microPublication BIOLOGY

8/7/2023 - *Open Access*

Computational analysis of variation in *C. elegans ugts*

Muhammad Zaka Asif[1,2*], Maci C. Benveniste[2,3*], Kyra D. Chism[2,3*], Ari L. Levin[2*], Deanna Lanier[2,4], Rockford E. Watkins[1,2], Rahil Taujale[2,4], Niyelle Tucker[2,3], Arthur S. Edison[1,2,4§]

[1]Department of Biochemistry & Molecular Biology, University of Georgia, Athens, Georgia, United States

[2]Complex Carbohydrate Research Center, University of Georgia, Athens, Georgia, United States

[3]Department of Genetics, University of Georgia, Athens, Georgia, United States

[4]Institute of Bioinformatics, University of Georgia, Athens, Georgia, United States

[§]To whom correspondence should be addressed: aedison@uga.edu

[*]These authors contributed equally.

Abstract

Caenorhabditis elegans are free-living nematodes with a relatively short life cycle and a wealth of genomic information across multiple databases. Uridine diphosphate-glycosyltransferases (UGTs) are a family of enzymes involved in Phase II modification of xenobiotics in *C. elegans*, which is the addition of a sizeable water-soluble molecule to a xenobiotic to allow for its excretion out of a cell. Little is known about the variation in UGTs across wild isolates and how that might affect their innate immune response. We analyzed the diversity in *ugt* genes across *C. elegans* isolates from different geographical locations from the *Caenorhabditis elegans* Natural Diversity Resource (CaeNDR) database. This was accomplished using whole genome data and data identifying genome regions as hyper-divergent for each isotype. We implemented three steps to identify *ugt* genes and make inferences based on their variation. First, we created a catalog of UGTs in the N2 reference strain and used them to create a phylogenetic tree that depicts the relationships between the UGT protein sequences. We then quantified *ugt* variation using the strains from the CaeNDR database and used their data to remove hyper-divergent *ugt* genes. The third step was to catalog the occurrence of minor allele frequency (MAF) > 0.05 for all the *ugts* to compare how that aligned with genes classified as hyper-divergent by CaeNDR. Of the 67 *ugt* genes analyzed, 18 were hyper-divergent. This research will help improve our understanding of *ugt* variation in *C. elegans*.

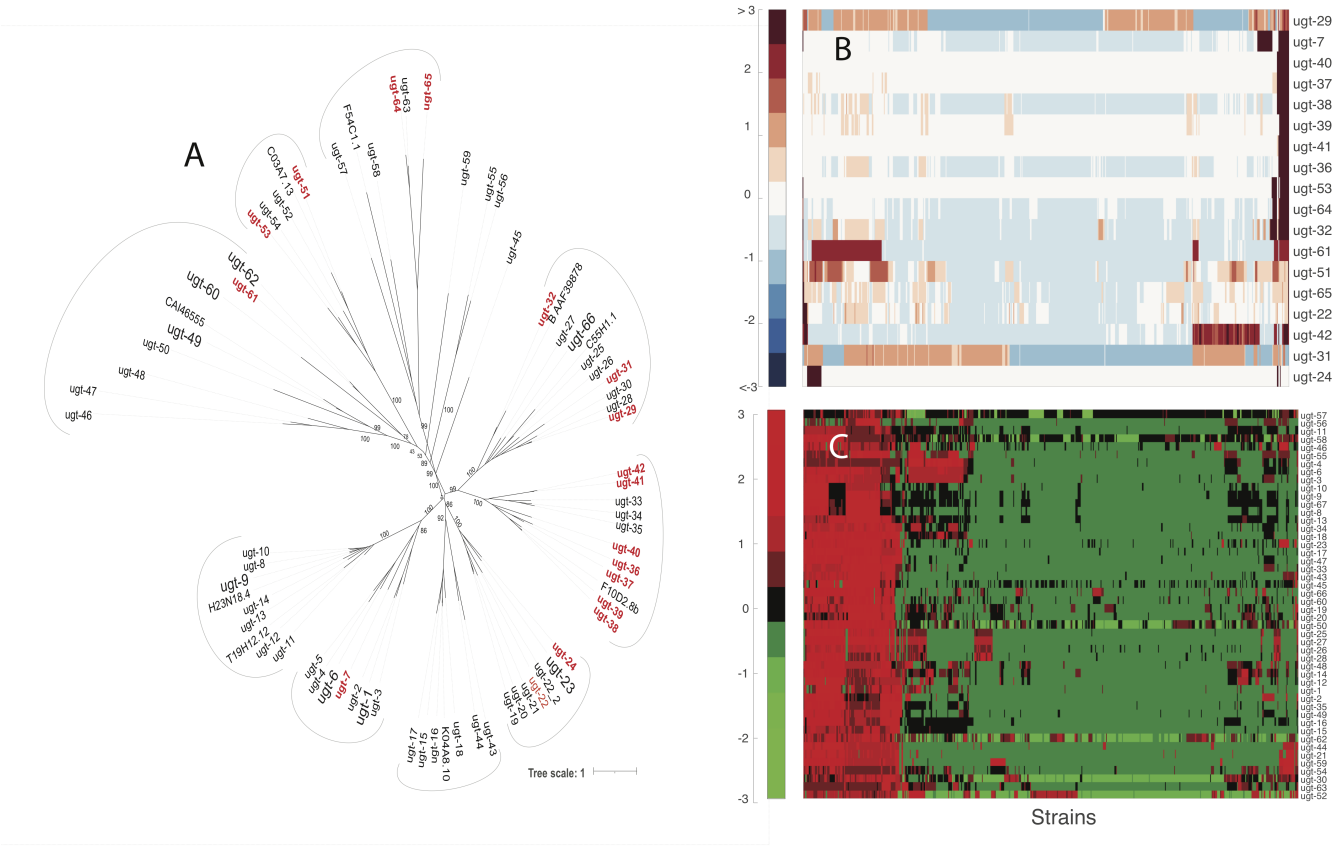

Figure 1. *ugt* variation in *C. elegans*:

A: Phylogenetic tree of the known UGTs in *C. elegans*. The hyper-divergent UGTs are enlarged and highlighted in red. **B**: Hierarchical cluster of the known hyper-divergent *ugt* genes. A heatmap of the z-score of the variation is plotted. **C**: Hierarchical cluster of the non-hyper-divergent *ugt* genes. A heatmap of the z-score of the variation is plotted.

Description

C. elegans has about 250 glycosyltransferases (Kellokumpu, Hassinen, & Glumoff, 2016), and the *ugt* family of 67 genes are responsible for the glycosylation of small molecule xenobiotics (Asif et al., In Preparation; Hartman et al., 2021; Laing et al., 2010; Liu, Samuel, Breen, & Ruvkun, 2014). We quantified the variation in 67 *ugt* genes across *C. elegans* isotypes using N2 as the reference strain (Cook, Zdraljevic, Roberts, & Andersen, 2017). Regions with higher-than-average concentrated genomic variation than N2 are hyper-divergent (Lee et al., 2021). According to Lee et al., hyper-divergent regions had nine consecutive bins of over 1kb equal to 16 single nucleotide variants (SNVs)/indels or lower than 35% read depth to the genome-wide average (Lee et al., 2021). These hyper-divergent regions in their respective UGTs were identified and removed from our analysis.

We used the Multiple Alignment using Fast Fourier Transform **(**MAFFT) tool first to align the amino acid sequences into a multiple sequence alignment which was then used to generate a phylogenetic tree via the iqtree tool (Fig.1A) (Nguyen, Schmidt, von Haeseler, & Minh, 2015). This phylogenetic tree groups evolutionarily related UGTs into clades that can be used to infer functional similarities. Ten clades were identified, providing an evolution-evidenced grouping of functionally related UGTs. Seven of the ten clades had at least one hyper-divergent UGT (shown in red in Figure 1A), and most were grouped into a single clade.

Figure 1B shows hyper-divergent genes that were removed from our analysis. We identified hyper-divergent genomic regions from Lee et al., which were defined as hyper-divergent regions in more than 5% of all isotypes. From that, we identified *ugt*s that lay within these regions. Those *ugt*s were defined as hyper-divergent. We verified hyper-divergent genes by looking at the minor allele frequency (MAF) > 0.05 for SNVs and found that the genes with the highest numbers of SNVs with MAF > 0.05

tended to be hyper-divergent. It is important to note that genes not classified as hyper-divergent by Lee et al. still have some isotypes with MAF > 0.05 bases.

Using the cluster gram function in MATLAB™, we visualized the number of mutations in each hyper-divergent *ugt*. Z-score normalization was performed on the *ugts*, and a Euclidean distance metric was employed to measure similarity or dissimilarity based on the magnitude of differences between *ugts* and isotypes. The resulting standardized and clustered data were represented as a heatmap in Figure 1B. This heatmap's dark colors represent values greater than three and lower than -3.

Figure 1C is the cluster gram for the non-hyper-divergent genes, as described above. The red-green colormap was chosen to contrast with Figure 1B of the hyper-divergent regions. The color bar indicates the z-score of the variation for the *ugts*. The non-hyper-divergent genes had a lower number of mutations than the hyper-divergent genes. The gene with the highest number of mutations was *ugt-12*. Furthermore, the isotypes with the highest frequency of mutations (shown in the red region to the left in Figure 1C) mostly are from Hawaii, indicating that isolation from other *C. elegans* isotypes allows for more divergent evolution.

Non-hyper-divergent *ugts* are an area of interest for future studies. Given the quantified genomic variation across isotypes from many locations, our results suggest that multiple environmental factors, such as climate, bacteria, pathogens, and environmental toxins, affect the variation.

Methods

Generating the Phylogenetic Tree: We collected 77 UGT amino acid sequences from the publicly available CAZy and Wormbase databases. Next, we used Multiple Alignment using Fast Fourier Transform (MAFFT Alignment) tool to align the amino acid sequences for the UGTs. Then, we generated the phylogenetic tree using the iqtree tool. We visualized the phylogenetic tree using the online tool called the Interactive Tree of Life (iTOL) (Fig. 1A) (Nguyen et al., 2015).

Identifying Genomic Variation of CaeNDR Strains Compared to N2: Using the information gathered above, we generated a Python script in Jupyter Notebook™ to parse CaeNDR's hard-filtered variants vcf file ***WI.20220216.impute.isotype.vcf.gz*** (released 20200815) and extracted the number of variants and location of mutations in *ugt* regions for 550 isotypes compared to the N2 reference genome. The genomes of the isotypes were aligned and compared to the N2 genome.

Removal of Hyperdivergent Regions from Analysis: Using the CaeNDR hyper-divergent region data file (20220216.bed), we created a Python script using Jupyter Notebook™ to determine which *ugt*s had hyper-divergent isotypes. Our data included regions that partially fell in a hyper-divergent range or had complete overlap. A table was created with our data. We further separated it into two Excel files containing non-hyper-divergent and hyper-divergent strains and UGTs with the number of base pair mutations across isotypes. If a gene from an isotype partially or fully fell into a hyper-divergent region, it was considered hyper-divergent for analysis purposes. All others were considered non-hyper-divergent.

Creation of Heatmap: Once the hyper-divergent regions were identified, a spreadsheet was created for both hyper-divergent and non-hyper-divergent genes. Both files contained the *ugt* names on the rows and the strain names in the columns. The total number of nucleotide variations in each strain for each *ugt* was also added. The non-hyper-divergent and hyper-divergent spreadsheet files were added to a working MATLAB® script. They were then used to create two cluster grams to help visualize variation trends, if any. (Fig. 1B and Fig. 1C). Figure 1B was given a red-blue color map, whereas 1C was given a red-green to help differentiate between the data. All scripts are available on GitHub.

Acknowledgements: We thank Prof. Erik Andersen, Orr Shalev, Jacob Salomon, Joshua Eli Mermelstein, Jacob "Slav" Slavkin, Benjamin Surasky, Megan McElroy, Sheza Mehdi, Eli Benveniste, Aleya Johnson, Bailey Nicolas, and Hao Nguyen for their assistance in our project. We would also like to acknowledge the Vertically Integrated Projects (VIP) Team under the Edison Lab from The University of Georgia.

References

Asif, M. Z., Nocilla, K. A., Ngo, L. T., Shah, M. K., Smadi, Y., Hafeez, Z. A., . . . Edison, A. S. (Unpublished Results). Role of ugt genes in detoxification and glycosylation of 1-hydroxyphenazine (1-HP) in Caenorhabditis elegans.

Cook DE, Zdraljevic S, Roberts JP, Andersen EC. 2017. CeNDR, the Caenorhabditis elegans natural diversity resource. Nucleic Acids Res 45: D650-D657. PubMed ID: 27701074

Hartman JH, Widmayer SJ, Bergemann CM, King DE, Morton KS, Romersi RF, et al., Meyer JN. 2021. Xenobiotic metabolism and transport in Caenorhabditis elegans. J Toxicol Environ Health B Crit Rev 24: 51-94. PubMed ID: 33616007

8/7/2023 - *Open Access*

Kellokumpu S, Hassinen A, Glumoff T. 2016. Glycosyltransferase complexes in eukaryotes: long-known, prevalent but still unrecognized. Cell Mol Life Sci 73: 305-25. PubMed ID: 26474840

Laing ST, Ivens A, Laing R, Ravikumar S, Butler V, Woods DJ, Gilleard JS. 2010. Characterization of the xenobiotic response of Caenorhabditis elegans to the anthelmintic drug albendazole and the identification of novel drug glucoside metabolites. Biochem J 432: 505-14. PubMed ID: 20929438

Lee D, Zdraljevic S, Stevens L, Wang Y, Tanny RE, Crombie TA, et al., Andersen EC. 2021. Balancing selection maintains hyper-divergent haplotypes in Caenorhabditis elegans. Nat Ecol Evol 5: 794-807. PubMed ID: 33820969

Liu Y, Samuel BS, Breen PC, Ruvkun G. 2014. Caenorhabditis elegans pathways that surveil and defend mitochondria. Nature 508: 406-10. PubMed ID: 24695221

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32: 268-74. PubMed ID: 25371430

Funding: Funding came from the Georgia Research Alliance.

Author Contributions: Muhammad Zaka Asif: conceptualization, formal analysis, investigation, methodology, project administration, validation, visualization, writing - original draft, writing - review editing. Maci C. Benveniste: data curation, formal analysis, investigation, methodology, validation, visualization, writing - original draft. Kyra D. Chism: data curation, formal analysis, investigation, methodology, validation, writing - original draft, writing - review editing. Ari L. Levin: data curation, formal analysis, investigation, methodology, validation, visualization, writing - review editing. Deanna Lanier: data curation, investigation, methodology, validation, writing - review editing. Rockford E. Watkins: data curation, formal analysis, investigation, methodology, writing - review editing. Rahil Taujale: conceptualization, data curation, formal analysis, investigation, methodology, project administration, validation, writing - review editing. Niyelle Tucker: investigation, methodology, writing - review editing, formal analysis. Arthur S. Edison: conceptualization, funding acquisition, project administration, supervision, writing - review editing.

Reviewed By: Anonymous

WormBase Paper ID: WBPaper00065839

History: Received March 25, 2023 **Revision Received** July 3, 2023 **Accepted** August 4, 2023 **Published Online** August 7, 2023 **Indexed** August 21, 2023

Citation: Asif, MZ; Benveniste, MC; Chism, KD; Levin, AL; Lanier, D; Watkins, RE; et al.; Edison, AS (2023). Computational analysis of variation in *C. elegans ugts*. microPublication Biology. 10.17912/micropub.biology.000819