NHA gene family phylogeny in *Eurytemora affinis* and Closely Related Arthropods

Patricia Zito

Abstract

In multiple separate events, populations of the *Eurytemora affinis* complex, a normally small marine copepod (crustacean), have adapted and invaded freshwater environments. When observing genetic differences between adapted and ancestral populations, multiple genes involved in ion transport show strong signals of natural selection, potentially indicating they have an important role in helping copepod populations adapt to new environments (Stern et al. 2020). One gene paralog among the highest signals of selection is NHA7, a member of the Sodium Hydrogen Antiporter (NHA) family. This result indicates that this gene family might play an important role in the adaption of copepod populations to freshwater environments. In this study, I will investigate the evolutionary history and patterns of molecular evolution of the NHA gene family to further elucidate the mechanism of salinity adaptation in copepod populations. Specifically, I will reconstruct a curated phylogeny of the NHA gene family in order to further infer its evolutionary history in arthropods. This information will be important for uncovering a clear mechanism for salinity adaptation.

Introduction

The development of the anthropocene has been the main cause of environmental changes, namely global warming, which has had broader consequences in rainfall patterns, melting of glaciers, and rise of the oceans (Lee et al., 2022). Salinity changes are deeply detrimental to fauna and flora around the world, as they impose a serious challenge for aquatic species physiological tolerances (Lee et al., 2022, Smyth, et al. 2023). In this study, I am using the copepod *Eurytemora affinis* complex as a model to examine the evolution of populations impacted by global warming, as they are abundant in nature, and their high capacity to adapt to new less-saline environments have made them invasive on multiple occasions. In fact, many

populations and clades in the *E. affinis* complex have independently invaded and adapted to freshwater (Stern and Lee, 2020).

In a recent study, Stern et al. (2022) simulated the effects of global warming by reducing salinity in replicate copepod lines in the lab. By pool-sequencing both lab and nature adapted populations, Stern et al. (2022) found that these copepods had undergone parallel evolution among replicate lines. In another study investigating these signatures of selection, Stern and Lee (2020) and Stern et al. (2022) found that most sites under selection located on ion transporter genes are parallel in replicate lines, suggesting that these ion transporter genes might contribute to adaptation of copepod populations to these new environments, and that these ion transporter genes might be co-adapting together to form these adapted phenotypes.

My research focuses on the ion transporter Sodium Hydrogen Antiporter (NHA) gene family, which is contained within the genomic region with the highest signals of selection across the copepod genome (Stern and Lee, 2020). This high signal of selection suggests their importance in adaptation from brackish to freshwater environments.

To answer the question of what allows copepod populations to survive and reproduce under salinity decay pressures, I will investigate the evolutionary history and patterns of molecular evolution of the NHA gene family. I will do this by reconstructing accurate phylogenies, from which I hope to later extract the order in which these mutations arise.

Materials and Methods

Collecting sequences

I had two major sources of data for collecting the sequences present in my analysis: (1) the in-lab genome and transcriptome data on *E. affinis* (Du, 2023 - unpublished), and (2) the ncbi Genbank database.

For the first source, I have used BLAST+ 2.6.0 (Camacho et al., 2009) to create a database of the recently assembled genome and transcriptome data, and then search through the database for potential matches. In total, I made 8 searches per database using as my queries 8 partial CDS sequences for NHA paralogs in *E. affinis* (Marthers, 2018 – unpublished). Then, for each of these paralogs, I used the best; longest transcript matches and aligned them with a 10kb window of the best-match genome result. From this, I was able to infer a more complete CDS sequence that did not contain the 5' cap or poly-A tail that are usually present in the transcript, or

the intron sequences, that are present in the genome (table 1). The final CDS sequences were aligned (figure 1) using T-Coffee Version_13.45.0.4846264 (Di Tommaso, 2011), and a simple preliminary tree was inferred using the PHYLIP Neighbor Joining building method, with the Dayhoff PAM distance matrix model (figure 2). Both alignments and tree were visualized using UGENE v46.0 (Okonechnikov, 2012).

Genes	gene size (bp)	# exons	Transcript	# isoforms	CDS size (bp)	Protein size (aa)
NHA1	2399	3	MSTRG.14342.3	9	1914	638
NHA2	3419	3	MSTRG.14344.1	2	1959	653
NHA3	4303	3	MSTRG.14347.2	4	3615	1205
NHA4	3425	3	MSTRG.14350.1	6	2361	787
NHA5	3685	5	MSTRG.14351.1	4	2262	754
NHA6	2803	3	MSTRG.14351.4	4	2127	709
NHA7	2093	6	MSTRG.14352.1	1	1770	590
NHA8	8682	6	MSTRG.15126.2	2	990	330
Averages	3851.125	4	-	4	2124.75	708.25

Table 1. Summary of *E. affinis* paralogs

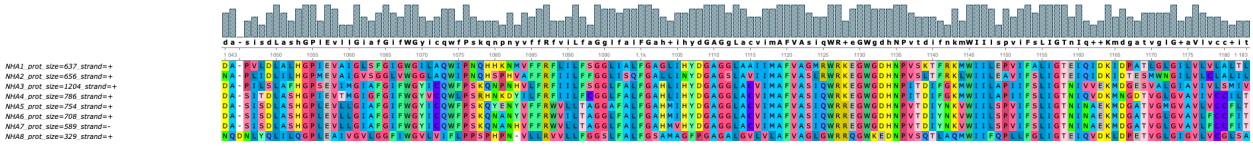

Figure 1. Conserved region of all 8 paralogue sequences alignment in *E. affinis*.

For my second source of sequences, I used my complete CDS sequences, as well as some validated Refseq NHA sequences in different organisms as queries to BLASTp (Morgulis, 2008) against the non-redundant BLAST database, using the scoring matrix BLOSUM62 and BLOSUM45, and excluding XP/XM sequences. The main reason for this latter criterion is because there is a lack of consistency in the annotation of proteins, such as some of them

containing or not containing information linked information and ids of their bioproject and assembly information, especially when comparing XP/XM sequences from other non-XP/XM sequences. Moreover, genome assemblies from XP sequences often do not show as potential results when looking in the NCBI database, even when directly looking up their name. Because of this, it became virtually impossible to automate acquiring the information necessary for my analysis.

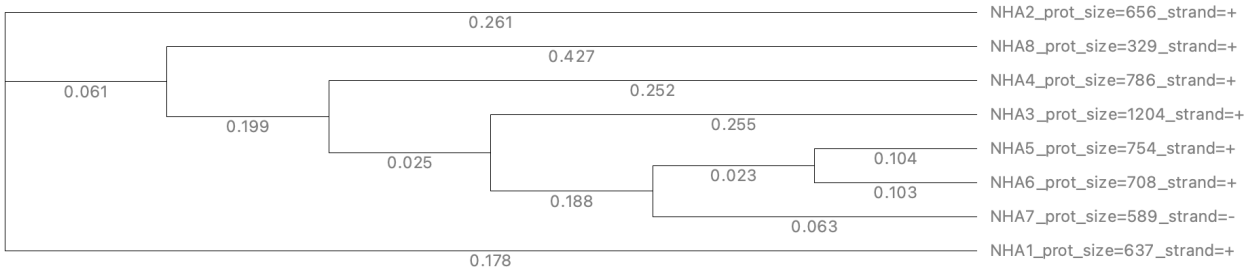

Figure 2. Preliminary tree of all *E. affinis* paralogs.

In summary, these extensive BLAST searches yielded a total of 214 non-XP/XM proteins. With this intial dataset, I then used the Biopython (Cock, 2009) library, and more specifically the Entrez module as a NCBI database API to recover in-depth information about these proteins, as well as their genome assembly.

Curation of orthologs

Because Genbank is a public database, it also contains several unreliable sequences that might have been reconstructed from poor sequencing or assembling techniques. For this reason, I used the pandas library (McKinney, 2010) available on python to filter all proteins results according to their respective genome assembly quality. To reinforce this quality check, I used the following criteria: minimum protein size = 200, minimum genome coverage = 50x, minimum scaffold N50 = 10000, and maximum scaffold count = 9000. With these filters, only 73 (34%) out of the 214 original protein results remained.

For a second curation step, given the recent divergence of the NHE and NHA gene families (Anderegg et al., 2022), I decided to align each of the 73 remaining sequences with a known NHA (*D. melanogaster* NHA2, NP_001247251.1) and a known NHE (*D. melanogaster* NHE3, AAF13702.1) sequences and check that the remaining sequences were more closely related to NHA than to NHE. To do this, I used the R/seqinR (Charif and Lobry, 2007) package to load and perform a simple global Pairwise alignment (Needleman and Wunsch, 1970) between

every sequence and NHE and NHA sequences. After this step, I was able to remove 26 potentially NHE sequences. With this, only 47 sequences, and 31 species remained.

barnacle	cockroach	copepod	crab	diptera	ephemera
1	3	2	1	9	6
hemiptera	**hymenoptera**	**lepidoptera**	**spider**	**tardigrade**	**thysanoptera**
10	12	27	3	5	1
tick					
3					

Table 2. Diversity of the dataset

For a third curation step, I used the conserved domain database analysis (Lu, 2020) available on NCBI to verify that the remaining sequences were contained within the NHA gene family. Using the *D. melanogaster* NHA2 (NP_001247251.1) as a reliable NHA containing a reliable NHA conserved domain, I re-checked every sequence and eliminated the sequences that did not have "NA_H_exchanger" as their conserved domain. Moreover, to ensure that I would obtain all paralogs available for each species, I also used BLAST+ to blast the sequences I had against their respective organism, thus obtaining all NHA paralogues for every species in my dataset. Additionally, at this stage, I also manually added separate species that had not originally passed the initial filtering step due to having XP/XM sequences, such as *Lepeophtheirus salmonis* and *Eriocheir sinensis*.

Finally, for the last curation step, I also added two species of tardigrades (*Ramazzottius varieornatus* and *Hypsibius exemplaris*), and all their paralogs as outgroup sequences. The finalized dataset had a total of 83 sequences, with 35 unique species: 3 belonging to crustacea, 2 belonging to chelicerates, 27 belonging to *hexapoda* and 2 belonging to *tardigrada* (table 2). On average, there were 2.37 NHA paralog per species (table 3).

As a safety measure, I have also kept a second "relaxed" dataset that was not filtered for their conserved domain, which contained 296 sequences, and 48 species, with a whopping average of 6.17 paralogs per species. This, however, did not yield good alignments or trees, so I proceeded future analyses with the more conservative dataset.

Acromyrmex echinatior	Amphibalanus amphitrite	Ampulex compressa
1	1	2
Anopheles gambiae	Apolygus lucorum	Arctia plantaginis
1	2	7
Brenthis ino	Ceratitis capitata	Cloeon dipterum
5	1	5
Cotesia congregata	Cotesia glomerata	Dendrolimus kikuchii
2	1	4
Diaphorina citri	Drosophila melanogaster	Eciton burchellii
6	2	1
Ephemera danica	Eriocheir sinensis	Glossina fuscipes
1	1	1
Helicoverpa armigera	Hypsibius exemplaris	Iphiclides podalirius
4	2	6
Lepeophtheirus salmonis	Lucilia cuprina	Megalurothrips usitatus
1	2	1
Nezara viridula	Oedothorax gibbosus	Ooceraea biroi
2	3	2
Oppia Nitens	Periplaneta americana	Polypedilum vanderplanki
3	3	2
Ramazzottius varieornatus	Spodoptera littoralis	Temnothorax longispinosus
3	1	2
Tigriopus californicus	Vespula vulgaris	
1	1	

Table 3. Paralogs across species

Alignment

For the alignment of the 83 sequences, I have used the MAFFT (Katoh, 2002) multiple sequence alignment tool. This yielded a clean alignment of all sequences with a clear conserved domain. To isolate it, I used GBlocks 0.91.1 (Castresana, et al., 2000, Talavera et al., 2007) available on the ngphylogeny.fr/tools/ server with the following parameters: minimum number of sequences for a conserved position= default, minimum number of sequences for a flank position= default, maximm number of contiguous non-conserved positions= 8, minimum length of a block= 10, allowed gap positions= with half. The resulting segmented alignment can be seen on the figure below.

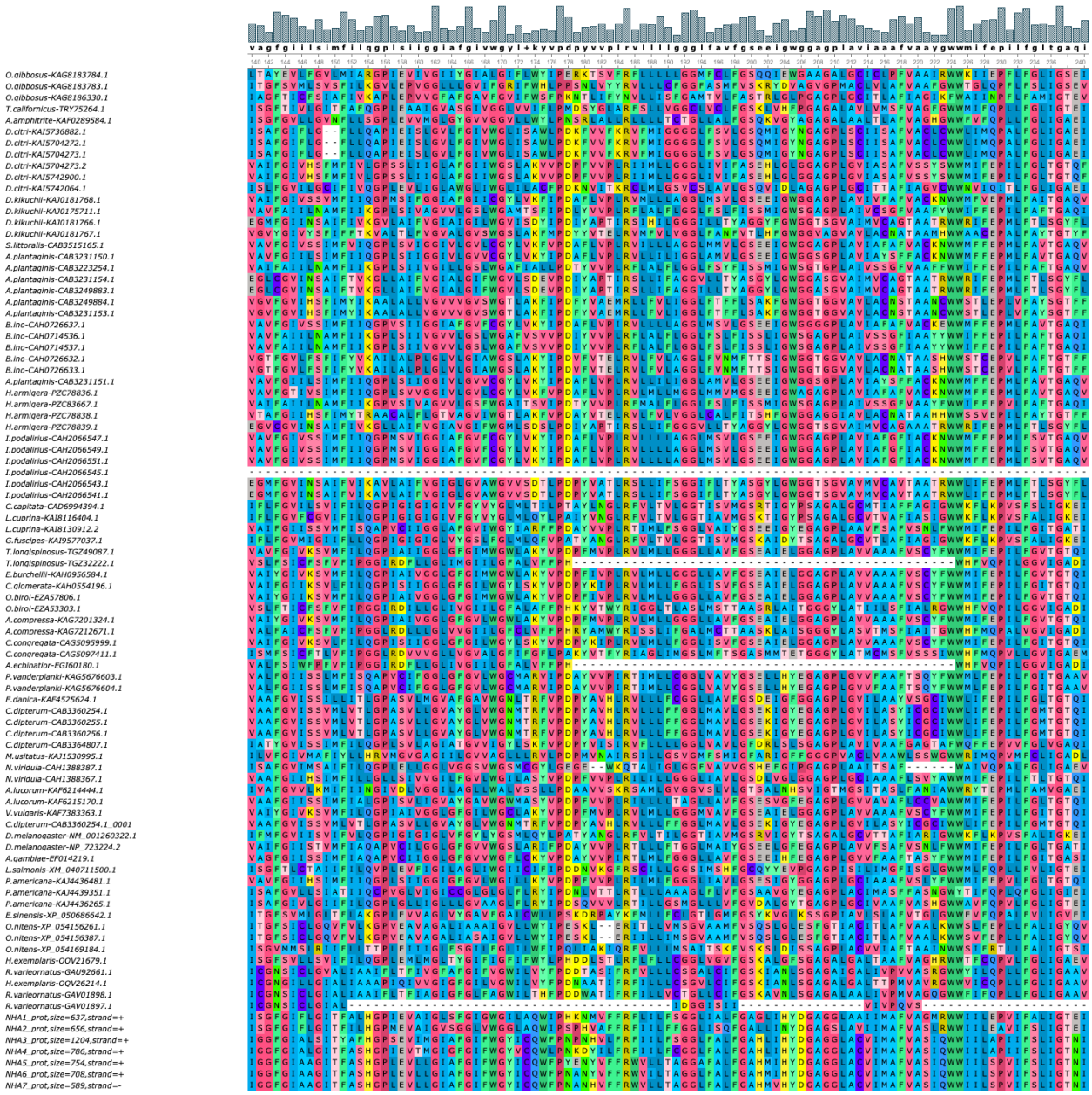

Figure 3 Contains a partial view of the multiple sequence alignment.

One of the main reasons for my choice in this alignment software is because it has shown to be as accurate as T-Coffee, but at a speed up to 10x faster (Katoh, 2002). Moreover, it was much mor friendly-user and easy to install.

Phylogeny Inference

To build the trees, I used two main approaches: (1) maximum likelihood, and (2) Bayesian. For my maximum likelihood approach, I decided to use IQTree ver 2.0.3 (Nguyen, 2014) because it is still considered one of the most accurate maximum likelihood models, it could also automatically perform tests and use best-fit evolutionary models, and it ran surprisingly fast on my trimmed sequences. Perhaps the greatest weakness to IQTree is that, like all other maximum likelihood approaches, it cannot guarantee a global optimum. However, given that no other software can, it is still considered a formidable tool.

For my parameters, I used the model LG+F+G4, which best fit the AIC criteria, and was automatically defined by IQTree. I also used specifically the OQV21679, *Hypsibius examplaris* paralogue, as my outgroup. This is because IQTree can only accept one sequence as an outgroup, and when building a neighbor-joining tree with all the paralog sequences of my two outgroups (*Hypsibius examplaris*, and *Ramazzottius varieornatus*), OQV21679 and GAU88015.1 were considered the most ancestral, however, only OQV21679 had the "NA_H_exchanger" domain.

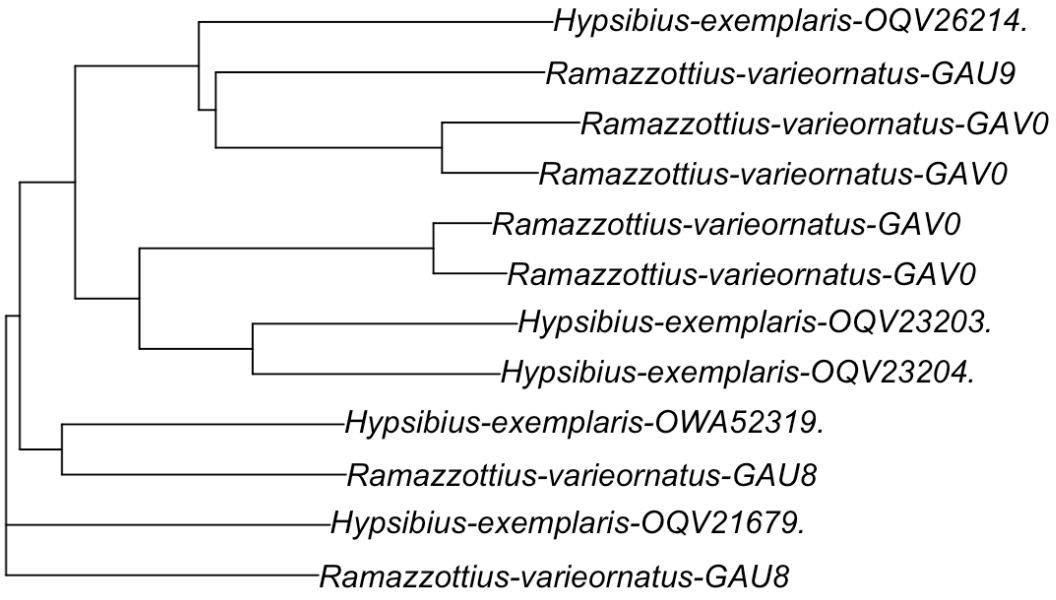

Figure 4 Neighbor-joining Tree of outgroup paralogues

For my Bayesian approach, I have attempted to download and install both MrBayes (Huelsenback and Ronquist, 2001) and BEAST (Drummond and Rambault, 2007), however, none of them worked due to an error in the installation of my Beagle library (Ayres et al., 2019). For this reason, I used the CIPRES server (phylo.org/portal2) which allowed me to submit Bayesian inferences as jobs. I have attempted both MrBayes and BEAST, however only BEAST worked with my set of inputs.

The advantages to using BEAST, besides being the only working software for Bayesian inferences on my dataset, are the introduction of a relaxed molecular clock model, which removes one of the large assumptions of tree building: that the mutation rate is the same across lineages.

To reconstruct the NHA phylogeny using BEAST (Drummond and Rambault, 2007), I used the following parameters: codon_partitioning=False, no_beagle_=False, nu_partitions=1, nu_patterns=307, path_sampling=False, runtime=10, save_everyval=100000, spec_seed=False, which_beast=104.

Results

The inferred IQTree generally had a good grouping of each taxonomic clades. For example, it clustered all crustacean, chelicerates, tardigrades and most of hexapods together. Curiously, however, it did not order these clades according to the expected by the normal arthropod species tree. For instance, while it did keep tardigrades as the outgroup clade to all other arthropods, it also grouped it together with hymenopterans, an order within hexapods, the most derived sub-phylum within the dataset. This, unfortunately, could be a sign of long-branch attraction. Likewise, crustaceans are shown to be more ancient than chelicerates.

More interestingly perhaps, when dividing the hexapod clades into orders, we may observe clades that are more ancient and more derived for the following orders: hemipterans, dipterans, hymenopterans, and lepidopterans. Although we should be skeptic about the quality of the tree, it seems to suggest that NHA duplication events are relatively new. For example, because there chelicerates, tardigrades and crustaceans form their own clades, it is likely that these duplication events occurred after speciation events. Moreover, because there are no clear clades containing different species (e.g. one clade containing one specific paralogue of a species of crustacean, a chelicerate and a dipteran), we cannot conclude there is a clear homology

between the arthropods. In other words, while many proteins are labeled as “NHA2” on the Genbank database, these proteins might not be necessarily directly orthologous to each other.

The Bayesian tree was, perhaps, even more unexpected than the resulting maximum likelihood phylogeny. Because I could not specify a specific outgroup, the Bayesian phylogeny assigned lepidopterans as the outgroup for the arthropods. Strangely, it also divided all taxonomic groups into different clades, for the exception of tardigrades. For example, lepidopterans, hymenopterans, crustaceans, chelicerates, dipterans and hemipterans.

Although the Bayesian inference had a smaller number of iterations than I originally intended, an analysis of the tracer plot of the joint probability showed that the MCMC algorithm most likely had a quick burn-in (figure 7, showed as translucent), a good mixing (see through the caterpillar pattern in blue), which show that it had most likely had time to converge. Additionally, the expected sample size (ESS) also had a value of 1036.2, which is much higher than the minimum recommended > 100.

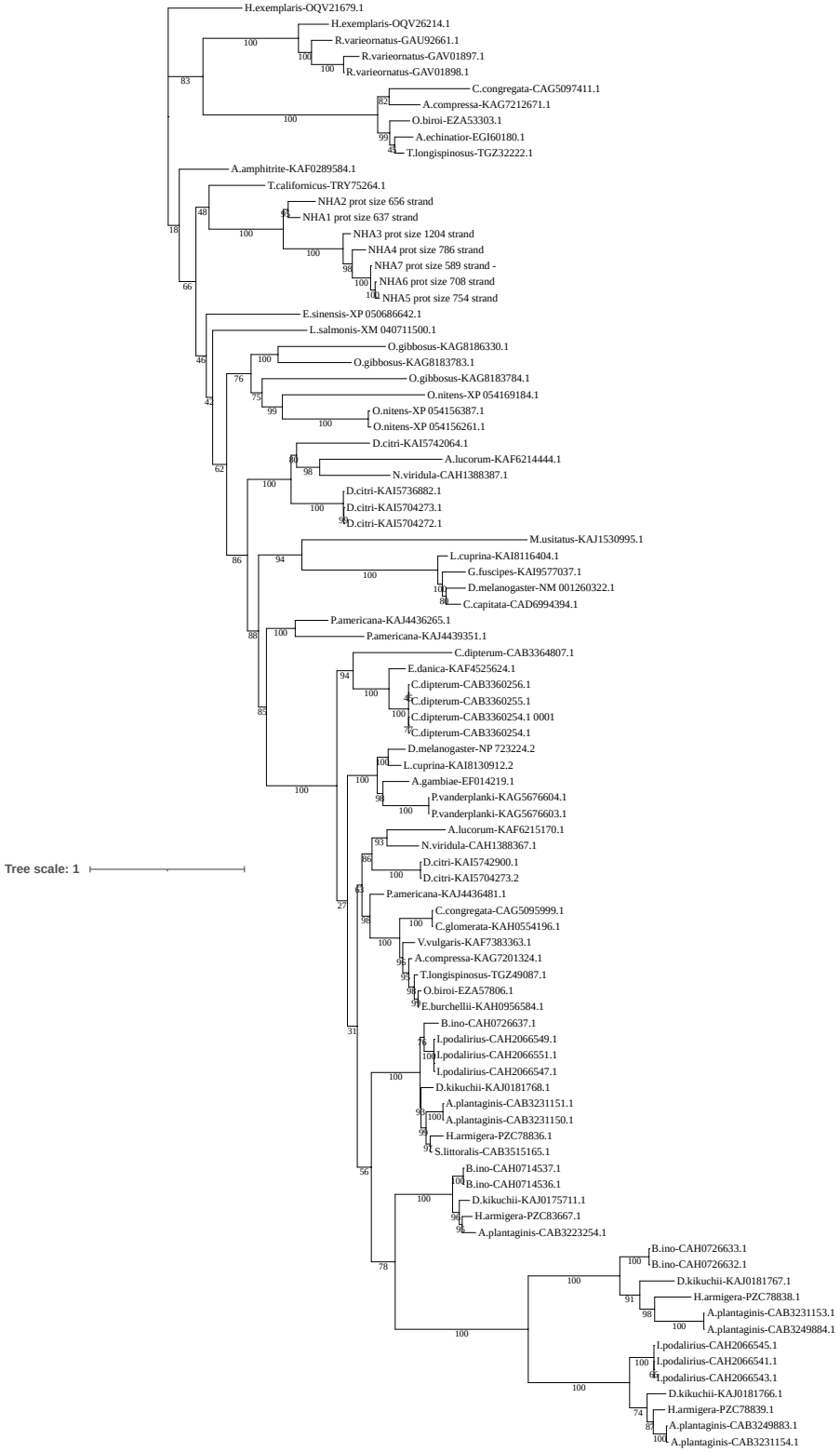

Figure 5. Maximum Likelihood phylogeny of NHA made using IQTree. It contains bootstrap values.

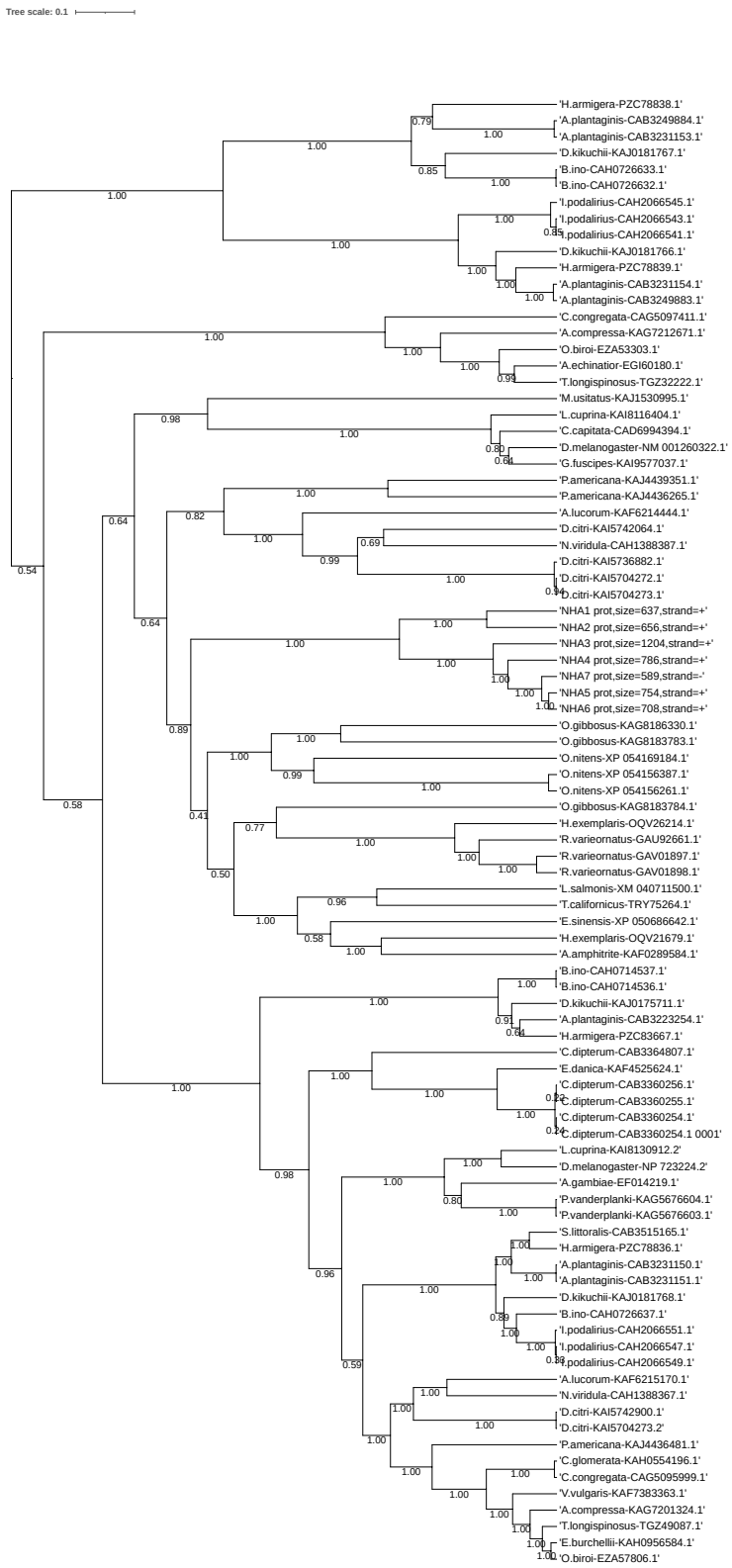

Figure 6. Bayesian-inferred phylogeny of NHA made using BEAST. It contains posterior probabilities to 2 decimals.

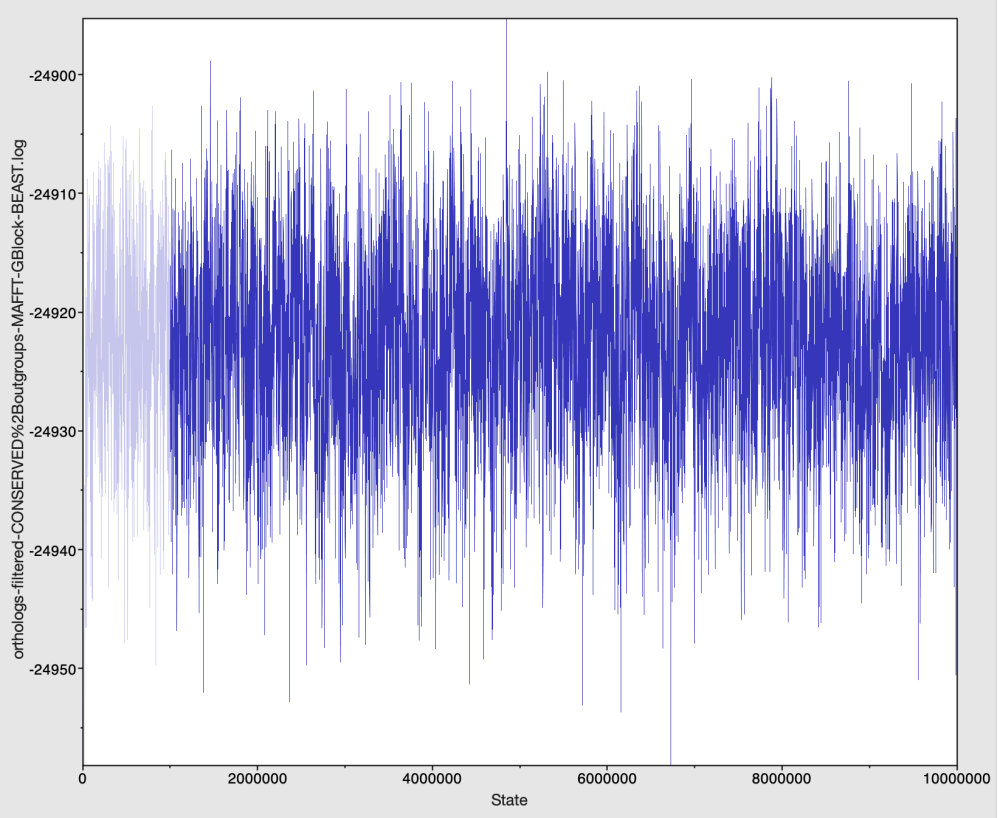

Figure 7. Tracer plot of the joint probability.

Discussion

In the future I would like to expand on the diversity of subphyla of my dataset. For example, due to my quality constraints and the availability of data, the majority of my current dataset is comprised of hexapods. Because I am currently investigating copepods, I would like to obtain more quality sequences of different crustaceans, chelicerates and potentially myriapods. I plan to do this by manually annotating potential NHA genes in two other copepod species which my lab is currently developing assemblies for. Likewise, it would be interesting to investigate the potential third gene within the CPA clade. Hemipterans and hymenopterans are divided into two.

Data collection and data curation are indeed the slowest step of the methodology used in this paper. Having said this, it is imperial that the curation step be so stringent, because future analyses regarding estimating ancestral sequences, order of mutations, protein modelling, and most importantly, signals of selection will depend on frequencies of synonymous and non-synonymous changes (Kryazhimskiy and Plotkin, 2008). As such, having assembly or

sequencing errors can have a drastic effect in the false positive rates. Moreover, a study by Wong et al. (2008) found a potential bias in phylogeny reconstructions. By comparing different MSA algorithms, they found that 46.2% of 1502 alignments of the same sequences produced one or more conflicting trees, implying a potential bias surging through the multiple sequence alignment step. Likewise, this error carried on to future steps, like inferring evolution parameters such as signals of selection, substitution rate, etc., that were also in conflict. Having this in mind for the future, I plan on reviewing and trying alternative curation techniques to optimize sequence quality, while retaining more crustacean and chelicerate species.

Literature Cited

Anderegg MA, Gyimesi G, Ho TM, Hediger MA, Fuster DG. The Less Well-Known Little Brothers: The SLC9B/NHA Sodium Proton Exchanger Subfamily-Structure, Function, Regulation and Potential Drug-Target Approaches. Front Physiol. 2022 May 25;13:898508. doi: 10.3389/fphys.2022.898508. PMID: 35694410; PMCID: PMC9174904.

Ayres DL, Cummings MP, Baele G, Darling AE, Lewis PO, Swofford DL, Huelsenbeck JP, Lemey P, Rambaut A, Suchard MA. BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics. Syst Biol. 2019 Nov 1;68(6):1052-1061. doi: 10.1093/sysbio/syz020. PMID: 31034053; PMCID: PMC6802572.

Camacho, C., Coulouris, G., Avagyan, V. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). https://doi.org/10.1186/1471-2105-10-421

Charif D, Lobry J (2007). "SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis." In Bastolla U, Porto M, Roman H, Vendruscolo M (eds.), *Structural approaches to sequence evolution: Molecules, networks, populations*, series Biological and Medical Physics, Biomedical Engineering, 207-232. Springer Verlag, New York. ISBN : 978-3-540-35305-8.

Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000 Apr;17(4):540-52. doi: 10.1093/oxfordjournals.molbev.a026334. PMID: 10742046.

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ. Biopython: freely available Python tools for computational

molecular biology and bioinformatics. Bioinformatics. 2009 Jun 1;25(11):1422-3. doi: 10.1093/bioinformatics/btp163. Epub 2009 Mar 20. PMID: 19304878; PMCID: PMC2682512.
Di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang JM, Taly JF, Notredame C. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W13-7. doi: 10.1093/nar/gkr245. Epub 2011 May 9. PMID: 21558174; PMCID: PMC3125728.
Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007 Nov 8;7:214. doi: 10.1186/1471-2148-7-214. PMID: 17996036; PMCID: PMC2247476.
Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001 Aug;17(8):754-5. doi: 10.1093/bioinformatics/17.8.754. PMID: 11524383.
Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002 Jul 15;30(14):3059-66. doi: 10.1093/nar/gkf436. PMID: 12136088; PMCID: PMC135756.
Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. PLoS Genet. 2008 Dec;4(12):e1000304. doi: 10.1371/journal.pgen.1000304. Epub 2008 Dec 12. PMID: 19081788; PMCID: PMC2596312.
Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 2020 Jan 8;48(D1):D265-D268. doi: 10.1093/nar/gkz991. PMID: 31777944; PMCID: PMC6943070.
McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).
Morgulis A., Coulouris G., Raytselis Y., Madden T.L., Agarwala R., Schaffer A.A. (2008) "Database indexing for production MegaBLAST searches." Bioinformatics 15:1757-1764. PubMed

Needleman, Saul B. & Wunsch, Christian D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology*. **48** (3): 443–53. doi:10.1016/0022-2836(70)90057-4. PMID 5420325.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015 Jan;32(1):268-74. doi: 10.1093/molbev/msu300. Epub 2014 Nov 3. PMID: 25371430; PMCID: PMC4271533.

Smyth, Katie, and Mike Elliott, 'Effects of changing salinity on the ecology of the marine environment', in Martin Solan, and Nia Whiteley (eds), *Stressors in the Marine Environment: Physiological and ecological responses; societal implications* (Oxford, 2016; online edn, Oxford Academic, 19 May 2016), https://doi.org/10.1093/acprof:oso/9780198718826.003.0009, accessed 27 Feb. 2023.

Stern DB, CE Lee. 2020. Evolutionary origins of genomic adaptations in an invasive copepod. *Nature Ecology & Evolution*. 4:1084-1094. doi: 10.1038/s41559-020-1201-y

Stern DB, NW Anderson, JA Diaz, CE Lee. 2022. Genome-wide signatures of synergistic epistasis during parallel adaptation in a Baltic Sea copepod. *Nature Communications*. 13:4024. doi: 10.1038/s41467-022-31622-8

Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007 Aug;56(4):564-77. doi: 10.1080/10635150701472164. PMID: 17654362.

Okonechnikov K, Golosova O, Fursov M; UGENE team. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012 Apr 15;28(8):1166-7. doi: 10.1093/bioinformatics/bts091. Epub 2012 Feb 24. PMID: 22368248.