copepod rRNA sequences.

Gene family expansions and contractions across the Arthropoda

Orthologous gene families in the *E. carolleeae* genome were identified by OrthoFinder v2.5.4 [134]. Protein sequences of 12 additional arthropod species with high-quality genomes, assembled with long-read sequences to the chromosome level, were downloaded from the GenBank database (Additional file 2: Table S18). These arthropod genomes included three chelicerates (*Hyalomma asiaticum*, *Hylyphantes graminicola*), one barnacle (Thecostraca: *Pollicipes pollicipes*), three copepods (*Caligus rogercressey*, *Lepeophtheirus salmonis*, *Tigriopus californicus*), four branchiopods (*Daphinia pulex*, *D. magna*, *D. pulicaria*, *D. sinensis*) and two hexapods (Insecta: *Drosophila melanogaster*, *Aphis gossypii*). We first filtered out alternative splice variants for each gene and only kept the longest transcript. We aligned proteins of our copepod and other arthropod species using BLASTP (e-value < 1e-5). Protein sequences of the identified single-copy genes were aligned by MAFFT v7.313 with the L-INS-i algorithm [135]. Gblocks v0.91b [136] was used to trim the alignment. A phylogeny was reconstructed using a Maximum Likelihood algorithm in RAxML v8.0.19 [137]. 100 bootstrap replicates were performed to assess statistical support for tree topology. We used MCMCTree from PAML v4.9 to estimate divergence times [131]. Three confidence time intervals retrieved from the TIMETREE v5 database [138] were applied in MCMCTree as calibrations for the divergence time (shown as red circles in Fig. 3). CAFÉ5 [139] was used to analyze the expansion and contraction of gene families among taxon in the phylogenetic tree. For gene families exhibiting expansion and contraction in the genome, GO and KEGG enrichment analyses were performed using TBtools v1.112 [140].

Syntenic relationships among three copepod species was analyzed using MCScan in JCVI [141]. We used the highest quality copepod genomes of *E. carolleeae*, *Tigriopus californicus*, and *Lepeophtheirus salmonis*, representing three different copepod orders, Calanoida, Harpacticoida, and Siphonostomatoida, respectively. Collinear gene blocks within the genome were identified using the longest coding sequence of each gene.

Genome-wide CpG$_{o/e}$ values in the *E. carolleeae* genome

To assess the patterns of historical methylation within gene bodies, genome-wide CpG$_{o/e}$ values were determined in the *E. carolleeae* genome. The CpG$_{o/e}$ value of each gene was computed as the observed frequency of CpG sites (f_{CpG}) divided by the product of C and G frequencies (f_C and f_G), i.e., f_{CpG}/f_C*f_G in the coding sequence (CDS) of each gene. The density of CpG$_{o/e}$ values for all genes was fitted and plotted in R. The distribution of CpG$_{o/e}$ values per gene was also plotted based on the order of gene locations on different chromosomes. To investigate the functional categories of the highest and lowest CpG$_{o/e}$ genes, we performed GO enrichments for the top 5% genes with the highest and lowest CpG$_{o/e}$ values using TBtools.

Localization of ion transporter genes across the *E. carolleeae* genome

A total of 490 genes with ion (cation and anion) transporting function were mapped onto the four chromosomes based on our genome annotation (shown as vertical light blue lines in Fig. 6a). In addition, 83 paralogs of key ion transporter genes that showed evolutionary shifts in gene expression and/or signatures of selection in prior studies [58] were manually annotated and separately mapped onto the chromosomes (shown as vertical lines and circles in other colors in Fig. 6a). These ion transporters are likely involved in hypothetical models of ion uptake (Fig. 6c). These include *Na^+/K^+ ATPase* α subunit (*NKA-α*), *Na^+/K^+ ATPase β* subunit (*NKA-β*), *Na^+/H^+ exchanger* (*NHE*), *Na^+/H^+ antiporter* (*NHA*), *$Na^+/K^+/Cl^-$ cotransporter* (*NKCC*), *Carbonic anhydrase* (*CA*), *Ammonia transporter* (*AMT*), *Rh protein* (*Rh*), *Vacuolar-type ATPase* (*VHA*), and Solute carrier family 4 of bicarbonate (HCO_3^-) transporters (*SLC4*) members, including *Anion exchanger* (*AE*), *$Na^+/HCO3^-$ cotransporter* (*NBC*), and *Na^+-driven Cl^- /HCO_3^- exchanger* (*NDCBE*) (Figs. 6b, c).

Distances between adjacent ion transporter genes were calculated and deviation of the distribution of these gene distances from a uniform distribution was tested using the Kolmogorov-Smirnov test in R. In addition, deviation of the distribution of these ion transporter genes from the distributions of the same number of functionally conserved genes was tested using the Chi-square goodness of fit test in R. For the functionally conserved genes, genes with the highest CpG$_{o/e}$ values identified in the prior section (Genome-wide CpG$_{o/e}$ values in the *E. carolleeae* genome) were used (Additional file 2: Table S16). This set of genes was enriched in RNA processing and DNA binding related functions, which tend to be functionally conserved housekeeping genes.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All sequencing reads generated in this study have been deposited in the NCBI Sequence Read Archive (SRA) database (PacBio CLR reads: XXX; PacBio CCS reads: XXX; Illumina reads: XXX; Hi-C reads: XXX; Transcriptome: XXX). The genome assembly was deposited in the i5k Workspace of the National Agricultural Library (US Department of Agriculture): https://i5k.nal.usda.gov/. The genome annotation files were deposited in XXX. Our previous whole genome sequencing data (NCBI Bioproject accession: PRJNA203087) from i5K Arthropod Genome Pilot Project (NCBI Bioproject accession: PRJNA163973) were downloaded and reanalyzed for genome size estimation. Our previous 49 transcriptome data (NCBI Bioproject accessions: PRJNA278152 and PRJNA275666) were downloaded and reanalyzed for genome annotation.

Competing interests

The authors declare that they have no competing interests.

Funding

This project was funded by the National Science Foundation grants OCE-1658517 and NSF DEB-2055356, and French National Research Agency ANR-19-MPGA-0004 to CEL (Macron's "Make Our Planet Great Again" award).

Authors' contributions

CEL designed and supervised this study. GWG, WM, and AT generated the inbred lines of copepods through 30 generations of full-sib mating. ZD performed the molecular experiments, data analyses and graphical illustrations. ZD and CEL interpreted the data and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank David Stern, Jesse Hunter, and Kim Oxendine for performing the karyotyping experiments.

Supplementary Information

Additional file 1: Supplementary Figures S1–S11 and online methods

Additional file 2: Supplementary Tables S1–S18.

References

1. Humes AG. How many copepods? In *Ecology and Morphology of Copepods: Proceedings of the 5th International Conference on Copepoda, Baltimore, USA, June 6–13, 1993*. Springer; 1994: 1-7.
2. Hardy A. *The Open Sea. The World of Plankton.* London: Collins; 1970.
3. Verity PG, Smetacek V. Organism life cycles, predation, and the structure of marine pelagic ecosystems. *Mar Ecol Prog Ser.* 1996;130:277-93.
4. Winkler G, Sirois P, Johnson LE, Dodson JJ. Invasion of an estuarine transition zone by *Dreissena polymorpha* veligers had no detectable effect on zooplankton community structure. *Can J Fish Aquat Sci.* 2005;62:578-92.
5. Heinle D, Flemer D. Carbon requirements of a population of the estuarine copepod *Eurytemora affinis*. *Mar Biol.* 1975;31:235-47.
6. Morgan CA, Cordell JR, Simenstad CA. Sink or swim? Copepod population maintenance in the Columbia River estuarine turbidity-maxima region. *Mar Biol.* 1997;129:309-17.
7. Peitsch A, Köpcke B, Bernát N. Long-term investigation of the distribution of *Eurytemora affinis* (Calanoida; Copepoda) in the Elbe Estuary. *Limnologica* 2000;30:175-82.
8. Gulati RD, Doornekamp A. The spring-time abundance and feeding of *Eurytemora affinis* (Poppe) in Volkerak-Zoommeer, a newly-created freshwater lake system in the Rhine delta (The Netherlands). *Hydrobiol Bull.* 1991;25:51-60.
9. Simenstad CA, Cordell JR. Structural dynamics of epibenthic zooplankton in the Columbia River delta. *SIL Proc 1922-2010* 2017;22:2173-82.
10. Shaheen PA, Stehlik LL, Meise CJ, Stoner AW, Manderson JP, Adams DL. Feeding behavior of newly settled winter flounder (*Pseudopleuronectes americanus*) on calanoid copepods. *J Exp Mar Biol Ecol.* 2001;257:37-51.

11. Viitasalo M, Flinkman J, Viherluoto M. Zooplanktivory in the Baltic Sea: a comparison of prey selectivity by *Clupea harengus* and *Mysis mixta*, with reference to prey escape reactions. *Mar Ecol Prog Ser.* 2001;216:191-200.

12. Winkler G, Dodson JJ, Bertrand N, Thivierge D, Vincent WF. Trophic coupling across the St. Lawrence River estuarine transition zone. *Mar Ecol Prog Ser.* 2003;251:59-73.

13. Kimmel DG, Miller WD, Roman MR. Regional scale climate forcing of mesozooplankton dynamics in Chesapeake Bay. *Estuar Coast.* 2006;29:375-87.

14. Livdāne L, Putnis I, Rubene G, Elferts D, Ikauniece A. Baltic herring prey selectively on older copepodites of *Eurytemora affinis* and *Limnocalanus macrurus* in the Gulf of Riga. *Oceanologia* 2016;58:46-53.

15. Simenstad CA, Small LF, Mcintire CD. Consumption processes and food web structure in the Columbia River estuary. *Prog Oceanogr.* 1990;25:271-97.

16. Viitasalo M, Vuorinen I, Saesmaa S. Mesozooplankton dynamics in the northern Baltic Sea: implications of variations in hydrography and climate. *J Plankton Res.* 1995;17:1857-78.

17. Viitasalo M, Katajisto T, Vuorinen I. Seasonal dynamics of *Acartia bifilosa* and *Eurytemora affinis* (Copepods: Calanoida) in relation to abiotic factors in the northern Baltic Sea. *Hydrobiologia* 1994;292-293:415-22.

18. Lee CE, Frost BW. Morphological stasis in the *Eurytemora affinis* species complex (Copepoda : Temoridae). *Hydrobiologia* 2002;480:111-28.

19. Lee CE. Global phylogeography of a cryptic copepod species complex and reproductive isolation between genetically proximate "populations". *Evolution* 2000;54:2014-27.

20. Lee CE. Evolutionary mechanisms of habitat invasions, using the copepod *Eurytemora affinis* as a model system. *Evol Appl.* 2016;9:248-70.

21. Alekseev VR, Souissi A. A new species within the *Eurytemora affinis* complex (Copepoda: Calanoida) from the Atlantic Coast of USA, with observations on eight morphologically different European populations. *Zootaxa* 2011;2767:41-56.

22. Sukhikh N, Souissi A, Souissi S, Winkler G, Castric V, Holl AC, Alekseev V. Genetic and morphological heterogeneity among populations of *Eurytemora affinis* (Crustacea: Copepoda: Temoridae) in European waters. *C R Biol.* 2016;339:197-206.

23. Lee CE. Rapid and repeated invasions of fresh water by the copepod *Eurytemora affinis*. *Evolution* 1999;53:1423-34.

24. Lee CE, Charmantier G, Lorin-Nebel C. Mechanisms of Na^+ uptake from freshwater habitats in animals. *Front Physiol.* 2022;13:1006113.

25. Lee CE, Remfert JL, Chang YM. Response to selection and evolvability of invasive populations. *Genetica* 2007;129:179-92.

26. Lee CE, Remfert JL, Gelembiuk GW. Evolution of physiological tolerance and performance during freshwater invasions. *Integr Comp Biol.* 2003;43:439-49.

27. Bradley BP. The anomalous influence of salinity on temperature tolerances of summer and winter populations of the copepod *Eurytemora affinis*. *Biol Bull.* 1975;148:26-34.

28. Devreker D, Souissi S, Winkler G, Forget-Leray J, Leboulenger F. Effects of salinity, temperature and individual variability on the reproduction of *Eurytemora affinis* (Copepoda; Calanoida) from the Seine estuary: A laboratory study. *J Exp Mar Biol Ecol.* 2009;368:113-23.

29. Gyllenberg G, Lundqvist G. The effects of temperature and salinity on the oxygen consumption of *Eurytemora hirundoides* (Crustacea, Copepoda). *Ann Zool Fenn.* 1979;16:205-8.

30. Stern DB, Lee CE. Evolutionary origins of genomic adaptations in an invasive copepod. *Nat Ecol Evol.* 2020;4:1084-94.

31. Mills EL, Leach JH, Carlton JT, Secor CL. Exotic species in the Great Lakes: a history of biotic crises and anthropogenic introductions. *J Great Lakes Res.* 1993;19:1-54.

32. Saunders JF. Distribution of *Eurytemora affinis* (Copepoda, Calanoida) in the southern Great Plains, with notes on Zoogeography. *J Crust Biol.* 1993;13:564-70.

33. De Beaufort LF. *Veranderingen in de Flora en Fauna van de Zuiderzee (thans IJsselmeer) na de Afsluiting in 1932.* Netherlands: C. de Boer Jr; 1954.

34. Diaz J, Stern D, Lee CE. Local adaptation despite gene flow in copepod populations across salinity and temperature gradients in the Baltic and North Seas. *Authorea* 2023; doi:10.22541/au.168311545.58858033/v1.

35. Lee CE, Kiergaard M, Gelembiuk GW, Eads BD, Posavi M. Pumping ions: rapid parallel evolution of ionic regulation following habitat invasions. *Evolution* 2011;65:2229-44.

36. Posavi M, Gulisija D, Munro JB, Silva JC, Lee CE. Rapid evolution of genome-wide gene expression and plasticity during saline to freshwater invasions by the copepod *Eurytemora affinis* species complex. *Mol Ecol.* 2020;29:4835-56.

37. Stern DB, Anderson NW, Diaz JA, Lee CE. Genome-wide signatures of synergistic epistasis during parallel adaptation in a Baltic Sea copepod. *Nat Commun.* 2022;13:4024.

38. Rotenberg D, Baumann AA, Ben-Mahmoud S, Christiaens O, Dermauw W, Ioannidis P, Jacobs CGC, Vargas Jentzsch IM, Oliver JE, Poelchau MF, et al. Genome-enabled insights into the biology of thrips as crop pests. *BMC Biol.* 2020;18:142.

39. Luo S, Tang M, Frandsen PB, Stewart RJ, Zhou X. The genome of an underwater architect, the caddisfly *Stenopsyche tienmushanensis* Hwang (Insecta: Trichoptera). *Gigascience* 2018;7:giy143.

40. Yuan JB, Yu Y, Zhang XJ, Li SH, Xiang JH, Li FH. Recent advances in crustacean genomics and their potential application in aquaculture. *Rev Aquac.* 2023. doi:10.1111/raq.12791.

41. Stillman JH, Colbourne JK, Lee CE, Patel NH, Phillips MR, Towle DW, Eads BD, Gelembuik GW, Henry RP, Johnson EA, et al. Recent advances in crustacean genomics. *Integr Comp Biol.* 2008;48:852-68.

42. Genome Database. NCBI. https://www.ncbi.nlm.nih.gov/genome. Accessed 1 April 2023.

43. Kang S, Ahn DH, Lee JH, Lee SG, Shin SC, Lee J, Min GS, Lee H, Kim HW, Kim S, Park H. The genome of the Antarctic-endemic copepod, *Tigriopus kingsejongensis*. *Gigascience* 2017;6:1-9.

44. Eyun SI, Soh HY, Posavi M, Munro JB, Hughes DST, Murali SC, Qu J, Dugan S, Lee SL, Chao H, et al. Evolutionary history of chemosensory-related gene families across the Arthropoda. *Mol Biol Evol.* 2017;34:1838-62.

45. Shao C, Sun S, Liu K, Wang J, Li S, Liu Q, Deagle BE, Seim I, Biscontin A, Wang Q, et al. The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights. *Cell* 2023;186:1279-94.e19.

46. Lee YH, Kim MS, Wang MH, Bhandari RK, Park HG, Wu RSS, Lee JS. Epigenetic plasticity enables copepods to cope with ocean acidification. *Nat Clim Change* 2022;12:918-27.

47. Joshi J, Flores AM, Christensen KA, Johnson H, Siah A, Koop BF. An update of the salmon louse (*Lepeophtheirus salmonis*) reference genome assembly. *G3* 2022;12:jkac087.

48. Simakov O, Marletaz F, Yue JX, O'Connell B, Jenkins J, Brandt A, Calef R, Tung CH, Huang TK, Schmutz J, et al. Deeply conserved synteny resolves early events in vertebrate evolution. *Nat Ecol Evol.* 2020;4:820-30.

49. Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P, Rastas P, Valimaki N, Paulin L, Kvist J, Wahlberg N, et al. The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nat Commun.* 2014;5:4737.

50. Bernot JP, Boxshall GA, Crandall KA. A synthesis tree of the Copepoda: integrating phylogenetic and taxonomic data reveals multiple origins of parasitism. *PeerJ.* 2021;9:e12034.

51. Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ, Barker MS. Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc Natl Acad Sci U S A.* 2018;115:4713-8.

52. Bird AP. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* 1980;8:1499-504.

53. Mattei AL, Bailly N, Meissner A. DNA methylation: a historical perspective. *Trends Genet.* 2022;38:676-707.

54. Ylla G, Nakamura T, Itoh T, Kajitani R, Toyoda A, Tomonari S, Bando T, Ishimaru Y, Watanabe T, Fuketa M, et al. Insights into the genomic evolution of insects from cricket genomes. *Commun Biol.* 2021;4:733.

55. Aliaga B, Bulla I, Mouahid G, Duval D, Grunau C. Universality of the DNA methylation codes in Eucaryotes. *Sci Rep.* 2019;9:173.

56. Manner L, Schell T, Provataris P, Haase M, Greve C. Inference of DNA methylation patterns in molluscs. *Philos Trans R Soc Lond B Biol Sci.* 2021;376:20200166.

57. Elango N, Hunt BG, Goodisman MA, Yi SV. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera. *Proc Natl Acad Sci U S A.* 2009;106:11206-11.

58. Lee CE. Ion transporter gene families as physiological targets of natural selection during salinity transitions in a copepod. *Physiology* 2021;36:335-49.

59. Lynch M, Conery JS. The origins of genome complexity. *Science* 2003;302:1401-4.

60. Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res.* 2009;19:1195-201.

61. Rasch EM, Lee CE, Wyngaard GA. DNA-Feulgen cytophotometric determination of genome size for the freshwater-invading copepod *Eurytemora affinis*. *Genome* 2004;47:559-64.

62. Provataris P, Meusemann K, Niehuis O, Grath S, Misof B. Signatures of DNA methylation across insects suggest reduced DNA methylation levels in Holometabola. *Genome Biol Evol.* 2018;10:1185-97.

63. Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. Polyploidy and genome evolution in plants. *Curr Opin Genet Dev.* 2015;35:119-25.

64. Sylvester T, Hjelmen CE, Hanrahan SJ, Lenhart PA, Johnston JS, Blackmon H. Lineage-specific patterns of chromosome evolution are the rule not the exception in Polyneoptera insects. *Proc Biol Sci.* 2020;287:20201388.

65. Mackintosh A, Vila R, Laetsch DR, Hayward A, Martin SH, Lohse K. Chromosome fissions and fusions act as barriers to gene flow between *Brenthis* fritillary butterflies. *Mol Biol Evol.* 2023;40:msad043.

66. Rieseberg LH. Chromosomal rearrangements and speciation. *Trends Ecol Evol.* 2001;16:351-8.

67. Grishanin A. Chromatin diminution in Copepoda (Crustacea): pattern, biological role and evolutionary aspects. *Comp Cytogenet.* 2014;8:1-10.

68. Drotos KHI, Zagoskin MV, Kess T, Gregory TR, Wyngaard GA. Throwing away DNA: programmed downsizing in somatic nuclei. *Trends Genet.* 2022;38:483-500.

69. Wyngaard GA, Rasch EM. Patterns of genome size in the copepoda. *Hydrobiologia* 2000;417:43-56.

70. Beermann S. The diminution of heterochromatic chromosomal segments in Cyclops (Crustacea, Copepoda). *Chromosoma* 1977;60:297-344.

71. Sun C, Wyngaard G, Walton DB, Wichman HA, Mueller RL. Billions of basepairs of recently expanded, repetitive sequences are eliminated from the somatic genome during copepod development. *BMC Genomics* 2014;15:186.

72. Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* 2014;26:577-90.

73. Jjingo D, Conley AB, Yi SV, Lunyak VV, Jordan IK. On the presence and role of human gene-body DNA methylation. *Oncotarget* 2012;3:462-74.

74. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 2010;466:253-7.

75. Wang Q, Xiong F, Wu G, Liu W, Chen J, Wang B, Chen Y. Gene body methylation in cancer: molecular mechanisms and clinical applications. *Clin Epigenetics* 2022;14:154.

76. Dixon G, Liao Y, Bay LK, Matz MV. Role of gene body methylation in acclimatization and adaptation in a basal metazoan. *Proc Natl Acad Sci U S A.* 2018;115:13342-6.

77. Kvist J, Goncalves Athanasio C, Shams Solari O, Brown JB, Colbourne JK, Pfrender ME, Mirbahai L. Pattern of DNA methylation in *Daphnia*: Evolutionary perspective. *Genome Biol Evol.* 2018;10:1988-2007.

78. Posavi M, Gelembiuk GW, Larget B, Lee CE. Testing for beneficial reversal of dominance during salinity shifts in the invasive copepod *Eurytemora affinis*, and implications for the maintenance of genetic variation. *Evolution* 2014;68:3166-83.

79. Lee CE, Gelembiuk GW. Evolutionary origins of invasive populations. *Evol Appl.* 2008;1:427-48.

80. Dodson SI, Skelly DA, Lee CE. Out of Alaska: morphological diversity within the genus *Eurytemora* from its ancestral Alaskan range (Crustacea, Copepoda). *Hydrobiologia* 2010;653:131-48.

81. Via S. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philos Trans R Soc Lond B Biol Sci.* 2012;367:451-60.

82. Feder JL, Gejji R, Yeaman S, Nosil P. Establishment of new mutations under divergence and genome hitchhiking. *Philos Trans R Soc Lond B Biol Sci.* 2012;367:461-74.

83. Yeaman S. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc Natl Acad Sci U S A.* 2013;110:E1743-51.

84. Lee CE, Bell MA. Causes and consequences of recent freshwater invasions by saltwater animals. *Trends Ecol Evol.* 1999;14:284-8.

85. Havel JE, Lee CE, Vander Zanden JM. Do reservoirs facilitate invasions into landscapes? *Bioscience* 2005;55:518-25.

86. Casties I, Seebens H, Briski E. Importance of geographic origin for invasion success: A case study of the North and Baltic Seas versus the Great Lakes-St. Lawrence River region. *Ecol Evol.* 2016;6:8318-29.

87. Thomas CD, Cameron A, Green RE, Bakkenes M, Beaumont LJ, Collingham YC, Erasmus BF, De Siqueira MF, Grainger A, Hannah L, et al. Extinction risk from climate change. *Nature* 2004;427:145-8.

88. Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, Raven PH, Roberts CM, Sexton JO. The biodiversity of species and their rates of extinction, distribution, and protection. *Science* 2014;344:1246752.

89. Durack PJ, Wijffels SE, Matear RJ. Ocean salinities reveal strong global water cycle intensification during 1950 to 2000. *Science* 2012;336:455-8.

90. Lee CE. Evolutionary genetics of invasive species. *Trends Ecol Evol.* 2002;17:386-91.

91. Lee CE. Evolution of invasive populations. In *Encyclopedia of Biological Invasions.* Edited by Simberloff D, Rejmanek M. Bekerley, CA: University of California Press; 2010.

92. Winkler G, Dodson JJ, Lee CE. Heterogeneity within the native range: population genetic analyses of sympatric invasive and noninvasive clades of the freshwater invading copepod *Eurytemora affinis*. *Mol Ecol.* 2008;17:415-30.

93. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;326:289-93.

94. i KC. The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered.* 2013;104:595-600.

95. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;34:i884-i90.

96. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 2017;33:2202-4.

97. Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, Li J, Sandoval JR, Cooper DN, Ye K, et al. An efficient error correction and accurate assembly tool for noisy long reads. *bioRxiv* doi:10.1101/2023.03.09.531669.

98. Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* 2020;36:2253-5.

99. Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol.* 2021;38:4647-54.

100. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* 2020;36:2896-8.

101. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* 2016;3:95-8.

102. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* 2017;356:92-5.

103. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* 2016;3:99-101.

104. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.

105. Animal Genome Size Database. http://www.genomesize.com. Accessed 10 January 2023.

106. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing.* 2016; https://www.R-project.org/.

107. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics.* 2004;Chapter 4:4.10.1-14.

108. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 2015;6:11.

109. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* 2021;12:2.

110. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A.* 2020;117:9451-7.

111. Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* 2002;12:1269-76.

112. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 1999;27:573-80.

113. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. *Bioinformatics* 2005;21 Suppl 1:i351-8.

114. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* 2008;9:18.

115. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;22:1658-9.

116. Ou S, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 2018;176:1410-22.

117. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 2011;12:491.

118. Poppe SA. Über eine neue Art der Calaniden-Gattung *Temora*, Baird. Abhandlungen des Naturwissenschaftlichen Vereins Zu Bremen. 1880;7:55-60.

119. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;12:357-60.

120. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* 2019;20:278.

121. Korf I. Gene finding in novel genomes. *BMC Bioinformatics* 2004;5:59.

122. Borodovsky M, Lomsadze A. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Curr Protoc Bioinformatics* 2011;Chapter 4:4.6.1-10.

123. UniProt C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 2019;47:D506-D15.

124. Gene Ontology C, Blake JA, Dolan M, Drabkin H, Hill DP, Li N, Sitnikov D, Bridges S, Burgess S, Buza T, et al. Gene Ontology annotations and resources. *Nucleic Acids Res.* 2013;41:D530-5.

125. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28:27-30.

126. Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019;47:D309-D14.

127. Cantalapiedra CP, Hernandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol.* 2021;38:5825-9.

128. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L, et al. InterPro in 2022. *Nucleic Acids Res.* 2023;51:D418-D27.

129. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 2013;41:e121.

130. Barker MS, Dlugosch KM, Dinh L, Challa RS, Kane NC, King MG, Rieseberg LH. EvoPipes. net: bioinformatic tools for ecological and evolutionary genomics. *Evol Bioinform Online* 2010;6:143-9.

131. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 2007;24:1586-91.

132. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* 2021;49:9077-96.

133. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* 2005;33:D121-4.

134. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019;20:238.

135. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30:772-80.

136. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol.* 2007;56:564-77.

137. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312-3.

138. Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. TimeTree 5: An expanded resource for species divergence times. *Mol Biol Evol.* 2022;39.

139. Mendes FK, Vanderpool D, Fulton B, Hahn MW. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* 2021;36:5516-8.

140. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R. TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol Plant* 2020;13:1194-202.

141. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. Synteny and collinearity in plant genomes. *Science* 2008;320:486-8.

Figure legends

Fig. 1. Chromosome-level genome assembly of the copepod *Eurytemora carolleeae (E. affinis* complex, Atlantic clade). (a) Circular diagram showing the genome landscape. I. Four chromosomes on the Mb scale. II. Density of protein-coding genes. III. Distribution of GC content (Mean GC = 32.5%). IV. Distribution of repetitive sequences. V. Distribution of LTR. All distributions were calculated in 100 kb non-overlapping sliding windows. **(b)** The proportion of repetitive sequences identified in the copepod genome. The circular diagram shows their relative proportions out of the total repetitive sequences (46.12% of the genome), and the numbers labelled on the diagram represent their percentage of occupied length in the genome assembly. **(c)** Well-isolated cell that shows the karyotype of the copepod (2n = 8) at metaphase. (**d)** The Hi-C contact map of the genome generated by Juicebox.

Fig. 2. Chromosome number and genome size evolution in the crustacean class Copepoda. (a) Phylogeny of copepod species from five copepod orders. The phylogenetic topology was obtained from the synthesis tree of copepods, which integrated 31 published phylogenies [50]. Chromosome numbers are shown within parentheses after the species names. Different colors of species names represent the ranges of chromosome numbers. Clades that occupy basal phylogenetic positions, but possess unknown karyotype, are shown in grey in the phylogeny. **(b)** Mean chromosome number of four copepod orders (see Additional file 2: Table S2 for details). Chromosome number differs significantly among the four orders (Kruskal-Wallis test, H = 35.52, DF = 3, P = 9.5e-8). **(c)** Mean genome size of four copepod orders. Calanoida mean genome size = 3993 Mb, Siphonostomatoida = 563 Mb, Harpacticoida = 315 Mb, and Cyclopoida = 667 Mb (see Additional file 2: Table S3 for details). Genome size differs significantly among the four orders (Kruskal-Wallis test, H = 49.58, DF = 3, P = 9.8e-11). Asterisks in (b–c) indicate the significance levels for Wilcoxon tests, where * refers to $P < 0.05$ and **** indicates $P <$ 1e-4. Nonsignificant P-values are not shown.

Fig. 3. Gene family expansions and contractions during the evolutionary history of the Arthropoda, with a focus on the Copepoda. Phylogenetic reconstruction of 13 high-quality arthropod genomes was performed using RAxML based on concatenated single copy ortholog genes. All nodes show bootstrap values of 100%, except for two nodes with green rectangles, which have values of 66% (left node) and 60% (right node). Red circles represent three calibrated nodes with confidence time intervals retrieved from the Timetree database and applied in MCMCTree. Mean estimated divergence times are shown at each node with brackets indicating 95% highest posterior densities. The divergence times are on a scale of millions of years ago (Mya). The numbers of expanded gene families (in blue) and contracted gene families (in red) are shown on the branch tips and next to each node.

Fig. 4. Significantly enriched of gene ontology (GO) terms in the expanded set of genes in the *Eurytemora carolleeae* genome. The GO terms were sorted by *P*-value (with higher *P*-value toward the right in each category). The complete list of enriched GO terms is shown in Additional file 2: Table S11. Only the top 20 GO terms of the Biological Process and Molecular Function categories, and top 15 GO terms of Cellular Component category are shown here.

Fig. 5. Patterns of genome-wide $\mathrm{CpG_{o/e}}$ values of gene bodies, corresponding to signatures of past gene methylation in the *E. carolleeae* genome. (a) The $\mathrm{CpG_{o/e}}$ values of the protein-coding gene sequences display a unimodal distribution. **(b)** The distribution of $\mathrm{CpG_{o/e}}$ values across the genome when the genes are arranged by their position on each chromosome. **(c)** GO enrichment of the 1013 genes with 5% lowest $\mathrm{CpG_{o/e}}$ values. The significance of GO enrichment is shown by the color of the circles and the enriched gene number is indicated by the size of the circles. The ion transporter genes tend to have the lowest $\mathrm{CpG_{o/e}}$ values, suggesting extremely high levels of methylation in the past [52].

Fig. 6. Localization of ion transporter genes on *E. carolleeae* chromosomes and hypothetical models of ion uptake from fresh water. (a) Ion transporter genes mapped onto the four *E. carolleeae*

chromosomes. The vertical light blue lines represent 490 genes with ion (cation and anion) transporting function based on the genome annotation. The vertical lines and circles in other colors represent 83 key genes that showed evolutionary shifts in gene expression and/or signatures of selection in prior studies and are likely involved in hypothetical models of ion uptake. The dashed lines marked with stars indicate the positions of centromeres based on the Hi-C contact map (Fig. 1d, Additional file 1: Fig. S11). **(b, c)** Hypothetical models of ion uptake from freshwater environments. **(b)** Model 1: VHA generates an electrochemical gradient by pumping out protons, to facilitate uptake of Na^+ through an electrogenic Na+ transporter (likely NHA). CA produces protons for VHA. **(c)** Model 2: An ammonia transporter Rh protein exports NH_3 out of the cell and this NH_3 reacts with H^+ to form NH_4^+. The deficit of extracellular H^+ concentrations cause NHE to export H^+ in exchange for Na^+. CA produces protons for NHE. These models are not comprehensive for all tissues or taxa and are not mutually exclusive.

Figures

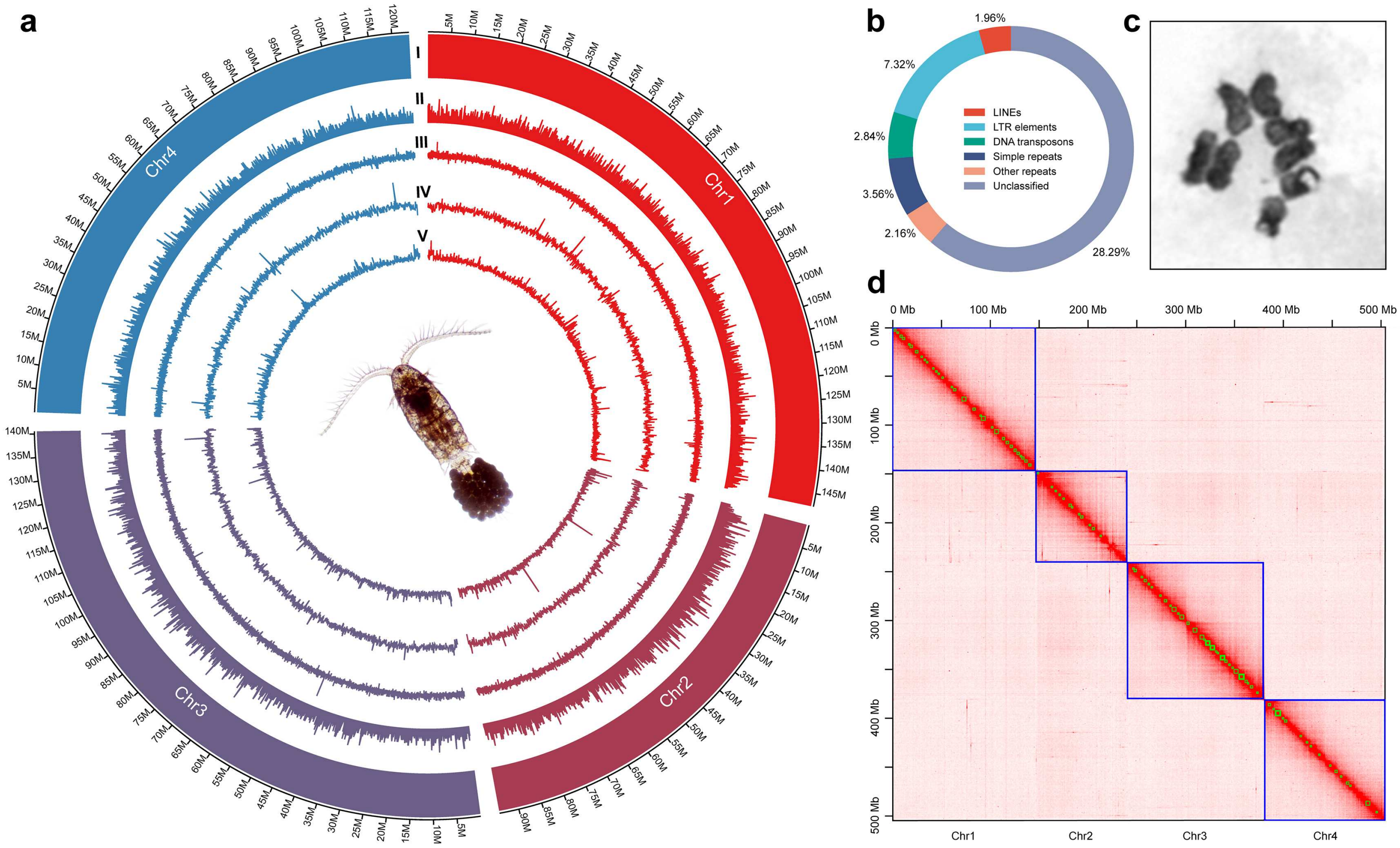

Figure 1

Chromosome-level genome assembly of the copepod *Eurytemora carolleeae* (*E. affinis*complex, Atlantic clade). (a) Circular diagram showing the genome landscape. I. Four chromosomes on the Mb scale. II. Density of protein-coding genes. III. Distribution of GC content (Mean GC = 32.5%). IV. Distribution of repetitive sequences. V. Distribution of LTR. All distributions were calculated in 100 kb non-overlapping sliding windows. **(b)**The proportion of repetitive sequences identified in the copepod genome. The circular diagram shows their relative proportions out of the total repetitive sequences (46.12% of the genome), and the numbers labelled on the diagram represent their percentage of occupied length in the genome assembly. **(c)** Well-isolated cell that shows the karyotype of the copepod (2n = 8) at metaphase. **(d)**The Hi-C contact map of the genome generated by Juicebox.

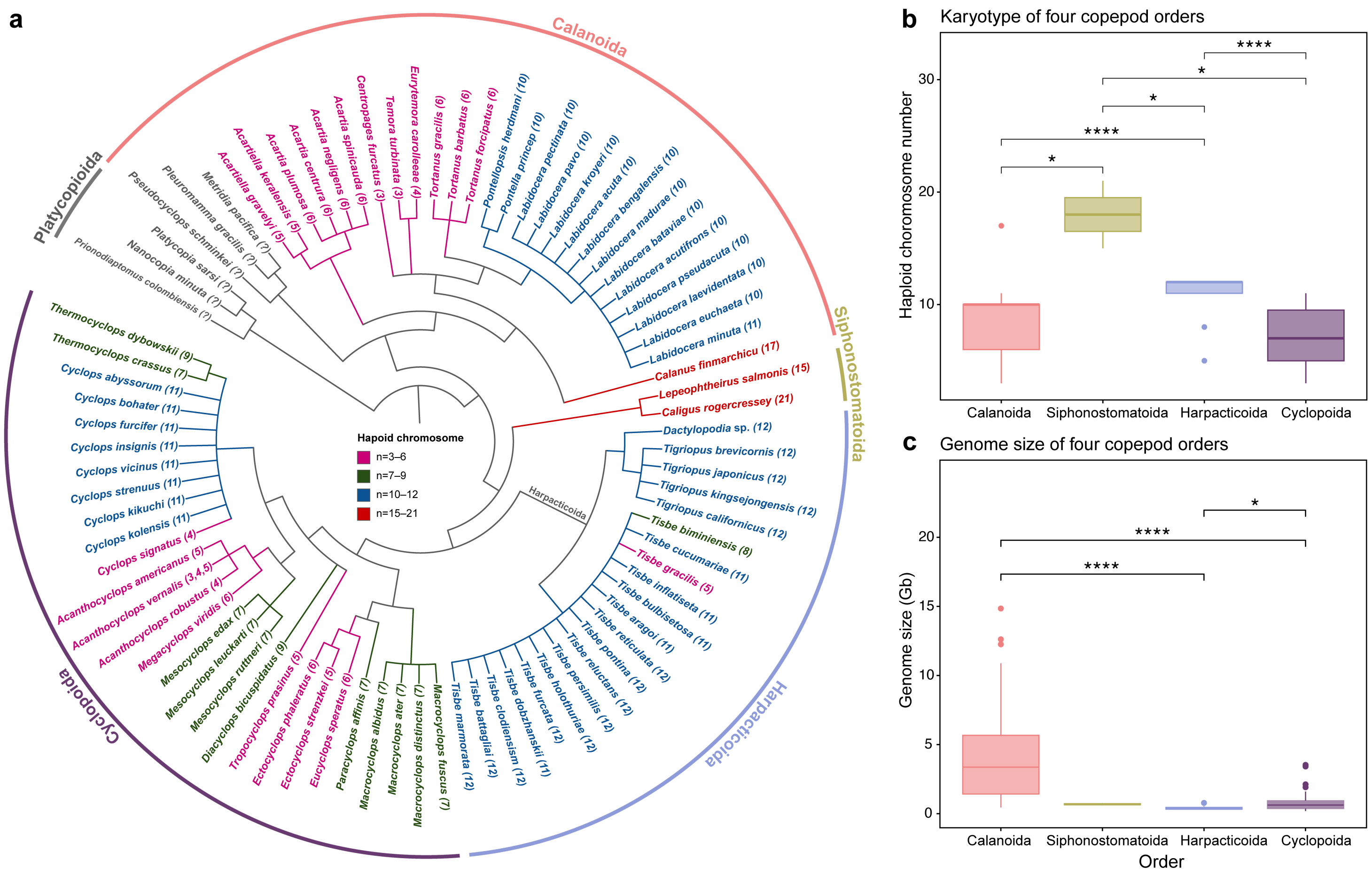

Figure 2

Chromosome number and genome size evolution in the crustacean class Copepoda. (a) Phylogeny of copepod species from five copepod orders. The phylogenetic topology was obtained from the synthesis tree of copepods, which integrated 31 published phylogenies [50]. Chromosome numbers are shown within parentheses after the species names. Different colors of species names represent the ranges of chromosome numbers. Clades that occupy basal phylogenetic positions, but possess unknown karyotype, are shown in grey in the phylogeny. **(b)** Mean chromosome number of four copepod orders (see Additional file 2: Table S2 for details). Chromosome number differs significantly among the four orders (Kruskal-Wallis test, H = 35.52, DF = 3, P= 9.5e-8). **(c)** Mean genome size of four copepod orders. Calanoida mean genome size = 3993 Mb, Siphonostomatoida = 563 Mb, Harpacticoida = 315 Mb, and Cyclopoida = 667 Mb (see Additional file 2: Table S3 for details). Genome size differs significantly among the four orders (Kruskal-Wallis test, H = 49.58, DF = 3, P = 9.8e-11). Asterisks in (b–c) indicate the significance levels for Wilcoxon tests, where * refers to P < 0.05 and **** indicates P< 1e-4. Nonsignificant P-values are not shown.

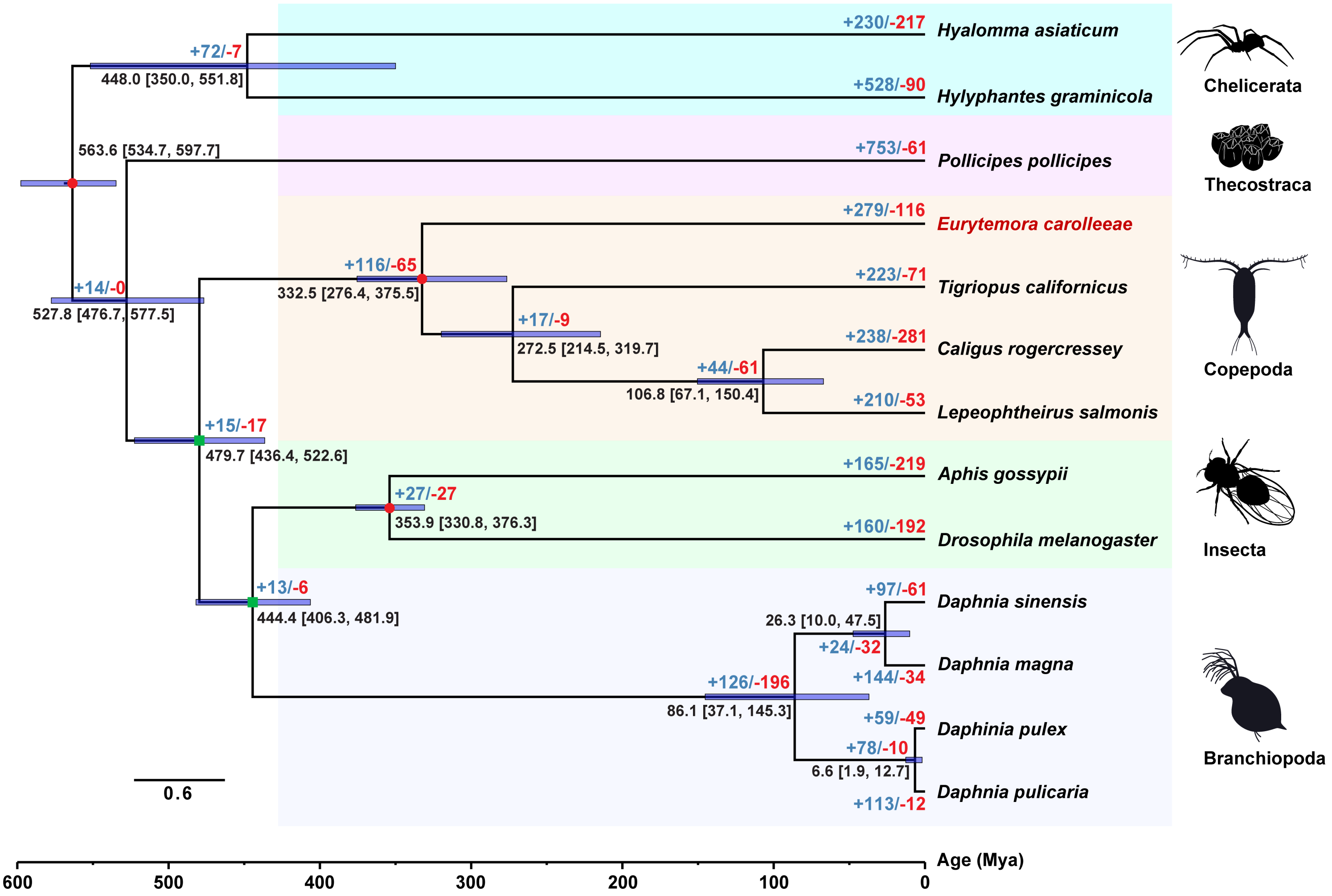

Figure 3

Gene family expansions and contractions during the evolutionary history of the Arthropoda, with a focus on the Copepoda. Phylogenetic reconstruction of 13 high-quality arthropod genomes was performed using RAxML based on concatenated single copy ortholog genes. All nodes show bootstrap values of 100%, except for two nodes with green rectangles, which have values of 66% (left node) and 60% (right node). Red circles represent three calibrated nodes with confidence time intervals retrieved from the Timetree database and applied in MCMCTree. Mean estimated divergence times are shown at each node with brackets indicating 95% highest posterior densities. The divergence times are on a scale of millions of years ago (Mya). The numbers of expanded gene families (in blue) and contracted gene families (in red) are shown on the branch tips and next to each node.

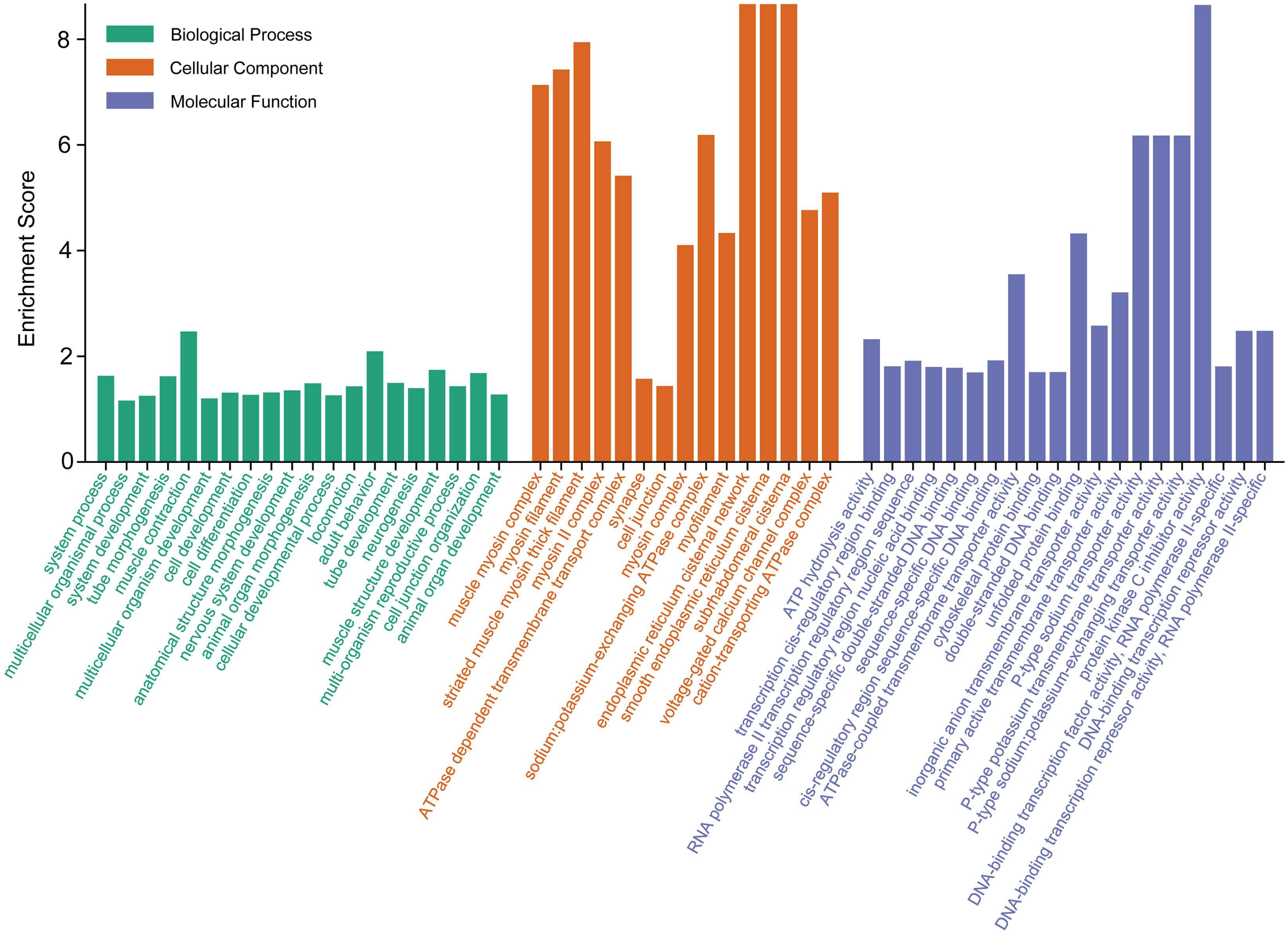

Figure 4

Significantly enriched of gene ontology (GO) terms in the expanded set of genes in the *Eurytemora carolleeae* genome. The GO terms were sorted by *P*-value (with higher *P*-value toward the right in each category). The complete list of enriched GO terms is shown in Additional file 2: Table S11. Only the top 20 GO terms of the Biological Process and Molecular Function categories, and top 15 GO terms of Cellular Component category are shown here.

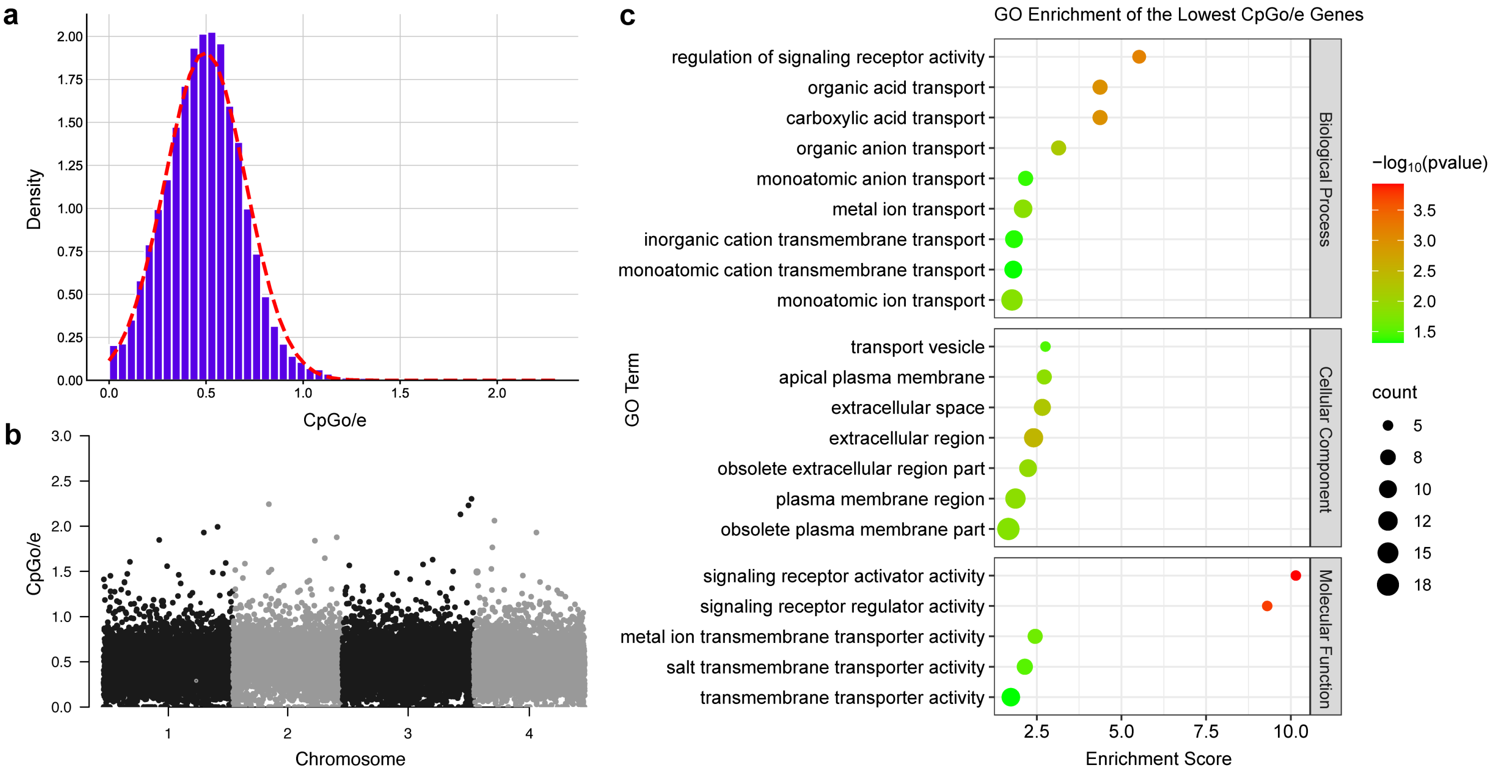

Figure 5

Patterns of genome-wide CpG$_{o/e}$ values of gene bodies, corresponding to signatures of past gene methylation in the *E. carolleeae* genome. (a) The CpG$_{o/e}$ values of the protein-coding gene sequences display a unimodal distribution. **(b)** The distribution of CpG$_{o/e}$ values across the genome when the genes are arranged by their position on each chromosome. **(c)** GO enrichment of the 1013 genes with 5% lowest CpG$_{o/e}$ values. The significance of GO enrichment is shown by the color of the circles and the enriched gene number is indicated by the size of the circles. The ion transporter genes tend to have the lowest CpG$_{o/e}$ values, suggesting extremely high levels of methylation in the past [52].

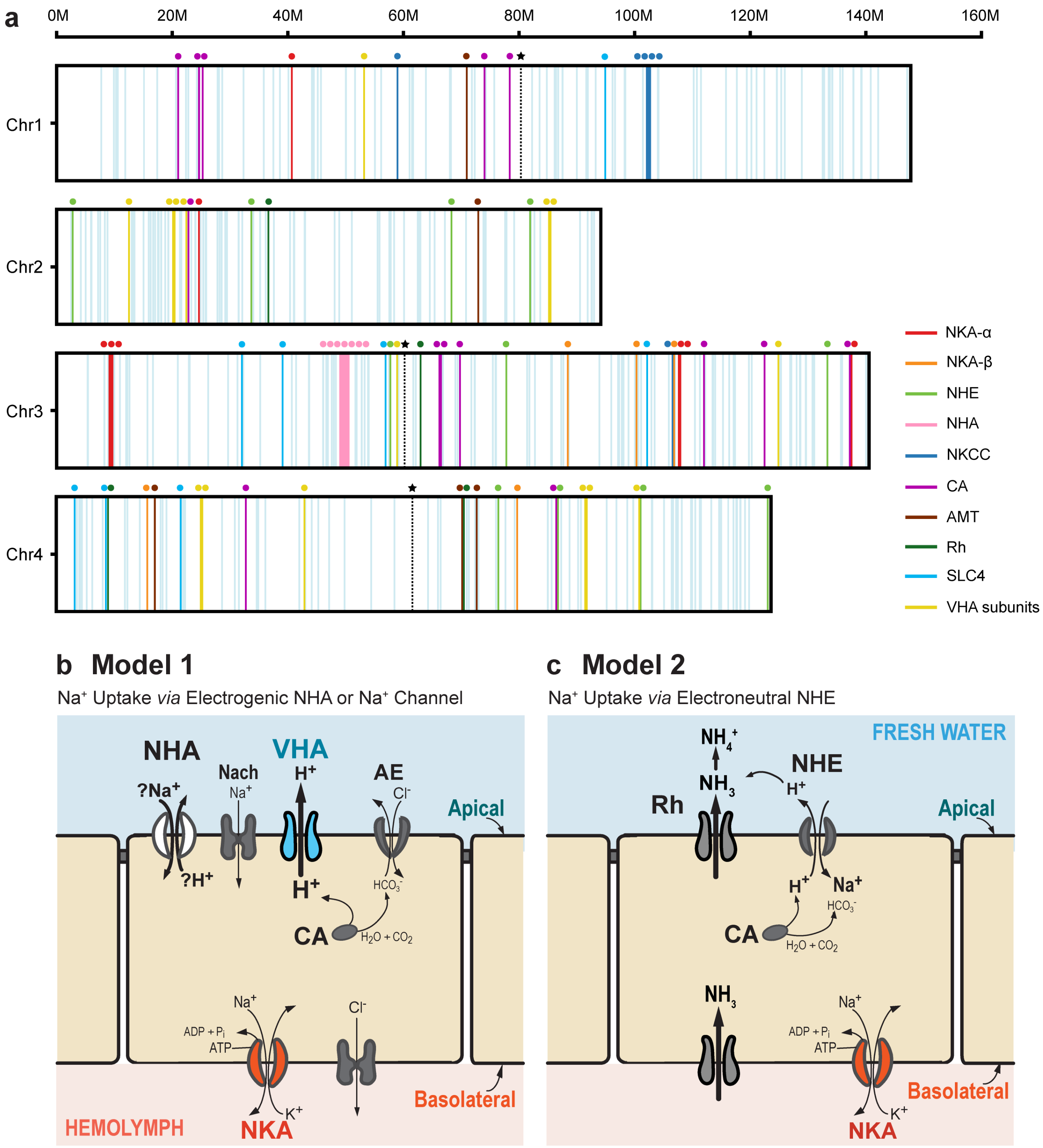

Figure 6

Localization of ion transporter genes on *E. carolleeae* chromosomes and hypothetical models of ion uptake from fresh water. (a) Ion transporter genes mapped onto the four *E. carolleeae* chromosomes. The vertical light blue lines represent 490 genes with ion (cation and anion) transporting function based on the genome annotation. The vertical lines and circles in other colors represent 83 key genes that showed evolutionary shifts in gene expression and/or signatures of selection in prior studies and are likely involved in hypothetical models of ion uptake. The dashed lines marked with stars indicate the positions of centromeres based on the Hi-C contact map (Fig. 1d, Additional file 1: Fig. S11). **(b, c)** Hypothetical models of ion uptake from freshwater environments. **(b)** Model 1: VHA generates an electrochemical gradient by pumping out protons, to facilitate uptake of Na^+ through an electrogenic Na^+ transporter (likely NHA). CA produces protons for VHA. **(c)** Model 2: An ammonia transporter Rh protein exports NH_3 out of the cell and this NH_3 reacts with H^+ to form NH_4^+. The deficit of extracellular H^+ concentrations cause NHE to export H^+ in exchange for Na^+. CA produces protons for NHE. These models are not comprehensive for all tissues or taxa and are not mutually exclusive.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1DuMay2023.docx
- AdditionalFile2DuMay2023.xlsx