Comparative Genomics of Mosquito Species:

Orthologous Groups

Vanesa Robledo (21041758)

December 8th, 2023

BIOL 469

Methods

Gene set comparison

Then we used OrthoFinder, a command line tool written in Python that determines orthologous groups of genes between species (Emms et al., 2019). OrthoFinder also creates gene trees and infers gene duplication events. OrthoFinder takes an input of FASTA files containing proteins for each species. This analysis used the protein files for the *Aedes aegypti*, *Anopheles moucheti*, and *Culex pipiens* genomes. The inferred groups of orthologous genes between species, or "orthogroups", found in `\Orthogroups\Orthogroups.tsv` of the results directory were used. This is a tab-delimited file that lists the RefSeq IDs (from the NCBI FASTA files) of proteins of each orthogroup for each species.

To analyze these orthogroups, these IDs were mapped using the UniProt ID Mapping tool to find more information about these proteins in the UniProt database. Specifically, gene ontology terms, or "GO terms", could be found in the UniProt database. We used a custom Python script to take the input file of orthogroups and automate this process to annotate gene ontology terms for each orthogroup. Lastly, these annotated orthogroup data and other OrthoFinder result data was analyzed using R with the `tidyverse` library, which is a collection of R packages for data analysis that includes `dplyr` and `ggplot`, to compare the gene ontology terms between orthogroups and species (see Appendix A3) for statistical analysis.

Results

Orthologous groups

Between *Aedes aegypti*, *Anopheles moucheti*, and *Culex pipiens*, there were 69,765 genes that were analyzed to find orthologs and paralogs. Orthofinder found 12,952 orthogroups between the three species. Most genes (93.9%) were assigned to an orthogroup. Of these groups, 1,616 were species-specific. All species were present in 9,118 orthogroups, with slightly more of the orthogroups being shared between *A. aegypti* and *A. culex*. There was an average of 5.1 genes in each orthogroup, with 3,676 orthogroups containing a single copy.

Among all of the orthogroups, 79.7% of listed proteins were mapped to proteins found in the UniProt Database. A subset of these proteins (22.5%) was listed as "uncharacterized protein". Gene duplications of transferrin and NOS were found in *A. aegypti* (Appendix A8 & A9). Orthologs for transferrin and NOS were not mapped to the UniProt database, so GO functions were analyzed. The top GO terms found for these proteins included ATP binding, metal ion binding, DNA binding and repair, and other characteristic shared function between eukaryotic species. GO terms that were related to disease transmission were also noted. GO terms related to transferrin include heme binding (GO:0020037), which was the 14th most common term with 83 annotations, and iron ion binding, which was the 16th most common term with 74 annotations. There were also orthologs related to the olfactory system. GO terms included odorant binding (GO:0005549), which was the 21st most common term with 50 annotations, and less commonly, olfactory receptor activity (GO:0004984) with 16 annotations. There were 14 proteins identified as "odorant receptor proteins".

Between species-specific orthogroups, there are more shared GO term functions between *A. aegypti* and *C. pipiens*, but a higher frequency of the same GO terms between *A. moucheti* and *C. pipiens* (Figure 2). In the latter group, there is a high frequency of the molecular GO terms structural constituent of cuticle (GO:0042302), serine-type endopeptidase activity (GO:0004252), RNA binding (GO:0003723),

olfactory receptor activity (GO:0004984), odorant binding (GO:0005549), metal ion binding (GO:0046872), and DNA binding (GO:0003677). There were no mapped functions found between *A. moucheti* and *C. pipiens*, which is most likely due to a lack of functional annotation and information in the database rather than a biologically significant result.

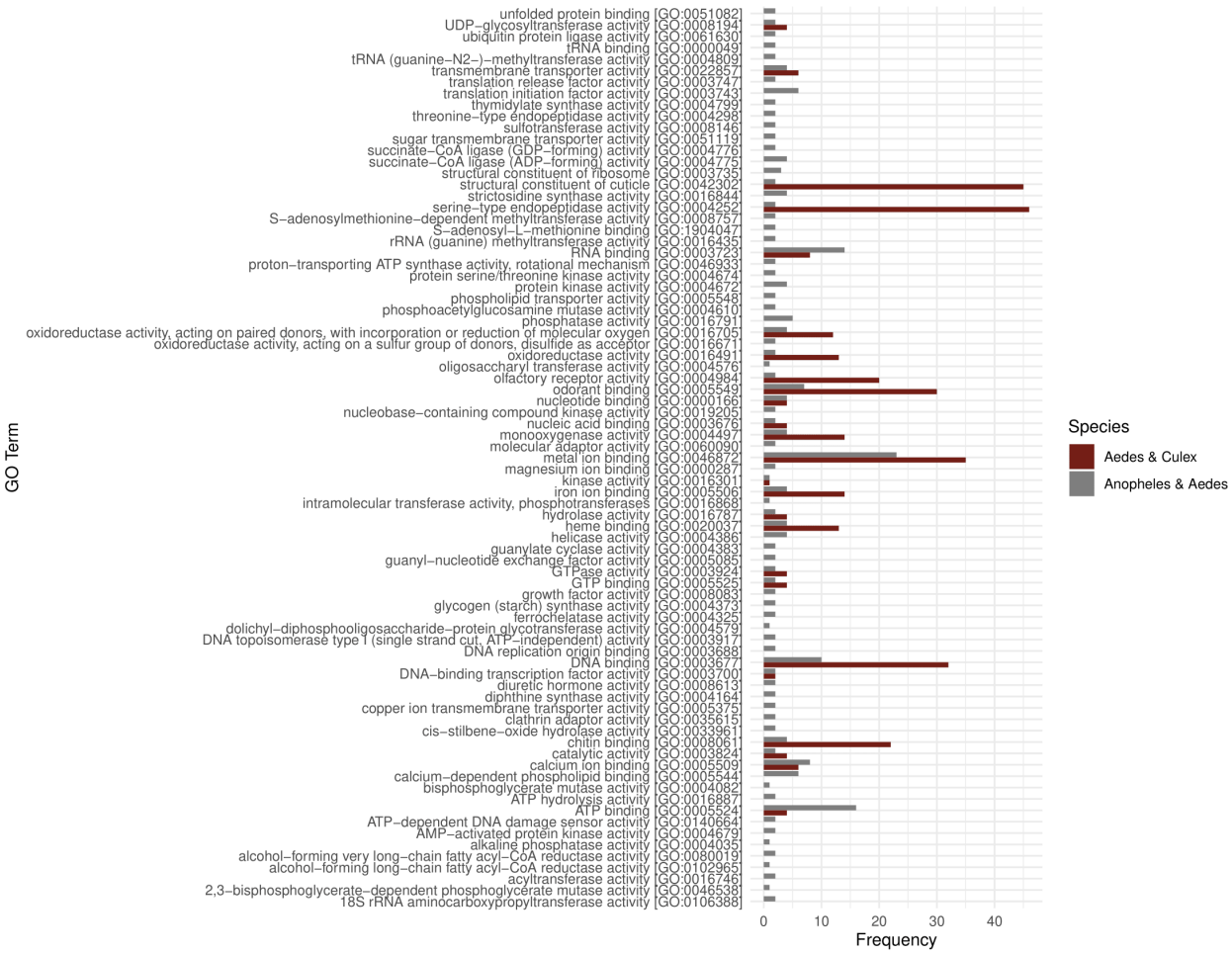


Figure 2. Molecular GO terms between species orthogroups: *Aedes* and *Culex* are in red, and *Anopheles* and *Aedes* GO terms are in grey.

Using our custom script, we were able to identify GO terms for 895 orthogroups. Of these groups, 760 had molecular GO term annotations and 508 had biological process GO term annotations, with 373 groups having both. 243 groups had one molecular GO term. Similar to looking at all of the orthogroups as a whole, the most frequent molecular GO terms in orthogroups related to ATP binding, metal ion binding, and DNA binding (Figure 3). There were 22 groups containing both iron ion and heme binding

terms, 18 groups related to odorant binding, and 9 related to olfactory receptor activity. Transferase activity (GO:0016740), related to transferrin, was also found as a top molecular GO term, occurring in 14 groups. There were less biological process annotations, most of which related to regular cell function.

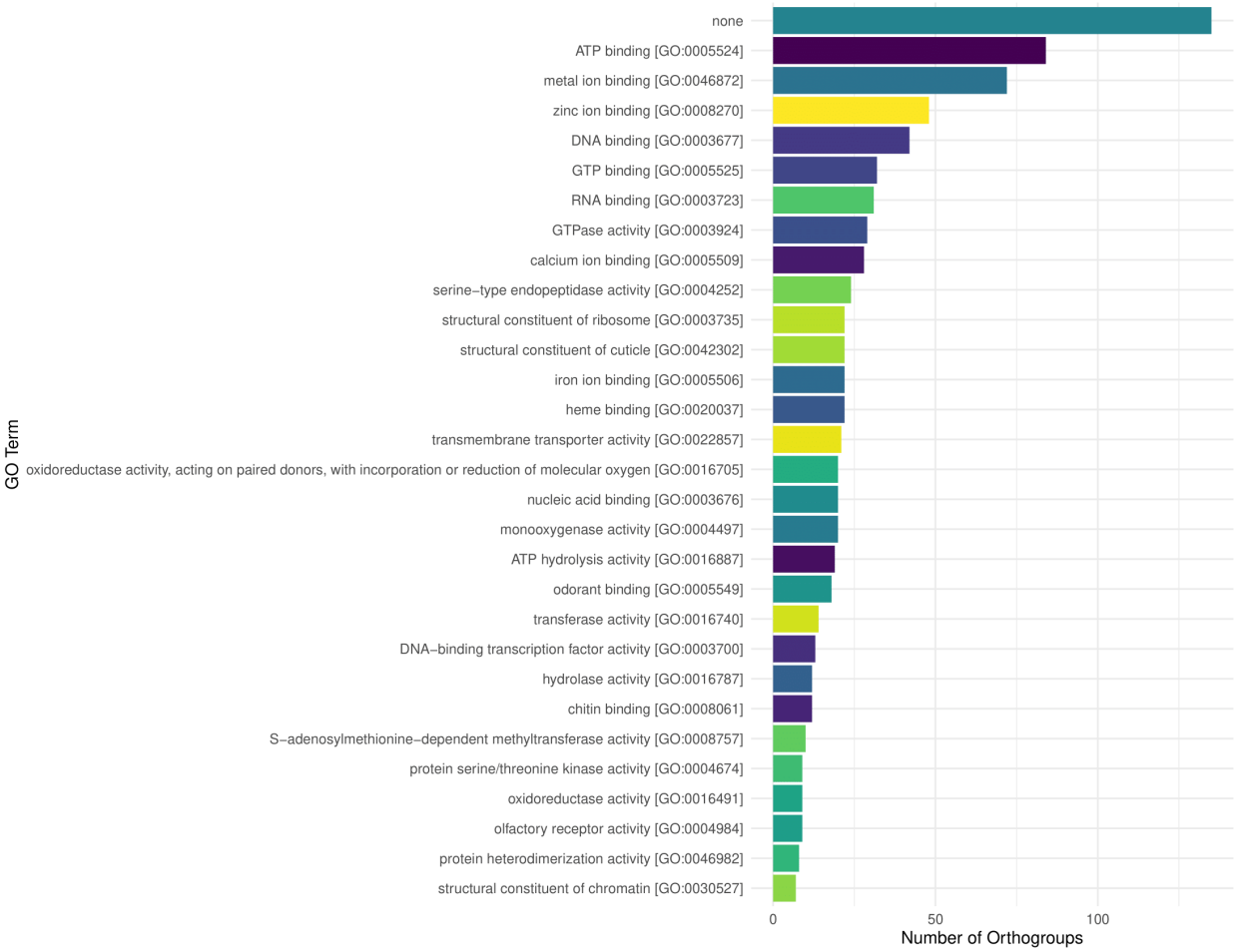


Figure 3. Top 30 molecular GO terms by number of orthogroups that contain them.

Discussion

Functional comparison of mosquitos

Many orthologs related to blood iron transport were found between the mosquitos. This may show that the role of transferrin may be very similar among all mosquito species regardless of whether they carry malaria, or other diseases such as yellow fever or West Nile. *A. aegypti* had many gene duplications for both transferrin and NOS that may reflect a difference in carrying more diseases than *Anopheles* and *Culex* species (Li et al., 2003). There were also many shared olfactory function GO terms and proteins between mosquito species, particularly between *A. aegypti* and *C. quinquefasciatus*. This suggests that the olfactory systems in disease-transmitting mosquitos are similar (Gene Ontology Consortium, 2004). Mosquitos rely on chemosensory cues such as odor for host-seeking. Mosquito's olfactory systems are highly complex and require more study to characterize (Wheelwright et al., 2021). The existence of these orthologs imply that there are many similarities regarding seeking human hosts and would require further analysis (Li et al., 2003).

However, the analysis of orthogroups was limited by mapping the RefSeq database to Uniprot and eggNOGG-Mapper. Many proteins were uncharacterized and lacked functional mapping. Less than 10% of shared orthogroups had functional annotations attached to them. More wet lab studies to confirm GO functional terms and database annotation would be needed.

References

Beerntsen, B. T., James, A. A., & Christensen, B. M. (2000). Genetics of Mosquito Vector Competence. *Microbiology and Molecular Biology Reviews*, 64(1), 115–137.

Emms, D.M. & Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20, 238. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1832-y

Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, *32*(suppl_1), D258-D261.

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178-2189.

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., ... & Sherry, S. T. (2022). Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 50(D1), D20.

The UniProt Consortium. (2023). UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1), D523-D531.

Wheelwright, M., Whittle, C. R., & Riabinina, O. (2021). Olfactory systems across mosquito species. *Cell and Tissue Research*, 383(1), 75-90.

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.4, https://github.com/tidyverse/dplyr, https://dplyr.tidyverse.org.

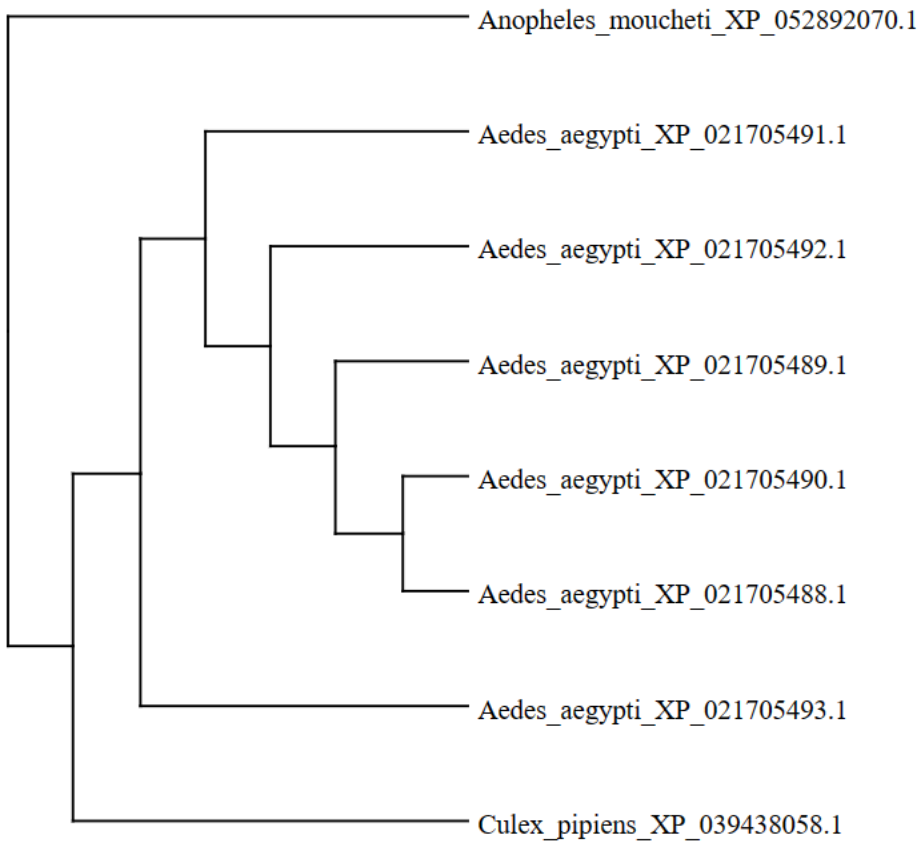
Appendix

A3. Sample of R script used for all data analysis, omitted sections indicated by […].

```
# LOAD LIBRARIES #
library(tidyverse)
library(viridis)


# LOAD DATASETS #
mapped_orthogroups_all <- read.delim("Orthogroups_3_shared_mapped.tsv", sep="\t")
mapped_orthogroups_anopheles_aedes <-
read.delim("Orthogroups_Aedes_Anopheles_mapped.tsv", sep="\t")
mapped_orthogroups_aedes_culex <- read.delim("Orthogroups_Aedes_Culex_mapped.tsv",
sep="\t")
#mapped_orthogroups_anopheles_culex - N/A - none were mapped to UniProt
mapped_orthogroups <- read.delim("Orthologs_shared_mapped.tsv", sep="\t")


[…]

# DATA FUNCTIONS #

# Frequency table: creates a frequency table of terms in a data frame column
delimited by semicolons and converts into data frame
frequency_table <- . %>% strsplit("; ") %>% unlist() %>% table %>%
sort(decreasing=T) %>% as.data.frame(stringsAsFactors = F) %>%
set_names(c("GO.term", "Frequency"))

# DATA PROCESSING #

# mapped_orthogroups_aedes_culex #
# GO terms frequency table
aedes_culex_go_terms_frequency <- frequency_table(aedes_culex_go_terms)

[…]

# mapped_orthogroups #

# Molecular GO Term Table of Orthogroups

mapped_orthogroups_m_terms <- mapped_orthogroups %>%
  select(Orthogroup, go_m, -job_id, -go_b, -go_id) %>%
  separate_rows(go_m, sep = "', |\\\", ") %>%
```

```
  mutate_at("go_m", str_replace, "\\[\\]", "none") %>%
  mutate_at("go_m", str_remove, "^\\['|^'|^\\\"") %>%
  mutate_at("go_m", str_remove, "'\\]$") %>%
  group_by(go_m) %>%
  mutate(orthogroup_list = map_chr(Orthogroup,
~toString(union(Orthogroup, .x)))) %>%
  select(go_m, orthogroup_list, -Orthogroup) %>%
  unique() %>%
  mutate(length = str_count(orthogroup_list, ",")+1) %>%
  arrange(desc(length))

mapped_orthogroups_m_terms
# GRAPHS #

# GO Terms by Species Bar Graph
species_molecular_frequency <- merge(anopheles_aedes_go_molecular_frequency,
aedes_culex_go_molecular_frequency, by="GO.term", suffix = c(".Anopheles.Aedes",
".Aedes.Culex"), all.x=T) %>%
  replace(is.na(.), 0) %>%
  gather(key="Species",value="Frequency", 2:3)

species_molecular_frequency_bargraph <- species_molecular_frequency %>%
  ggplot(aes(x=GO.term, y=Frequency, fill=Species)) +
  geom_bar(stat="identity", position = position_dodge()) +
  theme_minimal() +
  coord_flip() +
  labs(x = "GO Term", y = "Frequency") +
  scale_fill_manual(labels=c("Aedes & Culex", "Anopheles & Aedes"),
values=c('#742017', '#7F7F7F'))

species_molecular_frequency_bargraph

# GO Terms of Orthogroups Bar Graph
# Molecular Terms
mapped_orthogroups_m_frequency_bargraph <- head(mapped_orthogroups_m_terms, 30) %>%
  ggplot(aes(x=fct_reorder(go_m, length), y=length, fill=go_m)) +
  geom_bar(stat="identity", position = position_dodge(), show.legend=F) +
  labs(x = "GO Term", y="Number of Orthogroups") +
  scale_fill_viridis(discrete=T) +
  theme_minimal() +
  coord_flip()

mapped_orthogroups_m_frequency_bargraph
```

A8. Gene tree predicted by Orthofinder of transferrin.



A9. Gene tree predicted by Orthofinder of NOS.