In silico-driven analysis of the *Glossina morsitans morsitans* antennae transcriptome in response to repellent or attractant compounds

Consolata Gakii[1,2], Billiah Kemunto Bwana[3], Grace Gathoni Mugambi[2], Esther Mukoya[2], Paul O. Mireji[4] and Richard Rimiru[2]

[1] Department of Mathematics, Computing and Information Technology, University of Embu, Embu, Eastern, Kenya
[2] School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Nairobi, Kenya
[3] Department of Biological Sciences, University of Embu, Embu, Eastern, Kenya
[4] Biotechnology Research Center, Kenya Agricultural & Livestock Research Organization, Nairobi, Nairobi, Kenya

ABSTRACT

Background: High-throughput sequencing generates large volumes of biological data that must be interpreted to make meaningful inference on the biological function. Problems arise due to the large number of characteristics p (dimensions) that describe each record [n] in the database. Feature selection using a subset of variables extracted from the large datasets is one of the approaches towards solving this problem.

Methodology: In this study we analyzed the transcriptome of *Glossina morsitans morsitans* (Tsetsefly) antennae after exposure to either a repellant (δ-nonalactone) or an attractant (ε-nonalactone). We identified 308 genes that were upregulated or downregulated due to exposure to a repellant (δ-nonalactone) or an attractant (ε-nonalactone) respectively. Weighted gene coexpression network analysis was used to cluster the genes into 12 modules and filter unconnected genes. Discretized and association rule mining was used to find association between genes thereby predicting the putative function of unannotated genes.

Results and discussion: Among the significantly expressed chemosensory genes (FDR < 0.05) in response to Ɛ-nonalactone were gustatory receptors (GrIA and Gr28b), ionotrophic receptors (Ir41a and Ir75a), odorant binding proteins (Obp99b, Obp99d, Obp59a and Obp28a) and the odorant receptor (Or67d). Several non-chemosensory genes with no assigned function in the NCBI database were co-expressed with the chemosensory genes. Exposure to a repellent (δ-nonalactone) did not show any significant change between the treatment and control samples. We generated a coexpression network with 276 edges and 130 nodes. Genes CAH3, Ahcy, Ir64a, Or67c, Ir8a and Or67a had node degree values above 11 and therefore could be regarded as the top hub genes in the network. Association rule mining showed a relation between various genes based on their appearance in the same itemsets as consequent and antecedent.

Submitted 3 March 2021
Accepted 8 June 2021
Published 1 July 2021

Corresponding author
Consolata Gakii,
gakii.consolata@embuni.ac.ke

Academic editor
Kenta Nakai

Additional Information and Declarations can be found on page 15

DOI **10.7717/peerj.11691**

Subjects Bioinformatics, Computational Biology, Genomics, Data Mining and Machine Learning
Keywords Association rule mining, Co-expression network, RNASeq data, Discretization, In silico analysis

INTRODUCTION

Modern sequencing technologies generate large volumes of data from living cells under different physiological conditions and this ushered applied biology into the area of big data (*Marx, 2013*). As the dimensionality increases, the volume of data required for meaningful analysis also grows exponentially. This phenomenon was defined as a "curse of dimensionality" by *Bellman (1957)*. Problems arise due to the large number of characteristics p (dimensions) that describe each record [n] in the database, that is, large [n] and small [p]. Various approaches have been used to reduce the dimensionality of bigdata. Feature selection is one of the approaches used to extract a subset of features (or variables) from the large datasets while maintaining as much information as possible (*Yousef, Allmer & Khalifa, 2016*). Another dimensionality reduction technique is data discretization and association rule mining. Discretization is used to convert continuous variables such as read counts from RNASeq data to a discrete format (*Gallo et al., 2016*). This reduces data noise and computational resources (*Beleut et al., 2016*).

Differential gene expression analysis is also a dimensionality reduction strategy. It deals with analysis and interpretation of variation in transcription levels of various genes in a cell across different samples or conditions (*Nia et al., 2020*). The method relies on different expression metrics e.g., non-parametric generalized linear models, independent sample *t*-tests, and log2 fold changes. However, identification of genes or pathways involved is problematic because genes act in concert rather than alone (*Nacu et al., 2007*). Biological interactions among genes are referred to as gene networks. A gene co-expression network has been defined as an undirected graph where nodes (genes) and edges connect significantly correlated features (*Stuart et al., 2003*; *Gardner et al., 2003*). Interactions between genes form a functional module and eventually different gene modules show varying levels of interaction (*Gonzalez-Dominguez & Martin, 2017*; *Zhang et al., 2019*). Since biological networks are usually complex, algorithms that utilize network theory have been extremely useful in deciphering valuable molecular interactions at the cell level (*Fiscon et al., 2018*).

Weighted gene co-expression network analysis (WGCNA) is a popular algorithm that constructs a network based on the pairwise correlations between gene expression levels (*Zhang & Horvath, 2005*; *Langfelder & Horvath, 2008*). The algorithm uses power law distribution to generate a scale-free network (*Abbassi-Daloii, Kan & Raz, 2020*). Genes that are highly interconnected and probably share a similar biological function are defined as a gene module (*Roy, Bhattacharyya & Kalita, 2014*). Co-expressed genes are grouped into increasingly large modules using hierarchical clustering algorithms (*Abbassi-Daloii, Kan & Raz, 2020*). Network metrics such as node degree (*Deng, Zhu & Huang, 2015*), betweenness (*Wang, Hernandez & Van Mieghem, 2008*) and cluster coefficient (*Watts & Strogatz, 1998*) are then used to explain which proteins are most important and why (*Vella et al., 2017*).

Gene co-expression network analyses have been widely used to complement other methods used in gene expression studies. For example, WGCNA was used by *Morandin et al. (2016)* to explain that there exist conserved pathways amongst 16 species of ants that exhibit reproductive division of labor. Using network analysis, *Orsini et al. (2018)* highlighted the role of metabolic and cell signaling genes in relation to stress response in the crustacean *Daphnia magna*. *Kaur et al. (2019)* compared the behavior and brain transcriptomes of *Temnothorax longispinosus* and identified enrichment in the WGCNA module of genes positively correlated with parasite prevalence and negatively with host attacks. *Smith et al. (2018)* used WGCNA to identify five co-expression modules that indicated a correlation between pupal stage of *D. melanogaster and* Osiris genes that are essential for development and phenotypic plasticity. Clusters (modules) of highly correlated genes between larval and adult stages have also been identified by *Finch et al. (2020)* in an Antarctic midge using Weighted correlation network analysis. *Manfredini et al. (2017)* identified gene coexpression subnetworks that were responsible for each queen phenotype from the brain transcriptomes of *B. terrestris*. *Smith & Moran (2020)* used gene expression and WGCNA to show that there exists a metabolic tug-of-war between the aphid *Acyrthosiphon pisum* and *Buchnera aphidicola* which is a bacterial symbiont. In a study by *Chen et al. (2019)*, DEGs that may regulate wing dimorphism in the house cricket *Acheta domesticus* were detected using WGCNA. More recently, *Mwangi, Macharia & Bargul (2021)* identified significantly enriched modules for genes that play key roles during the development of *Trypanosoma brucei* in tsetse flies.

Tsetse flies belong to the genus *Glossina* which has twenty-three species and eight sub-species (*Leak, 1999*; *Krafsur, 2009*). They use olfaction to seek for hosts, locate oviposition sites, search for mates, as well as detecting and escaping from potential predators. They have been shown to exhibit unique behavioral responses towards volatile short-range allomones (*Gikonyo et al., 2002*). Responses towards various synthetic blends of these compounds elicits an avoidance behavior in the flies meaning that the compounds behave as repellents (*Gikonyo et al., 2003*). δ-octalactone is the most effective repellent in the blends (*Bett, Saini & Hassanali, 2015*). Exposure of male *G. m. morsitans* to ε-nonalactone (attractant) or δ-nonalactone (repellent) elicits antennal molecular responses that includes canonical and non-canonical chemosensory as well as novel odor specific transcripts (*Kabaka et al., 2020*).

The aim of this study was to use an in silico driven analysis to elucidate the putative function of some of the non-chemosensory genes that are co-expressed with the chemosensory genes. Differential expression analysis and coexpression network analysis were applied on the RNAseq data to identify the genes that were upregulated or downregulated due to exposure to either a repellant (δ-nonalactone) or an attractant (ε-nonalactone) in compassion to the untreated controls. Association rule mining (ARM) was thereafter used to identify itemset patterns/associations in the differentially expressed gene sets. ARM is a useful market basket analysis algorithm described by *Agrawal & Srikant (1994)* whereby datasets are presented in a transaction format whereby a transaction $\boldsymbol{t}\ \boldsymbol{T} \in \boldsymbol{D}$ contains itemset $X \in I$ if $X \subseteq I$. Using this approach, we were able to

predict the potential function of uncharacterized genes based on their association with those with an assigned biological function.

MATERIALS & METHODS

Sample collection, RNA extraction and sequencing

The male *G. m. morsitans* flies used in this study were from a colony reared at Yale University insectary as described by *Bateta et al. (2017)*. Feeding, exposure to the chemicals and RNA extraction was done as described (*Kabaka et al., 2020*). A pair of antennae were carefully hand-dissected from fifty flies in each treatment (attractant, repellent, or control) and replicate (1, 2 or 3) as described (*Menuz et al., 2014*). PCR amplicons were generated using tsetse fly specific *beta-tubulin* gene primers to confirm removal of the gDNA from total RNA (*Bateta et al., 2017*). Sequencing was done on Illumina HiSeq 2500 at Yale University Center of Genome Analysis (YCGA), New Haven, CT, USA. The raw transcriptome sequences have been deposited at the Sequence Read Archive (SRA) under study accession number PRJNA343267.

Data preprocessing and differential gene expression analysis

Quality of the FastQ files was checked using FastQC v0.11.5 (*Babraham Bioinformatics, 2013*) software and the output cleaned using Trimmomatic software v0.38 (*Bolger, Lohse & Usadel, 2014*). Any contaminating rRNA reads in the data were removed using SortMeRna v2.0 (*Kopylova, Noé & Touzet, 2012*). The clean paired reads from different treatments and replicates were then separately mapped onto *G. m. morsitans* transcripts gene-set version 1.9 from Vectorbase or genome version 1.0 (*Giraldo-Calderón et al., 2015*) using STAR software v2.7.3a (*Dobin et al., 2013*). VectorBase provides reliable datasets that are community reviewed and regularly updated. Chemosensory proteins in these datasets have been annotated (*Obiero et al., 2014*; *Macharia et al., 2016*; *Liu et al., 2012*). Transcript mapping provided information on transcript specific mapping abundance (*Mortazavi et al., 2008*). Mapping reads to the genome was an additional quality control procedure that would get rid non-*G. m. morsitans* contaminants. The BAM files contained information on sequences aligned onto the transcripts, sorted by respective coordinates to facilitate downstream analyses. The number of reads (counts) aligned onto each transcript in the respective BAM files were quantified using Salmon software v1.2.1 (*Patro et al., 2017*). This analysis provided data on relative abundance of read from different treatments and replicates that mapped onto *G. m. morsitans transcripts*.

Differential expression analysis was done on the gene count matrix from Salmon using the DESeq2 R-package version 1.28.0 (*Love, Huber & Anders, 2014*). Default parameters for count data normalization as recommended (*Conesa et al., 2016*) were used to allow for control of $\log_2$ fold change shrinkage, custom p-value and fold change cut-offs. Genes were considered differentially expressed and retained for further analysis if the test statistics *p-value* (adjusted for false detection rate) (FDR) was less than 0.05 according to the method from *Benjamini & Hochberg (1995)*. Since the antennae is functionally specialized for olfaction, and potentially enriched with associated canonical chemosensory gene transcripts (*Kabaka et al., 2020*), we separately probed for expression profiles of these

transcripts, and isolated those with at least two-fold change in difference between the attractant or repellent and control treatments. We generated heatmaps for visualizing the relationships between the different treatments using the package Pheatmap *v*1.0.12 (*Kolde, 2012*) in R software (*R Core Team, 2017*).

Co-expression analysis using WGCNA

WGCNA package is designed for clustering genes based on their expression profiles and therefore we used the gene lists generated from Salmon for weighted co-express analysis. Normalization was done by filtering out genes with counts less than 10 in more than 90% of samples since they are not informative and tend to introduce noise (*Farhadian et al., 2021*). Co-expression networks were generated using WGCNA using the Bioconductor R package v3.5.1 (*Langfelder & Horvath, 2008*). This algorithm calculates a similarity co-expression matrix using correlation for all genes defined as: $s_{ij} = \left|\mathrm{cor}\left(x_i, x_j\right)\right|$. An adjacency matrix was calculated by raising the co-expression similarity to a soft thresholding power beta β defined as $a_{ij} = p_{ower(S_{ij}}, \beta) = \left|s_{ij}\right|^{\beta}$, where a_{ij} represents the resulting adjacency which is a measure of the connection strengths. A β of 12 was selected for construction of a gene co-expression network based on estimated scale-free topology as described (*Langfelder & Horvath, 2008*). A topological overlap matrix (TOM) was computed, converted into dissimilarity matrix and then hierarchical clustering used to generate a tree (dendrogram). Co-expression modules were detected using Dynamic Tree Cut (DTC) algorithm based on an edge height cut-off of 0.25, and a minimum module size of 30 genes. Relationship between different co-expression was explored using FlashClust function (*Langfelder & Horvath, 2012*) and a heatmap was used to visualize the correlation between the modules.

Network analysis

Network visualization and centrality measure analysis were done using Cytoscape v3.7.2 (*Shannon et al., 2003*). The degree, betweenness centrality and clustering coefficient of the network were analyzed using network analyzer (a Cytoscape plugin) as described by *Assenov et al. (2008)*. Degree $d(v)$ of a vertex v, in a network $G = (V, E)$, counts the number of edges in E incident upon v. Given G, define f (d) to be the fraction of vertexes $v \in V$ with degree $d(v) = d$. Genes having large degrees are referred to as hubs genes, indicating that they hold multiple genes/proteins together and have the highest potential to regulate the node *v*. Betweenness is a measure of the number of shortest paths that are connecting any two nodes (j, k). On the other hand, closeness is a measure of the ability of a node to interact with all other nodes, including the indirectly connected nodes. It is defined as:

$$j_1 = \frac{1}{N} \sum_{J} h(i, i)$$

with $h(i,i)$ being the shortest distance between gene i and j. Finally, clustering coefficient is defined as the edge density of the neighborhood of node i which is calculated as:

$$u_{i=} \frac{m_i(j,k)}{m_i}$$

where $m_i(j,k)$ is the number of edges connecting nodes (j, k) neighboring node i *and* m_i is the total number of visible edges of all the neighboring nodes of i that are fully connected.

Discretization and association rule mining

Discretization is a data pre-processing step used in machine learning in order to transform continuous or numerical attributes into discrete ones (*Kotsiantis & Kanellopoulos, 2006*; *Hacibeyoglu & Ibrahim, 2018*). In this study, we used the equal frequency discretization method (*Dougherty, Kohavi & Sahami, 1995*) implemented in GEDPROTOOLS (*Gallo et al., 2016*) to transform the data from continuous to discrete values. Equal frequency discretization reduces the effect of outliers and collects similar values in the same interval (*Hacibeyoglu & Ibrahim, 2018*). Genes that were co-expressed in the network were retrieved for discretization using the steps outlined below:

Input: the continuous values of attribute and number if intervals $\boldsymbol{A} = \{a_1, a_2, \ldots, a_{n-1}, a_n\}$ and number of intervals $\boldsymbol{k}$, where $\boldsymbol{k} > 0$.
Step 1: Sort all values of in ascending order,
Step 2: Divide $\boldsymbol{A}$ by k intervals,
Step 3: Create bins according to number of elements in each interval,
Step 4: Determine boundaries of each interval by calculating the average value of the Maximum value of the current bin and the minimum value of the next bin,
Step 5: The continuous values of $\boldsymbol{A}$ are transformed into discrete ones by determiningthe interval that they belong to,
Output: $\boldsymbol{A}$ with discrete values

Two bins were created in step 3, with the final output being discretized measurements whereby a value of zero represented a gene that was under-expressed, while a value of one represented a gene that was overexpressed. In this context, the discretized data was used to identify frequent itemsets using the Apriori algorithm (*Agrawal & Srikant, 1994*) implemented in the R package arules ver. v1.6-4 (*Hahsler et al., 2014*). In association rule mining, a rule is typically described by three measures: support, confidence, and lift. These three represent the significance and interest of a rule. Support of a rule $X \Rightarrow Y$ is equal to the support of the itemset $X \cup Y$ and is defined as the probability of finding all the genes in sets X. Support of an itemset X is calculated as:

$$Support_D(x) = \frac{|\{T \in Dx \subseteq T\}|}{|D|}$$

The confidence of rule $X \Rightarrow Y$ is the probability of finding all the differentially expressed genes in set Y as compared with the differentially expressed genes in set X. The confidence is calculated as:

$$Confidence_D(x \Rightarrow y) = \frac{Supp_D(X \cup Y)}{supp_D(X)}$$

Lift measures the strength of the rule and varies in the interval [0, ∞]. Lift is defined as:

$$lift(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) * supp(Y)}$$

We used a support value of 0.5 and a confidence value of 0.99. Support values greater than 0.5 gave zero rules while support values of less than 0.5 and confidence of less than 0.99 resulted in too many rules. We therefore filtered and retained only the rules that had a lift value ≥ 2.

Gene set enrichment analysis

Gene Set Enrichment Analysis (GSEA) was done using WEB-based Gene SeT AnaLysis Toolkit (WebGestalt) as described by *Wang et al. (2017)*. Using *Drosophila melanogaster* homologs as proxy to assess gene enrichment since the database of WebGestalt does not have *G. morsitans*, homologs. FDR corrected and p-value ranked *D. melanogaster* gene homologs of the differentially expressed *G. m. morsitans* genes were selected as input for the analysis using default parameters (5–2,000 Entrez Gene IDs, FDR < 0.05, 1,000 permutations and 20 categories with the outputted leading-edge genes) as described by *Dębski et al. (2016)*. We separated significantly enriched non-redundant biological processes, cellular components, and molecular function Gene Ontology (GO) terms as identified by GSEA (*Ashburner et al., 2000*; *Stark et al., 2006*; *The Gene Ontology Consortium, 2017*).

RESULTS

Differential expression in response to repellants (δ-nonalactone) or attractant (Ɛ-nonalactone)

Raw reads per sample ranged between 23 and 73 million. Ribosomal RNA contamination levels were below 5% in all the samples indicating a near-perfect ribosomal RNA depletion before sequencing was done. The number of clean reads in the samples after preprocessing and quality filtering ranged between 11.1 M and 38.3 M. Reads that mapped to the genome were between 78% and 92%, while the remainder mapped to multiple locations, were not unique or had no features. During differential gene expression analysis, 2,097 low-count genes were filtered out leaving a total 10,921. Genes with a False discovery rate (FDR) < 0.05 were considered as being of biological significance (*Von Der Weid et al., 2015*). A set of 308 genes were identified as differentially expressed after exposure to either repellent (δ-nonalactone) or attractant (Ɛ-nonalactone) as compared to the control (no treatment). Figure 1A shows the top 50 significantly expressed chemosensory genes (FDR < 0.05) in response to Ɛ-nonalactone. Nine genes showed upregulation in flies

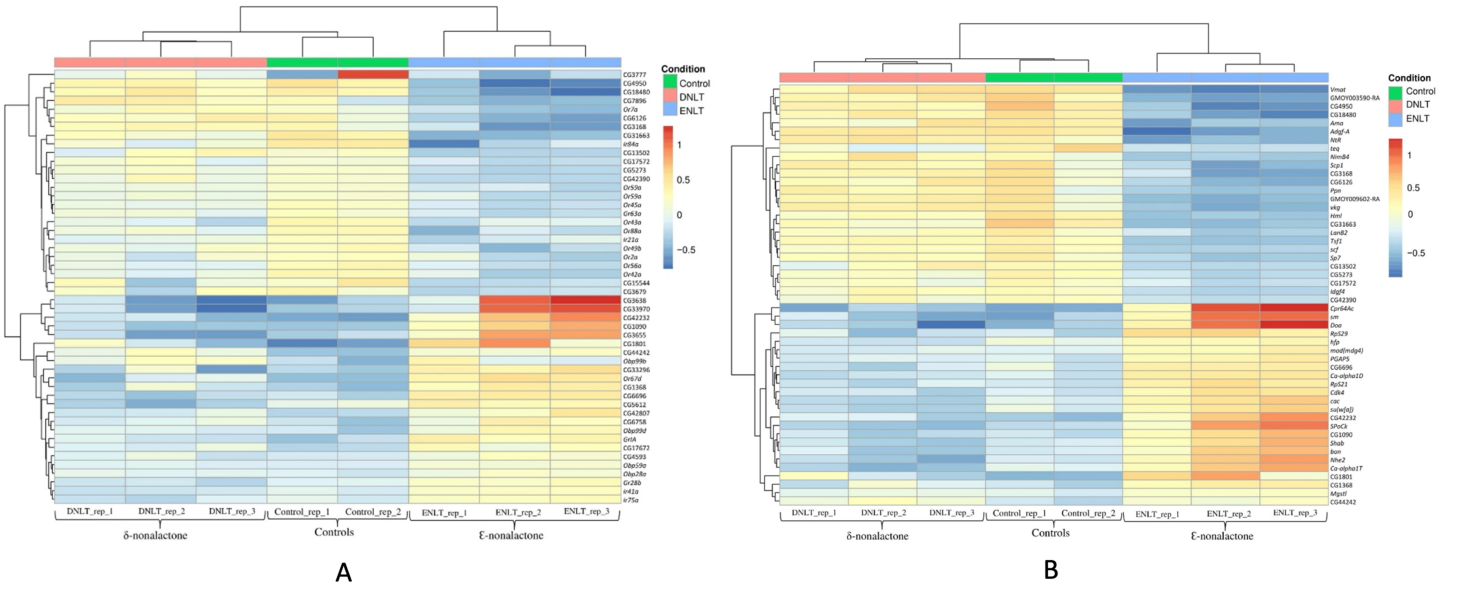

Figure 1 Top 50 chemosensory genes. Top 50 differentially expressed genes after exposure to repellant (δ-nonalactone) or attractant (Ɛ-nonalactone). (A) Top 50 chemosensory genes and associating genes with no assigned function. (B) Top 50 non-chemosensory genes and associating genes with no assigned function.
Full-size DOI: 10.7717/peerj.11691/fig-1

exposed to an attractant (Ɛ-nonalactone). These are the Gustatory receptors (GrIA and Gr28b), ionotrophic receptors (Ir41a and Ir75a), odorant binding proteins (Obp99b, Obp99d, Obp59a and Obp28a) and the odorant receptor (Or67d). In addition, several non-chemosensory genes with no assigned function in the NCBI database were co-expressed with the chemosensory genes. In contrast, exposure to repellent (δ-nonalactone) did not show any significant change between the treatment and control samples. We identified 24 non-chemosensory that showed significant upregulation in response to Ɛ-nonalactone as shown in Fig. 1B. Six of these genes have no assigned function biological function and were also upregulated in the chemosensory geneset.

Co-expression network analysis and hub genes identification

A scale-free topology weighted gene network was constructed using WGCNA based on a soft thresholding power (β). From candidate powers of between 1–20, β = 12 returned a scale-free topology fit index of –0.1. An adjacency matrix based on the criterion of approximate scale-free topology is shown in Figs. 2A and 2B. Using the dynamic tree cutting algorithm, all the 11,808 genes were grouped into 12 modules, which ranged in size from 42 to 9,325 genes per modules (Figs. 2C and 2D).

We identified 12 modules (Fig. 3) with the major modules being turquoise ($n = 9{,}325$ genes), blue ($n = 1{,}040$ genes), and brown ($n = 377$ genes).

Correlation between modules

We performed cluster analysis to identify whether chemosensory genes were evenly distributed with the 12 co-expressed modules. Using an edge height cut-off of 0.25, all modules were below the 75% similarity hence no modules were merged. Interestingly,

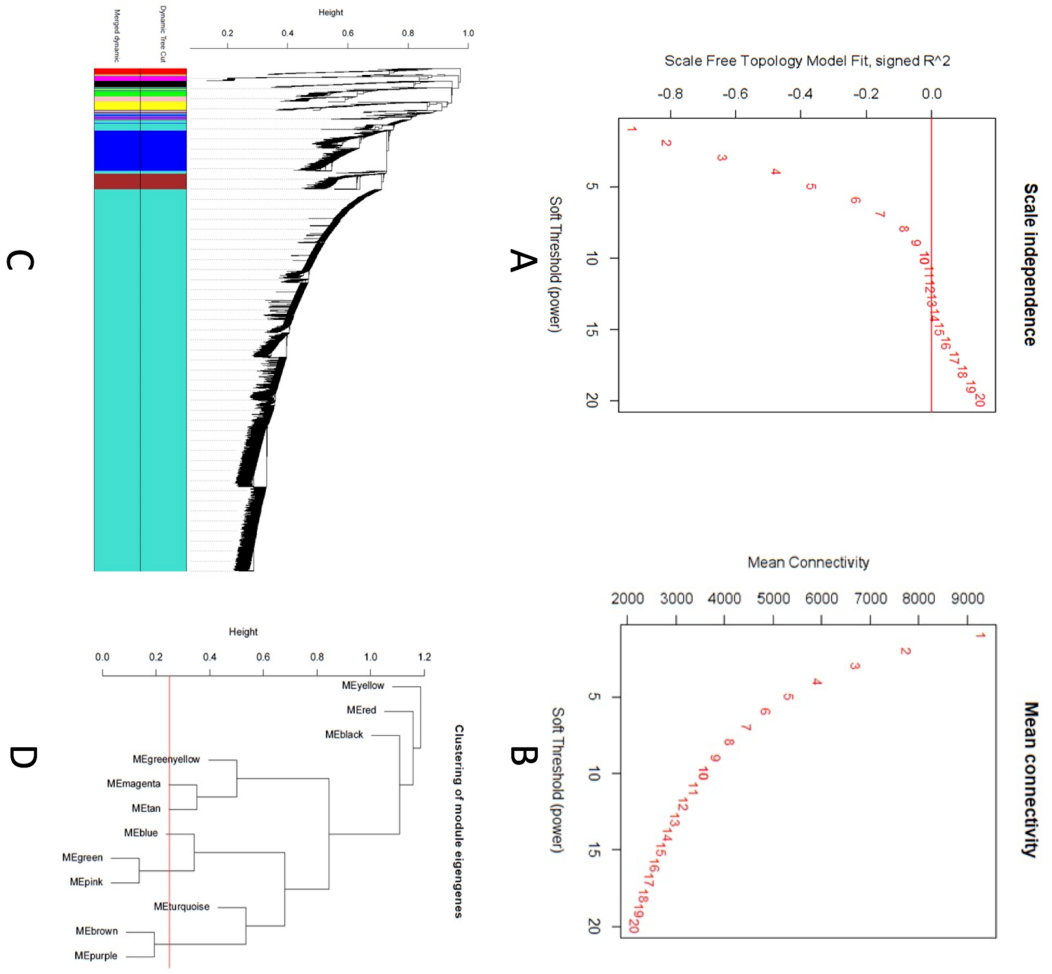

Figure 2 Co-expression network analysis using WGCNA. (A) Scale-free fit index versus soft--thresholding power. (B) Mean connectivity versus soft-thresholding power of 12. (C) Cluster dendrogram based on dissimilarity measures (1-TOM). The branches correspond to modules of highly interconnected groups of genes. Colors in the horizontal bar represent the modules. A total of 12 modules were identified. (D) Module network dendrogram constructed by clustering module eigengene distances. The red line shows the merging threshold. Full-size DOI: 10.7717/peerj.11691/fig-2

most of the chemosensory genes (n = 83) were in the turquoise module which also had 77 non-chemosensory genes as well as 46 genes that have no assigned function yet. In the other modules, genes Or67d was identified in the blue module, Obp19b & Obp83g in the brown module, and Obp73a in the yellow module. Therefore, the turquoise module was interesting from the chemosensation point of view. We proceeded to probe if there exist relationships between genes in the turquoise, blue, brown and yellow modules by visualizing only the functionally annotated genes in a network (Figs. 4A and 4B). Chemosensory genes are depicted in green, non-chemosensory in yellow and those with unknown function in grey. We filtered out genes with degree value of less than 5 to reduce the size of the graph and got a final network with 51 nodes and 148 edges (Fig. 4B). Filtering changed the network density from 0.05 to 0.116 while overall clustering coefficient increased from 0.475 to 0.587 (Figs. 4C and 4D).

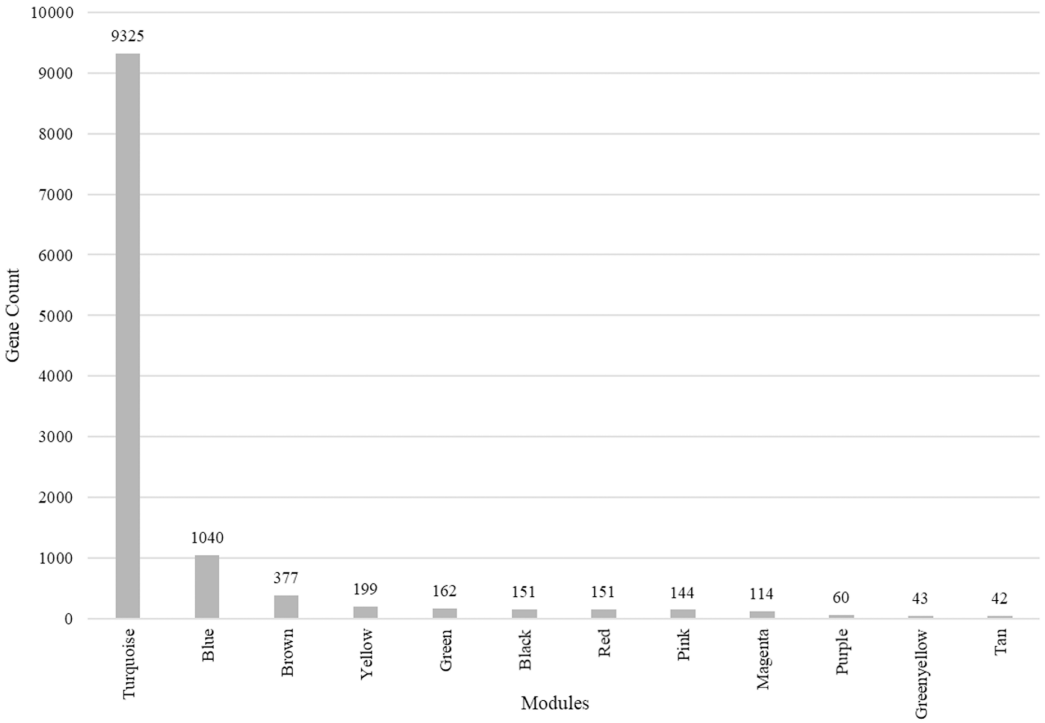

Figure 3 Modules identified along with the number of genes in each module. Total number of modules identified along with the number of genes in each module.
Full-size DOI: 10.7717/peerj.11691/fig-3

The degree, average shortest path length and clustering coefficient for the top genes with a node greater than 8 are shown in Table 1. Genes *CAH3*, *Ahcy*, *Ir64a*, *Or67c*, *Ir8a* and *Or67a* can be regarded as the top hub genes since they had degree values above 10. Fourteen of the top 20 genes are associated with chemosensation, which is an important biological function in insects. Clustering coefficient of hub gene nodes ranged between 0.29 and 0.68 and this is an indication that some parts of the network were more intricately connected than others.

Association rule mining

Discretized and transformed gene expression data is presented in a transaction format where the samples represent transaction IDs and genes represent items. A minimum support of 0.5 and a confidence of 0.99 was used to generate 801 rules, which we further filtered using a lift of ≥ 2 to empirically produce the best results. Lift values lower than 2 or support values less than 0.5 generated too many rules whereas when we used support values greater than 0.5, no rules were generated. Genes with no assigned biological function were of interest since we wanted to find out if association rule mining could help predict their function. We therefore arrowed down the rules that implied an association between known genes and those with no known function. Twenty-two representative rules are shown in Table 2.

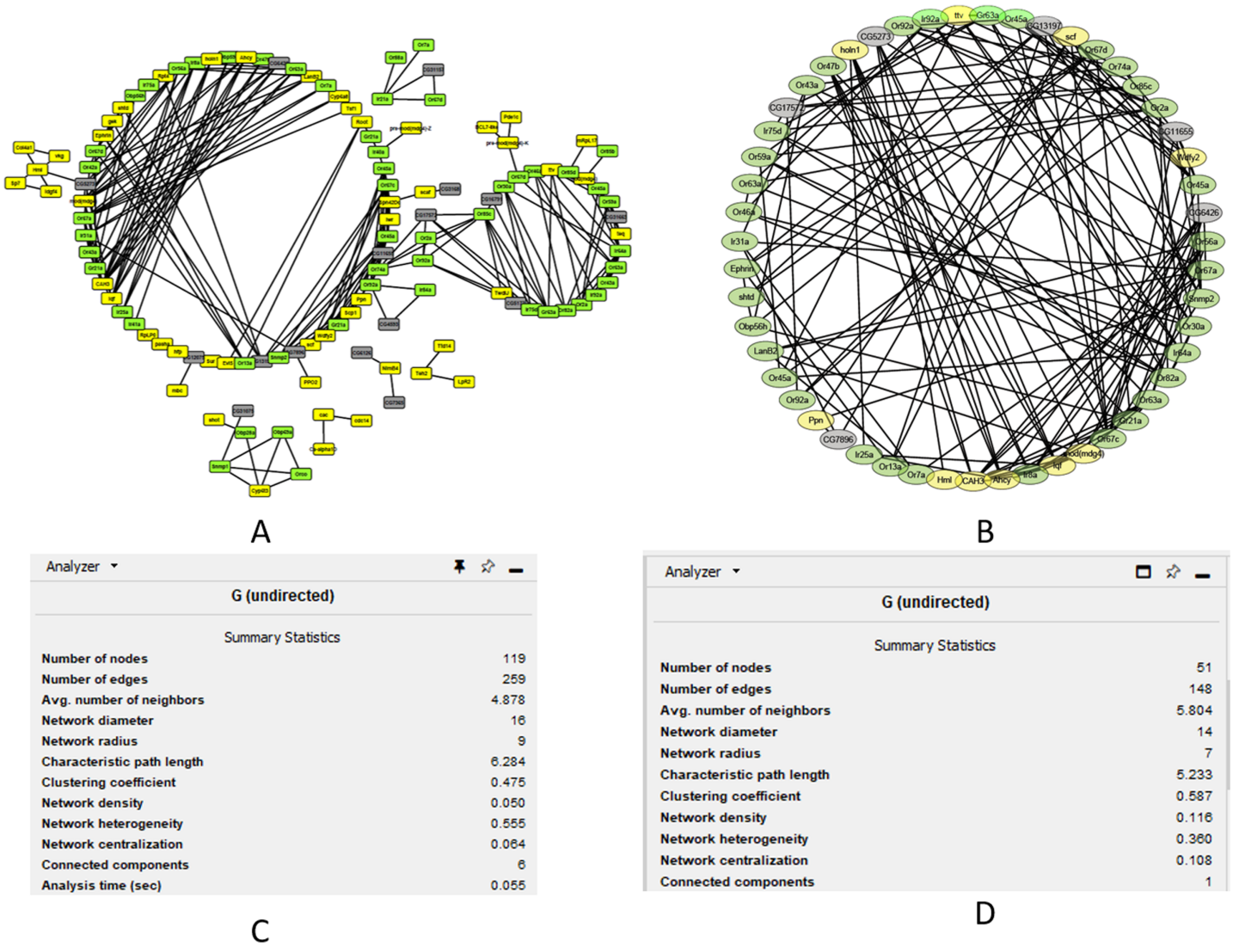

Figure 4 Co-expression networks for the genes in the turquoise, blue, brown and yellow modules. Co-expression networks analysis. (A) Sub-network for genes in the turquoise, blue, brown and yellow modules. (B) Filtered co-expression network for the genes with a degree value greater than five. (C) Network summary statistics before filtering. (D) Network summary statistics after filtering.

Full-size DOI: 10.7717/peerj.11691/fig-4

Gene set enrichment analysis (GSEA)

GSEA of the antennae transcripts revealed several enriched processes involved in detection of stimulus. GoSlim GO analysis component of the GSEA assigned most of the significantly expressed genes to various biological, cellular components and molecular function as shown in Fig. 5.

The top Enriched biological processes included biological regulation, response to stimulus, metabolic processes, and cell communication. Enriched cellular components were mainly associated with the membrane. Most enriched molecular functions were associated with ion binding, molecular transducer activity and protein binding. As we

Table 1 Network topology for the top genes.

No.	Gene symbol	Clustering coefficient	Degree	Betweenness centrality
1	*Ahcy*	0.47	11	0.06
2	*CAH3*	0.53	11	0.07
3	*Ir64a*	0.38	10	0.1
4	*Or67c*	0.44	10	0.2
5	*Ir8a*	0.6	10	0.03
6	*Or67a*	0.29	10	0.08
7	*lqf*	0.56	9	0.02
8	*mod(mdg4)*	0.58	9	0.01
9	*Or30a*	0.31	9	0.24
10	*Or45a*	0.5	9	0.11
11	*Or85c*	0.36	9	0.32
12	*CG11655*	0.57	8	0.18
13	*Wdfy2*	0.61	8	0.03
14	*Or82a*	0.39	8	0.05
15	*Or2a*	0.54	8	0
16	*Or67d*	0.32	8	0.22
17	*Or45a*	0.46	8	0.2
18	*Or63a*	0.68	8	0.02
19	*Or56a*	0.46	8	0.09
20	*Gr21a*	0.43	8	0.14

expected, overrepresentation analysis as well as the network expansion analysis showed that most of the significantly expressed genes were associated with sensory responses.

DISCUSSION

We analyzed RNAseq data generated from the antennae of *G. m. morsitans* to understand gene regulation in response to either repellent (δ-nonalactone) or attractant (Ɛ-nonalactone) as compared to the untreated controls. Downstream analysis of the entire gene set (13,019 genes) revealed that there was differential expression of various genes in response to the two treatments. We focused on a set of 308 genes that showed significant upregulation due to exposure to an attractant or repellent. The identified chemosensation genes such as gustatory receptors, ionotrophic receptors, odorant binding proteins and Odorant receptors give the fly a situational awareness of the surrounding environment. For example, Gr21a is a potential CO_2 receptor (*Jones et al., 2007*) while the highly conserved Ir25a is a co-receptor for Ir21a and together they mediate thermotransduction (*Ni et al., 2016*). Putrescine and spermidine which are foul smelling amines generated by bacterial degradation of arginine are detected by Ir41a (*Silbering et al., 2011*). Their differential expression could therefore be attributed to either the repellent (δ-nonalactone) or the attractant (Ɛ-nonalactone) used as a treatment in the experiment.

Table 2 Association rules among genes that showed significant upregulation.

No.	Association rule	Set
1.	{Ir84a,Or2a,Or42a,Or56a} => {CG3679}	A
2.	{Or2a,Or42a,Or56a,Or49b} => {CG3679}	
3.	{Ir84a,Or42a,Or56a,Or49b} => {CG3679}	
4.	{Or88a,Gr63a,CG5273,CG17572} => {CG18480}	
5.	{CG4950,Or88a,Gr63a,CG5273} => {CG18480}	
6.	{Or88a,Gr63a,CG17572,CG31663} => {CG18480}	
7.	{Or88a, Gr63a, CG5273,CG17572} => {CG31663}	
8.	{CG4950, Or88a, Gr63a,CG5273} => {CG31663}	
9.	{CG4950,Or88a,CG18480,Gr63a} => {CG31663}	
10.	{CG4950,Or88a,Gr63a,CG31663} => {CG17572}	
11.	{Or88a,Gr63a,CG5273,CG31663} => {CG17572}	
12.	{Or88a,Gr63a,CG17572,CG31663} => {CG5273}	
13.	{CG4950,Or88a,Gr63a,CG17572} => {CG5273}	
14.	{CG4950,Or88a,CG18480,Gr63a} => {CG5273}	
15.	{Tsf1,Scp1,vkg,Adgf.A} => {CG6126}	B
16.	{Ppn,Sp7,NtR,Adgf.A} => {CG6126}	
17.	{vkg,Sp7,NtR,Adgf.A} => {CG6126}	
18.	{LanB2,Sp7,NtR,Adgf.A} => {CG6126}	
19.	{Sp7,NtR,Adgf.A,CG6126} => {CG3168}	
20.	{Ppn,NtR,Adgf.A,CG6126} => {CG3168}	
21.	{Idgf4,NtR,Adgf.A,CG6126} => {CG3168}	
22.	{Ppn,Sp7,Adgf.A,CG6126} => {CG3168}	

WGCNA is statistical model that is data and works even on non-model organisms (*Degli Esposti et al., 2019*). Therefore, when we used WGCNA on *G. m. morsitans* data, we were able to reduce the dimensionality of the data and to extract meaningful patterns using network centrality measures. The resulting co-expression network was useful in selecting genes with significant connectivity patterns that are biologically meaningful. Node degrees helped us identify two genes (CAH3 and Ahcy) that had a degree of 11 and therefore these can be regarded as the top hub genes. In *Drosophila melanogaster*, CAH3 is a carbonate dehydratase involved in generation of protons and bicarbonate from carbonic acid (*Overend et al., 2016*). *Ahcy* is involved in methionine biosynthesis and metabolism (*Brosnan & Brosnan, 2006*). The observed betweenness centrality measures mean that some of the genes such as *Teh2* and *Ir21a* which had values of 1 and 0.8 respectively would have more control over the network as compared to those with lower values. Therefore genes/nodes with high betweenness centrality values are more biologically informative in a module (*Riquelme Medina & Lubovac-Pilav, 2016*). Closeness centrality measures for majority of the nodes were between 0.1 and 1, which means the resulting network was closely connected, while twenty-seven of the nodes in the network had a clustering coefficient of 1 an indication of complete node connection (*Liu, 2018*).

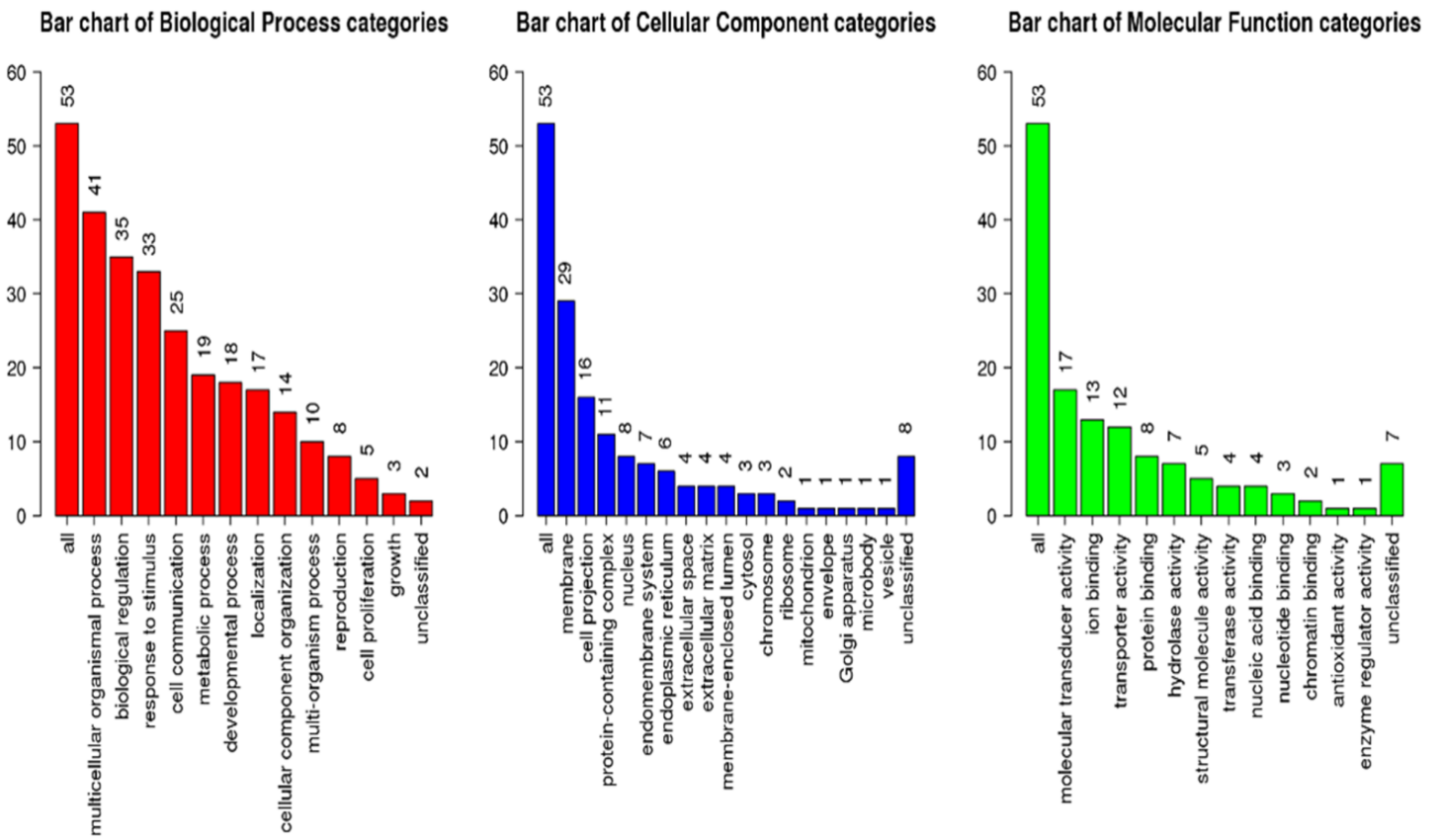

Figure 5 **Geneset Enrichment Analysis (GSEA) of differentially expressed transcripts in *G. m. morsitans* antennae.**
Full-size DOI: 10.7717/peerj.11691/fig-5

Association rule mining enabled us to identify itemset patterns based on the RNAseq genes expression patterns. The first 22 rules indicate that there is a relationship among the genes (itemsets) expression in each condition (transaction) with the following genes *CG18480, CG31663, Ir84a, CG17572, CG5273, Gr63a, Or88a, Or49b, Or2a, Or56a, Or42a, CG3679, Adgf-A, NtR, teq, NimB4, Scp1*, CG3168, CG6126, *Ppn, vkg, LanB2, Tsf1, scf, Sp7* and *Idgf4* always being up or downregulated in response to either repellent (δ-nonalactone) or attractant (Ɛ-nonalactone). The identified rules in this study are biologically significant based on the concept that similar items as in market basket analysis appear together is clearly shown in our results. For example, where the genes CG18480, CG31663, CG17572, CG5273 and CG3679 referred to as consequents (right side) were up (highly expressed), all the genes on the rule antecedent (left side) were also up. The rest of the rules can be interpreted in a similar manner. Genes *Ir84a, Gr63a, Or88a, Or49b, Or2a, Or56a* and *Or42a* are involved in chemosensation in insects. However, genes CG18480, CG31663, CG17572, CG5273 and CG3679 that are co-expressed with the chemosensation genes have no assigned biological function. We therefore hypothesize that these genes also play a role in chemosensation due to their co-expression and association with chemosensory genes. Only two genes (CG3168 and CG6126) were associated with the top non-chemosensory genes that were upregulated due to exposure to an attractant.

We used a lift value greater than 2 in generating the rules because values greater than 1 indicate that consequent and antecedent are dependent on one another (*Hahsler, Grün & Hornik, 2007*).

CONCLUSIONS

We analyzed RNA-Seq data from the antennae of *Glossina morsitans morsitans* (Tsetsefly) after exposure to either a repellant (δ-nonalactone) or an attractant (ε-nonalactone). After pre-processing the data using several methods and software tools to filter out low quality reads and low counts, we proceeded to carry out co-expression network analysis which facilitated the feature reduction by filtering out genes based on the node degree. We found a set of genes occurring in the two treatments and then attempted to find out if these genes could assist us predict the function on uncharacterized genes based on association rule mining. Results from rule mining showed that some of the genes were related based on their appearance in the same itemsets as consequent and antecedent. Most interesting is that we were able to predict the function of some uncharacterized genes in the network. Genes CG18480, CG31663, CG17572, CG5273, CG3679, CG3168 and CG6126 are designated as uncharacterized proteins in the NCBI database. From the association rule analysis, we presume that genes CG18480, CG31663, CG17572, CG5273 and CG3679 have a potential function in chemosensation due to their association with characterized chemosensory genes. Therefore, application of WGCNA and association rule mining is a powerful approach for dimensionality reduction and selecting informative features based on their relationship in the network and eventually the rules. This can assist in predicting the role of biological features with no known function.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Consolata Gakii performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Billiah Kemunto Bwana performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Grace Gathoni Mugambi analyzed the data, prepared figures and/or tables, and approved the final draft.
- Esther Mukoya analyzed the data, prepared figures and/or tables, and approved the final draft.
- Paul O. Mireji conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

- Richard Rimiru conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The sequences are available at Sequence Read Archive (SRA): PRJNA343267.

REFERENCES

Abbassi-Daloii T, Kan HE, Raz V. 2020. Recommendations for the analysis of gene expression data to identify intrinsic differences between similar tissues. *Genomics* **112(5)**:3157–3165 DOI 10.1016/j.ygeno.2020.05.026.

Agrawal R, Srikant R. 1994. Fast algorithms for mining association rules. In: *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB.* Vol. 1215. 487–499.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* **25(1)**:25–29 DOI 10.1038/75556.

Assenov Y, Ramírez F, Schelhorn SE, Lengauer T, Albrecht M. 2008. Computing topological parameters of biological networks. *Bioinformatics* **24(2)**:282–284 DOI 10.1093/bioinformatics/btm554.

Bateta R, Wang J, Wu Y, Weiss BL, Warren WC, Murilla GA, Aksoy S, Mireji PO. 2017. Tsetse fly (Glossina pallidipes) midgut responses to Trypanosoma brucei challenge. *Parasites & Vectors* **10(1)**:1–12 DOI 10.1186/s13071-017-2569-7.

Beleut M, Soeldner R, Egorov M, Guenther R, Dehler S, Morys-Wortmann C, Moch H, Henco K, Schraml P. 2016. Discretization of gene expression data unmasks molecular subgroups recurring in different human cancer types. *PLOS ONE* **11(8)**:e0161514.

Bellman R. 1957. A Markovian decision process. *Journal of Mathematics and Mechanics* **6(5)**:679–684.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57(1)**:289–300.

Bett MK, Saini RK, Hassanali A. 2015. Repellency of tsetse-refractory waterbuck (Kobus defassa) body odour to Glossina pallidipes (Diptera: Glossinidae): Assessment of relative contribution of different classes and individual constituents. *Acta Tropica* **146**:17–24.

Babraham Bioinformatics. 2013. FastQC a quality control tool for high throughput sequence data. *Available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30(15)**:2114–2120 DOI 10.1093/bioinformatics/btu170.

Brosnan JT, Brosnan ME. 2006. The sulfur-containing amino acids: an overview. *The Journal of Nutrition* **136(6)**:1636S–1640S.

Chen Q, Wen M, Li J, Zhou H, Jin S, Zhou JJ, Wang Y, Ren B. 2019. Involvement of heat shock protein 40 in the wing dimorphism of the house cricket Acheta domesticus. *Journal of Insect Physiology* **114(559)**:35–44 DOI 10.1016/j.jinsphys.2019.02.007.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biology* **17(1)**:1–19.

Degli Esposti D, Almunia C, Guery MA, Koenig N, Armengaud J, Chaumot A, Geffard O. 2019. Co-expression network analysis identifies gonad-and embryo-associated protein modules in the sentinel species Gammarus fossarum. *Scientific Reports* **9(1)**:1–10 DOI 10.1038/s41598-019-44203-5.

Deng SP, Zhu L, Huang DS. 2015. Predicting hub genes associated with cervical cancer through gene co-expression networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **13(1)**:27–35 DOI 10.1109/TCBB.2015.2476790.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29(1)**:15–21 DOI 10.1093/bioinformatics/bts635.

Dougherty J, Kohavi R, Sahami M. 1995. Supervised and unsupervised discretization of continuous features. In: *Machine learning proceedings 1995*, Morgan Kaufmann, 194–202.

Dębski KJ, Pitkanen A, Puhakka N, Bot AM, Khurana I, Harikrishnan KN, Ziemann M, Kaspi A, El-Osta A, Lukasiuk K, Kobow K. 2016. Etiology matters-genomic DNA methylation patterns in three rat models of acquired epilepsy. *Scientific Reports* **6(1)**:1–14 DOI 10.1038/srep25668.

Farhadian M, Rafat SA, Panahi B, Mayack C. 2021. Weighted gene co-expression network analysis identifies modules and functionally enriched pathways in the lactation process. *Scientific Reports* **11(1)**:1–15.

Finch G, Nandyal S, Perretta C, Davies B, Rosendale AJ, Holmes CJ, Gantz JD, Spacht DE, Bailey ST, Chen X, Oyen K. 2020. Multi-level analysis of reproduction in an Antarctic midge identifies female and male accessory gland products that are altered by larval stress and impact progeny viability. *Scientific Reports* **10(1)**:1–27.

Fiscon G, Conte F, Farina L, Paci P. 2018. Network-based approaches to explore complex biological systems towards network medicine. *Genes* **9(9)**:437 DOI 10.3390/genes9090437.

Gallo CA, Cecchini RL, Carballido JA, Micheletto S, Ponzoni I. 2016. Discretization of gene expression data revised. *Briefings in Bioinformatics* **17(5)**:758–770.

Gardner TS, Di Bernardo D, Lorenz D, Collins JJ. 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301(5629)**:102–105 DOI 10.1126/science.1081900.

Gikonyo NK, Hassanali A, Njagi PG, Gitu PM, Midiwo JO. 2002. Odor composition of preferred (buffalo and ox) and nonpreferred (waterbuck) hosts of some savanna tsetse flies. *Journal of Chemical Ecology* **28(5)**:969–981 DOI 10.1023/A:1015205716921.

Gikonyo NK, Hassanali A, Njagi PG, Saini RK. 2003. Responses of Glossina morsitans morsitans to blends of electroantennographically active compounds in the odors of its preferred (buffalo and ox) and nonpreferred (waterbuck) hosts. *Journal of Chemical Ecology* **29(10)**:2331–2345 DOI 10.1023/A:1026230615877.

Gonzalez-Dominguez J, Martin MJ. 2017. MPIGeneNet: parallel calculation of gene co-expression networks on multicore clusters. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **15(5)**:1732–1737.

Giraldo-Calderón GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, Ho N, Gesing S, VectorBase Consortium Madey G, Collins FC, Lawson D. 2015. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Research* **43(D1)**:D707–D713.

Hacibeyoglu M, Ibrahim MH. 2018. EF_Unique: an improved version of unsupervised equal frequency discretization method. *Arabian Journal for Science and Engineering* **43(12)**:7695–7704 DOI 10.1007/s13369-018-3144-z.

Hahsler M, Grün B, Hornik K. 2007. Introduction to arules-mining association rules and frequent item sets. *SIGKDD Explorations* **2(4)**:1–28.

Hahsler M, Buchta C, Gruen B, Hornik K. 2014. arules: Mining Association Rules and Frequent Itemsets. R package version 1.3-1. *Available at https://CRAN.Rproject.org/package=arules*.

Jones WD, Cayirlioglu P, Kadow IG, Vosshall LB. 2007. Two chemosensory receptors together mediate carbon dioxide detection in Drosophila. *Nature* **445(7123)**:86–90 DOI 10.1038/nature05466.

Kabaka JM, Wachira BM, Mang'era CM, Rono MK, Hassanali A, Okoth SO, Oduol VO, Macharia RW, Murilla GA, Mireji PO. 2020. Expansions of chemosensory gene orthologs among selected tsetse fly species and their expressions in Glossina morsitans morsitans tsetse fly. *PLOS Neglected Tropical Diseases* **14(6)**:e0008341 DOI 10.1371/journal.pntd.0008341.

Kaur R, Stoldt M, Jongepier E, Feldmeyer B, Menzel F, Bornberg-Bauer E, Foitzik S. 2019. Ant behaviour and brain gene expression of defending hosts depend on the ecological success of the intruding social parasite. *Philosophical Transactions of the Royal Society B* **374(1769)**:20180192 DOI 10.1098/rstb.2018.0192.

Kolde R. 2012. Pheatmap: pretty heatmaps. R package v. 1.0. 8. *Available at https://CRAN.R-project.org/package=pheatmap*.

Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28(24)**:3211–3217 DOI 10.1093/bioinformatics/bts611.

Kotsiantis S, Kanellopoulos D. 2006. Discretization techniques: a recent survey. *GESTS International Transactions on Computer Science and Engineering* **32(1)**:47–58.

Krafsur ES. 2009. Tsetse flies: genetics, evolution, and role as vectors. *Infection, Genetics and Evolution* **9(1)**:124–141 DOI 10.1016/j.meegid.2008.09.010.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9(1)**:1–13 DOI 10.1186/1471-2105-9-559.

Langfelder P, Horvath S. 2012. Fast R functions for robust correlations and hierarchical clustering. *Journal of Statistical Software* **46(11)**:i11 DOI 10.18637/jss.v046.i11.

Leak SG. 1999. Tsetse biology and ecology: their role in the epidemiology and control of trypanosomosis. In: *Tsetse Biology and Ecology: their role in the epidemiology and control of tryppanosomosis,* 568.

Liu BH. 2018. Differential Coexpression network analysis for gene expression data. In: *Computational Systems Biology*. New York: Humana Press, 155–165.

Liu R, He X, Lehane S, Lehane M, Hertz-Fowler C, Berriman M, Field LM, Zhou JJ. 2012. Expression of chemosensory proteins in the tsetse fly Glossina morsitans morsitans is related to female host-seeking behaviour. *Insect Molecular Biology* **21(1)**:41–48 DOI 10.1111/j.1365-2583.2011.01114.x.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15(12)**:1–21 DOI 10.1186/s13059-014-0550-8.

Macharia R, Mireji P, Murungi E, Murilla G, Christoffels A, Aksoy S, Masiga D. 2016. Genome-wide comparative analysis of chemosensory gene families in five tsetse fly species. *PLOS Neglected Tropical Diseases* **10(2)**:e0004421 DOI 10.1371/journal.pntd.0004421.

Manfredini F, Romero AE, Pedroso I, Paccanaro A, Sumner S, Brown MJ. 2017. Neurogenomic signatures of successes and failures in life-history transitions in a key insect pollinator. *Genome Biology and Evolution* **9(11)**:3059–3072 DOI 10.1093/gbe/evx220.

Marx V. 2013. The big challenges of big data. *Nature* **498(7453)**:255–260.

Menuz K, Larter NK, Park J, Carlson JR. 2014. An RNA-seq screen of the Drosophila antenna identifies a transporter necessary for ammonia detection. *PLOS Genetics* **10(11)**:e1004810 DOI 10.1371/journal.pgen.1004810.

Morandin C, Tin MM, Abril S, Gómez C, Pontieri L, Schiøtt M, Sundström L, Tsuji K, Pedersen JS, Helanterä H, Mikheyev AS. 2016. Comparative transcriptomics reveals the conserved building blocks involved in parallel evolution of diverse phenotypic traits in ants. *Genome Biology* **17(1)**:1–19.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5(7)**:621–628 DOI 10.1038/nmeth.1226.

Mwangi KW, Macharia RW, Bargul JL. 2021. Gene co-expression network analysis of Trypanosoma brucei in tsetse fly vector. *Parasites & Vectors* **14(1)**:1–11.

Nacu Ş, Critchley-Thorne R, Lee P, Holmes S. 2007. Gene expression network analysis and applications to immunology. *Bioinformatics* **23(7)**:850–858 DOI 10.1093/bioinformatics/btm019.

Ni L, Klein M, Svec KV, Budelli G, Chang EC, Ferrer AJ, Benton R, Samuel AD, Garrity PA. 2016. The ionotropic receptors IR21a and IR25a mediate cool sensing in Drosophila. *Elife* **5**:e13254 DOI 10.7554/eLife.13254.

Nia AM, Chen T, Barnette BL, Khanipov K, Ullrich RL, Bhavnani SK, Emmett MR. 2020. Efficient identification of multiple pathways: RNA-Seq analysis of livers from 56 Fe ion irradiated mice. *BMC Bioinformatics* **21(1)**:1–12 DOI 10.1186/s12859-020-3446-5.

Obiero GF, Mireji PO, Nyanjom SR, Christoffels A, Robertson HM, Masiga DK. 2014. Odorant and gustatory receptors in the tsetse fly Glossina morsitans morsitans. *PLOS Neglected Tropical Diseases* **8(4)**:e2663 DOI 10.1371/journal.pntd.0002663.

Orsini L, Brown JB, Shams Solari O, Li D, He S, Podicheti R, Stoiber MH, Spanier KI, Gilbert D, Jansen M, Rusch DB. 2018. Early transcriptional response pathways in Daphnia magna are coordinated in networks of crustacean-specific genes. *Molecular Ecology* **27(4)**:886–897 DOI 10.1111/mec.14261.

Overend G, Luo Y, Henderson L, Douglas AE, Davies SA, Dow JA. 2016. Molecular mechanism and functional significance of acid generation in the Drosophila midgut. *Scientific Reports* **6(1)**:1–11 DOI 10.1038/srep27242.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14(4)**:417–419 DOI 10.1038/nmeth.4197.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. *Available at https://www.R-project.org/*.

Riquelme Medina I, Lubovac-Pilav Z. 2016. Gene co-expression network analysis for identifying modules and functionally enriched pathways in type 1 diabetes. *PLOS ONE* **11(6)**:e0156006.

Roy S, Bhattacharyya DK, Kalita JK. 2014. Reconstruction of gene co-expression network from microarray data using local expression patterns. *BMC Bioinformatics* **15(7)**:1–14 DOI 10.1186/1471-2105-15-S7-S10.

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 Research* **13(11)**:2498–2504 DOI 10.1101/gr.1239303.

Silbering AF, Rytz R, Grosjean Y, Abuin L, Ramdya P, Jefferis GS, Benton R. 2011. Complementary function and integrated wiring of the evolutionarily distinct Drosophila

olfactory subsystems. *Journal of Neuroscience* **31(38)**:13357–13375 DOI 10.1523/JNEUROSCI.2360-11.2011.

Smith TE, Moran NA. 2020. Coordination of host and symbiont gene expression reveals a metabolic tug-of-war between aphids and Buchnera. *Proceedings of the National Academy of Sciences of the United States of America* **117(4)**:2113–2121 DOI 10.1073/pnas.1916748117.

Smith CR, Morandin C, Noureddine M, Pant S. 2018. Conserved roles of Osiris genes in insect development, polymorphism and protection. *Journal of Evolutionary Biology* **31(4)**:516–529 DOI 10.1111/jeb.13238.

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. 2006. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research* **34(suppl_1)**:D535–D539 DOI 10.1093/nar/gkj109.

Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302(5643)**:249–255 DOI 10.1126/science.1087447.

The Gene Ontology Consortium. 2017. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research* **45**:D331–D338.

Vella D, Zoppis I, Mauri G, Mauri P, Di Silvestre D. 2017. From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. *EURASIP Journal on Bioinformatics and Systems Biology* **2017(1)**:1–16 DOI 10.1186/s13637-017-0059-z.

Von Der Weid B, Rossier D, Lindup M, Tuberosa J, Widmer A, Dal Col J, Chenda K, Carleton A, Rodriguez I. 2015. Large-scale transcriptional profiling of chemosensory neurons identifies receptor-ligand pairs in vivo. *Nature Neuroscience* **18(10)**:1455–1463.

Wang H, Hernandez JM, Van Mieghem P. 2008. Betweenness centrality in a weighted network. *Physical Review E* **77(4)**:046105 DOI 10.1103/PhysRevE.77.046105.

Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. 2017. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Research* **45(W1)**:W130–W137 DOI 10.1093/nar/gkx356.

Watts DJ, Strogatz SH. 1998. Collective dynamics of 'small-world'networks. *Nature* **393(6684)**:440–442 DOI 10.1038/30918.

Yousef M, Allmer J, Khalifa W. 2016. Feature selection for microRNA target prediction-comparison of one-class feature selection methodologies. In: *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOINFORMATICS, (BIOSTEC 2016).* 216–225.

Zhang B, Horvath S. 2005. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* **4(1)**:17 DOI 10.2202/1544-6115.1128.

Zhang M, Li Q, Yu D, Yao B, Guo W, Xie Y, Xiao G. 2019. GeNeCK: a web server for gene network construction and visualization. *BMC Bioinformatics* **20(1)**:1–7 DOI 10.1186/s12859-018-2560-0.