Title: NHA gene family in Arthropods

Author: 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, and other ion transporter families, 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. As such, starting this summer, I plan to sequence and analyze single copepods from populations currently undergoing selection in response to salinity and temperature change in the laboratory evolution experiment. The goal is to observe frequency shifts of alleles of NHA gene paralogs under selection. I plan to determine the protein structural changes of the beneficial alleles, relative to the alleles that go extinct in the experiment, and functional consequences of the mutations favored by selection. I will also determine which alleles are favored by selection across different ion transporter paralogs.

This study will address fundamental questions of molecular evolution, including the topic of evolutionary history and the effect of mutations on protein evolution. This project also has real world applications regarding the biology of invasive populations and the future of population survival under global warming.

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. Because of their high level of parallel signal of selection, they are a likely candidate to explain the copepod population’s abilities to adapt into freshwater.

Finally, from these phylogenetic data I will also extract the order in which these mutations arise. Determining the temporal order of mutations across the phylogeny would enable us to understand how permissive mutations might explain the evolution of high specificity between ion channels and substrates necessary in physiological pathways (Ortlund et al., 2006, Bridgham et al., 2007). Permissive mutations are mutations that do not have a direct impact on the protein, but they stabilize the structure of the protein, which allows for future direct functional changes.

Materials and Methods

Collecting sequences

For collecting the sequences present in my analysis I had two major sources of data: (1) genome and transcriptome data on *E. affinis* (Du, 2023 - unpublished), and (2) the ncbi Genbank database. I have used BLAST+ 2.6.0 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

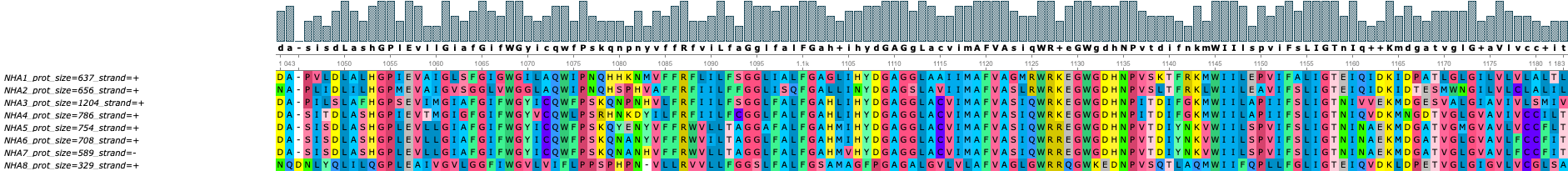


Figure 1. Conserved region of all 8 paralogue sequences alignment in *E. affinis*.

Diagram

Description automatically generated with medium confidence

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

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 against the non-redundant BLAST database (Morgulis, 2008), using the scoring matrix BLOSUM62 and BLOSUM45, and excluding XP/XM sequences. The main reason for this latter criterion is because not only 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, but they are also erroneous, especially when comparing XP/XM sequences. (They may be labeled as NHA, when they are a completely different gene). Entrez.entries also do not contain the CDS sequence, making it impossible to be obtained through the API. Assemblies from XP sequences often do not show as potential results when looking in the NCBI database. So if this information is not contained within the first protein entry page (it is often not present), it is impossible to obtain this information through an automated process.

In summary, these extensive BLAST searches yielded a total of 214 results. I then used the Biopython (Cock, 2009) library, and more specifically the Entrez module as a NCBI database API. By using this tool, I was able to recover in-depth information about these proteins, as well as their genome assembly.

Curation of orthologs

Although the sequences available were already filtered before being made available on Genbank, the public platform still contains numerous inconsistencies that may compromise the study (SEE IF THERE’S ANY OFFICIAL REPORT ON THIS?). Moreover, annotations of proteins and genes are often erroneous. Moreover, because Genbank is a public database, it also contains several unreliable sequences. For this reason, I used the pandas library (McKinney, 2010) to filter all proteins results according to their genome assembly quality. To reinforce this quality check, I used the following criteria: minum 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 (CITATION??????), 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.

For a third curation step, I used the conserved domain database analysis (Lu, 2020) available on NCBI to verify that the remaining sequences were truly NHA. Using the *D. melanogaster* NHA2 (NP\_001247251.1), as a reliable NHA, 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. I then selected and obtained all remaining confirmed high quality NHAs for every species. Additionally, 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*.

Finally, for the last curation step, I also added two species of tardigrades, 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. On average, there were 2.37 NHA paralog per species.

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 2. Paralogs across species

Finally, 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.

Alignment

For the alignment of the 83 sequences, I have used MAFFT (Katoh, 2002). This yielded a clear 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. I used 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.

TALK ABOUT WHY MAFFT – STRENGHTS, WEAKNESSES AND ASSUMPTIONS.

A picture containing graphical user interface

Description automatically generated

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

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 because it is still considered one of the most accurate maximum likelihood models, it was easy to install, and it could also automatically perform tests and use best-fit evolutionary models.

IQTree (Nguyen, 2014), uses a combination of hill climbing algorithms, random perturbation and a large range of starting trees. Ever since its release, IQTree is still considered one of the most, if not the most, accurate maximum likelihood phylogenetic tool available. It also was easy to install, and it actually 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 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, and I also used *Hypsibius examplaris* (OQV21679.1) 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. As such, after filtering these paralogs for the NA\_H\_exchanger domain, only some remained.

Maximum Likelihood approach

A picture containing table

Description automatically generated

Figure 4 Outgroups

Bayesian approach

For Bayesian, I have honestly 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 (CITATION). 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. Like all tree building methods, it still assumes… necessary to compute and estimate mutation rates from substitution and evolution models.

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



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.

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.

A picture containing background pattern

Description automatically generated

Figure 7. Tracer plot of the joint probability.

Discussion

For both trees, lepidopterans seem to be most divergent. Could be a sign of rapid divergence. Hemipterans, Hymenopterans and dipterans. In both trees, chelicerates

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 (CITATION?). As such, having assembly or sequencing errors can have a drastic effect in the false positive rates (CITATION?). 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. Moreover, they also warned of the potential danger of trimming poor aligned regions from the alignments, claiming that they can have an additive effect when dealing with insertion and deletions, causing later SNPs to be excluded from analysis. They finally show that even after trimming alignments, discordant trees are still produced.

Literature Cited

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, 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](https://www.ncbi.nlm.nih.gov/pubmed/18567917?dopt=Citation)

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](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1016/0022-2836(70)90057-4](https://doi.org/10.1016%2F0022-2836%2870%2990057-4). [PMID](https://en.wikipedia.org/wiki/PMID_(identifier)) [5420325](https://pubmed.ncbi.nlm.nih.gov/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.

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.