Differential genomic profiling of *Solenopsis invicta* Buren subtypes via gene transcript counter-regulation and functional annotation

Honors Project

In fulfillment of the Requirements for

The Esther G. Maynor Honors College

University of North Carolina at Pembroke

By

Department of Biology

9 May 2015

Marcus D. Sherman — 30 April 2015
Honors College Scholar — Date

Conner Sandefur, Ph.D. — 4/30/2015
Co-Faculty Mentor — Date

Robert Poage, Ph.D. — 4/30/2015
Co-Faculty Mentor — Date

Mark Milewicz, Ph.D. — 4/30/15
Dean, Esther G. Maynor Honors College — Date

ACKNOWLEDGEMENTS

I would first like to thank God and my wife. Without their support, this work and everything would have been impossible.

I would also like to say specific thanks to Drs. Conner Sandefur and Robert Poage. Their steadfast support and mentoring made me not only capable, but also this work possible.

A special thank you to Dr. Mark Milewicz and M. Gordon Byrd for their logistical support in tackling my Honors thesis. Thank you also Joanna Cole, Sonda Rogers, and Sailaja Vallabha.

A special thanks to John Wang of the Keller group at the Université de Lausanne for his correspondence, to which this project would not have been possible.

This work was supported by both grant #5R25GM077634-04 from the NIGMS (National Institute of General Medical Sciences) supporting the UNCP RISE Program.

This work was also supported by the SPIRE grant, #GM000678, from the Training, Workforce Development, & Diversity division of the National Institute of General Medical Sciences

TABLE OF CONTENTS

List of Figures

ABSTRACT

DIFFERENTIAL GENOMIC PROFILING OF *SOLENOPSIS INVICTA* BUREN SUBTYPES VIA GENE TRANSCRIPT COUNTER-REGULATION AND FUNCTIONAL ANNOTATION:

by,

© Copyright by Marcus D. Sherman, 2015
Bachelor of Science in Biology with Biomedical Emphasis
The University of North Carolina at Pembroke
9 May 2015

Solenopsis invicta (red imported fire ant) poses a significant ecological threat to the southeastern United States by way of outcompeting native species and disturbing native ecological communities. The two social forms of red imported fire ants, polygyne (multiple reproducing queens per colony) and monogyne (one reproductive queen per colony), have major morphological and behavioral differences. Polygyne colonies tend to have populations with much greater density, whereas monogyne queens tend to be much larger than polygyne queens. It was hypothesized that a distinct genomic profile could be ascertained linking development with social form. Using publicly available microarray datasets, the data were interrogated using a custom Python pipeline. The pipeline was developed to identify differentially expressed gene transcripts ($p<0.001$) across social forms and developmental stages and characterize the genetic profiles using Gene Ontology (GO). Differentially expressed gene transcripts were found across both social forms and for each age class (pupa, 1 day virgin, 11 day virgin, and fully reproductive). Coupling differentially expressed gene transcripts with GO annotation, relationships were identified to discern a link between social form and age class. Examples of these trends were an enrichment for signal transduction in 11 day virgin queens that was not present elsewhere, and polygyne queens of both 11 day and reproductive age classes up-regulating gene transcripts associated with lipid metabolism and odorant binding while monogyne queens down-regulated these same transcripts. These data, while subject to lack of annotation due to the model organism, indicate relationships between developmental age class and social form.

Introduction

Since its introduction to the southeast United States in the 1930s-1940s from regions in Argentina (Caldera et al., 2008), *Solenopsis invicta* Buren (red imported fire ant hereafter RIFA) has posed a significant ecological threat to the southeastern United States by way of outcompeting native species and disturbing native ecological communities. Outcompeting native invertebrate species is particularly deleterious since native communities play an important role in nutrient cycling, pollination, and seed dispersal behaviors such as myrmecochory conducted by local ant species (Epperson & Allen, 2010). Likewise, the introduction of an invasive species such as the RIFA potentially leads to a trophic simplification of the food web due to ecological disturbance and dominance (Epperson & Allen, 2010).

There are two social forms that characterize RIFA: polygyne (multiple reproducing queens per colony) and monogyne (one reproductive queen per colony). Each social form has major morphological and behavioral differences. Morphologically, monogyne queens tend to be much more massive than their polygyne counterparts. This suggests that gene regulated lipid metabolism pathways are differentially expressed between the two social forms. Behaviorally, polygyne colonies tend to have populations with much greater density, which is only exacerbated by tolerance for non-colony polygyne ants—leading to colony fission, whereas monogyne colonies are characterized by independent colony founding events (LeBrun, Plowes, & Gilbert, 2012). Due to the differing colony founding strategies, both social forms can have profound effects on local ecology. Polygyne colonies have a greater likelihood of severe disturbance in local communities, meanwhile monogyne colonies have a much greater chance of invading intact and undisturbed ecosystems due to independent colony founding behaviors (LeBrun, Plowes, &

Gilbert, 2012). Therefore it is imperative to not only understand the species itself, but also the genetic differences between these two social forms.

It is already known that the genomic profiles of *S. invicta* change based on developmental stage (Wurm, Wang, & Keller, 2010), social class (i.e. queen, male and worker) (Manfredini et al., 2013; Nipitwattanaphon et al., 2014), and social form (i.e. monogyne or polygyne) (Wang et al., 2013). Social events tend to have a profound effect on the gene expression of RIFA in a variety of ways to include development. One such example is that of an orphaning event occurring in a monogyne colony. As the sole reproductive queen dies, the mature virgin queens undergo gene expression changes that present physiologically and phenotypically (Wurm, Wang, & Keller, 2010). Likewise, when monogyne queens leave their colonies on their nuptial flight, if they leave in a group (pleometrosis), the expression profiles of the dominant queen in the group becomes considerably different than that of the 'losing' subordinate queens as the colony founding event proceeds (Manfredini et al., 2013).

In an interesting type of feedback loop, gene expression in RIFA also plays a major part in the generation of the two ant social forms, affecting both physiology and behavior. This is evidenced by recent research that suggests that a large inversion on the 'social' chromosome of RIFA—characterized by the presence of heterozygous or homozygous allelic variants of GP-9 (a general protein linked to odorant/pheromone binding) (Wang et al., 2013). This non-recombining region is termed a "supergene" that has over 600 associated genes (to include GP-9) linked to factors that give rise to the two social forms (Wang et al., 2013).

However, as of yet, we lack an understanding of how social form and development are linked in RIFA. We hypothesized that there are distinct developmental changes in gene expression between monogyne and polygyne social forms. A major hindrance to investigating gene function changes on a system-wide scale is the lack of functional annotation for *S. invicta.* To test this hypothesis, an open-sourced pipeline was developed to investigate gene expression differences in publically available microarray data. The pipeline then searched for differentially expressed genes between and across developmental state and social form, which were annotated based on protein homology. This research could potentially re-inform ongoing RIFA research with regard to pest management strategies and/or basic science pertaining to RIFA as a social organism.

Methods & Materials

Experimental Data

We used the *S. invicta* gene transcript expression levels from the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) data sets series GSE42062 and SuperSeries GSE42786 (Nipitwattanaphon et al., 2012). We isolated an equal amount of sample expression

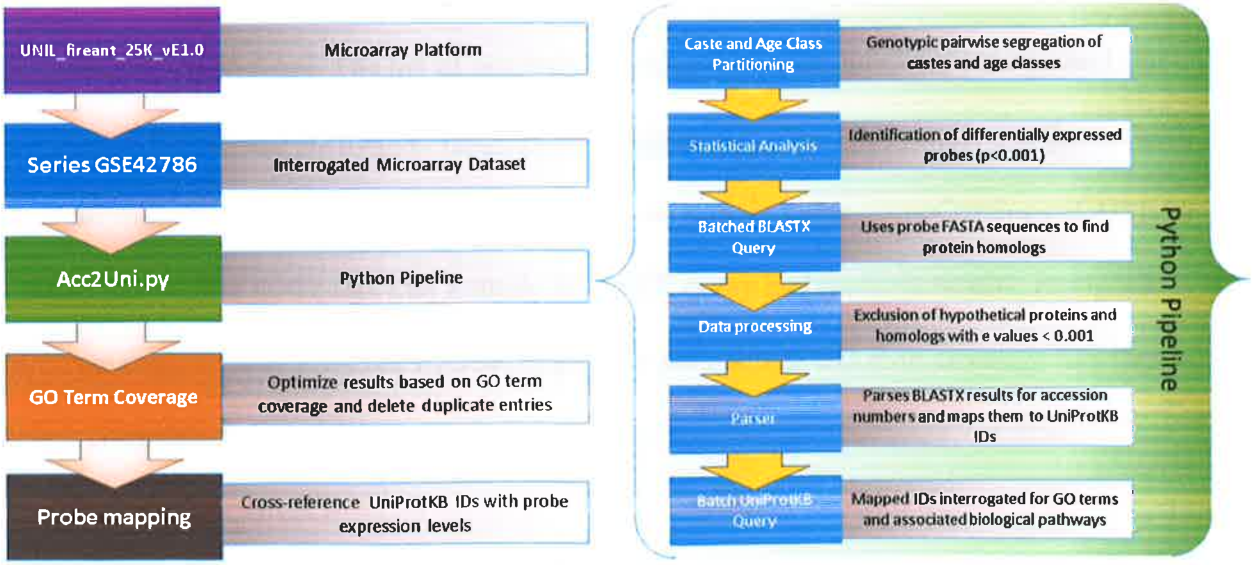

UNIL_fireant_25K_vE1.0
Microarray Platform
Series GSE42786
Interrogated Microarray Dataset
Acc2Uni.py
Python Pipeline
GO Term Coverage
Optimize results based on GO term coverage and delete duplicate entries
Probe mapping
Cross-reference UniProtKB IDs with probe expression levels
Caste and Age Class Partitioning
Genotypic pairwise segregation of castes and age classes
Statistical Analysis
Identification of differentially expressed probes (p<0.001)
Batched BLASTX Query
Uses probe FASTA sequences to find protein homologs
Data processing
Exclusion of hypothetical proteins and homologs with e values < 0.001
Parser
Parses BLASTX results for accession numbers and maps them to UniProtKB IDs
Batch UniProtKB Query
Mapped IDs interrogated for GO terms and associated biological pathways
Python Pipeline

FIGURE 1: Project workflow. *(Left) General workflow of the project. (Right) Custom Python pipeline components integrated into the workflow.*

levels of polygyne and monogyne *S. invicta* Buren queens of the following age classes: pupa (n = 10), 1 day old (n = 16), 11 day old (n = 15), and fully reproductive queens (n = 8). All samples were compared to a common reference made up of multiple whole organism samples of all social and age classes.

Mapping transcript ID to homologous protein

We developed a custom Python pipeline to process the microarray data (Fig.1). Paired sets of expression level data for both social forms data was then segregated by age class. Significantly differentially expressed gene transcripts were identified by a statistical threshold of p-values < 0.001. If the gene transcript had a corresponding GenBank (http://www.ncbi.nlm.nih.gov/genbank/) accession number, the program would pull the resulting FASTA sequence in batches of 50. These sequences were then entered in batches of 15 into a BLASTX (http://blast.ncbi.nlm.nih.gov) query, performed by the Bio.Blast.NCBIWWW package in the Biopython module (http://www.biopython.org). BLASTX query results were constrained to nine results per query, an E-value threshold to <0.001, and proteins not listed as predicted or hypothetical. The list of resultant homolog accession numbers were parsed and passed to a UniProtKB ID mapper (http://www.uniprot.org). The UniProtKB IDs list for each age class was then submitted to UniProt to obtain the protein name, GeneOntology (GO) IDs, and GO terms. Due to the fact that a single BLASTX query had multiple results, retrograde duplicate removal was achieved by selecting firstly for optimal GO term coverage and then by lowest corresponding E-value.

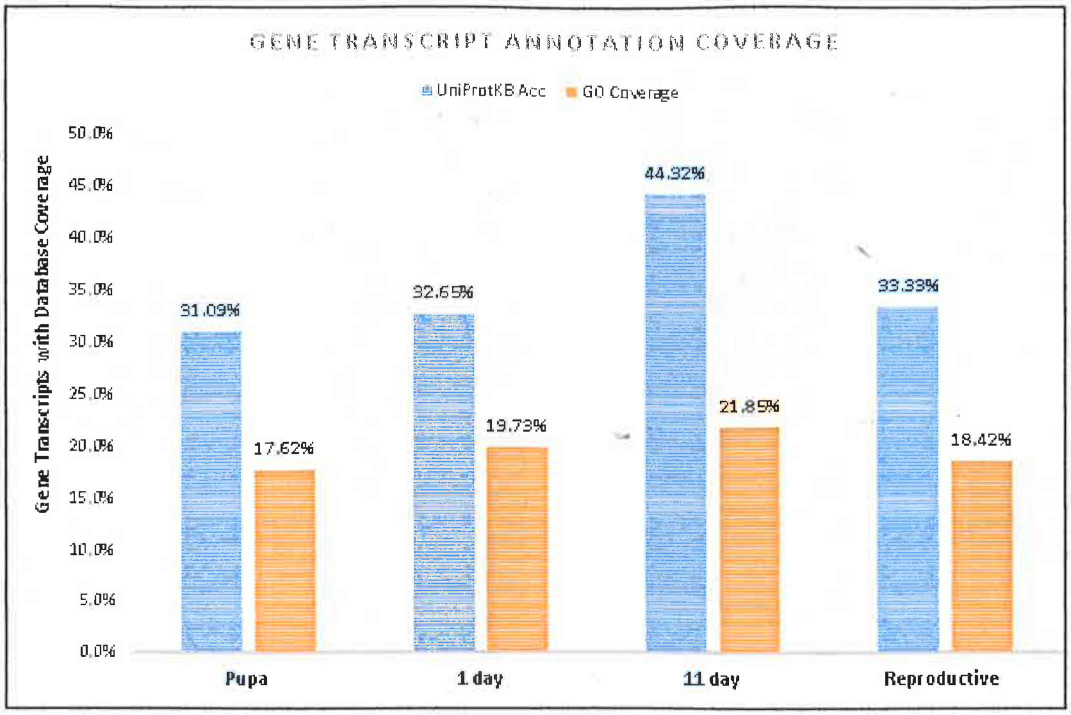
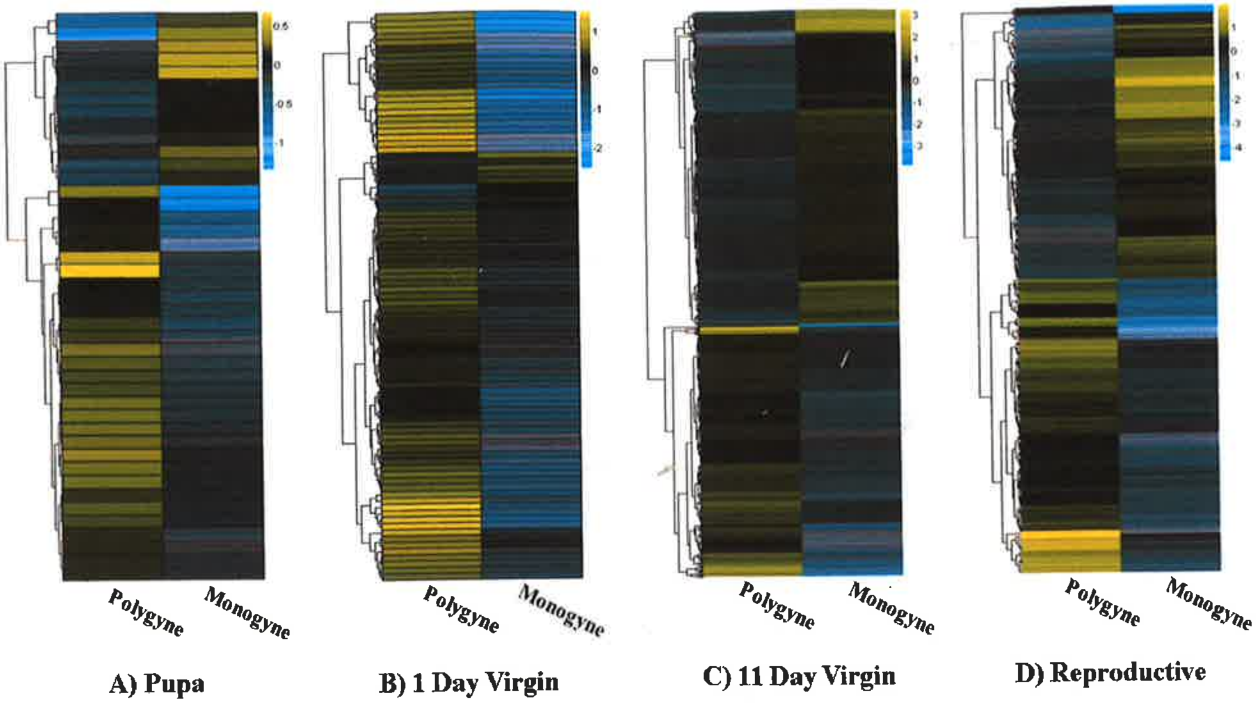


FIGURE 2: Relative proportion of gene transcript annotation. *The number of differentially expressed gene transcripts by age class of S. invicta [pupa (n= 193 gene transcripts), 1 day virgin (n= 147 gene transcripts), 11 day virgin (n = 1135 gene transcripts), and Reproductive (n = 657 gene transcripts)] divided by the number of gene transcripts with a given accession number or GO annotation.*

Identification of Gene Ontology (GO) terms

A list of UniProtKB IDs of differentially expressed gene transcripts was uploaded to UniProtKB. Utilizing the UniProtKB Gene Ontology tools, a top down approach was used to identify greatest common GO term factors all the way to least common GO term factor. This allowed for proof of concept via identification of proportionally constant 'anchor' terms. An anchor term is defined as an original ancestral GO term, such as Biological Process (GO:0008150) or Molecular Function (GO:0003674), that ought to be proportionally expressed throughout age classes of a given organism.

Results



A) Pupa **B) 1 Day Virgin** **C) 11 Day Virgin** **D) Reproductive**

FIGURE 3: Counter-regulated gene transcripts across developmental stage. *The four heatmaps are a graphic representations of the number of gene transcripts of a given developmental age class that are counter-regulated. Gene transcript expression levels are log_2 transformed and illustrated by a blue (down-regulated) and yellow (up-regulated) color scale to the right of each heatmap. A. Pupa (43 gene transcripts), B. 1 day virgin (89 gene transcripts), C. 11 day virgin (609 gene transcripts), and D. Reproductive (175 gene transcripts).Expression levels are log_2 normalized and color bars are calibrated for each age class.*

Identification of differentially expressed gene transcripts

Using the gene expression datasets GSE42062 and SuperSeries GSE42786 obtained from NCBI GEO, both social forms were paired by age class. These data were then pushed through our custom Python pipeline (Fig.1). The pipeline first identified the significantly ($p<0.001$) differentially expressed gene transcripts. The differentially expressed gene transcripts for Pupa (n=193), 1 day virgin (n=147), 11 day virgin (n=1135), and fully reproductive (n=657) developmental age classes were identified. These differentially expressed gene transcripts were then passed to NCBI to obtain FASTA sequences that were later used for BLASTX queries. As the differentially expressed gene transcripts were matched to significant (E-value < 0.001) homologs, the results were then mapped to UniProtKB for further interrogation. Of those differentially expressed gene transcripts, the following had positive matches to UniProtKB IDs: Pupa, 60; 1 day virgin, 48; 11 day virgin, 503; and reproductive, 219. The number of gene transcripts with GO annotation were 34 (pupa), 29 (1 day virgin), 248 (11 day virgin), and 121 (reproductive). While annotation for *S. invicta* is currently incomplete, we found that proportional coverage (number of gene transcripts with a given characteristic divided by total number of gene transcripts) was maintained throughout all developmental age classes (Fig. 2).

To illustrate the change in the number of differentially expressed gene transcripts through development we generated a heatmap displaying counter-regulated gene transcripts (Fig. 3). This figure also explores the changes in expression levels between the different age classes. A counter-regulated gene transcript was defined as any gene transcript up-regulated in one social form and down-regulated in the other. This exemplifies the maximal differentiation observed between the two social forms at any given age class: Pupa (n = 43), 1 day virgin (n=89), 11 day virgin (n=609), and reproductive (n=175) queens.

Gene Ontology Analysis

To determine the functional changes between the developing age classes, gene transcripts were further analyzed for gene ontology. Due to the diffuse nature of *S. invicta* annotation, GO terms had to be mapped via ancestral terms in order to elucidate pertinent information. Using the Perl module Circos (http://circos.ca), the relationships between a given age class and selected list of GO terms could be ascertained (Fig. 4). Using this rendering, the functional profiles of all age classes could be interrogated at once. With regard to age class only, all age classes had relatively equal enrichment for the terms Metabolic Process, Binding, and Catalytic Activity. 11 day and

Reproductive queens were the only age classes enriched for both Odorant Binding and Lipid Metabolic Process. Pupa queens were disproportionately under-enriched for Nitrogen Compound Metabolic Process. Likewise, 1 day queens were disproportionately under-enriched for Primary Metabolic Process. The 11 day developmental age class was the only age class that was highly enriched for Signal Transduction and Transporter Activity (not shown) while these same terms were not present or significantly enriched in the other age classes.

When coupled with counter-regulated gene transcript analysis, we observed that polygyne queens tend to up-regulate gene transcripts linked to Metabolic Process, Catalytic Activity, and Oxidoreductase Activity throughout all developmental age classes, whereas monogyne queens tend to down-regulate these same gene transcripts. Monogyne queens, however, tend to up-regulate gene transcripts linked to Ligase Activity (not shown) throughout all developmental age classes, whereas polygyne queens down-regulate this same gene transcripts. Lastly, polygyne queens tend to up-regulate the gene transcripts linked to Lipid Metabolic Activity and Odorant (pheromone) Binding, whereas monogyne queens tend to down-regulate these same gene transcripts.

Discussion

In this study, gene expression microarray data of RIFA was analyzed to discern a distinct relationship between age class and social form. To do this, a custom Python pipeline was developed to identify significantly differentially expressed gene transcripts. The pipeline then automated BLASTX searches for protein homologs and interrogated UniProtKB for their associated Gene Ontology annotation.

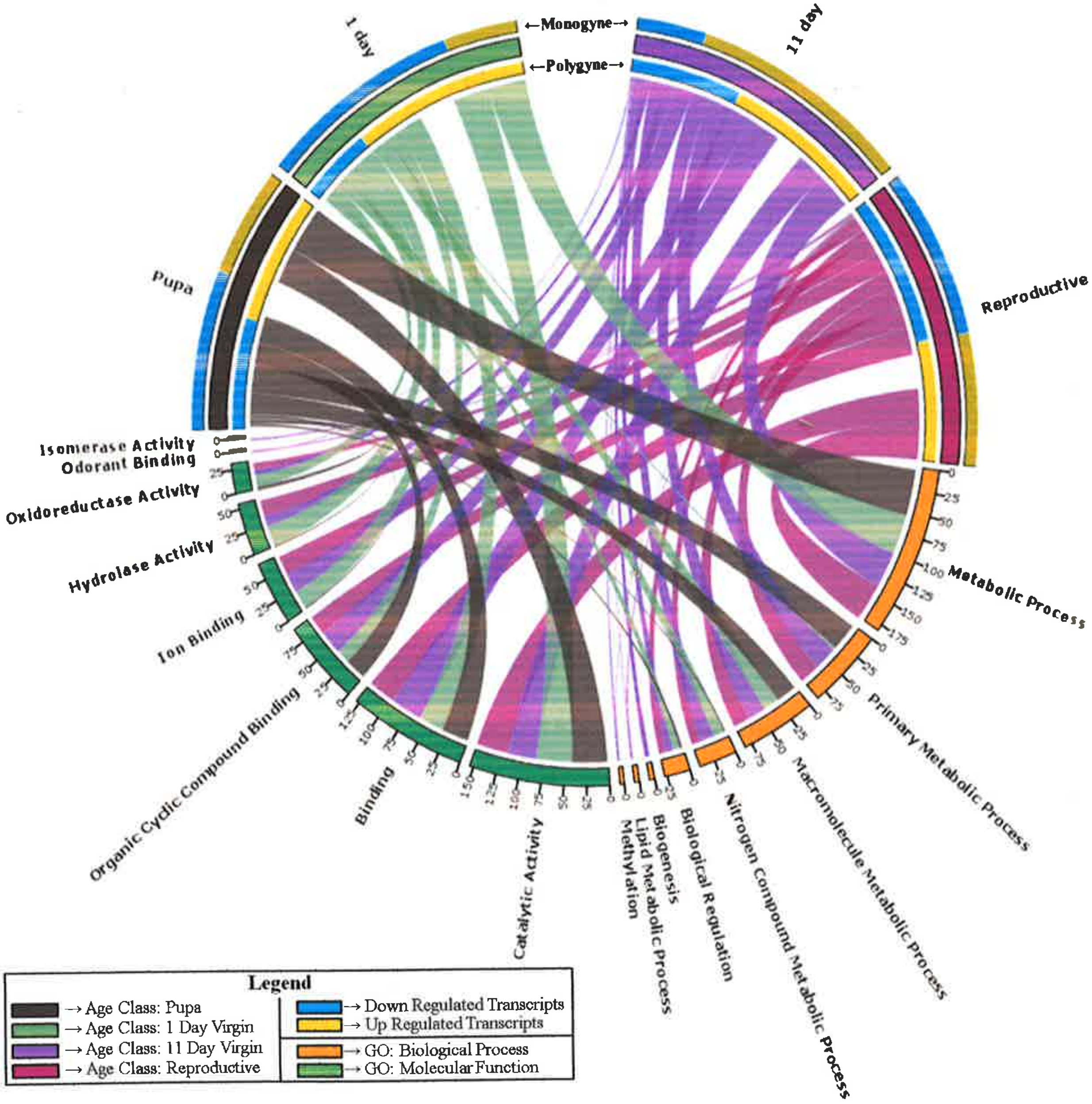


FIGURE 4: GO annotation of developmental age classes. *This Circos ideogram represents the functional profiles of each developmental age class with respect to GO terminology. The blue and yellow bars represent the proportional amount of gene transcripts of either monogyne (top) or polygyne (bottom) that are both up-(yellow)/under-(blue)regulated and have GO annotation. The bars with no tick marks are developmental age classes. The bars with tick marks are various GO terms of either biological process or molecular function ancestry. The width of the ribbon between an age class and GO term is proportional to the amount of gene transcripts of a given age class that have that embedded GO term divided by the number of gene transcripts that have a UniProtKB accession entry.*

Internal validity was established via relatively equal representation throughout all age classes of anchor terms like Metabolic Process and Catalytic Activity. This is due to the fact that all organisms display these processes as a part of respiration. Therefore, any divergent representation of GO enrichment suggests changes to an organism's functional profile through development.

Counter-regulation and GO coupled analysis indicated that polygyne queens are more highly enriched for up-regulated gene transcripts linked to Oxidoreductase Activity involved in the electron transport chain. This may indicate that polygyne queens have a much higher energy requirement compared to monogyne queens. Linking this finding with that of polygyne queens tending to up-regulate the gene transcripts linked to Lipid Metabolic Process supports this hypothesis. Due to the increased need for energy, polygyne queens may up-regulate these lipid metabolic processes to meet this requirement. This increased energy utilization may lead to a diminished mass compared to monogyne queens. Due to the fact that there are multiple reproductive queens in polygyne colonies, the rank as a reproductive queen is therefore dependent on reproductive capabilities. Therefore, the evolutionary advantage of increased utilization of fat could potentially allow for increased reproductive abilities at the loss of energy stores meant to sustain the queen.

Meanwhile, the gene transcripts linked to Odorant Binding that polygyne queens express to a greater degree than monogyne queens is counter-intuitive. Since monogyne ants tend to be less tolerant of immigrant ants entering the colony and have a stricter response protocol with regard to social and reproductive cues, it was posited that transcripts linked to these processes would be up-regulated in monogyne queens—not polygyne queens. At this time, however, it is impossible for us to interrogate the effect of these gene transcripts at the functional level due to incomplete annotation of the *S. invicta* genome.

As an important note, the inability to ascertain a functional profile with greater resolution lies directly with the lack of existing homolog and/or ortholog data and annotation. Since RIFA is not studied nearly as much as other model organisms (e.g. *Drosophila melanogaster* or *Apis mellifera*), much of this data is missing. This is evidenced in other studies regarding RIFA (Wang et al., 2013; Wurm, Wang, & Keller, 2010). With the existing genome being only a draft sequence (Wurm et al., 2011) and the most widely used microarray platform for RIFA using clones that do not encompass long enough ORFs (optimal reading frames) or truncated sequences (Wang, Jemielity, Uva, Wurm, Gräff, & Keller, 2007), this annotation disparity is likely to be apparent in all similar studies. Similarly, the microarray expression level datasets did not contain more than these four age classes, therefore resolution is hindered. This is due to the fact that monogyne queens undergo different genetic changes pre-/post-nuptial flight, and it is posited that mature polygyne queens that are not reproductive would present with different genomic profiles than that of fully reproductive queens.

That being said, future derivations of this research are still viable. The methodologies developed can be utilized in labs that are interrogating better model organisms. Since the framework of the bioinformatic pipeline does not take into account the organism, the same type of genomic interrogation is possible. Similarly, this pipeline is a much more cost-effective alternative with regard to more mainstream software. This is particularly pertinent to smaller labs or campuses that do not have the funds to pay for such software packages. Likewise, in conjunction with the region that this project was undertaken and ongoing research in this region, RIFA projects can still utilize these methodologies with relative ease of access to samples with regard to filling the annotation gap missing on RIFA.

In summary, this project aimed to test if a bioinformatics-based approach could be utilized to ascertain an explicit link between the social forms and developmental stages of RIFA by way of differential genomic and functional profiling. We found that while queens of both social forms were more similar at early developmental stages, later developmental stages showed drastic differentiation in both the number of differentially expressed gene transcripts, but also the degree in which those transcripts were expressed. Moreover, functional changes in catalytic, metabolic, and response to stimulus were evident. This data supports the hypothesis that social form and developmental stage of RIFA are indeed linked.

REFERENCES

REFERENCES

Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. 1990. Basic local alignment search tool. *J. Mol. Biol. 215*: 403-410.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. 2013. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res. Jan*(41)(Database issue):D991-5.

Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J., and Sayers, E.W. 2011. GenBank. *Nucleic Acids Res. Dec*: 1-6 doi:10.1093/nar/gkr1202

Caldera, E.J., Ross, K.G., DeHeer, C.J., and Shoemaker, D.D. 2008. Putative native source of the invasive fire ant *Solenopsis invicta* in the USA. *Biol Invasions 10*: 1457-1479. doi 10.1007/s10530-008-9219-0

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, and de Hoon MJ. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics 25*(11) 1422-3. doi:10.1093/bioinformatics/btp163 pmid:19304878

Epperson, D.M.and Allen, C.R. 2010. Red Imported Fire Ant Impacts on Upland Arthropods in Southern Mississippi. *Am. Midl. Nat. 163*: 54-63.

Krzywinski, M. et al. 2009. Circos: an Information Aesthetic for Comparative Genomics. *Genome Res 19*:1639-1645

LeBrun, E.G., Plowes, R.M., and Gilbert, L.E. 2012. Imported fire ants near the edge of their range: Disturbance and moisture determine prevalence and impact of an invasive social insect. *Journal of Animal Ecology 81*: 884-895. doi: 10.1111/j.1365-2656.2012.01954.x

Manfredini et al. 2013. Sociogenomics of cooperation and conflict during colony founding in the fire ant *Solenopsis invicta. PLoS Genet 9*(8): e1003633. doi:10.1371/journal.pgen.1003633

Nipitwattanaphon M, Wang J, Ross KG, Riba-Grognuz O, Wurm Y, Khurewathanakul C, Keller L. 2014. Effects of ploidy and sex-locus genotype on gene expression patterns in the fire ant *Solenopsis invicta. Proc Biol Sci. 281*(1797). doi: 10.1098/rspb.2014.1776

The UniProt Consortium. 2015. UniProt: a hub for protein information. *Nucleic Acids Res. 43*: D204-D212.

Wang et al. 2013. A Y-like social chromosome causes alternative colony organization in fire ants. *Nature 493*(7434): 664-668. doi:10.1038/nature11832

Wang, J., Jemielity, S., Uva, P., Wurm, Y., Gräff, J., and Keller, L. 2007. An annotated cDNA library and microarray for large-scale gene-expression studies in the ant *Solenopsis invicta*. *Genome Biology 8*(1): R9. doi:10.1186/gb-2007-8-1-r9

Wurm et al. 2011. The genome of the fire ant *Solenopsis invicta*. *PNAS 108*(14): 5679–5684. doi: 10.1073/pnas.1009690108

Wurm, Y., Wang, J., and Keller, L. 2010. Changes in reproductive roles are associated with changes in gene expression in fire ant queens. *Molecular Ecology 19*: 1200-1211. doi: 10.1111/j.1365-294X.2010.04561.x