12-2-2022 10:00 AM

Gene Regulatory Context of Honey Bee Worker Sterility

Rahul Choorakkat Unnikrishnan, *The University of Western Ontario*

Supervisor: Graham, Thompson, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Biology

Recommended Citation

Choorakkat Unnikrishnan, Rahul, "Gene Regulatory Context of Honey Bee Worker Sterility" (2022). *Electronic Thesis and Dissertation Repository*. 9257.
https://ir.lib.uwo.ca/etd/9257

Abstract

Honey bee workers deactivate their ovaries and are functionally sterile when a queen is present in the colony. I adopt a bioinformatics approach to up-date a model transcriptional regulatory network (TRN) to study gene-regulatory processes that regulate fecundity in workers. On splitting the network, I obtained nine clusters and each cluster conformed to properties associated with real-world networks. Two of the nine clusters are enriched for 'sterility genes' and contained single well-connected hub genes (GB44769, *ftz-f1*). The genes in the two clusters were functionally enriched for nucleic acid binding (GO:0003676) and nucleotide binding (GO:0000166). I identified homologous genes for my two clusters of interest and constructed corresponding gene regulatory networks for *D. melanogaster*. In these clusters, I found genes enriched for properties like embryo development in *D. melanogaster* such as *arm*, *kay* and *r-l* whose homologues in *A. mellifera* could be tested for their role in honey bee reproduction.

Keywords

Apis mellifera, trancriptional regulatory network, network analysis, gene ontology, homologues, bioinformatics

Summary for Lay Audience

Honey bees have an estimated total contribution of $2.57 billion in direct additional harvest value to the Canadian economy according to Agriculture and Agri-food Canada. Moreover, bees play a vital role in pollinating various plants that serve as a source of food to many species both large and small. Hence, the importance of bees cannot be disputed and is the subject of multiple research projects each year. An area of interest is the differences in behaviour displayed by the caste members of a bee colony. The queen is the most reproductive member of the colony influencing the behaviour of both the workers and drones. One behaviour change that has been noticed is the deactivation of worker ovaries in the presence of queen pheromones. These workers forgo their egg laying capabilities and take care of other duties around the hive like caring for the young, foraging and guarding the hive, etc. Social insect researchers have identified the suite of genes that may play a role in worker ovary de-activation in response to pheromone. However, a full understanding of how these genes interact to regulate the reproductive division of labor in colonies could be clarified. Over the last couple of decades, researchers have begun to realise that visualizing and studying a collection of genes in the form of a network – that is, a graph showing how individual genes interact with each other – makes it easier to identify key genes and their functionality, which is in part dependant on their position in the network. The Thompson lab has previously utilized network analysis to identify key genes involved in the pathway that regulates honey bee worker sterility. My thesis study attempts to advance this work by taking advantage of new information to up-date and re-analyze earlier work. Through this project, I aim to solidify our knowledge of the genes involved in the reproductive pathway of honey bees and provide an avenue for future research.

Co-Authorship Statement

All the analysis, figures and methodology were developed and conducted in conjunction with Dr. Graham J. Thompson. Any subsequent publications that I write from this thesis will be co-authored with Dr. Thompson.

Acknowledgements

First, I would like to thank Dr. Thompson for the confidence that he bestowed in me when he selected me to helm this diverse and challenging subject. Despite having no background in social biology, Dr. Thompson believed that I would be a fit for this project and I am glad that he took the risk. Despite having our differences about the direction and scope of the project, Dr. Thompson continued to work with me showing due diligence and offered me the support that I needed in times of hardship. I will always be indebted to him for this.

I have had a wonderful advisory committee in Dr. Vera Tai and Dr. Mark Daley. They have both been insightful about the key elements I have missed in my research at each point, providing me with direction and purpose. Dr. David Smith has been a huge influence on my research from the start with his varied thought and questions. Last but not the least, Dr. Ben Rubin has been amazing with his unwavering support and patience in answering all the questions related to statistics that I faced.

I would be remiss if I did not acknowledge my wonderful labmates, Christine, Anthony, Anna, Vonica, Tian and Justine who shared the joys, the pain and the mischief with me over the last two years. I am deeply indebted to them for striving to be the best at what they do and in turn making me bring out the best in me. I wish them all the best in the professional careers and will look back on the times spend with them with utmost fondness.

I would also like to thank my beautiful wife Parvathy Krishna for her support, help and belief in me through all the trying times that I faced over the last couple of years. She is the rock that tethered my ship during the storms and the dark cloudy nights and I am extremely grateful to her for that. I am sure that I would have long lost hope and probably quit if it wasn’t for the strength

that she offered me. I would further like to express my gratitude to my in-laws for being a beacon of support, offering me my daily bread and a place to rest my weary bones at night. They have been strict and forgiving at the proper levels to keep me unwavering from my path. Finally, I am appreciative to my parents for the prayers and the encouragement in my times of need.

Table of Contents

List of Tables

List of Figures

List of Supplemental Materials

List of Abbreviations

Amel: *Apis mellifera*

ARACNE: Algorithm for the Reconstruction of Gene Regulatory Networks

BLASTN: Basic local alignment search tool nucleotides

DOR: Dense Overlapping Regulons

GLay: Community clustering and Graph Layout algorithm

GO: Gene Ontology

HDA: hydroxy-2-decenoic acid

HVA: 4-hydroxy-3-methoxyphenylethanol

KS Test: Kolmogorov-Smirnov test

ODA: oxo-2-(E)-decenoic acid

QMP: Queen mandibular pheromone

TRN: Transcriptional regulatory network

SIM: Simple Input Module

TF: Transcription Factor

1. Introduction

1.1 Eusociality and the evolution of altruism

Eusocial species are characterized by several key features such as an overlap of multiple adult generations, cooperative brood care, and a clear-cut division of labour between reproductive and non-reproductive helper castes (Batra 1966; Michener 1974; Wilson 1971). In eusocial insect colonies, the reproductive and non-reproductive castes are physically and functionally specialized in their roles. For example, queens typically activate their ovaries and lay large numbers of eggs. Workers, by contrast, are normally rendered reproductively altruistic and will de-activate their ovaries and specialize on alloparental care (Plowes 2010). In this manner, non-reproductive workers, despite having few or no offsprings of their own, can nonetheless gain indirect fitness, which would be greater than the fitness that they would have gained from making their own offspring (Queller and Strassmann 1998). If the indirect fitness gain to workers is sufficient, then altruism can evolve, as predicted by Hamilton's rule (1964). Moreover, the reproductive division of labour between a queen and her workers can remain evolutionarily stable, provided caste interests are aligned or, if not, that reproductive conflicts are somehow resolved into a compromise (Ratnieks et al. 2006).

Though the concept of indirect selection is widely accepted as an explanation for the evolution of reproductive altruism (Abbott et al. 2011; Bourke 2011; West & Gardner 2013), it depends on certain assumptions regarding the nature of genes that regulate its expression. Consider, for example, a gene causing altruism in its carrier. If this gene were constitutively expressed in all carriers, there would be no gene-carrying beneficiaries and the gene would quickly go extinct (Crozier and Pamilo 1996). As such, genes underlying the evolution and expression of altruism must be conditionally expressed such that some carriers express altruism (i.e., workers) while others do not (i.e., queens) (Charlesworth 1978; Parker 1989; Queller and Strassmann 1998; Seger 1981; Thompson 2013;). Altruism evolves when reproductive beneficiaries pass-on unexpressed copies of the altruist's genes. The conditional expression of genes underlying altruism can be investigated as a means of capturing them via gene expression technologies, for example, microrarrays or RNAseq (Thompson et al. 2006).

1.2 Different castes in honey bees

The European honey bee (*Apis mellifera*) colonies typically consist of three castes of bees, namely, queens, workers and drones. Workers are the most populous caste handling everyday functions around the hive like nursing the young, guarding the hive, foraging for pollen and nectar, clearing out dead bees, among other non-reproductive tasks (Moore et al. 1987). The second most populous caste is the drone caste, who’s main function is to mate with a receptive queen from anther hive (Ruttner 1966). A typical hive will only consist of one egg-laying queen. A hive without a queen usually doesn’t survive for long (Winston et al. 1991). The queen can be identified visually by her large abdomen that contains well developed ovaries and a sperm storage organ (a spermatheca).

1.3 The regulatory control of worker reproduction in honey bees

Honey bee workers typically de-activate their ovaries in the presence of a fertile queen (Butler 1957; Oldroyd and Osborne 1999). However, workers are not obligately sterile; they can activate their small ovaries and lay unfertilized (haploid) eggs that can develop into males (Bell 1982). Although exceedingly rare in queenright colonies, ovary-active workers are more common in queen-less colonies (Ratnieks and Visscher 1989). Egg-laying workers are also known from some mutant honey bee lines, including the 'anarchy' line (Oldroyd et al. 1994), in which workers appear to ignore the queen signal and lay eggs, even in her presence (Barron and Oldroyd 2001). Worker reproductive success is therefore a combination of direct and indirect fitness, depending on the social circumstance. The mechanism by which individual workers can respond to the presence or absence of queen pheromone and switch their ovaries 'OFF' or 'ON' is not fully understood, but presumably evolved as a genetic pathway that is triggered by environmental – namely, pheromonal – cues (Barron and Oldroyd 2001; Ronai et al. 2016; Tan et al. 2015; Thompson et al. 2006). If so, there is merit in discovering the components of this pathway.

Apis mellifera queen pheromones have multiple roles to play in a colony, including the inhibition of worker egg-laying, the formation of a worker retinue around the queen and attracting drones to mating congregation areas (Keeling et al. 2003). Queen mandibular pheromone (QMP) is considered a key signal emitted by queens with both long term (e.g., worker ovary inactivation) and short-term (e.g., worker retinue formation) effects (Winston et al. 1991). The phromoen has multiple components, which are: (R)- and (S)-(E)-9-hydroxy-2-decenoic acid (9HDA), 10-hydroxy-2 (E)-decenoic acid (10HDA), methyl p-hydroxybenzoate (HOB) and 4-hydroxy-3-methoxyphenylethanol (HVA) and 9-oxo-2-(E)-decenoic acid (9-ODA), which function with some redundancy (Maisonnasse et al. 2010). One component that has long been understood to inhibit worker ovaries is 9-ODA (Butler and Fairey 1963); a function that may be evolutionarily conserved (Van Oystaeyen et al. 2014), as evidenced by similar ovary-inhibiting effects of QMP on other species and sub-species. For example, queen pheromone from subspecies *Apis mellifera scutellata* can de-activate worker ovaries in *A. mellifera capensis* (Mumoki et al. 2018). Further, QMP can inhibit ovary activation in a totally un-related and non-social insect, *Drosophila* *melanogaster*. Camiletti et al. (2013) found that virgin female flies had fewer, smaller eggs when compared to untreated control flies that were not exposed to bee pheromone. This bizarre cross-species effect is apparently transferable to more than one species of fly (Nayar et al. 1963), as well as to termites (Müller et al. 1959) and ants (Carlisle et al. 1956). Each of these studies are quite different in how they sampled, treated and measured their samples, but together they provide some disparate evidence that genes involved in reproductive regulation may be conserved beyond the Hymenoptera (Croft et al. 2017).

Two different approaches have been used to identify genes that could potentially regulate ovary activity in *A. mellifera*. One approach takes advantage of caste differences to compare ovary-active queens with ovary-inactive workers (Evans and Wheeler 1999; Lago et al. 2016). These studies have revealed widespread differences in expression between castes. One experimental confound of these inter-caste comparisons is, however, that the resultant gene lists are only partially associated with ovary activation *per se*. The lists are likewise enriched for genes associated in their expression with any caste differences, of which there would be many related to size, physiology, behaviour and anatomy (Evans and Wheeler 1999; Lattorff and Moritz 2013). An alternative approach is to focus on a single caste and directly compare ovary-active with de-active workers. By comparing ovary-active and in-active workers in queenright colonies of the 'anarchist' strain, Thompson et al. (2006) found few genes differentially expressed between reproductive and non-reproductive workers, but these genes were highly relevant to reproduction (e.g., major royal jelly proteins, vitellogenin). This more-focussed approach potentially circumvents the confounds associated with inter-caste comparisons. Similarly, other studies have identified genes implicated in the regulation of worker sterility (e.g., Cardoen et al. 2011; Grozinger et al. 2007). Based on the importance of these discoveries, further effort must be made to test how single genes from these lists interact with each other or within a broader regulatory context.

1.4 Genes for honey bee worker altruism: a network analysis

Gene action within individuals is often coordinated across loci, as evidenced by a generation of microarray, RNA-seq and other expression-based screens that typically reveal an abundance of of gene co-regulation. One common output of genome-wide expression screens is the ubiquitous 'gene list' – that is, the tally of gene names or accession numbers that often accompany -omics papers. A survey of three leading journals in the field of genomics, *BMC Genomics* (BioMed Central), *Genome Biology and Evolution* (Oxford University Press) and *Genome Research* (Cold Spring Harbor Press), for example, reveals that ~15% of recent (2014-2018) papers feature at least one gene list, which are often long (100s or even 1000s of entries), sorted by order of implied importance, and sometimes relegated to supplementary material (Figure 1). This tabular approach to deciphering gene function is a necessary first step in functional genomic studies but, in my view, is unsatisfying as an end-point because the gene ID information is static and does not in itself explain how individual genes on a list might interact with each other, or how these interactions change in real-time with social or environmental circumstances.

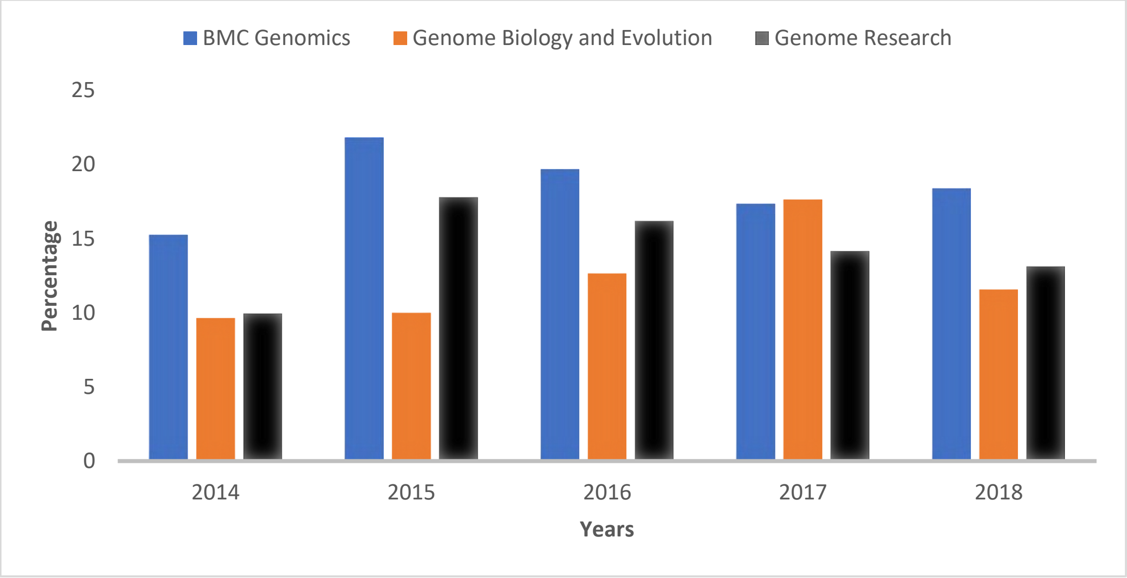

Figure 1 Survey of journals for the number of gene lists from 2014-2018. Survey of three -omics journals revealed that between 9 - 21% of recent (2014 – 2018) papers feature at least one gene list (including in associated supplementary files, and regardless of whether the list was from genetic or genomic study). Over this five year period, *BMC Genomics* had a total of 5578 published articles while *Genome Biology and Evolution* and *Genome Research* had a total of 1374 and 909 published articles respectively, for a total 7861 papers. I visually checked to see if the publications had any gene lists associated with them.

A gene function is often determined by examining when and where it is expressed in a cell or an organism. However, experimentally deducing the function of each gene individually is a time-consuming process. Here, I adopt a view that is common in systems biology in which gene lists are converted into or transposed onto a gene-regulatory framework (e.g., Chouvardas et al. 2016; Segal et al. 2003; Verfaillie et al. 2014). This is done through the prioritization of genes based on their interactions with each other. A gene regulatory network can be best defined as a set of genes that interact with each other to fulfill a specific cell function. One approach to developing a gene-regulatory context within which to infer the coordinated function of otherwise-unknown genes is to construct gene co-expression networks (Schlitt et al. 2003). Using gene-expression information it may be possible to build a gene network and from it test to its overall function, for example, through gene enrichment analysis (Brazhnik et al. 2002; Davidson and Levin 2005).

Pathway information of each gene can also be inferred through network analysis. Hence, I predict that *A. mellifera* genes with a role to play in fertility will display similar network properties and be situated near each other on the network. By situating prior listed genes onto a network topology, we can potentially re-prioritize individual genes based on their connectivity, as opposed to their stand-alone expression value (Ramsahai et al. 2017; Rapaport et al. 2007). Further, by situating individual genes within a multi-gene environment we can better interpret their functional roles, as might be revealed, for example, by multi-gene regulatory modules, motifs and clusters.

Among insects, some of the earliest comparative genomic studies helped transform gene lists into gene networks for *D. melanogaster* (Costello et al. 2009; Haye et al. 2014). Early network comparison studies between sterile and fertile castes involved research on ants where a group of genes were found to be differentially expressed between winged (fertile) and sterile (wingless) castes (Abouheif and Wray 2002). For honey bees, gene network studies have been used in multiple instances, like comparing the aggressive behaviour between different species (Alaux et al. 2009), analyzing gene expression differences in knockout studies (Ament et al. 2012) and caste comparison (Barchuk et al. 2007).

The Thompson lab has contributed to identifying genes responsible for fecundity in *A. mellifera* with a study done by Mullen et al. (2014) identifying nine small networks enriched for genes involved in reproductive functions. More recently, a study by Sobotka et al. (2016) helped to position sterility genes on the honey bee gene regulatory network for *A. mellifera,* which was computationally predicted in a previous study (Chandrasekaran et al. 2011). Here, 'sterility' is short hand for a suite of behavioural, physiological and anatomical changes to worker reproduction upon exposure to queen pheromone (Ronai et al. 2016). This reproductive trait likely evolved under indirect selection (Thompson et al. 2013) and is commonly measured as a function of ovary de-activation (Backx et al. 2012). A 'sterility gene' – or, a *gene for sterility* – is a term that I and others use to describe any genetic difference that explains phenotypic differences in egg-laying behavior observed among worker honey bees within a colony. The genetic effect may be due to a difference in nucleotide sequence – for example, a mutation or polymorphism – or it may be due to a difference in gene expression, however realized. In my case, I use the term to describe genes that co-vary in their expression with ovary activation. My working hypothesis is that worker ovary activation is regulated in response to genetic and environmental cues and, if so, I should be able to reconstruct this network from a list of genes that I know to be co-expressed with worker ovaries. That, in fact, is the main goal of my thesis, to reconstruct the gene network regulating worker sterility and to examine its properties.

Sobotka et al. (2016) used the GLay clustering algorithm of Su et al. (2010) to first divide the network into clusters of well-connected nodes, of which there are two types: genes and transcription factors. A *gene cluster* is a group of interacting or potentially interacting genes that function in concert to regulate the expression of a trait. A cluster is a network but when the latter has subcomponents (or subnetworks) that are themselves interacting, then we refer to those subnetworks as gene clusters. One criterion for recognizing clusters is via the high degree of interconnectedness: clusters, by definition, have more inward (within-cluster) connections than outward (between cluster) connections, as typically inferred by optimizing algorithms implemented by network analysis software. Further, gene clusters often contain one or more so-called hub genes that have accumulated an exceptionally large number of inward connections and are thus presumably very important to the subnetwork's function.

Their best-fit clustering model used by Slobotka et al. (2016) revealed that the honey bee TRN (transcriptional regulatory network) is composed of as few as eight sub-networks. Sobotka et al. (2016) then tested the distribution of published sterility gene sets across the sub-networks. They reasoned that if subsets of genes co-regulate worker sterility, then the genes should form interconnected modules within the honey bee TRN. They found that most gene sets examined (3 of 4) did tend to cluster into a particular region of the 2,382-gene network, which Sobotka et al. (2016) dubbed 'Cluster 3', the third-largest cluster of the TRN. Furthermore, in this study, two genes that were identified as potentially having a key role to play in the pheromonal regulation of worker reproduction were *fushi-tarazu factor-1* (*ftz-f1*) and *fruitless* (*fru*). These two genes were centrally located in the network and the *ftz-f1* gene has previously been found to have a role in honey bee maturation (Cardoen et al. 2011). The genes *fru*, meanwhile, was found to be downregulated in workers in the presence of a queen (Grozinger et al. 2003).

The discovery of this subnetwork, apparently associated with the expression of worker sterility, would not have been possible from an analysis of the individual gene lists. Their network analysis revealed, for the first time, that the candidate genes for sterility identified from microarrays, are functionally connected to each other. Studying this and other subnetworks within the honey bee TRN could, therefore, potentially reveal how changes to a worker's social environment – i.e., the presence or absence of queen pheromone (Backx et al. 2012) – can alter the state of the network to de-activate worker ovaries and render them sterile.

1.5 Objectives

The overall goal of my research effort is two-fold: I will use bioinformatic tools to update the edge list and network developed by Chandrasekharan et al. (2011) and use this new construct to up-date the cluster analysis performed by Sobotka et al. (2016) to infer the of functional organization of the network as it relates to honey bee worker sterility. I elaborate on both objectives immediately below.

1.5.1 Update the TRN and draw structural and functional inferences

The gene set used by Chandrasekaran et al. (2011) for the computational prediction of the *A. mellifera* network was outdated, because they used an older Gene set v1.0 (Honeybee Genome Sequencing Consortium 2006). There is now a newer Gene Set v3.2 (Elsik et al. 2014) available for *A. mellifera* genes, which would contain information on newly discovered genes, have up-dated annotation information of previously identified genes as well as deleted erroneously identified genes included in the previous gene set. The up-dated TRN (after removal of obsolete genes) provides an opportunity to re-analyze the structure of the network, particularly with reference to any clusters that might be enriched for sterility genes. Since Sobotka et al. (2016)

inferred that the original TRN constructed by Chandrasekaran was not random based on its structured and highly interconnected nature, I reasoned that the updated TRN should also display similar properties. Specifically, I test if the honey bee model TRN and its sub-networks show a typical scale-free *degree distribution* in which a plot of the frequency distribution of node degrees reveals one or a few 'hub' genes are disproportionately connected to the remaining majority of genes with relatively few connections (Liseron-Monfils and Ware 2015). I also estimate the *eigenvector centrality* of each node (a measure of node connectedness in a network) in each subnetwork. I look for other measures of adaptive complexity in the form of logical patterns or 'motifs' (to reveal the pattern or arrangement of interconnections, like feed-forward loops, etc.) that are not expected of random or otherwise non-evolved biological networks (McDonnell et al. 2014). Finally, I use homology-based enrichment analyses (Jonsson et al. 2006) whereby I use sequence similarity to genes with known function to infer the most likely biological function of each subnetwork.

1.5.2 Infer the function of genes in subnetworks via homology analysis

With most of the genes in *A. mellifera* being uncharacterized (of unknown function), I decided to examine the function of these genes by finding their homologues in better studied organisms. Rather than simply conducting a gene ontology analysis of the homologous genes to find functional enrichment, I also decided to view these genes in a gene regulatory context to identify key genes (based on degree) from among a group participating in a similar function. I reasoned that there could be some related biological pathways involving reproduction between *A. mellifera* and *D. melanogaster* based on evidence from multiple research studies about similar effects of ovary inhibition found in *D. melanogaster* from exposure to queen pheromones (Camiletti and Thompson 2016; Camiletti et al. 2013; Croft et al. 2017). This would also allow me to test the hypothesis that social reproduction in *A. mellifera* is derived from gene pathways that once regulated individual reproduction in solitary ancestors, as represented by the fly. Specifically, I will identify the subnetworks in the *A. mellifera* transcriptional regulatory network that are enriched for sterility genes (if any) and construct their corresponding networks for *Drosophila*.

For identifying clusters in the *A. mellifera* network I used the same set of sterility genes compiled by Sobotka et al. (2016) based on a literature survey of differentially expressed *A. mellifera* genes (Table 1). Sobotka's list of genes is compiled from comparable studies, which used QMP for deactivating the ovary and gene expression analysis from microarray studies between ovary active and inactive *A. mellifera* workers (Cardoen et al. 2011; Grozinger et al 2003; Grozinger et. al 2007). These older-generation transcriptome screens either sampled brain tissue, abdomen tissue, or both, or whole body tissue (Table 1). There is no certain 'best practice' and each study has its own merits, discoveries and limitations. For my thesis, I take it as a given that these studies produced valid gene lists (they are published) and I regard my analyses as a type of meta-analysis whereby I use the output from other studies as input for my own.

Since one or more of the subnetworks I would be using are potentially enriched for sterility genes, I expected the ontology analysis of the *Drosophila* homologues to likewise be enriched for functions related to reproduction. I predicted that genes involved in the reproductive functions of *Drosophila*, which is a solitary insect, could potentially be performing a similar function in *A. mellifera.* Through this methodology, I hope to identify well connected genes lying along key reproductive functional pathways of *A. meliifera,* which could serve as novel targets for future gene silencing or knockdown studies in *A. mellifera* with verifiable changes in the morphological and genetical changes of reproductive and non-reproductive *A. mellifera* workers.

Table 1 Microarray datasets that I assembled from the literature. Table shows the number of differentially expressed genes (DEGs) reported for each study, together with a very brief description of the experimental design and tissue sampled. This same meta-dataset was used by Sobotka et al. (2016) and I here use them again to test the if they cluster together on a honey bee transcriptional gene regulatory network.

Study	Experimental Design	Tissue Type	Total number of DEGs
Grozinger et al. (2003)	QMP-treated versus untreated workers in cages	Brain	1607
Thompson et al. (2006)	Wild-type versus anarchist workers in colonies	Brain Abdomen	20 20
Grozinger et al. (2007)	QMP-treated versus untreated workers in cages	Brain	94
Thompson et al. (2008)	Wild-type versus anarchist workers in colonies	Brain Abdomen	7 5
Cardoen et al. (2011)	Ovary-active versus ovary-inactive workers in colonies	Whole Body	1292

2. Methods

2.1 Re-constructing the regulatory network and its sub-networks

The honey bee transcriptional regulatory network was constructed computationally in the Hood-Price lab and is available publicly for download in an Excel format (https://hood-price.systemsbiology.org/research/honeybee-transcriptional-regulatory-network/). The network is based upon the transcription profiles of 853 individual honey bees exhibiting 48 distinct behavioural phenotypes, as described in detail in Chandrasekaran et al. (2011). The bee brain transcriptional network was created using an Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE) and consists of transcription factors (n = 205) and their target genes (n = 2176). Since the gene data used for the TRN construction is based on the first sequencing project (Honeybee Genome Sequencing Consortium 2006) and first official gene set (amel_OGSv1.0), the edge list needed to be up-dated to accommodate changes associated with the most recent sequencing up-grade (Elsik et al. 2014) and new official gene set (amel_OGSv3.2). A previous study (Molodtsova et al. 2014) mapped (using Blastn v. 2.2.28+) the original TRN's oligonucleotide probe IDs to the new gene set. I used their cross-referenced information that I obtained from the corresponding author to establish an up-dated edge list of my own that was consistent with the latest assembly. I used this up-dated edge list in all my subsequent analyses.

To reveal any underlying substructure to the network, I first imported the TRN into the network visualization software package CYTOSCAPE (Version 3.6.0; Shannon et al. 2003). I tested different clustering algorithms (Affinity Propagation Clustering Algorithm, Markov Clustering Algorithm, Fuzzy C-means Clustering Algorithm and the GLay Clustering Algorithm) and, for each, varied the number of clusters tested. I had no *a priori* expectation for how any of these algorithms might partition my data set so I did 'sample then', so to speak, to see if they generated any big differences. I used the ‘modularity score’ of Peterman et al. (2016) to gauge best fit between any one model and, simultaneously, between different parameters (i.e., no. of clusters) for a given model. Specifically, I interpreted a high modularity score to signify a dense connectivity between nodes present in a sub-network and weak connectivity between nodes belonging to different sub-networks. I selected the GLay clustering algorithm, which utilizes the Girvan-Newman fast-greedy algorithm for my partitions because it 1 – allowed the sharing of nodes between clusters so as to prevent the potential breaking up of larger clusters into smaller ones, 2 – did not ask the user to predict the number of clusters and, 3 – has a higher modularity score when compared to the other clustering algorithms.

2.2 Estimating the structural parameters of each network

Networks present in the real world, be it social, technological, biological or information networks, are typically scale-free i.e., the number of connections to each node (theirs 'degree') is highly heterogenous and follows a power law frequency distribution. So, I decided to test if each sub-cluster of my updated *A. mellifera* network was also scale free. In scale free networks, the probability distribution of the number of nodes and their corresponding degrees follows the condition $P(X) = C.X^{-\alpha}$ where C is a constant, α is the rate of decay or the slope of the distribution, X is a degree and P(X) is the number of nodes with the degree X (Newman 2005). The coefficient of determination (R^2) is defined as the square of the correlation coefficient between log(P(X)) and log(X) (Zhang and Horvath 2005). I computed the R^2 values for each cluster to test if the slope was between a value of 2-3, which is consistent with a scale-free distribution (Albert 2005).

I used a two-sample Kolmogorov-Smirnov test (KS test) to compare my observed degree distributions to those of Erdos–Renyi random networks (Erdös and Rényi 1959). The KS test determines whether the observed real-world distribution is different from the null (random) distribution. I used the IGRAPH package in R to generate the Erdos–Renyi random networks for each cluster, holding the number of edges and nodes constant.

I plotted the degree distribution of each sub-network and identified the genes with highest degree. I calculated the mean degree of each network and, following Sporns et al. (2007), I considered nodes with a degree of more than one standard deviation above the mean degree as a 'hub'. I also computed the eigen centrality score (Tang et al. 2015) for each gene. This measures the influence of a node in the network by assigning a high score to a node when connected to high degree nodes and a low score when connected to low degree nodes. This provides an indication of how connected a node is in the network through its first- and second-degree neighbours.

2.3 Motif Analysis

I examined the distribution of two types of three-node motifs that could be found in a bipartite network. The first motif, which is termed the Single Input Module (SIM; Domedel-Puig et al. 2010), consists of a TF regulating two genes - 'one into two'. The second type of motif termed the Dense Overlapping Regulons (DOR) is where more than one regulator influences a gene - 'two into one'. I calculated the number of three-node motifs in each cluster using the G-tries algorithm (Ribeiro and Silva 2010), as implemented in CYTOSCAPE under the application name Motif Discovery to find the number of three node motifs in each sub-cluster. To test if the observed number of motifs in my clusters was different from random expectation, I first generated a population of 500 Erdos–Renyi random subnetworks in which the number of nodes and edges are held constant but the connections are shuffled (using IGRAPH package in R; Csardi and Nepusz 2006). I then used a one sample *t*-test to compare observed versus expected number of motifs.

2.4 Testing for the distribution of sterility genes

Since the study by Sobotka et. al (2016) concluded that a significant number of sterility genes (compiled from sterility genes reported in the literature) fell on a specific cluster, we tested whether the up-dated TRN likewise supported a single cluster enriched for sterility genes. I conducted a chi-square test for independence to test if the distribution of sterility genes over the clusters was biased – that is, I tested if a significant number of the genes mapped onto any specific cluster, while controlling for cluster size. I also used a custom-made Python script to conduct a randomization test to test which specific clusters have a significant higher number of sterility genes mapping onto them. I also mapped genes from each cluster obtained by Sobotka et al. (2016) during her study onto my current clustering arrangement.

2.5 Gene Ontology Analysis

To determine any predominantly expressed functions by a group of genes in each cluster, I obtained the Gene Ontology (GO) terms associated with each gene ID (if available) using the Ensembl Metazoa database, which contains genomic data of different metazoan species including *A. mellifera* (Kersey et al. 2017). To test for enrichment of GO terms in each cluster, I performed a GO analysis using the Database for Annotation, Visualization and Integrated Discovery (DAVID) platform (Huang et al. 2007). I set the GO parameters in DAVID to yield gene enrichment pathways specific to *A. mellifera*, but otherwise used default search criteria (i.e., enrichment score *P*-values less than 0.1 with a minimum of three genes per GO category). I used

Ensembl BioMart to individually check for the functional information of each gene present in each cluster.

2.6 Analysis using homologous D. melanogaster genes.

I used the data dissemination search tool in BioMart (Kinsella et al. 2011) to mine for homologous FlyBase IDs of *D. melanogaster* for the genes in the *A. mellifera* clusters of interest. Homologues in Ensembl are inferred through the construction of gene trees designed to predict the evolutionary history of a family of genes. For this part of the analysis, I used the FlyBase IDs as an input on the STRING database (database containing gene and protein interaction information) to build a network based on gene interaction information present in the database. The gene interaction information contained in the STRING database includes computational predictions, indirect and direct physical interactions as well as interaction information contained in different databases (Szklarczyk et al. 2016). I used a CYTOSCAPE application named ClueGO (Bindea et al. 2009) to functionally annotate the *D. melanogaster* clusters. I set the network specificity option in ClueGO to 'global', which enables the most general annotations. I optimized ClueGO to detect any gene ontology terms (biological process, cellular component and molecular function) that have been detected experimentally in previous studies rather than computationally predict the function of a gene. Finally, I set the rest of the parameters in ClueGO to their default values for the analysis.

3. Results

A total of 1,839 of the 2,382 nodes in the network had corresponding gene identification numbers in the new honey bee gene set. Only three genes were 'orphaned' and thus did not have any connections (i.e., were not attached to any other gene). An additional two genes were connected only to each other and thus disconnected from the rest of the network. After removing these five nodes, the modified version of the honey bee TRN consisted of 1,834 nodes (195 TFs and 1639 genes) and 5,085 connections. The number of connections I inferred for the up-dated TRN is smaller than the original TRN (with 6,756 connections), but presumably this number is more accurate (Figure 2). Upon clustering the up-dated network, the GLay clustering algorithm fit our network the best with nine clusters and a modularity score of 0.609. The total number of transcription factors, the genes they regulate and the number of edges in each cluster is summarised in Table 2.

3.1 Degree Distribution of the clusters

The coefficient of determination values (R^2) were positive for all nine clusters. The correlation coefficient values for the clusters ranged from 0.475 to 0.949. The slope of clusters ranged from 0.87 (Cluster 8) to 1.37 (Cluster 1). The degree distributions of each cluster were significantly different from the degree distribution of their random counterparts with a *P*-value of less than 0.001 in each case (Table 3).

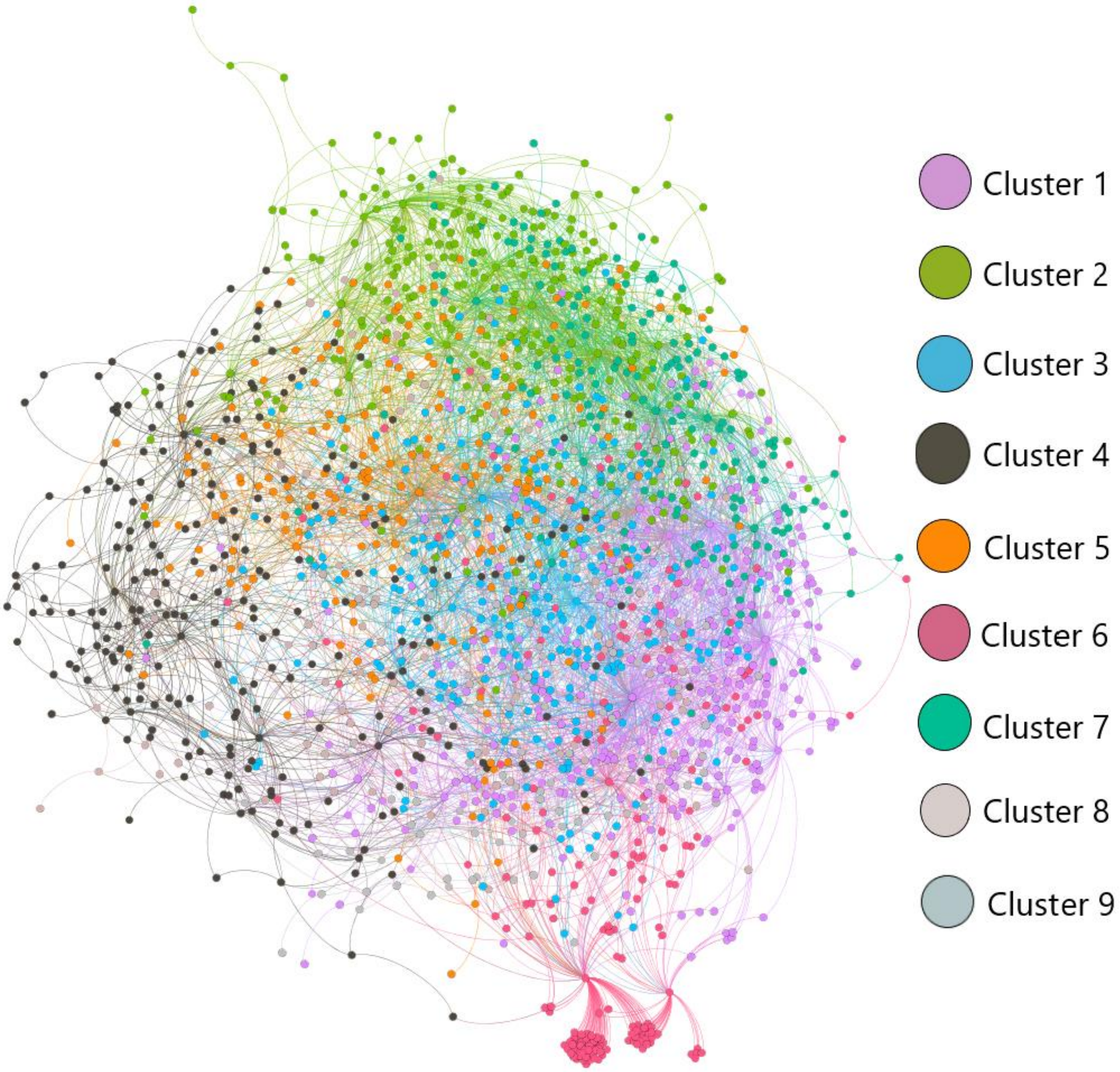

Figure 2 Honey bee brain transcriptional regulatory network, as visualized using the software GEPHI. The best score was obtained when the network was partitioned into nine clusters using its in-build Louvain method of community detection. Each colour represents a different partition or subcluster and they are numbered from largest to smallest. The two smallest clusters are difficult to see.

Table 2 The total number of nodes along with the number of transcription factors and genes in each cluster. The number of transcription factors (TFs) and genes in each sub-network arranged from the largest to the smallest cluster in terms of the total number of nodes. By comparing the total number of TFs and genes for each cluster, the largest cluster has an average of four genes per TF while Cluster 6 has an average of more than 10 genes per TF.

Cluster Number	Total Number of nodes	Number of TFs	Number of Genes	Number of Edges	Ratio of Genes/TFs
1	305	62	243	630	3.91
2	261	32	229	550	7.15
3	252	32	220	466	6.87
4	212	14	198	374	14.14
5	209	16	193	349	12.06
6	179	11	168	215	15.27
7	173	14	159	245	11.35
8	153	6	147	191	24.5
9	90	8	82	109	10.25

Table 3 The degree distribution study done on each cluster. The correlation coefficient values and R^2 for each cluster obtained using the Network Analyzer tool in CYTOSCAPE. The Kolmogorov-Smirnov D value calculated for each cluster is provided along with the KS test result (all *P*-values less than 0.001).

Cluster Number	Correlation coefficient	Coefficient of determination (R^2)	Slope Measurement	Kolmogorov Smirnov D	Significantly different distribution from a random distribution?
Cluster 1	0.518	0.782	1.372	0.1180	Yes
Cluster 2	0.475	0.607	0.978	0.2069	Yes
Cluster 3	0.66	0.754	1.221	0.1667	Yes
Cluster 4	0.827	0.632	0.978	0.1321	Yes
Cluster 5	0.839	0.606	1.007	0.1388	Yes
Cluster 6	0.923	0.698	0.939	0.2067	Yes
Cluster 7	0.963	0.684	1.086	0.1734	Yes
Cluster 8	0.942	0.532	0.877	0.2353	Yes
Cluster 9	0.949	0.621	0.99	0.1889	Yes

3.2 Hubs of the network.

Figure 3 shows the degree for each TF and target gene within each cluster. The size of the cluster ranged from 305 nodes in the largest to just 90 nodes in the smallest cluster. The average degree for all the nodes of each cluster varied from 4.22 in Cluster 2 to 2.42 in Cluster 9. Each cluster had at least one transcription factor with a degree of 1 except Clusters 7 and 8. Table 4 has the list of all the hub nodes (node with the highest degree and eigenvector centrality value). The node with the highest degrees also had the highest eigenvector centrality values in each cluster. *Myb*, CG9932, *ftz-f1* and CG17912 were also found to be nodes with the highest degrees in different clusters (clusters numbering 1, 2, 3 and 6) in the study done by Sobotka et. al (2016). The largest hub node of each cluster along with its first and second step neighbours have been marked in Figure 4. From a visual examination, it is evident that the hubs and their neighbours in each cluster encompass most nodes in each cluster except for the largest cluster which is Cluster 1.

3.3 Motifs

Figure 5A shows the relationship between the SIM (single input module) motifs in the random networks against the total number of such motifs generated by the nine clusters. A *t*-test comparison between the number of motifs in each cluster and their random counterparts provided evidence that each cluster had significantly more SIM motifs. When each cluster was checked for the number of DOR (dense overlapping regulons) motifs, I could not find evidence for there being more DOR motifs present when compared to corresponding random networks. Figure 5B provides a comparison of the number of SIM motifs in each cluster.

3.4 Distribution of sterility genes

Sterility genes were not distributed randomly among clusters (χ 2 = 28.39, df = 8, *P*-value < 0.001). A custom-made Python script was used to compare the actual distribution of sterility genes over all the clusters with what the distribution of sterility genes would be over similarly sized random clusters for a total of 10^4 iterations. From this comparison, it was discovered that significantly more number of sterility genes mapped onto Cluster 1 (*P*-value < 0.001) and Cluster 3 (*P*-value < 0.001). On mapping genes from the clusters found by Sobotka et al. (2016) onto our clustering arrangement, the majority of genes from Cluster 3 (sterility cluster), fall onto our current Cluster 3 and Cluster 1 (Figure 6).

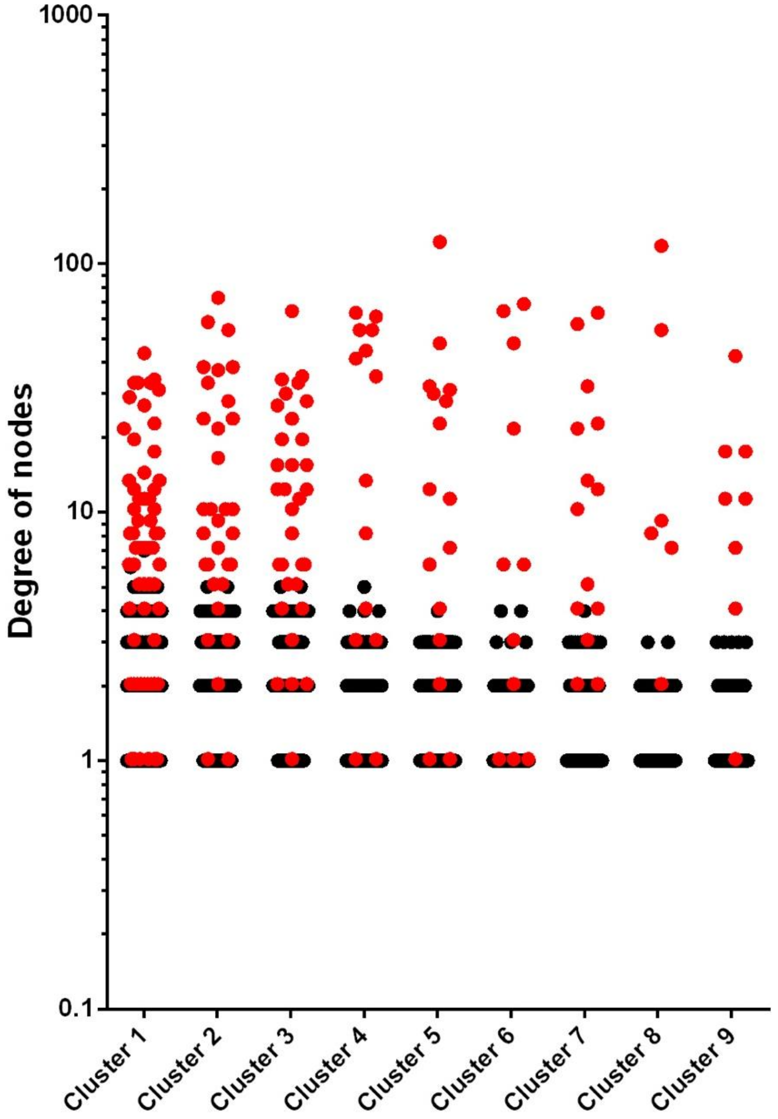

Figure 3 Degrees of transcription factors and genes in each cluster. Red indicates transcription factors and black indicates the genes they potentially regulate. Y-axis depicts the node degrees with the smallest value of '1' and the highest observed degree being 117 (Cluster 5). Note: there are no values for degree below '1' but I have created my graph from 0.1 on a log scale for better visualization).

Table 4 Hub nodes in each cluster. The hub nodes in each cluster arranged from the smallest to the largest cluster. The most significant node has been selected based on its degree being more than one standard deviation above the mean degree of each cluster and eigen centrality values. Gene names have been provided for the nodes if available from homologous *Drosophila* genes. and, if not, I use the Uniprot ID.

Cluster No:	**Probe ID**	**Degree**	**Bee Base ID**	**Eigenvector centrality values**	**Gene name**
1	AM06919	42	GB44769	0.387	*CG32121*
2	AM09526	70	GB45259	0.3832	*CG9932*
3	AM09450	62	GB42142	0.5643	*ftz-f1*
4	AM05115	61	GB44791	0.3654	*Myb*
5	AM04747	117	GB51429	0.6381	*Lag1*
6	AM08033	66	GB54118	0.525	*Rotund*
7	AM04205	61	GB46492	0.5023	*CrebB-17A*
8	AM09018	113	GB51757	0.6534	*CG17912*
9	AM06097	41	GB55012	0.6707	*Dp*

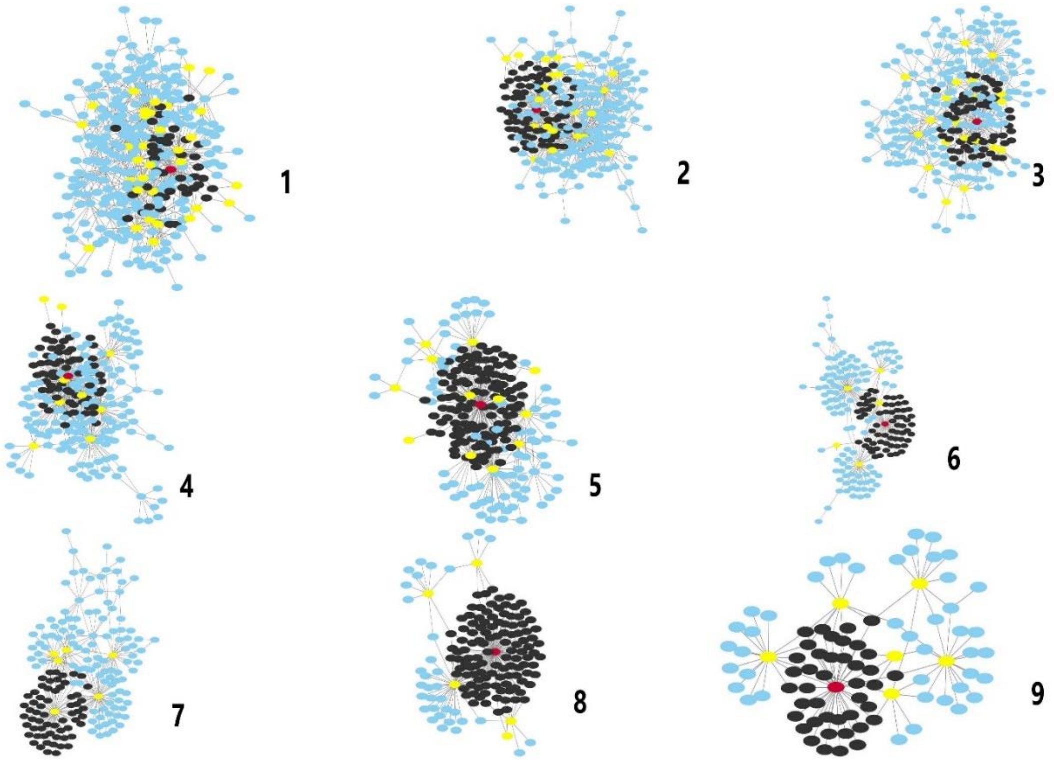

Figure 4 The hub of each cluster with its first and second step neighbours. Each cluster arranged from largest to smallest (1 to 9) with the hub node marked in red, its first neighbour nodes marked in black and it's second neighbour nodes marked in yellow. All the other nodes in the network are blue coloured.

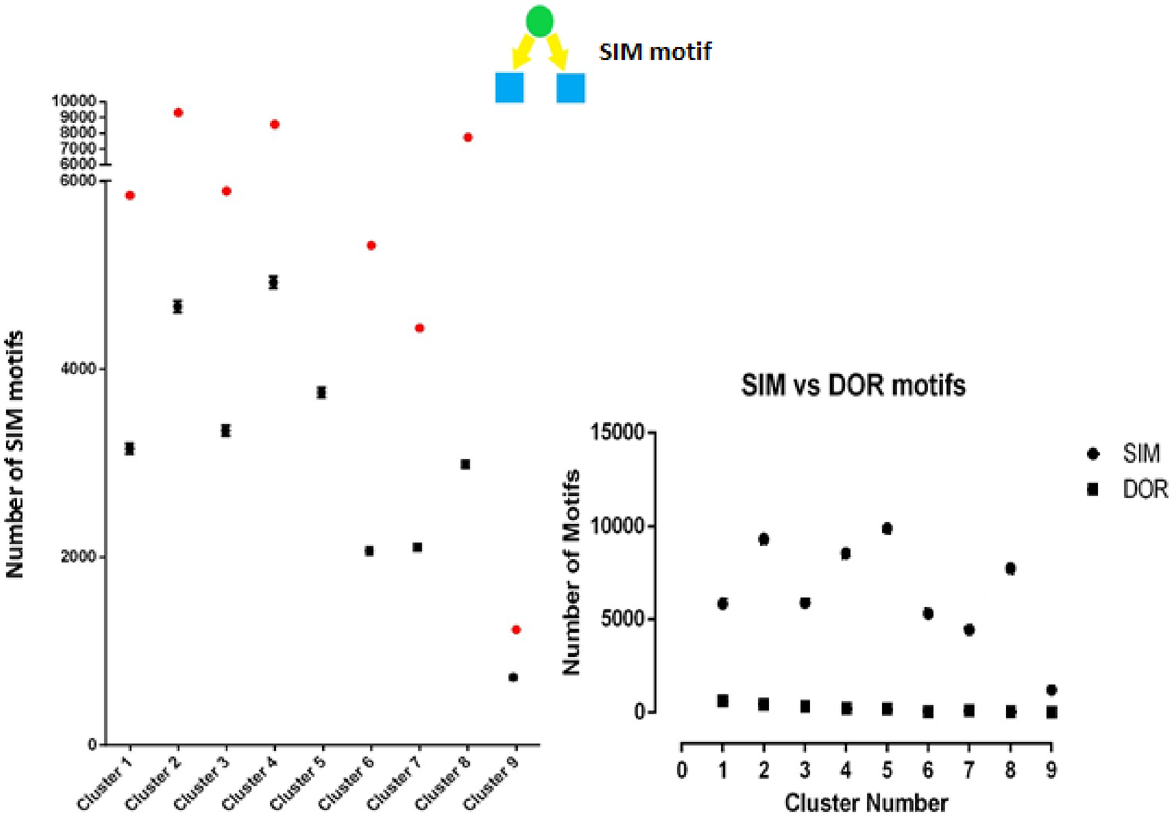

Figure 5 A comparison of the number of motifs against their corresponding random networks. (**A)** Box and whisker plots showing the range in number of SIM (Single Input Module; graphically shown at top) motifs from 500 random networks generated for each cluster. Red dot depicts the observed number of SIM motifs from my honey bee gene regulatory clusters. (**B)** Comparison plot between the number of SIM and DOR (Dense Overlapping Regulons) motifs in each cluster.

Table 5 The number of SIM (Single Input Module) and DOR (Dense Overlapping Regulon) motifs in each cluster. A one sample *t*-test with the mean of the number of SIM motifs in random networks against the number of motifs found in our clusters. The number of DOR motifs discovered in each cluster is also noted.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Real cluster SIM motif number	5846	9312	5892	8555	9876	5315	4437	7740	1229
Real cluster DOR motif number	629	450	337	226	189	56	104	46	32
Random network SIM motif mean	3177.39	4699.51	3370.11	4958	3777.53	2080.24	2119.39	3007.67	727.32
Discrepancy	2668.61	4612.49	2521.89	3597	6098.47	3234.76	2317.61	4732.33	501.68
P-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

3.5 Gene Ontology analysis

Out of 1,834 genes, a total of 1,526 genes in the whole network were mapped to different *D. melanogaster* Gene Ontology IDs by DAVID. Table 6 contains information on the significantly enriched gene ontology terms in each cluster. All the clusters provided evidence of being enriched for gene ontology terms except Cluster 5.

A total of 34 gene ontology enrichment categories are present when all the clusters were analyzed using DAVID. The number of enrichment terms between clusters ranged from 0 in Cluster 5 to a total of twelve in Cluster 1. An analysis on Ensembl BioMart showed that most of the genes in each cluster has corresponding functional information (through both computational and experimental studies). However, the majority of genes did not pass the minimum threshold in DAVID to be considered enriched for a particular gene ontological function. Several of our clusters showed gene ontology terms associated with the molecular function of binding like ATP binding (GO:0005524) and nucleotide binding (GO:0000166). Our clusters of interest with a statistically significant number of sterility genes mapping onto them, i.e., Cluster 1 and Cluster 3 were found to have genes enriched for the ontology terms transcription factor binding activity (GO:0003700) and nucleotide binding (GO:0000166) respectively. However, functional ontology analysis did not give any enrichment terms pertaining specifically to reproduction for either cluster.

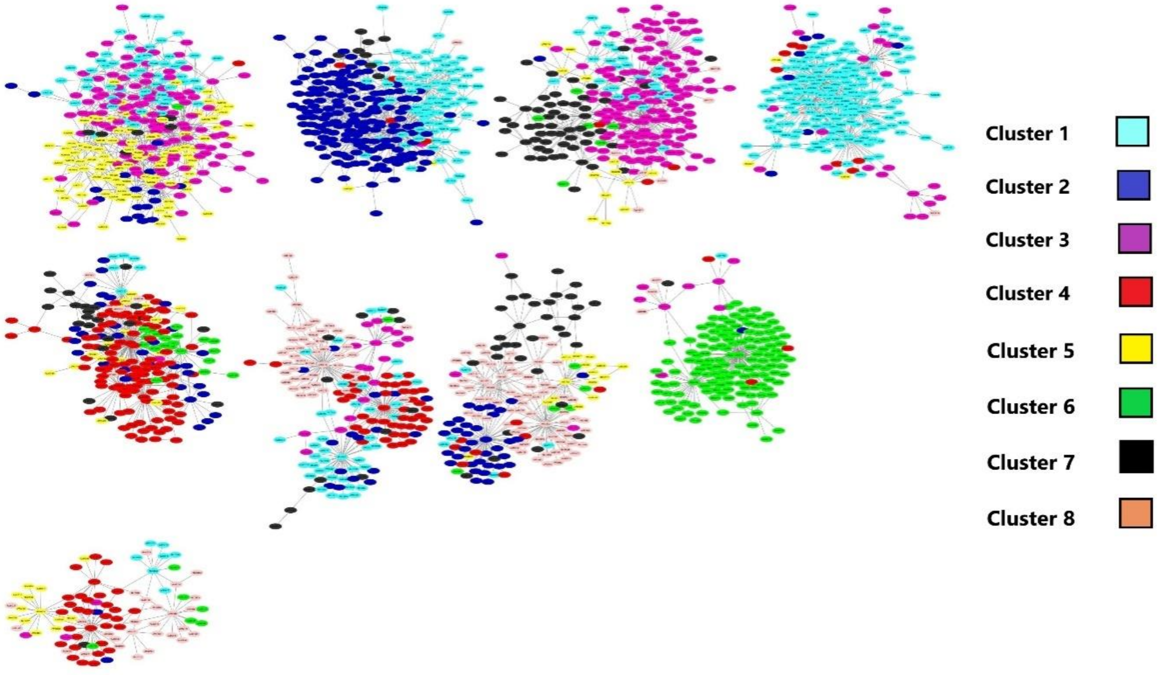

Figure 6. Mapping the arrangement of nodes from the study of Sobotka et al. (2016) onto my new clustering arrangement. Each colour shown in legends (1-8) represents a cluster from Sobotka et al. (2016) that I mapped onto the nine clusters of the updated transcriptional regulatory network (unmatched genes are left out). The new clusters are arranged based on size (left to right, top to bottom) from the largest Cluster 1 to the smallest Cluster 9.

Table 6 Gene ontology analysis data of each cluster. Gene enrichment information obtained using homologous *D. melanogaster* genes in the DAVID (Database for Annotation, Visualization, and Integrated Discovery) database for each cluster. The number of genes in each cluster with a role to play in each function as well as the *P*-value which measures the significance of the enrichment for each function is provided. The *P*-values have been corrected for multiple tests using Benjamini correction.

Gene Ontology Process	*P*-Value	Number of Genes
Cluster 1		
GO:0003700 transcription factor activity, sequence-specific DNA binding	0.003	10
GO:0043565 sequence-specific DNA binding	0.019	10
GO:0005249 voltage-gated potassium channel activity	0.039	3
GO:0003676 nucleic acid binding	0.046	16
GO:0030130 clathrin coat of trans-Golgi network vesicle	0.049	2
GO:0030132 clathrin coat of coated pit	0.049	2
GO:0016192 vesicle-mediated transport	0.070	4
GO:0005634 nucleus	0.075	17
GO:0006099 tricarboxylic acid cycle	0.075	3
GO:0006355 regulation of transcription, DNA-templated	0.086	8
GO:0051287 NAD binding	0.096	3
GO:0005786 signal recognition particle, endoplasmic reticulum targeting	0.096	2
Cluster 2		
GO:0005737 cytoplasm	9.23E-04	15
GO:0003700 transcription factor activity, sequence-specific DNA binding	0.063	7
GO:0015991 ATP hydrolysis coupled proton transport	0.070	3
GO:0009331 glycerol-3-phosphate dehydrogenase complex	0.076	2
GO:0005524 ATP binding	0.086	20
Cluster 3		
GO:0000166 nucleotide binding	0.019	8
GO:0005886 plasma membrane	0.032	7
Cluster 4		
GO:0005634 nucleus	1.23E-05	20
GO:0005524 ATP binding	0.002	23
GO:0003676 nucleic acid binding	0.005	16
GO:0000166 nucleotide binding	0.080	7
GO:0016569 covalent chromatin modification	0.081	2

GO:0004525 ribonuclease III activity	0.088	2
Cluster 5		
No significant enrichment of genes		
Cluster 6		
GO:0008270 zinc ion binding	0.047	11
GO:0019509 L-methionine biosynthetic process from methylthioadenosine	0.069	2
Cluster 7		
GO:0016272 prefoldin complex	0.002	3
GO:0003899 DNA-directed RNA polymerase activity	0.044	3
GO:0005885 Arp2/3 protein complex	0.097	2
GO:0005815 microtubule organizing center	0.097	2
Cluster 8		
GO:0008168 methyltransferase activity	0.047	3
Cluster 9		
GO:0016787 hydrolase activity	0.084	3

3.6 Analysis on Clusters 1 and 3 using D. melanogaster homologues

Since Cluster 1 and Cluster 3 had a significantly higher number of sterility genes I decided to only reconstruct these two clusters with homologous *D. melanogaster* genes. Of the 305 genes in Cluster 1, 216 of them had corresponding homologues for *D. melanogaster*. The network that I reconstructed using STRING included all these genes with a total of 238 edges. A total of 94 genes have no connections and I considered them orphans. Furthermore, 12 nodes were found to be connected to just one other node and disconnected from the main component of the network. I considered all the homologous genes for the functional analysis. Of the 216 nodes in the cluster, ClueGO identified 209 genes while 8 genes were not recognized. Only 136 genes out of this list were found to have gene ontology terms associated with them. A total of 87 genes were found to be associated with 45 gene ontology terms/pathways and pass the parameters I had set for functional enrichment on ClueGO, thus being significant. Forty-nine genes were found to be associated with gene ontology terms related to reproduction (14 biological processes in total). The top three genes with the highest number of connections are Armadillo (*arm*), Kayak (*kay*) and Jun-related antigen (*Jra*) with degrees of 23, 18 and 16 respectively.

From the 252 nodes present in *A. mellifera* Cluster 3, only 163 had homologous *D. melanogaster* genes and hence I used only these on STRING for network construction. The constructed network had a total of 117 edges with one giant component made up of 67 nodes. A total of 83 nodes were orphaned with a further 13 nodes forming smaller connected components among themselves. On analyzing with ClueGO, I found a total of 10 ontological terms to be significant with 6 terms being related to reproduction. Based on the parameters set, all gene ontological terms were broadly clustered into three groups with the first two groups containing terms related to different biological processes while the third group consisted of genes enriched for biological terms like anatomical structure development (GO:0048856) and multicellular organism development (GO:0007275). Seventy-three genes are found to be involved in this broad category. The genes with the highest number of connections in Cluster 3 are RE73195p (*r-l*), Actin-C (*Act5C*) and Something about silencing protein 10 (*Sas10*) with degrees of 18, 13 and 9 respectively.

Through gene ontology analysis using ClueGO, total of 122 of the homologous *D. melanogaster* genes had functional roles closely associated with reproduction/embryo development from both the clusters. I found that a total of 45 genes from this list are also present in the sterility gene lists of Sobotka et al. (2016). The other 78 genes which do not yet have a known function in *A. mellifera*, provide us with viable targets that could be analyzed through expression studies to investigate their effects in the *A. mellifera* reproductive pathway.

4. Discussion

In this study, I used a conversion file obtained from Molodtsova et al. (2014) to up-date my copy of the adjacency matrix that previously defined the honey bee brain TRN (Chandrasekaran et al. 2011). Specifically, I updated it to reflect updates that had occurred in the honey bee official gene set v1 to v3.2 (Elsik et al. 2014). This updated network was smaller: it had a total of 1,839 nodes and 5,085 connections, as opposed to 2,382 nodes and 6,756 connections in the original network. I used the GLay clustering algorithm to partition the TRN into nine clusters. These nine clusters followed properties associated with real world networks like a different degree distribution and significantly greater number of motifs when compared with their random counterparts. My gene ontology analysis of all the clusters did not provide evidence for any specific cluster displaying functions related to reproduction or sterility. However, Cluster 1 and Cluster 3 had a significantly greater number of sterility genes (procured and assembled by Sobotka et al. 2016) mapping onto them when compared with the remaining seven clusters. I found the corresponding homologous genes in *D. melanogaster* for the *A. mellifera* genes in Clusters 1 and 3 and constructed corresponding *D. melanogaster* networks. Gene ontology analysis of these clusters helped me identify some key genes that may have important roles in the reproduction/sterility pathway of *D. melanogaster*. Their counterparts in the *A. mellifera* gene set might play a similar role and provides an avenue for further analysis and investigation.

4.1 Structural analysis of the network

The TRN is bipartite in nature with a set of transcription factors regulating a set of genes. The bipartite nature of the network essentially means that there is no interaction (no edges) among the transcription factors or among the genes themselves. Studies on bipartite graphs range across a wide variety of fields including bacterial complexity (Corel et al. 2016), diseases (Goh et al.

2007) and social networks (Borgatti 2009). A network can be constructed to visualize all the interactions occurring between different biological units like genes or proteins, which increases the complexity and diversity of gene networks making it arduous to analyze the network as a whole. It is easier to scrutinize smaller portions of the network (or sub-networks) to better understand the underpinning functions and interconnectivity between related genes (D'haeseleer et al. 2000). Even-though clustering is not a compulsory mechanism to study a network, it can help to see key network features (de Oliveira et al. 2008). Based on the default clustering criteria of each clustering algorithm (available on CYTOSCAPE), I obtained a different number of clusters for my updated *A. mellifera* network (2 clusters to 127 clusters). The number of nodes in our largest to smallest cluster obtained from our TRN after clustering ranged from 305 to 90 (Figure 2; Table 2), but the number of transcription factors ranged from 62 in Cluster 1 to just eight in Cluster 9 (Figure 3). Transcription factors have many regulatory connections and thus have a consistently higher degree when compared to the downstream target genes that they regulate. I only obtained one more cluster when compared to Sobotka et al. (2016) using the same clustering algorithm, which is not a considerable difference.

To test if each cluster conformed to properties displayed by real world networks (and not a random collection of edges), I decided to test the clusters for specific properties associated with real world networks. One feature of real-world networks is the decaying nature of its degree distribution graphs due to the presence of nodes with both small and significantly larger degrees thus differing from random networks (networks generated by forming connections between nodes without considering any node characteristics), which have a low degree heterogeneity value (a network with all the nodes having a single degree *k* is a homogenous network). The degree distribution of random networks have a Poisson distribution with a bell curve (Strogatz

2001). Some of the properties that characterize real-world networks are, 1- small-world property, 2 - high clustering co-efficient values and 3- large, connected component (Wadhwa and Bhatia 2013). Similarly, an oft accepted property of a real-world network is that real-world networks are scale-free in nature, which typically means that the degree distribution of the network follows a power law with a decay constant between 2 and 3. However, recent research has disproved this notion, stating that biological networks are rarely scale free and do not necessarily need to follow this rule (Broido and Clauset 2018). The slopes of none of my nine clusters fell between 2 and 3 (Table 3), however, as stated, this isn't conclusive evidence to disregard my clusters and state that they are random. Hence, I used the KS test to show that there is a difference between the degree distribution of my clusters and their random counterparts.

Another feature of real-world networks is that they have a significantly greater number of motifs when compared to their analogous random networks (Song et al. 2005). There are multiple types of motifs that can be detected in bipartite networks (Saracco et al. 2016) but I focussed on just two types, each with just three nodes: the so-called SIM and DOR motifs (Shen-Orr et al. 2002). In each of the nine clusters (Table 5), I found that the number of SIM motifs were much more (in some cases even double) than the number of SIM motifs generated by their equivalent random counterparts (with a significant *P*-value of less than 0.0001). The TFs that are part of SIM motifs are generally auto-regulated, with a majority of them repressing themselves when a threshold of expression is reached (Ali et al. 2020). The TFs also regulate gene expression in these motifs based on the activation threshold of the genes that are a part of it (Shen-Orr et al. 2002). This gives rise to a cascading effect or the sequential activation on genes. SIM network motifs generally show broad biological functions like carbon utilization (Alon 2007). SIM motifs are present in significant numbers in the *Escherichia coli* and *Saccharomyces cerevisiae* gene networks (Lee et al. 2002; Shen-Orr et al. 2002).

4.2 Comparison with the study conducted by Sobotka et al. (2016)

In the study conducted by Sobotka et al. (2016), the largest cluster was a total of 431 nodes while the smallest cluster was 197 nodes. After updating and re-clustering, the *A. mellifera* TRN during our analysis, we got an additional cluster for a total of nine. All the clusters were comparatively smaller, but the reduction in sizes of the clusters is expected due to the removal of obsolete nodes in the *A. mellifera* TRN (more than 500 nodes) that was used by Sobotka et al. (2016) (Figure 6). On comparing the hubs in the clusters between both the studies, five hubs from the eight clusters found in the Sobotka et. al (2016) study were also found to be hubs in five of the nine clusters in my study. The difference in hub genes between some clusters found in the study done by Sobotka et. al (2016) and my analysis could be attributed to the decrease in connections between genes due to the removal of obsolete genes from the network. Also, Sobotka et. al (2016) had one cluster that she defined as the "sterility cluster" while I had two. However, there was close to 80% overlap between Sobotka's "sterility cluster" and my two clusters of interest leading to the conclusion that my clusters were composed mostly of the same nodes that were present in the single Sobotka cluster. Hence, both studies effectively predicted a similar set of genes present in the "sterility pathway of *A. mellifera* with my analysis having the advantage of being more streamlined with an updated gene set and the removal of obsolete and redundant genes.

All the *A. mellifera* genes that occupied hub positions in the clusters were of unknown function (Figure 4), hence their homologous *D. melanogaster* genes have been used to predict their functions. Cluster 1 and Cluster 3 had the hub genes *CG32121* and *ftz-f1* respectively (Table 4).

The first gene of interest *CG32121* is in the *D. melanogaster* geneset, and through sequence similarity analysis, its function has been predicted to be sequence specific DNA binding (GO:0043565). Nuclear hormone receptor (*ftz-f1*) is a well-studied gene and works as a co-factor to the fushi tarazu (*ftz*) gene facilitating its binding to DNA. Fushi tarazu is a homeotic protein and plays a role in the segmentation of *D. melanogaster* embryos. A mutation (or its absence) in the *ftz-f1* gene causes the same defect as a lack of the *ftz* gene (*ftz* though present is unable to activate its target genes) and could lead to cuticular defects (Yu et al. 1997).

Ontology analysis of each of the clusters for enriched functional roles of the genes in each cluster identified very few genes from each cluster taking part in functional roles, which was expected since most genes in each cluster are yet to be identified and functionally classified (Table 6). Almost all the clusters showed functional enrichment for biological processes involved in binding like DNA binding (GO:0043565), NAD binding (GO:0051287), ATP binding (GO:0005524) etc. This leads me to believe that the clusters are not mutually exclusive and some of the genes may be part of pathways with the genes present in other clusters. This is to be expected since all the clusters are part of the single TRN. Different clusters had genes specific to different biological functions (supplementary materials), which could be indicative of the localized nature (since the same biological functions were not observed in other clusters) of some of the biological processes and molecular functions found enriched in each cluster. Even though I observed most genes from the sterility gene list converging onto two specific clusters i.e., Cluster 1 and Cluster 3, I did not obtain any significant enrichment for processes related to *A. mellifera* reproduction in these two clusters. I deduce that this is either due to not enough genes being present in each individual cluster to give functional enrichment based on DAVID’s algorithm or due to the lack of functional information being available on these genes.

4.3 An analysis of the D. melanogaster clusters 1 and 3

Even though the orders containing the species *D. melanogaster* and *A. mellifera* diverged ~350 million years ago (Lovegrove et al. 2020), the potential to use *D. melanogaster* as a model to study social insect behaviour has been explored in different studies (Brenman-Suttner et al. 2020; Camiletti and Thompson 2016; Reaume and Sokolowski 2011). Furthermore, various studies have made observations on the tendency of female fruit flies to decrease their ovary activity from a treatment of QMP. This suggested the presence of similar genes and pathways regulating the process of reproduction in both the species (Croft et al. 2017).

The clusters with more sterility genes i.e., Cluster 1 and Cluster 3 provided a total 122 genes that could have a potential role to play in the reproduction of *D. melanogaster*. When elucidating the role of these genes in *A. mellifera*, it was inferred that a significant number of these genes (a total of 44) have already been identified in the literature survey conducted by Sobotka et al. (2016) as having a role to play in the sterility/reproduction of *A. mellifera*. However, it is currently unknown if the remaining 78 genes have a role to play in the reproduction of *A. mellifera* and could be further investigated.

The *Armadillo* segment polarity gene (*arm*, gene with the highest degree in the homologous *D. melanogaster* Cluster 1) has been identified to have a vital role in the development of the nervous system (Loureiro and Peifer 1998). The next two genes with the highest number of connections are transcription factors that work in conjunction in specific cells for embryo development (Perkins et al. 1990; Zhang et al. 1990). *Arm* and *kay* have been identified to be differentially expressed in female bees with active and inactive ovaries (Niño 2012). Though forager bees have been known to be enriched with the *Jra* gene (Vannette et al. 2015), its role in reproduction is yet to be investigated.

The *r-l* gene identified as the gene with the highest degree (number of connections) in Cluster 3, is a gene involved in pyrimidine synthesis activity in *D. melanogaster* (Eisenberg et al. 1990). The same gene covers a similar role in *Homo sapiens* and its mutation/silencing has been known to cause Pyrimidine Metabolic Disorder that causes developmental problems (Nyhan 2005). In *D. melanogaster*, a study of the mutations in different genes involved in the pyrimidine synthesis pathway found shortened wings and differences in cuticle pigmentation in the mutants when compared to the control (Rawls 2006). The second gene with the highest degree in Cluster 3, *Act5C*, is one of two isoforms of the actin gene found in *D. melanogaster*. The gene is essential for the development of the cytoskeleton and its loss has been shown to be lethal (Wagner et al. 2002). *Sas10* (the gene with the third highest number of connections in Cluster 3) is involved in DNA silencing and has a role to play in development (Peters et al. 2003). In *Mus musculus*, the gene has been found to be essential for brain development (Sakuma et al. 2001).

4.4 Conclusion and future direction

Recent advances in genetic studies have focused interest on the construction of gene networks from expression arrays (Rapaport et al. 2007). Gene networks have enabled the prediction of novel gene functions to progress at a significantly higher rate through the construction of gene co-expression networks in which gene relationships may reflect their involvement in common biological pathways (Hwang et al. 2011). However, network construction using microarray expression information suffers from a defect in that networks constructed off the expression of genes in a single time point may not sufficiently demonstrate the inter-connectivity of genes in an organism. Even though *A. mellifera* has a genome of more than 10,000 genes (Honeybee Genome Sequencing Consortium 2006), our network was comprised of only 1839 genes, hence, the network is incomplete to deduce the complete sterility pathway in *A. mellifera*. Also, in most cases, there is only enough information to create fragments of the network (Rapaport et al. 2007). The same genes may function differently based on factors like the physical condition, age, physiological characteristics, time of the day etc. in the same organism. Utilizing data repositories like STRING can significantly diminish this drawback by using gene interaction information procured over multiple studies and conditions. Another key feature that needs to be well investigated while working on networks is the type of clustering algorithm that is being used. There are a multitude of clustering algorithms available, each with its own set of advantages and disadvantages (Emmons et al. 2016) providing a different number of clusters for the same network. This can affect the results of the same study and hence careful consideration needs to be exercised.

The network studies performed here offer insights into how new network models can be analyzed to glean essential information. The network features and the key genes identified in this study can be further targeted to check for changes in the *A. mellifera* phenotype through gene knock-out studies. In one previous study, genes playing a role in *D. melanogaster* segmentation were identified and their orthologues knocked down in *A. mellifera*. Patterning defects were noticed, implying that these genes played a role in the segmentation of *A. mellifera* embryos too (Wilson and Dearden 2012). Hence, knocking down orthologous genes in *A. mellifera* with known functions in *D. melanogaster* is a viable methodology to help pinpoint gene functions. For future studies, it would be more convenient to work with the functional annotation information specific to *A. mellifera* without using *D. melanogaster* as a homologous model. This would minimize errors that could have crept in due to gene conversion using the various repositories available.

References

Abbot P, Abe J, Alcock J, Alizon S, Alpedrinha JA, Andersson M, Andre JB, Van Baalen M, Balloux F, Balshine S, Barton N (2011) Inclusive fitness theory and eusociality. Nature 471:p.E1.

Abouheif E, Wray GA (2002) Evolution of the gene network underlying wing polyphenism in ants. Science 297:249-52.

Alaux C, Sinha S, Hasadsri L, Hunt GJ, Guzmán-Novoa E, DeGrandi-Hoffman G, Uribe-Rubio JL, Southey BR, Rodriguez-Zas S, Robinson GE (2009) Honey bee aggression supports a link between gene regulation and behavioral evolution. Proc Natl Acad Sci USA 106:15400-5.

Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118:4947-4957.

Ali MZ, Parisutham V, Choubey S, Brewster RC (2020). Inherent regulatory asymmetry emanating from network architecture in a prevalent autoregulatory motif. eife 9: e56517.

Ament SA, Wang Y, Chen CC, Blatti CA, Hong F, Liang ZS, Negre N, White KP, Rodriguez-Zas SL, Mizzen CA, Sinha S (2012) The transcription factor Ultraspiracle influences honey bee social behavior and behavior-related gene expression. PLoS genet 8:e1002596.

Backx A, Guzman-Novoa E, Thompson G (2012) Factors affecting ovary activation in honey bee workers: a meta-analysis. Insectes Soc 59:381-388.

Barchuk AR, Cristino AS, Kucharski R, Costa LF, Simões ZL, Maleszka R (2007) Molecular determinants of caste differentiation in the highly eusocial honeybee *Apis mellifera*. BMC developmental biology 7:70.

Barron AB, Oldroyd BP (2001) Social regulation of ovary activation in ’anarchistic’ honey-bees (*Apis mellifera*). Behav Ecol Sociobiol 49:214-219.

Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. Third international AAAI conference on weblogs and social media 8:361-362.

Batra SWT (1966) Nests and social behavior of halictine bees of India (Hymenoptera: Halictidae). Indian J Entomol 28:375– 393.

Bell G (1982). The Masterpiece of Nature: The Evolution and Genetics of Sexuality. Cambridge University Press, Cambridge, UK.

Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman W, Pagès F, Trajanoski Z, Galon J (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 25:1091-1093.

Blondel VD, Guillaume J, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech: Theory Exp 2008:P10008.

Borgatti SP (2009) 2-Mode concepts in social network analysis. Encyclopedia of complexity and system science 6: 8279-8291.

Bourke AF (2011) The validity and value of inclusive fitness theory. Proc Biol Sci 278:3313-3320.

Brazhnik P, de la Fuente A, Mendes P (2002) Gene networks: how to put the function in genomics. Trends Biotechnol 20:467-472.

Brenman-Suttner DB, Yost RT, Frame AK, Robinson JW, Moehring AJ, Simon AF (2020) Social behavior and aging: A fly model. Genes, Brain and Behavior, 19(2), e12598.

Broido AD, Clauset A (2018) Scale-free networks are rare. Nat Commun 10:1017

Butler CG (1957) The process of queen supersedure in colonies of honeybees (*Apis mellifera Linn*.). Insectes sociaux 4:211-223.

Butler CG, Fairey EM (1963) The role of the queen in preventing oogenesis in worker honeybees. J Apic Res 2:14-8.

Camiletti AL, Percival-Smith, A, Thompson GJ (2013) Honey bee queen mandibular pheromone inhibits ovary development and fecundity in a fruit fly. Entomol Exp Appl 147:262-268.

Camiletti AL, Thompson GJ (2016) *Drosophila* as a genetically tractable model for social insect behavior. Ecol Evol 4:40.

Cardoen D, Ernst UR, Van Vaerenbergh M, Boerjan B, De Graaf DC, Wenseleers T, Schoofs L, Verleyen P (2011) Differential proteomics in dequeened honeybee colonies reveals lower viral load in hemolymph of fertile worker bees. PLoS One 6:e20043

Cardoen D, Wenseleers T, Ernst UR, Danneels EL, Laget D, De Graaf DC, Schoofs L, Verleyen P (2011) Genome-wide analysis of alternative reproductive phenotypes in honeybee workers. Mol Ecol 20:4070-4084.

Carlisle DB, Butler CG (1956) The ‘Queen-Substance’ of honeybees and the ovary-inhibiting hormone of crustaceans. Nature 177:276.

Chandrasekaran S, Ament SA, Eddy JA, Rodriguez-Zas SL, Schatz BR, Price ND, Robinson GE (2011) Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. Proc Natl Acad Sci USA 108:18020-18025.

Charlesworth B (1978) Some models of the evolution of altruistic behaviour between siblings. J Theor Biol 72:297-319.

Chouvardas P, Kollias G, Nikolaou C (2016) Inferring active regulatory networks from gene expression data using a combination of prior knowledge and enrichment analysis. BMC Bioinformatics 17:181.

Corel E, Lopez P, Méheust R, Bapteste E (2016) Network-thinking: Graphs to analyze microbial complexity and evolution. Trends Microbiol 24:224-237.

Costello JC, Dalkilic MM, Beason SM, Gehlhausen JR, Patwardhan R, Middha S, Eads BD, Andrews JR (2009) Gene networks in *Drosophila melanogaster*: integrating experimental data to predict gene function. Genome Biol 10:R97.

Croft JR, Liu T, Camiletti AL, Simon AF, Thompson GJ (2017) Sexual response of male *Drosophila* to honey bee queen mandibular pheromone: implications for genetic studies of social insects. J Comp Physiol A 203:143-149.

Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal, Complex Syst. 1695:1-9.

Crozier RH, Pamilo P (1996) Evolution of social insect colonies. Oxford University Press, Oxford, UK.

Curators F (2008) Assigning gene Ontology (GO) terms by sequence similarity in FlyBase. FlyBase analysis.

Davidson E, Levin M (2005) Gene regulatory networks. Proc Natl Acad Sci USA 102:4935.

D'haeseleer P, Liang S, Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16:707-726.

de Oliveira TB, Zhao L, Faceli K, de Carvalho AC (2008) Data clustering based on complex network community detection. IEEE Congress on Evolutionary Computation 2121-2126.

Domedel-Puig N, Pournara I, Wernisch L (2010) Statistical model comparison applied to common network motifs. BMC Syst Biol 4:18.

Eisenberg M, Gathy K, Vincent T, Rawls J (1990) Molecular cloning of the UMP synthase gene rudimentary-like from *Drosophila melanogaster*. Mol Gen Genet 222:1-8.

Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, de Graaf DC, Debyser G, Deng J, Devreese B (2014) Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics 15:86.

Emmons S, Kobourov S, Gallant M, Börner K (2016) Analysis of network clustering algorithms and cluster quality metrics at scale. PloS one 11:e0159161.

Evans JD, Wheeler DE (1999) Differential gene expression between developing queens and workers in the honey bee, *Apis mellifera*. Proc Natl Acad Sci USA 96:5575-5580.

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL (2007) The human disease network. Proc Natl Acad Sci USA 104:8685-8690.

Grozinger CM, Fan Y, Hoover SE, Winston ML (2007) Genome-wide analysis reveals differences in brain gene expression patterns associated with caste and reproductive status in honey bees (*Apis mellifera*). Mol Ecol 16:4837-4848.

Grozinger CM, Sharabash NM, Whitfield CW, Robinson GE (2003) Pheromone-mediated gene expression in the honey bee brain. Proc Natl Acad Sci U S A 100 Suppl 2:14519-14525.

Hamilton WD (1964) The genetical evolution of social behaviour. II. J Theor Biol 7:17-52.

Haye A, Albert J, Rooman M (2014) Modeling the *Drosophila* gene cluster regulation network for muscle development. PloS one 9(3):e90285.

Honeybee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443:931-949.

Hwang S, Rhee SY, Marcotte EM, Lee I (2011) Systematic prediction of gene function in *Arabidopsis thaliana* using a probabilistic functional gene network. Nat Protoc 6:1429.

Jonsson PF, Cavanna T, Zicha D, Bates PA (2006) Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. BMC Bioinformatics 7:2.

Keeling CI, Slessor KN, Higo HA, Winston ML (2003) New components of the honey bee (*Apis mellifera L.*) queen retinue pheromone. Proc Natl Acad Sci USA 100:4486-4491.

Kersey PJ, Allen JE, Allot A, Barba M, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Grabmueller C (2017) Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res 46:D802-D808.

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database 11: bar049.

Lago DC, Humann FC, Barchuk AR, Abraham KJ, Hartfelder K (2016) Differential gene expression underlying ovarian phenotype determination in honey bee, *Apis mellifera L.*, caste development. Insect Biochem Mol Biol 79:1-12.

Lattorff HM, Moritz RF (2013) Genetic underpinnings of division of labor in the honeybee (*Apis mellifera*). Trends Genet 29:641-8.

Liseron-Monfils C, Ware D (2015) Revealing gene regulation and associations through biological networks. Current Plant Biology 3:30-39.

Loureiro J, Peifer M (1998) Roles of *Armadillo*, a *Drosophila* catenin, during central nervous system development. Curr Biol 8:622-633.

Maisonnasse A, Alaux C, Beslay D, Crauser D, Gines C, Plettner E, Le Conte Y (2010) New insights into honey bee (*Apis mellifera*) pheromone communication. Is the queen mandibular pheromone alone in colony regulation? Front Zool 7:18.

McDonnell MD, Yaveroğlu ÖN, Schmerl BA, Iannella N, Ward LM (2014) Motif-role-fingerprints: the building-blocks of motifs, clustering-coefficients and transitivities in directed networks. PloS one 9:e114503.

Michener CD, Brothers DJ (1974) Were workers of eusocial Hymenoptera initially altruistic or oppressed? Proc Natl Acad Sci USA 71:671-674.

Molodtsova D, Harpur BA, Kent CF, Seevananthan K, Zayed A (2014) Pleiotropy constrains the evolution of protein but not regulatory sequences in a transcription regulatory network influencing complex social behaviors. Front Genet 5:431.

Moore AJ, Breed MD, Moor MJ (1987). The guard honey bee: ontogeny and behavioural variability of workers performing a specialized task. Animal Behaviour, 35(4), 1159-1167.

Mullen EK, Daley M, Backx AG, Thompson GJ (2014) Gene co-citation networks associated with worker sterility in honey bees. BMC Syst Biol 8:38.

Müller HJ (1961) The Ontogeny of Insects. Acta symposii de evolutione insectorum, Praha 1959. Entomologia Experimentalis et Applicata 4:334

Mumoki FN, Pirk CW, Yusuf AA, Crewe RM (2018) Reproductive parasitism by worker honey bees suppressed by queens through regulation of worker mandibular secretions. Sci Rep 8:7701.

Nayar JK (1963) Effect of Synthetic 'Queen Substance'(9-oxodec-trans-2-enoic acid) on Ovary Development of the House-fly, *Musca domestica L*. Nature 197(4870):923.

Newman ME (2005) Power laws, Pareto distributions and Zipf's law. Contemp Phys 46:323-351.

Nyhan WL (2005) Disorders of purine and pyrimidine metabolism. Mol Genet Metab 86:25-33.

Perkins KK, Admon A, Patel N, Tjian R (1990) The *Drosophila* Fos-related *AP-1* protein is a developmentally regulated transcription factor. Genes Dev 4:822-834.

Oldroyd BP, Osborne KE (1999) The evolution of worker sterility in honeybees: the genetic basis of failure of worker policing. Proc R Soc Lond B Biol Sci 266:1335.

Oldroyd BP, Smolenski AJ, Cornuet JM, Crozler RH (1994) Anarchy in the beehive. Nature 371:749

Parker GA (1989) Hamilton's rule and conditionality. Ethol Ecol Evol 1:195-211.

Peterman WE, Ousterhout BH, Anderson TL, Drake DL, Semlitsch RD, Eggert LS (2016) Assessing modularity in genetic networks to manage spatially structured metapopulations. Ecosphere 7:e01231.

Peters NT, Rohrbach JA, Zalewski BA, Byrkett CM, Vaughn JC (2003) RNA editing and regulation of *Drosophila* 4f-rnp expression by sas-10 antisense readthrough mRNA transcripts. RNA 9:698-710.

Plowes N (2010) An introduction to eusociality. Nature Education Knowledge 3:7.

Queller, DC, Strassmann, JE (1998) Kin selection and social insects. Bioscience 48:165-175.

Ramsahai E, Walkins K, Tripathi V, John M (2017) The use of gene interaction networks to improve the identification of cancer driver genes. PeerJ 5:e2568.

Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert J (2007) Classification of microarray data using gene networks. BMC Bioinformatics 8:35.

Ratnieks FL, Foster KR, Wenseleers T (2006). Conflict resolution in insect societies. Annu Rev Entomol 51:581-608.

Ratnieks FL, Visscher PK (1989) Worker policing in the honeybee. Nature 342:796.

Rawls JM Jr (2006) Analysis of pyrimidine catabolism in *Drosophila melanogaster* using epistatic interactions with mutations of pyrimidine biosynthesis and beta-alanine metabolism. Genetics 172:1665-1674.

Reaume CJ, Sokolowski MB (2011) Conservation of gene function in behaviour. Philos Trans R Soc Lond B Biol Sci 366:2100-2110.

Ribeiro P, Silva F (2010) G-tries: an efficient data structure for discovering network motifs. ACM Symposium on Applied Computing 1559-1566.

Ronai I, Oldroyd B, Vergoz V (2016) Queen pheromone regulates programmed cell death in the honey bee worker ovary. Insect Mol Biol 25:646-652.

Ruttner F (1966). The life and flight activity of drones. Bee World, 47(3), 93-100.

Sanchez C, Lachaize C, Janody F, Bellon B, Röder L, Euzenat J, Rechenmann F, Jacq B (1999) Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an Internet database. Nucleic Acids Res 27:89-94.

Sakuma T, Li Q, Jin Y, Choi L, Kim E, Ito K, Ito Y, Nomura S, Bae S (2001) Cloning and expression pattern of a novel PEBP2β-binding protein (charged amino acid rich leucine zipper-1 [*Crl-1*]) in the mouse. Mech Dev 104:151-154.

Sannasi A (1969) Inhibition of ovary development of the fruit-fly, *Drosophila melanogaster* by synthetic “queen substance”. Life Sci 8:785-789.

Saracco F, Di Clemente R, Gabrielli A, Squartini T (2016) Detecting early signs of the 2007–2008 crisis in the world trade. Sci Rep 6:30286.

Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, Ukkonen E, Brazma A (2003) From gene networks to gene function. Genome Res 13:2568-76.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34:166-176.

Seger J (1981) Kinship and covariance. J Theor Biol 91:191-213.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498-2504.

Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31:64.

Sobotka JA, Daley M, Chandrasekaran S, Rubin BD, Thompson GJ (2016) Structure and function of gene regulatory networks associated with worker sterility in honeybees. Ecol Evol, 6:1692-1701.

Song S, Sjöström PJ, Reigl M, Nelson S, Chklovskii DB (2005) Highly nonrandom features of synaptic connectivity in local cortical circuits. PLoS biol 3:e68.

Sporns O, Honey CJ, Kötter R (2007) Identification and classification of hubs in brain networks. PloS one 2:e1049.

Strogatz SH (2001) Exploring complex networks. Nature 410:268.

Su G, Kuchinsky A, Morris JH, States DJ, Meng F (2010) GLay: community structure analysis of biological networks. Bioinformatics 26:3135-3137.

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P (2016) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res 45:D362-D368.

Tan K, Liu X, Dong S, Wang C, Oldroyd BP (2015) Pheromones affecting ovary activation and ovariole loss in the Asian honey bee *Apis cerana*. J Insect Physiol 74:25-29.

Tang Y, Li M, Wang J, Pan Y, Wu F (2015) CytoNCA: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks. BioSystems 127:67-72.

Thompson GJ, Hurd PL, Crespi BJ (2013) Genes underlying altruism. Biol Lett 9:20130395.

Thompson GJ, Kucharski R, Maleszka R, Oldroyd BP (2008) Genome-wide analysis of genes related to ovary activation in worker honey bees. Insect Mol Biol 17:657-665.

Thompson GJ, Kucharski R, Maleszka R, Oldroyd BP (2006) Towards a molecular definition of worker sterility: differential gene expression and reproductive plasticity in honey bees. Insect Mol Biol 15:537-644.

Vannette RL, Mohamed A, Johnson BR (2015) Forager bees (*Apis mellifera*) highly express immune and detoxification genes in tissues associated with nectar processing. Sci Rep 5:16224.

Van Oystaeyen A, Oliveira RC, Holman L, van Zweden JS, Romero C, Oi CA, d'Ettorre P, Khalesi M, Billen J, Wäckers F, Millar JG (2014) Conserved class of queen pheromones stops social insect workers from reproducing. Science 343(6168):287-90.

Verfaillie A, Imrichová H, Van de Sande B, Standaert L, Christiaens V, Hulselmans G, Herten K, Sanchez MN, Potier D, Svetlichnyy D (2014) iRegulon: from a gene list to a gene regulatory network using large motif and track collections. PloS Comput Biol 10:e1003731.

Wadhwa P, Bhatia MPS (2013) An insight into properties of real world networks. International Conference on Advances in Computing, Communications and Informatics 1930-1935.

Wagner CR, Mahowald AP, Miller KG (2002) One of the two cytoplasmic actin isoforms in *Drosophila* is essential. Proc Natl Acad Sci U S A 99:8037-8042.

Wilson EO (1971) The insect societies. The insect societies, Belknap Press

Wilson MJ, Dearden PK (2012) Pair-rule gene orthologues have unexpected maternal roles in the honeybee (*Apis mellifera*). PLoS One 7:e46490.

Winston ML, Higo HA, Colley SJ, Pankiw T, Slessor KN (1991) The role of queen mandibular pheromone and colony congestion in honey bee (*Apis mellifera L.*) reproductive swarming (Hymenoptera: Apidae). J Insect Behav 4:649-660.

West SA, Gardner A (2013) Adaptation and inclusive fitness. Curr Biol 23:R577-R584.

Yu Y, Li W, Su K, Yussa M, Han W, Perrimon N, Pick L (1997) The nuclear hormone receptor *Ftz-F1* is a cofactor for the *Drosophila* homeodomain protein *Ftz*. Nature 385:552.

Zhang K, Chaillet JR, Perkins LA, Halazonetis TD, Perrimon N (1990) Drosophila homolog of the mammalian *jun* oncogene is expressed during embryonic development and activates transcription in mammalian cells. Proc Natl Acad Sci U S A 87:6281-6285.

Supplemental Materials

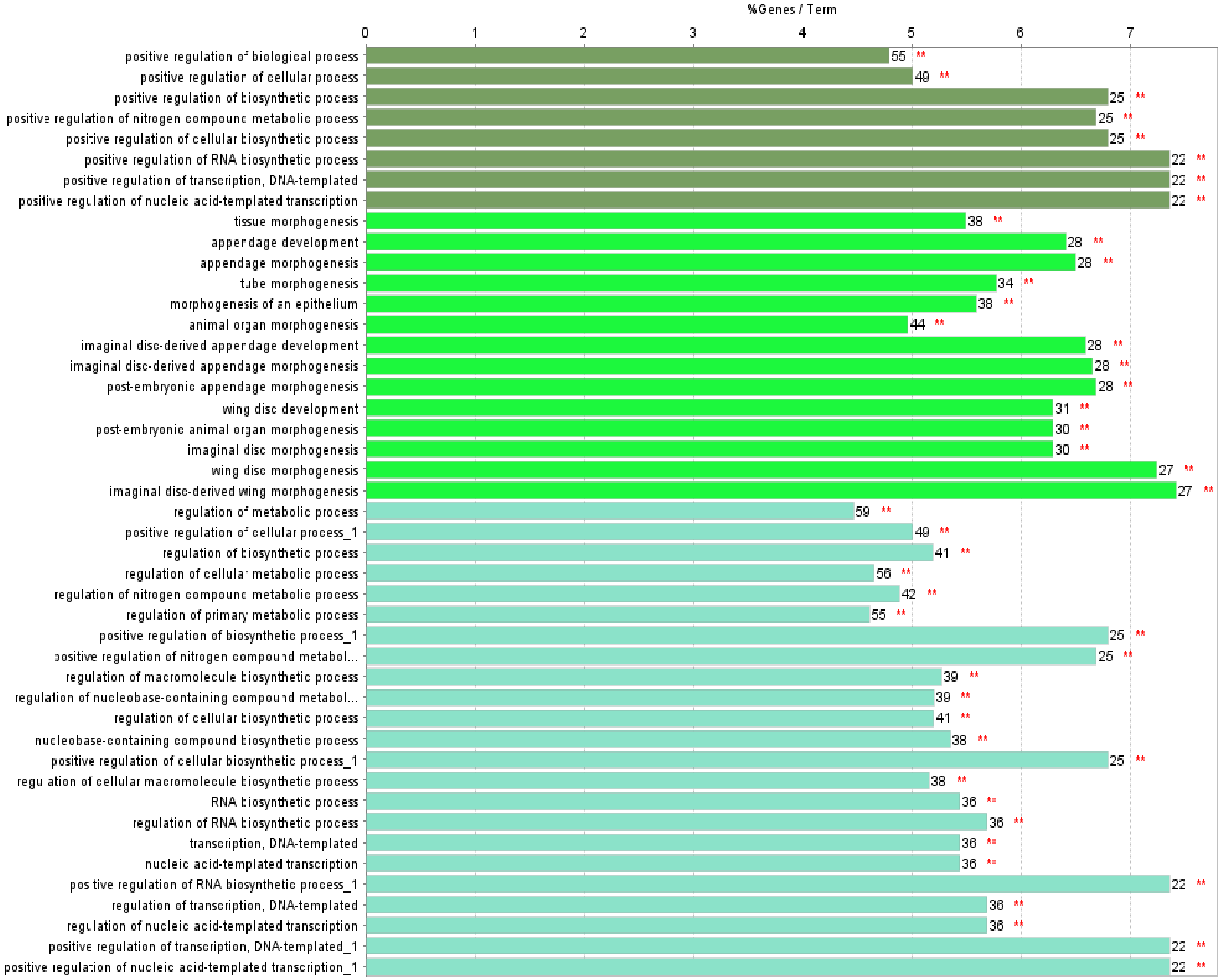

Figure S1 Gene Ontology terms identified from homologous *D. melanogaster* Cluster 1 using ClueGO. Gene Ontology terms identified from the homologous *D. melanogaster* cluster 1 created through STRING. The three colours depict the three groups into which the ontology terms have been segregated based on how related the ontology terms are to each other.

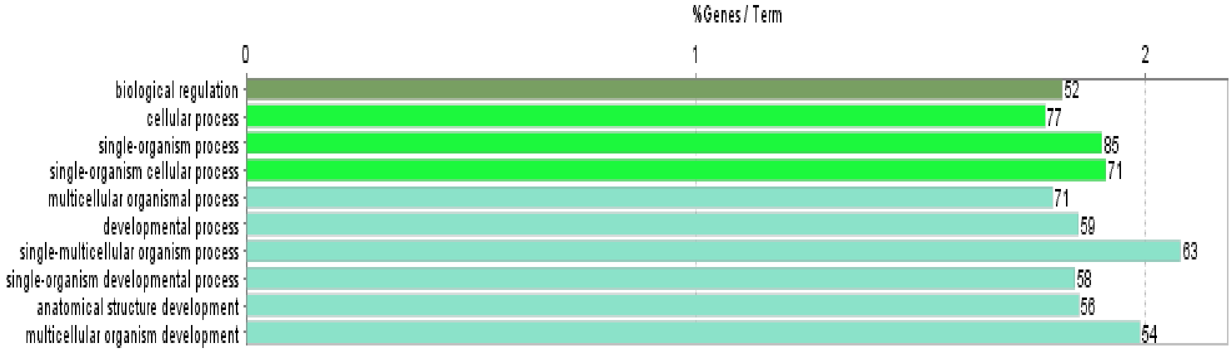

Figure S2 Gene Ontology terms identified from homologous *D. melanogaster* Cluster 3 using ClueGO. Gene ontology terms identified from the D. melanogaster homologous genes for cluster 3. There are three groups of ontology terms based on how related each term is to the other. The ordering is each group of the ontology terms is based on significance.

Table S1 Homologous *D. melanogaster* genes identified from Cluster 1 enriched for terms related to reproduction. Genes of interest found in Cluster 1 with the *D. melanogaster* homologous genes after ontology analysis. The genes have been grouped in the table based on their degree in the *D. melanogaster* clusters. It has been noted on if the *A. mellifera* genes from which the *D. melanogaster* homologues were derived have previously been implicated in any sterility studies.

Flybase ID	Gene Name	Beebase ID	Degree	Present in the list of sterility genes we compiled?
FBgn0000117	*arm*	GB54774	23	Yes Grozinger et al. (2003)
FBgn0001297	*kay*	GB42049	18	No
FBgn0001291	*Jra*	GB53318	16	Yes Grozinger et al. (2003)
FBgn0262733	*Src64B*	GB46371	15	Yes Grozinger et al. (2003)
FBgn0010341	*Cdc42*	GB45657	13	No
FBgn0004101	*bs*	GB47234	9	No
FBgn0000319	*Chc*	GB50357	9	No
FBgn0001624	*dlg1*	GB40648	8	No
FBgn0039227	*polybromo*	GB42921	7	Yes Cardoen et al. (2011b)
FBgn0011655	*Med*	GB50071	6	No
FBgn0261885	*osa*	GB44899	6	No
FBgn0003345	*sd*	GB54841	6	No
FBgn0086357	*Sec61alpha*	GB41886	4	Yes Grozinger et al. (2003)
FBgn0000097	*aop*	GB45540	4	No
FBgn0037555	*Ada2b*	GB52323	4	No

FBgn0020496	*CtBP*	GB43266	4	No
FBgn0010909	*msn*	GB51134	4	No
FBgn0004569	*aos*	GB42377	3	Yes Cardoen et al. (2011b)
FBgn0041111	*lilli*	GB55387	3	No
FBgn0003870	*ttk*	GB47057	3	No
FBgn0010470	*Fkbp14*	GB48497	3	No
FBgn0020386	*Pdk1*	GB43004	2	Yes Grozinger et al. (2003)
FBgn0025879	*Timp*	GB40700	2	Yes Cardoen et al. (2011b)
FBgn0053193	*sav*	GB48671	2	Yes Cardoen et al. (2011b)
FBgn0041604	*dlp*	GB42671	2	No
FBgn0000543	*ecd*	GB42321	2	No
FBgn0086655	*jing*	GB55576	2	No
FBgn0034876	*wmd*	GB43989	2	No
FBgn0260798	*Gprk1*	GB51749	1	Yes Cardoen et al. (2011b)
FBgn0259789	*vfl*	GB52047	1	Yes Cardoen et al. (2011b)
FBgn0035993	*Nf-YA*	GB50732	1	No
FBgn0261064	*Rbsn-5*	GB51395	1	No
FBgn0025571	*SF1*	GB47816	1	No
FBgn0039509	*bigmax*	GB55103	1	No
FBgn0001108	*DCTN1-p150*	GB50038	1	No
FBgn0262582	*cic*	GB43462	1	No
FBgn0004198	*ct*	GB55715	1	No
FBgn0005558	*ey*	GB50342	1	No
FBgn0016081	*fry*	GB42489	1	No
FBgn0027108	*Inx2*	GB45399	0	Yes Cardoen et al. (2011b)
FBgn0266756	*btsz*	GB52239	0	Yes Grozinger et al. (2003)
FBgn0015600	*toc*	GB44180	0	Yes Cardoen et al. (2011b)
FBgn0010473	*tutl*	GB53012	0	Yes Grozinger et al. (2003)

FBgn0010620	*CG10939*	GB55482	0	No
FBgn0030174	*CG15312*	GB55840	0	No
FBgn0266369	*Mtp*	GB49869	0	No
FBgn0052529	*Hers*	GB54625	0	No
FBgn0004449	*Ten-m*	GB48972	0	No
FBgn0004607	*zfh2*	GB54030	0	No

Table S2 Homologous *D. melanogaster* genes identified from Cluster 3 enriched for terms related to reproduction. Genes of interest found in Cluster 3 with the *D. melanogaster* homologous genes after ontology analysis. The genes have been grouped in the table based on their degree in the *D. melanogaster* cluster. It has been noted on if the *A. mellifera* genes from which the *D. melanogaster* homologues were derived have previously been implicated in any sterility studies.

Flybase ID	Gene Name	Beebase ID	Degree	Present in the list of sterility genes we compiled?
FBgn0003257	*r-l*	GB54166	18	No
FBgn0000042	*Act5C*	GB44311	13	Yes Cardoen et al. (2011b)
FBgn0029755	*Sas10*	GB54371	9	No
FBgn0038235	*CG8461*	GB42039	8	No
FBgn0031050	*Arp10*	GB44879	8	No
FBgn0041210	*HDAC4*	GB43234	7	Yes Grozinger et al. (2003)
FBgn0038275	*CG3817*	GB50375	7	No
FBgn0003429	*slo*	GB47138	5	Yes Cardoen et al. (2011b)
FBgn0264607	*CaMKII*	GB49535	4	No
FBgn0003744	*trc*	GB53582	4	No
FBgn0004516	*Gad1*	GB40118	3	Yes Cardoen et al. (2011b)

FBgn0266411	*sima*	GB44532	3	Yes Grozinger et al. (2003)
FBgn0053051	*CG33051*	GB47465	3	No
FBgn0004837	*Su(H)*	GB45655	3	No
FBgn0000289	*cg*	GB44656	3	No
FBgn0263239	*dar1*	GB45841	3	No
FBgn0004656	*fs(1)h*	GB49193	3	No
FBgn0004168	*5-HT1A*	GB48005	2	Yes Grozinger et al. (2003)
FBgn0040508	*ACXC*	GB48102	2	Yes Cardoen et al. (2011b)
FBgn0002921	*Atpalpha*	GB42054	2	Yes Cardoen et al. (2011b)
FBgn0025352	*Thiolase*	GB53132	2	Yes Grozinger et al. (2003)
FBgn0043364	*cbt*	GB45040	2	Yes Cardoen et al. (2011b)
FBgn0261873	*sdt*	GB43138	2	Yes Cardoen et al. (2011b)
FBgn0264075	*tgo*	GB44259	2	Yes Grozinger et al. (2003)
FBgn0046114	*Gclm*	GB40955	2	No
FBgn0016977	*spen*	GB47009	2	No
FBgn0031762	*CG9098*	GB45036	1	Yes Grozinger et al. (2003)
FBgn0015609	*CadN*	GB45972	1	Yes Grozinger et al. (2003)
FBgn0000568	*Eip75B*	GB47224	1	Yes Grozinger et al. (2003)
FBgn0266084	*Fhos*	GB43054	1	Yes Grozinger et al. (2003)
FBgn0003380	*Sh*	GB43660	1	No
FBgn0001078	*ftz-f1*	GB42142	1	Yes Cardoen et al. (2011b)
FBgn0013755	*Bro*	GB45157	1	No
FBgn0023143	*Uba1*	GB55847	1	No
FBgn0000536	*eas*	GB48085	1	No
FBgn0266465	*GckIII*	GB50672	1	No
FBgn0041781	*SCAR*	GB47014	1	No
FBgn0040285	*Scamp*	GB40151	1	No
FBgn0004652	*fru*	GB44836	1	No

FBgn0266672	*Sec8*	GB44781	1	No
FBgn0002524	*lace*	GB42666	1	No
FBgn0035272	*mRpL46*	GB46986	1	No
FBgn0003435	*sm*	GB51622	1	No
FBgn0261238	*Alh*	GB41753	0	Yes Cardoen et al. (2011b)
FBgn0051140	*CG31140*	GB50415	0	Yes Grozinger et al. (2003)
FBgn0031081	*Nep3*	GB41659	0	Yes Cardoen et al. (2011b)
FBgn0034070	*SP2353*	GB40908	0	Yes Grozinger et al. (2003)
FBgn0011481	*Ssdp*	GB45216	0	Yes Grozinger et al. (2003)
FBgn0263352	*Unr*	GB44291	0	Yes Cardoen et al. (2011b)
FBgn0033015	*d4*	GB40564	0	Yes Cardoen et al. (2011b)
FBgn0016794	*dos*	GB55584	0	Yes Cardoen et al. (2011b)
FBgn0000611	*exd*	GB51904	0	Yes Cardoen et al. (2011b)
FBgn0260499	*qvr*	GB47508	0	Yes Cardoen et al. (2011b)
FBgn0041723	*rho-5*	GB49046	0	Yes Cardoen et al. (2011b)
FBgn0261041	*stj*	GB44648	0	Yes Grozinger et al. (2003)
FBgn0030406	*CG1463*	GB51911	0	No
FBgn0050069	*CG30069*	GB55591	0	No
FBgn0052698	*CG32698*	GB42541	0	No
FBgn0031068	*Alr*	GB42690	0	No
FBgn0038037	*Cyp9f2*	GB43728	0	No
FBgn0001079	*fu*	GB54742	0	No
FBgn0031381	*Npc2a*	GB42887	0	No
FBgn0016970	*l(2)k10201*	GB46387	0	No
FBgn0265296	*Dscam2*	GB45774	0	No
FBgn0038402	*Fer2*	GB40407	0	No
FBgn0024963	*GluClalpha*	GB43543	0	No
FBgn0028688	*Rpn7*	GB55528	0	No

FBgn0013334	*Sap47*	GB52438	0	No
FBgn0000449	*dib*	GB47901	0	No
FBgn0010877	*l(3)05822*	GB54950	0	No
FBgn0263594	*lost*	GB48933	0	No
FBgn0033476	*oys*	GB51188	0	No
FBgn0266848	*wap*	GB54449	0	No

Code S1: Python Algorithm to identify Sterility Clusters. Custom Python script used to identify the clusters with most number of sterility genes mapping onto them.

```
import numpy
import scipy
import pandas
import sklearn.preprocessing
import networkx
import copy


# Preserve cluster size
def resample_clusters(clustcols):
clust_size=[261,252,212,209,179,173,153,90]

flat=[]
for i,c in enumerate(clust_size):
flat+=(clustcols[:,i][:c].tolist())
flat=numpy.random.permutation(numpy.array(flat))

rs = numpy.zeros_like(clustcols)
offset=0
for i,c in enumerate(clust_size):
rs[:c,i]=flat[offset:offset+c]
offset+=c

return rs

# Return a list with one entry each gene in 'genelist', telling us which column (cluster)
# it appears in in 'clustcols'
def find_clusters(genelist,clustcols):
gene_is_in_cluster=[]
for g in genelist:
#print g
try:
gene_is_in_cluster.append(numpy.where(clustcols==g[:7])[1][0]+1)
except:
#print g, 'NOT FOUND!'
pass
return numpy.array(gene_is_in_cluster)

# Load clusters
xl_file=pandas.ExcelFile('cluster_ids3.xlsx')
dfs = {sheet_name: xl_file.parse(sheet_name)
for sheet_name in xl_file.sheet_names}
```

```
TBIdf = dfs['Sheet1']

clustcols=TBIdf.values

# Load sterility list
filelist=['Cardoen Sterility Genes .xlsx','Baxck Sterility Genes.xls','Emma Hub Genes
.xlsx','Grozinger 2003 Sterility Genes .xlsx','Grozinger 2007 Sterility Genes .xlsx']
#,'Baxck Sterility Genes.xls','Emma Hub Genes .xlsx','Grozinger 2003 Sterility Genes
.xlsx','Grozinger 2007 Sterility Genes .xlsx'

for f in filelist:
xl_file=pandas.ExcelFile(f)
dfs = {sheet_name: xl_file.parse(sheet_name)
for sheet_name in xl_file.sheet_names}
try:
sheet=dfs['Sheet1'].values
except:
sheet=dfs['Bakc.csv'].values
try:
cardoen = numpy.vstack([cardoen,sheet])
except:
cardoen = sheet

# This would store the clusters from the file
#cardoen_clust=cardoen[:,1].astype(int)
#% hist(cardoen_clust)



# Find which clusters genes belong to in the raw (observed) data
cardoen_clust=find_clusters(cardoen[:,0],clustcols)
observed_dist=numpy.bincount(cardoen_clust)
observed_max=numpy.max(observed_dist)

print (observed_dist)
# Now let's resample the clustcols
null_dist = [] # Full null
max_dist=[] # Distribution of 'largest number of sterility genes found in one cluster
ns=10 # How many times to resample
for i in range(ns):


resampcols = resample_clusters(clustcols)
reclust = find_clusters(cardoen[:,0],resampcols)
null_dist.append(numpy.bincount(reclust))
```

```
max_dist.append( numpy.max(numpy.bincount(reclust)))
null_dist=numpy.array(null_dist)
max_dist=numpy.array(max_dist)
print(null_dist)
print ("This is a break")
print (max_dist)
print (max_dist.shape[0])
print ("This is a break")
print (observed_max)

#hist(max_dist)

p = max_dist[max_dist>=observed_max].shape[0]/float(max_dist.shape[0])
print(max_dist.shape[0])
print (filelist)
print (p)
```

Rahul Choorakkat Unnikrishnan B.Tech MSc

-

Education

2012-2013 Post Graduate Diploma – Bioinformatics and Drug Designing, Siddhaganga University, Karnataka, India

2008-2009 Master of Science - Bioinformatics, University of Leicester, Leicester, UK

2003-2007 Bachelor of Technology - Bioinformatics, Bharath Institute of Science and Technology (BIHER), Tamil Nadu, India

PRESENTATION, PAPERS AND POSTERS

Rahul, C.U. (2016, April). Conserved miRNA detection in the ESTs of *Ganoderma lucidum.* Presentation at the National Conference on Frontiers in Genetics & Genomics, Department of Genomic Science, Central University of Kerala.

Rahul, C.U. (2015, October). Computational Prediction of the Secretome of *Ganoderma lucidum*. Presentation at the International Conference on Advanced IT, Engineering and Management, Mangalore.

Rahul, C.U. (2014, October). Detection of mature miRNAs in the mitochondrial genome of *Phytophthora spp*. Poster presented at the National seminar on New Horizons and Challenges in Biotechnology and Bioinformatics, Kasaragod.

Rahul, C.U (2014, August). Discovery of miRNAs in genomes of *Phytophthora spp.* using computational tools. Poster presented at the National conference on sustainability of coconut, arecanut and cocoa farming-technological advances and way forward, Kasaragod.

Rahul, C.U. (2013, November). An analysis of the differential expression of rat genes when treated with common pesticidal compounds using microarray data. Paper presented at the National Seminar on Applications of Bioinformatics in Agriculture, Kasaragod.

Rahul C.U. (2004, September). Biodegradation: A critical study of threats and solutions. Paper presented at the National level symposium titled Bioinfo-2004, Mannargudi.