The genome architecture of a copepod invading novel habitats

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 (carolle@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: September 5th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3002580/v3

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Running title: Genome architecture of a copepod

The genome architecture of a copepod invading novel habitats

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

With climate change, salinity is rapidly changing in marine and freshwater habitats throughout the globe. In addition, many of the most destructive aquatic invaders are crossing salinity boundaries. Populations of the copepod *Eurytemora affinis* species complex are numerically dominant and highly invasive, with the exceptional capacity to rapidly invade and adapt to novel salinities. Prior studies have found striking patterns of parallel adaptation in *E. affinis* complex populations, with selection acting on the same sets of ion transporter genes during independent saline to freshwater invasions. Our chromosome-level genome of *Eurytemora carolleeae* (Atlantic clade of the *E. affinis* complex) revealed a peculiar genome architecture that might contribute to its remarkable capacity to acclimate and evolve during salinity invasions. We assembled the highest quality copepod genome to date, using high-coverage PacBio and Hi-C sequencing of an inbred line generated through 30 generations of full-sib mating. Our new genome consisted of 529.3 Mb (contig N50 = 4.2 Mb, scaffold N50 = 140.6 Mb) anchored onto four chromosomes. Of its predicted 20,262 protein-coding genes, we found an extraordinary expansion of ion transporter gene families based on a comparative analysis of 13 arthropod genomes, with many of these expansions being recent. Notably, we found extreme CpG depletion of the ion transporter gene bodies, indicating high rates of methylation at these genes and suggesting transcriptional robustness of ion transporter gene regulation. This high-quality genome provides an invaluable resource that could help yield fundamental insights into the capacity of populations to expand their ranges into novel habitats.

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

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].

The species complex *E. affinis* has been the focus of intense ecological and evolutionary interest because of its extraordinary ability to invade a wide range of salinities on ecological time scales [18-22]. This copepod has the rare ability for an invertebrate to cross salinity boundaries ranging from hypersaline to completely freshwater habitats [18, 23-28]. Within a few decades, saline populations of this species complex have invaded freshwater habitats multiple times independently on three continents through human activity [18, 29, 30]. For instance, with the opening of the St. Lawrence Seaway, estuarine populations of the Atlantic clade of the *E. affinis* complex (aka. *E. carolleae*, Alekseev & Souissi, 2011) [31] were 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 [18, 32]. 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 Mississippi drainage system over a time period of ~60 years [18, 30]. 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 [18, 33]. These freshwater invasions by saline *E. affinis* complex populations were accompanied by the rapid evolution of freshwater tolerance, along with evolutionary changes in life history and ion regulatory function [23, 24, 34, 35]. Natural selection experiments in the laboratory revealed that this rapid freshwater adaptation could occur within 6–10 generations [23, 34, 36].

Across these independent evolutionary transitions to novel salinities, selection has repeatedly acted on ion transporter genes in the *E. affinis* complex populations [28, 34, 37]. Multiple prior studies...
have found that ion transporter genes form the largest functional category under selection during salinity change [29, 35-37]. We found that the same loci (and alleles) are targets of natural selection across salinity gradients in wild populations on different continents and in replicate selection lines in the laboratory [29, 36]. In addition, these same ion transporter genes show coordinated evolutionary shifts in gene expression between saline and freshwater populations [35]. Most notably, parallel selection acting on the same alleles might be driven by positive synergistic epistasis among beneficial alleles and selection on standing genetic variation in the native range populations [29, 36]. These results suggest that in response to salinity change, a set of cooperating ion transporters undergoes selection and evolves together as a unit.

Given the extraordinary evolutionary capacity of *E. affinis* complex populations during invasions, their particular genome architecture might be contributing to this rapid evolutionary response. For instance, specific gene family expansions in a genome could provide the genomic substrate for selection [38-42]. Genome-wide epigenetic signatures, such as the extent and localization of DNA methylation, would impact patterns of gene expression [43-46]. Additionally, the distribution of critical genes on chromosomes could affect their patterns of coordinated expression [47, 48]. Despite their potential importance, the role of genome architecture in affecting responses to selection have remained understudied and poorly understood [49, 50].

Thus, our goal was to investigate the genome architecture that might underlie the exceptionally invasive and adaptive capacities for this copepod species complex. Here, we present the first chromosome-level reference genome for a calanoid copepod, specifically for *E. carolleeae*, the Atlantic clade of the *E. affinis* species complex [18, 31, 51]. We produced 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 [29], 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, which was based only on Illumina sequencing of the same inbred line [52].
We find in this study that the genome sequence of *E. carolleeae* displays an unusual genome architecture underlying the targets of natural selection during salinity change in *E. affinis* complex populations, particularly at the ion transporter genes. In addition, this genome provides a valuable resource, as one of the most contiguous copepod genomes to date and among the highest quality of marine invertebrate genomes. Only four chromosome-level genome assemblies are available for copepods in the Genome database [53], namely for two parasitic copepods (Siphonostomatoida) and two species of the intertidal copepod *Tigriopus* (Harpacticoida), and none for the copepod orders Calanoida and Cyclopoida. The deficit of genomic resources for copepods has been quite striking, given their enormous ecological roles as grazers of the sea and their contribution of ~70% of the total zooplankton biomass [1, 54]. In addition, copepod invasions could have profound detrimental impacts by displacing the native food source of local fisheries [55-57]. 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). Moreover, dissecting the peculiar genome architecture of this copepod provides novel insights into its incredible capacity to invade novel environments.

**Results**

**High-quality chromosome-level genome assembly of *E. carolleeae***

The genome assembly we generated for *E. carolleeae* (Atlantic clade of the *E. affinis* complex) [31] had a much higher degree of completeness and contiguity than other available copepod genomes (Supplementary 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 (Supplementary 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 of 32.5% (Fig. 1a). This GC content was comparable to those of other calanoid copepods, but lower than those of harpacticoid copepods (Supplementary 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[42]. 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.

**Fig. 1.** Chromosome-level genome assembly of the copepod *Eurytemora carolleeae* (Atlantic clade of the *E. affinis* complex). (a) Circular diagram showing the genome landscape, including: 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 with 100 kb non-overlapping sliding windows. (b) Circular diagram showing proportion of different categories of repetitive sequences identified in the copepod genome, with the numbers on the diagram indicating their percentage of occupied length in the genome assembly. Repetitive sequences comprise 46.12% of this copepod genome. (c) Well-isolated cell showing the karyotype of the copepod (2n = 8) at metaphase. (d) Hi-C contact map of the genome generated by Juicebox.
This new genome is vastly superior to our prior assembly based on only Illumina sequencing of the same inbred line [52]. In this new genome, the contig N50 was greatly improved (from 67.7 kilobases [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 copepod species in Genome database [53]. 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 [58, 59]. The contig N50 of our genome was also longer than 151 out of 154 published genome assemblies of marine invertebrates in a recent survey [42]. Thus, this genome is one of the most contiguous copepod genomes to date and also among the highest quality of marine invertebrate genomes.

E. carolleeae genome size and karyotype in the context of copepod evolutionary history

Among copepods, E. carolleeae of the E. affinis species complex has a relatively 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 (Supplementary Table S2). 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) (Supplementary Tables S1 and S3), with significant differences among the four copepod 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, Supplementary 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; Supplementary Table S3) 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) (Supplementary 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. This lack of synteny among the three 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 [60, 61].

**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 [62]. 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 Supplementary Table S3 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 Supplementary Table S2 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 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; Supplementary 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 (Supplementary 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 Genome database [53].

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 (Supplementary Table S6). Among these genes, almost all genes (20,259) were functionally assigned based on at least one of eight functional annotation databases (Supplementary 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 (Supplementary 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% (Supplementary 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)[63]. Based on the Ks plot, we found no evidence of ancient WGD in the E.
carolleeae genome (Supplementary Fig. S4). Interestingly, the largest proportions of gene duplication
events occurred quite recently (Ks = 0–0.04, Supplementary 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 in E. carolleeae and across the Arthropoda,
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.
A phylogeny was reconstructed using a matrix of 101 concatenated single copy ortholog genes (Supplementary Table S8). This phylogeny supported the topology of (((Hexapoda + Branchipoda) + Copepoda) + Thecostraca) + Chelicerata); although, the relationships between Hexapoda, 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 (Supplementary 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, Supplementary Fig. S6, Supplementary Tables S9–S12). Relative to other arthropod genomes, this copepod genome displayed the expansion of 279 ortholog groups (aka. gene families), containing 1162 genes (Supplementary Table S9), and the contraction of 116 gene families, comprising 224 genes (Fig. 3, Supplementary 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 Supplementary 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, Supplementary Fig. S6, Supplementary 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. 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). Similarly, 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). 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).

The significantly expanded gene families in the *E. carolleeae* genome (Fig. 3, Supplementary Table S9) included ion transporter gene families that were found to be repeatedly under natural selection during salinity change in *E. affinis* complex populations in previous studies\(^5,26,34,80\). These gene families included Na\(^+\)/K\(^+\) ATPase \(\alpha\) subunit (*NKA-\(\alpha\)*), Na\(^+\)/K\(^+\) ATPase \(\beta\) subunit (*NKA-\(\beta\)*), and Solute carrier 4 (SLC4) of the bicarbonate (HCO\(_3^-\)) transporter gene families, including Anion exchanger (AE), Na\(^+\)/HCO\(_3^-\) cotransporter (NBC), and Na\(^+\)-driven Cl\(^-\)/HCO\(_3^-\) exchanger (NDCBE).
Of the ion transporter gene families under selection during salinity change⁵,26,34,80, many have
greater numbers of paralogs than other gene families in the genome. For instance, the ion transporter gene
families under selection [29, 35-37] have a mean gene paralog number of 7.6 (Supplementary Table S13),
which is higher than the mean paralog number of 4.2 for all expanded gene families in the *E. carolleae*
genome.

Interestingly, many of the ion transporter gene families show signatures of very recent
duplications. In particular, the ion transporter gene families *NKA-α*, *NKA-β*, *AMT*, and *VHA-a* are
contained within the most recent category of gene duplications (Ks = 2e-5–0.09 substitutions per
synonymous site) based on a Ks distribution analysis (Supplementary Fig. S4).

**Genome-wide CpG<sub>o/e</sub> depletion in gene bodies as signatures of gene body methylation**

The *E. carolleae* genome exhibited striking genome-wide signatures CpG deficiencies in gene bodies,
indicating high levels of gene body methylation across the genome. Notably, the most CpG deficient
genes tended to be ion transporter genes (Fig. 5). DNA methylation of gene bodies is typically associated
with increased expression levels and/or greater regulation of expression of genes [43, 64-69].

We determined the genome-wide distribution of CpG sites, to determine genome-wide signatures
of historical DNA methylation of our protein-coding genes. We calculated CpG<sub>o/e</sub> 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]). Typically, genes with lower CpG<sub>o/e</sub> values (lower numbers of observed CpG
sites than expected) likely have undergone higher levels of DNA methylation in the past (see Discussion)
[45, 70, 71].

The CpG<sub>o/e</sub> values across all genes displayed a unimodal distribution, with a very low mean
CpG<sub>o/e</sub> value of 0.5 in the *E. carolleae* genome (Fig. 5a). This unimodal distribution and low mean
CpG<sub>o/e</sub> value represents an extreme case of CpG depletion, indicating genome-wide signatures of high
levels of historical gene body methylation [72]. Most of the genes (19,960 out of 20,262) had CpG<sub>o/e</sub>
values lower than 1 (Fig. 5a). The distribution of CpG<sub>o/e</sub> values was not biased by the positions of genes
on different chromosomes (Fig. 5b). The mean CpG<sub>o/e</sub> value of our genome was much lower than the
unimodal distribution of <i>Drosophila melanogaster</i> (mean CpG<sub>o/e</sub> value around 1) [44] and the unimodal
distribution differed from the bimodal distributions found in many molluscs [72] and insects [44, 70].
Based on our genome annotation, the DNA methyltransferases (DNMTs) required for methylation are
present in the <i>E. carolleeae</i> genome. In contrast, these genes are lacking in the genomes of <i>Drosophila</i>
melanogaster and some other model organisms, such as yeast <i>Saccharomyces cerevisiae</i> and the
nematode worm <i>Caenorhabditis elegans</i> [44, 73, 74].

GO enrichment analysis to determine functions of genes with the 5% lowest and 5% highest
CpG<sub>o/e</sub> values (1013 genes) revealed very different sets of gene functions in the two groups. Notably,
genes with the lowest CpG<sub>o/e</sub> values were significantly enriched with GO terms related to ion
transmembrane transport functions (Fig. 5c, Supplementary Table S14). 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<sub>o/e</sub> values for ion transporter genes suggest that these
genes had extremely high levels of gene body methylation in the past [75].

Specifically, the mean CpG<sub>o/e</sub> value for 490 ion transporter genes was 0.47, similar to the low
genome-wide average of 0.5 in the <i>E. carolleeae</i> genome (Fig. 5a). Likewise, for 83 candidate ion
transporter paralogs that were identified as targets of natural selection during salinity transitions in <i>E.
affinis</i> complex populations [29, 35-37], the CpG<sub>o/e</sub> values ranged between 0.09–0.81 (Mean = 0.46, SD =
0.17; Supplementary Table S13). Of these 83 key ion transporter paralogs, 63.9% (53 out of 83)
possessed a CpG<sub>o/e</sub> value lower than the mean CpG<sub>o/e</sub> value of 0.5 in the <i>E. carolleeae</i> genome (Fig. 5a),
indicating even higher levels of historical gene body methylation for many of the ion transporter genes.

In contrast, genes with the highest CpG<sub>o/e</sub> values were enriched with conserved cellular functions,
such as "nucleic acid binding" (GO:0003676), "RNA processing" (GO:0006396), and "RNA metabolic
process” (GO:0016070) (Supplementary Table S15). These GO terms represent housekeeping genes, with our results suggesting relatively low levels of past methylation. In contrast, these genes were identified as hypermethylated in previous studies on insects [44, 70].

Fig. 5. Patterns of genome-wide CpG\textsubscript{o/e} values of gene bodies, corresponding to signatures of past gene methylation in the \textit{E. carolleeae} genome. (a) Genome-wide CpG\textsubscript{o/e} values of protein-coding gene sequences display a unimodal distribution with a low mean value of 0.5. (b) Distribution of CpG\textsubscript{o/e} values across the genome when the genes are arranged by their position on each chromosome. (c) GO enrichment of 1013 genes with 5% lowest CpG\textsubscript{o/e} values. Significance levels of GO enrichment are shown by the color of circles and numbers of enriched genes are indicated by the size of circles. The ion transporter genes tend to have the lowest CpG\textsubscript{o/e} values, suggesting extremely high levels of historical gene body methylation [75].

Localization of ion transporter genes on the four chromosomes

Given that ion transport-related genes were the most enriched GO category in the \textit{E. carolleeae} genome, we manually annotated and localized the ion transporter gene paralogs on the four chromosomes (Fig. 6a, Supplementary Table S13). We found that these ion transporter gene paralogs and subunits were distributed unevenly on the different chromosomes. The distribution of ion transporter genes on the chromosomes deviated significantly from a uniform distribution and tended to be more clustered than expected (Supplementary Fig. S7, see below). In addition, the distributions of ion transporter genes
differed significantly from those of functionally conserved housekeeping genes (Supplementary Tables S15 and S16, see below for details).

We focused heavily on the ion transporter paralogs that were identified as targets of natural selection during salinity transitions in *E. affinis* complex populations[29, 35-37] 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 *NKA-α, NKA-β, NHE, NHA, NKCC, CA, AMT, Rh, SLC4 (AE, NBC, NDCBE), and VHA* subunits[37]. We found unequal numbers of ion transporters on each chromosome, with 14, 14, 33, and 22 paralogs 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.

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 (Supplementary Fig. S7), for both 83 key ion transporter genes found under selection [37] (Fig. 6a, colored vertical lines) (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 (Supplementary Tables S15 and S16) and showed a higher frequency of closely spaced genes (Supplementary Fig. S8), both for 83 key ion transporter genes (Supplementary 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 (Supplementary 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, Supplementary Figs. S9 and S10). Although, the set of gene paralogs clustered around the centromeres are not the specific ones that show coordinated gene expression or parallel evolution across multiple studies [37].
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 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 (b, c). The dashed lines marked with stars indicate the positions of centromeres based on the Hi-C contact map (Fig. 1d, Supplementary 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. NKA-\(\alpha\) = Na\(^+\)/K\(^+\) ATPase \(\alpha\) subunit, NKA-\(\beta\) = Na\(^+\)/K\(^+\) ATPase \(\beta\) subunit, NHE = Na\(^+\)/H\(^+\) exchanger, NHA = Na\(^+\)/H\(^+\) antiporter, NKCC = Na\(^+\)/K\(^+\)/Cl\(^-\) cotransporter, CA = Carbonic anhydrase, AMT = Ammonia transporter, Rh = Rh protein, SCL4 = Solute carrier family 4 of bicarbonate transporters, including Cl\(^-\)/HCO\(_3^-\) exchangers (AE) and Na\(^+\)-driven HCO\(_3^-\) transporters (NBC, NDCBE), VHA = Vacuolar-type H\(^+\) ATPase.
Discussion

Copepods form the largest biomass of animals on the planet and contribute to the majority of total zooplankton biomass in aquatic habitats [1, 54]. However, despite their critical roles for ecosystem functioning and maintenance of fisheries of the planet, high-quality genomic resources had been relatively lacking. Among copepods, populations of the *Eurytemora affinis* species complex are notable for their extremely high abundance and critical importance as grazers in coastal ecosystems [12-15]. Intriguingly, some populations have greatly expanded their ranges in recent years into novel habitats, particularly into novel salinities [18, 28, 37].

Our results on the copepod *E. carolleae* (Atlantic clade of the *E. affinis* complex) reveal a genome architecture that likely enables its populations to be particularly responsive to changes in habitat salinity. The genome architecture underlying responses to salinity includes the extensive and recent expansions of ion transporter gene families, the extremely high signatures of historical methylation of ion transporter gene bodies, and the physical clustering of ion transporter genes. Such genomic features might contribute to the extraordinary ability of populations of the *E. affinis* species complex to invade biogeographic boundaries into novel salinities [18, 76].

Features of the calanoid copepod reference genome

This project presents the first chromosome-level calanoid copepod reference genome, specifically for *Eurytemora carolleae* (Atlantic clade of the *E. affinis* species complex) [18, 31, 51], with the highest level of completeness and continuity relative to other copepod genomes [53]. Moreover, this genome ranks among the highest quality among all marine invertebrate genomes [42]. 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, Supplementary Tables S2 and S3) [53]. We also found extremely low synteny with genomes of other copepod species (Supplementary Fig. S3). The relatively small genome size of *E. carolleae* might be a result of its large effective population size in
nature [77]. The effective population size of *E. carolleae* is approximately $10^6$ in the St. Lawrence estuary, based on our previous estimates of Watterson's theta (0.0131) [29] and assuming a mutation rate of $3.46 \times 10^{-9}$ based on *Drosophila melanogaster* [78].

The *E. carolleae* genome size ($1C = 529.3$ Mb) is within a similar range as our previous estimates for the same inbred line or population (L’Isle Verte). Our earlier draft genome sequence assembled from Illumina sequences [52], based on DNA exclusively from egg sacs (embryonic tissue) from the same inbred line, yielded a similar genome size of $\sim 510$ Mb (Supplementary Fig. S1). Our estimates based on DNA cytophotometry of embryonic cells from the same population yielded a 2C genome size of 0.6–0.7 pg DNA/cell or $1C = 318$ Mb [79].

In general, we found that genome size and chromosome number among copepods are not conserved but highly variable (Fig. 2, Supplementary Tables S2 and S3). For instance, chromosome number variation in copepods is on par with the levels of variation found in vertebrates and insects [60, 61]. The high variance in chromosome number in copepods suggests an evolutionary history of chromosomal fusions and fissions [80] and associated genomic rearrangements [81]. Such genomic rearrangements might explain the low levels of synteny we found among copepod genomes (Supplementary Fig. S3). The relatively large genome sizes (> 1 Gb) of some copepod species, especially in the Cyclopoida (Supplementary Table S2), reflect only the germline genome and not the somatic genome [82–84]. 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 [82, 85, 86]. There is no evidence of chromatin diminution in *E. carolleae* [79].

**Massive expansions of ion transporter gene families in the *E. carolleae* genome**

Based on a comparative analysis of 13 high-quality arthropod genomes, we found substantial gene family expansions in the *E. carolleae* genome (Figs. 3 and 4, Supplementary Tables S9 and S11). The expanded gene families were significantly enriched with ion transporter gene categories, comprising 29.2% of Molecular Function GO terms. Ion transporter genes have been found repeatedly to be the largest
functional (GO) category under selection during salinity change in previous evolutionary and physiological studies of *E. affinis* complex populations (colored bars in Fig. 6a)[28, 29, 37, 87]. These ion transporter gene families with signatures of selection tended to have greater numbers of paralogs (Mean = 7.6) than other expanded gene families throughout the *E. carolleeae* genome (Mean = 4.2). Moreover, low divergence times among duplicates (Ks = 2e-5–0.09; Supplementary Fig. S4) indicates that many of the of ion transporter gene family expansions (i.e., *NKA*-α, *NKA*-β, *AMT*, and *VHA*-a) have been quite recent, within roughly $10^4$–$10^8$ generations.

These massively expanded ion transporter gene families likely provide the potential for functional differentiation among the ion transporter paralogs. Given the negative genetic correlations between saline and freshwater tolerance in *E. affinis* complex populations [88], it is quite possible that different ion transporter paralogs are functioning optimally at different salinities. Such functional differentiation could provide greater versatility in acclimatory response and valuable genetic substrate for natural selection in the face of salinity change. In fact, a previous study had found that different ion transporter paralogs (e.g., *NHA*, *NKA*, *CA*, *NKCC*) show considerable variation in acclimatory and evolutionary shifts in gene expression in response to salinity change [35]. In addition, ion transporter paralogs vary in their signatures of selection across salinity gradients in wild populations and during salinity decline in laboratory selection lines [29, 36, 37, 87]. More studies are required to determine functional differences among the ion transporter paralogs and how acclimatory and selection responses differ among them.

**High rates of methylation of ion transporter gene bodies**

In the *E. carolleeae* genome, we found a genome-wide signature of CpG$_{oe}$ depletion, with a very low mean value of 0.5, indicating genome-wide signatures of high historical gene body methylation (Fig. 5). In fact, 98.5% of genes showed a pattern of CpG depletion, with CpG$_{oe}$ values lower than 1. Genes with lower CpG$_{oe}$ values (lower numbers of observed CpG sites than expected) indicate that they likely have undergone higher levels of DNA methylation in the past (and might be prone to ongoing methylation). Most DNA methylation events occur at CpG sites and result in the production of 5-methylcytosine.
Subsequently, spontaneous deamination of 5mC leads to C to T conversion\cite{45, 75}. Thus, high levels of DNA methylation will eventually cause the depletion of CpG sites associated with genes \cite{45, 70, 71}.

The genome-wide low CpG\textsubscript{o/e} values observed here tend to be rare in invertebrate species, but more common in vertebrates\cite{44, 71-73, 89}. The mean CpG\textsubscript{o/e} value of 0.5 in the \textit{E. carolleeae} genome was lower than those of 152 out of 154 arthropod species\cite{89}. Based on this survey, the mean CpG\textsubscript{o/e} value 0.5 for \textit{E. carolleeae} was comparable to the lowest CpG\textsubscript{o/e} value of 0.47 for two species, the fiddler crab \textit{Celuca pugilator} and the remipede crustacean \textit{Xibalbanus tulumensis} \cite{89}. This CpG depletion likely contributes to the low GC content of this genome (32.5\% GC).

In addition, the CpG\textsubscript{o/e} values across all genes in the \textit{E. carolleeae} genome displayed a characteristic unimodal distribution, which is typical of gene body methylation \cite{71}. Such a unimodal distribution of CpG\textsubscript{o/e} values is common in vertebrates, but extremely rare in invertebrates, which tend to have a mosaic pattern of both low and high CpG\textsubscript{o/e} genes \cite{44, 65, 71-73, 89}. This unimodal distribution of low CpG\textsubscript{o/e} values in the \textit{E. carolleeae} genome indicates an unusual pattern of mostly low CpG\textsubscript{o/e} genes (98.5\% lower than 1), reflecting high levels of methylation of most genes.

Most notably, the ion transporter genes exhibited the lowest CpG\textsubscript{o/e} values (Fig. 5c, Supplementary Table S13 and S14), indicating complete and nearly complete depletion of CpG sites. This result suggests that the ion transporter genes have experienced extremely high levels of historical gene body methylation. This pattern might be consistent with the critical roles of ion transporter genes and the need for controlled transcriptional regulation during the evolutionary history of environmental fluctuations of this species complex and perhaps of the genus \textit{Eurytemora} \cite{28, 29, 37, 90-96}.

DNA methylation of gene bodies was found to be positively correlated with increased levels of gene expression, in contrast to the suppression of gene expression by DNA methylation of gene promoter sequences \cite{43, 64-69}. Gene body methylation has been proposed to facilitate responses to environmental change and assist in acclimation by modulating gene expression \cite{58, 93, 97-99}. Gene body methylation has also been suggested to maintain the transcriptional robustness by preventing aberrant transcription or
regulating splicing efficiency, which might contribute to long-term stress adaptation [94, 95, 100]. In the
E. carolleeae genome, the extremely low CpG_{o/e} value distribution (Fig. 5a, b) suggests high levels of past
(and possibly ongoing) genome-wide gene body methylation of most genes (with the highest levels of
methylation at the ion transporter genes). This preponderance of low CpG_{o/e} genes suggests genome-wide
global responses to changing environments.

Clustering of ion transporter genes on the four chromosomes

The greater spatial clustering than expected of ion transporter paralogs on the four chromosomes (Fig. 6)
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 [47, 48]. The close physical linkage of
beneficial alleles might be favored by selection due to reduced recombination [36, 101-103], 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 [29, 35, 36] were not necessarily clustered together in the
genome. 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 [29, 35-37, 87]. 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 [29, 36, 87]. These shared targets of selection included
paralogs of the ion transporters NHA, NKA, VHA, CA, NKCC and Rh[37]. While many of the ion
transporter paralogs under selection were localized on Chromosome #3, many others resided on the other
three chromosomes (Supplementary Fig. S9). For instance, the specific ion transporter paralogs that we
found near any of the centromeres (Chromosomes 1, 3, and 4; Supplementary Figs. S9 and S10) were not
the ones that showed parallel evolution in the previous studies [37].
The significant clustering of ion transporter paralogs might be a byproduct of neutral processes, such as the recent expansions of ion transporter genes in the *E. carolleae* genome (previous section; Supplementary 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 functional or 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.

**Concluding Remarks**

The genomic characteristics 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 [76, 92, 104, 105]. 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, with rapid salinity declines in high-latitude coastal regions [106-108]. The high-quality genome sequence generated by this study provides an invaluable resource for gaining novel insights into genomic mechanisms that enable rapid responses to environmental change and rapid invasions into novel habitats [109, 110].

**Materials and Methods**

**Sampling and laboratory inbreeding of *E. carolleae***

A population from the Atlantic clade (*E. carolleae*) 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 [111]. 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...
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 times a week with water changed weekly. The inbred line VA-1 was used for this study.

**Sequencing of the *E. carolleae* 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 (Supplementary 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 protocol [112] from 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 [52, 113] 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 [114] was used to trim the raw sequencing reads with default parameters. Genome size was estimated based on the k-mer distribution using JELLYFISH [115] (count -m 21/25 -C -s 1G -F 2, histo -h 1,000,000) GenomeScope v2.0 [116] was used to estimate the genome size, heterozygosity, and proportion of repetitive sequences 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 [117] (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 [118] (-x ava-pb -t 8 -k17 -w17) and Nextgraph in NextDenovo (-a 1) were used to generate the primary genome assembly. NextPolish [119] 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 [120]. These two assemblies based on different datasets showed very similar quality with respect to continuity and completeness (shown in Supplementary Table S17). This assembly (#1) with higher contig N50 was further used in the following analyses (Supplementary Table S17). Purge_dups [121] was applied to remove heterozygous duplicates in the genome assembly.

For chromosome scaffolding, Juicer [122] and 3D-DNA [123] were used to scaffold the genome assembly to the chromosome level. Juicebox v1.91 [124] 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 using BLAST v2.8.1 [125].

**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[53] and from published cytophotometric and karyological investigations (Supplementary Tables S2 and S3). We also retrieved records for copepod species from the Animal Genome Size Database [126]. The chromosome numbers were mapped onto a synthesis tree of the Copepoda that integrated 31 published phylogenies [62]. 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 [127].

**Genome annotation of *E. carolleeae***

RepeatMasker v4.07 [128] was used to identify repetitive sequences and transposable elements in the genome based on searching in Repbase v202101 [129], Dfam v3.7 [130], a *de novo* repeat library built by RepeatModeler v1.0.8[131], the integrated tools RECON [132], TRF v4.09 [133], and RepeatScout [134]. Long terminal repeat (LTR) searches were also performed with dependent LtrHarvest [135], CD-HIT [136], and Ltr_retriever [137] installed. We applied the MAKER v3.01 [138] 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*, *Lepeophthereus 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 that were sequenced in our previous gene expression study under various salinity treatments [35], three that were sequenced in our previous i5K genome sequencing project [52, 113], and two that were sequenced in this present study using samples from two other clades in the *E. affinis* species complex (Europe [*E. affinis* proper (Poppe, 1880)] [139] and Gulf of Mexico, Supplementary Online methods). These transcriptomic data sets were collected and reassembled based on our new reference genome, using HISAT v2.0.4 [140] and StringTie v2.2.1 [141]. Regarding *ab initio* gene prediction, we trained the gene predictor SNAP [142] with the
gene models predicted with the above evidence. The self-trained predictor GeneMark-ES [143] 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 databases of invertebrates and using a separate self-established database with all gene sequences of *E. affinis* in RefSeq. GO [145], KEGG [146], COG, and eggNOG [147] databases were searched using eggNOG-mapper v2.1.9 [148]. The Pfam database in InterPro [149] was also searched by HMMER v3.2 [150].

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 [151]. 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 [152].

Transfer RNAs (tRNAs) were identified using tRNAscan-SE v2.0 [153] with default parameters. MicroRNA and small nuclear RNA were identified with BLASTN against the Rfam database v12.0 [154] 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 [155]. 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 (Supplementary Table S18). These arthropod genomes included three chelicerates (*Hyalomma asiaticum*, *Hylyphantes graminicola*), one barnacle (*Thecostraca: Pollicipes pollicipes*), three copepods (*Caligus rogercressey*, *Caligus rogercressey*,...
Lepeophtheirus salmonis, Tigriopus californicus), four branchiopods (Daphnia pulex, D. magna, D. pulicaria, D. sinensis) and two hexapods (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 [156]. Gblocks v0.91b [157] was used to trim the alignment. A phylogeny was reconstructed using a Maximum Likelihood algorithm in RAxML v8.0.19 [158]. 100 bootstrap replicates were performed to assess statistical support for tree topology. We used MCMCTree from PAML v4.9 to estimate divergence times [152]. Three confidence time intervals retrieved from the TIMETREE v5 database [159] were applied in MCMCTree as calibrations for the divergence times (shown as red circles in Fig. 3). CAFÉ5 [160] was used to analyze the expansion and contraction of gene families among taxa in the phylogenetic tree. For gene families exhibiting expansions and contractions in the E. carolleeae genome, GO and KEGG enrichment analyses were performed using TBtools v1.112 [161].

Syntenic relationships among three copepod species were analyzed using MCScan in JCVI [162]. We used the highest quality available copepod genomes, E. carolleeae, Tigriopus californicus, and Lepeophtheirus salmonis, representing three different copepod orders, Calanoida, Harpacticoida, and Siphonostomatoida, respectively. Collinear gene blocks within the genomes were identified using the longest coding sequence of each gene.

Genome-wide CpG\textsubscript{o/e} values in the E. carolleeae genome

To assess the patterns of historical methylation within gene bodies, genome-wide CpG\textsubscript{o/e} values were determined in the E. carolleeae genome. The CpG\textsubscript{o/e} value of each gene was computed as the observed frequency of CpG sites ($f_{\text{CpG}}$) divided by the product of C and G frequencies ($f_C$ and $f_G$), i.e., $f_{\text{CpG}}/f_C * f_G$ in the coding sequence (CDS) of each gene [70]. The density of CpG\textsubscript{o/e} values for all genes was fitted and plotted in R. The distribution of CpG\textsubscript{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\textsubscript{o/e} genes, we performed GO enrichments for the top 5% genes with the highest and lowest CpG\textsubscript{o/e} values using TBtools [161].

**Localization of ion transporter genes across the *E. carolleae* 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 [37] were manually annotated and mapped separately 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 \( \text{Na}^{+}/\text{K}^{+} \text{ATPase} \) subunit (\textit{NKA-}\(\text{\alpha}\)), \( \text{Na}^{+}/\text{K}^{+} \text{ATPase} \) subunit (\textit{NKA-}\(\text{\beta}\)), \( \text{Na}^{+}/\text{H}^{+} \) antiporter (\textit{NHA}), \( \text{Na}^{+}/\text{H}^{+} \) exchanger (\textit{NHE}), \( \text{Na}^{+}/\text{K}^{+}/\text{Cl}^{-} \) cotransporter (\textit{NKCC}), \textit{Carbonic anhydrase} (\textit{CA}), \textit{Ammonia transporter} (\textit{AMT}), \textit{Rh protein} (\textit{Rh}), \textit{Vacuolar-type H}^{+} \text{ATPase} (\textit{VHA}), and \textit{Solute carrier family 4} (\textit{SLC4}) of bicarbonate (HCO\(_3\)\textsuperscript{-}) transporters, including \textit{Anion exchanger} (\textit{AE}), \( \text{Na}^{+}/\text{HCO}^{3}\textsuperscript{-} \) cotransporter (\textit{NBC}), and \( \text{Na}^{+}\text{-driven Cl}^{-}/\text{HCO}^{3}\textsuperscript{-} \) exchanger (\textit{NDCBE}) (Figs. 6b, c; Supplementary Figures S9 and S10).

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 gene distances 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\textsubscript{o/e} values identified in the prior section (Genome-wide CpG\textsubscript{o/e} values in the *E. carolleae* genome) were used (Supplementary Table S16). This set of genes was enriched in RNA processing and DNA binding related functions, which tend to be functionally conserved housekeeping genes.

**Data availability**
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.

**Acknowledgements**

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

**Supplementary Information**

Supplementary Figures S1–S11 and Supplementary Methods.

Supplementary Tables S1–S18.
References


31. 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(2767):41-56.
32. Mills EL, Leach JH, Carlton JT, Secor CL. Exotic species in the Great Lakes: a history of biotic
33. De Beaufort LF. Veranderingen in de Flora en Fauna van de Zuiderzee (thans IJsselmeer) na de
34. Lee CE, Kiergaard M, Gelembiuk GW, Eads BD, Posavi M. Pumping ions: rapid parallel
expression and plasticity during saline to freshwater invasions by the copepod \textit{Eurytemora affinis}
36. Stern DB, Anderson NW, Diaz JA, Lee CE. Genome-wide signatures of synergistic epistasis
37. Lee CE. Ion transporter gene families as physiological targets of natural selection during salinity
38. Luo S, Tang M, Frandsen PB, Stewart RJ, Zhou X. The genome of an underwater architect, the
caddisfly \textit{Stenopsyche tienmushanensis} Hwang (Insecta: Trichoptera). Gigascience.
2018;7(12):giy143.
ecoresponsive genome of \textit{Daphnia pulex}. Science. 2011;331(6017):555-61.
43. Seymour DK, Gaut BS. Phylogenetic shifts in gene body methylation correlate with gene


Supplementary Files

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

- Supplementalfiguresmethods.docx
- Supplementaltables.xlsx