Machine Learning models identify gene predictors of waggle dance behaviour in honeybees

Marcell Veiner[1], Juliano Morimoto[1], Elli Leadbeater[2], and Fabio Manfredini[1]

[1]University of Aberdeen
[2]Royal Holloway University of London Faculty of Science

September 25, 2021

Abstract

The molecular characterisation of complex behaviours is a challenging task as a range of different factors are often involved to produce the observed phenotype. An established approach is to look at the overall levels of expression of brain genes – known as 'neurogenomics' – to select the best candidates that associate with patterns of interest. This approach has relied so far on a set of powerful statistical tools capable to provide a snapshot of the expression of many thousands of genes that are present in an organism's genome. However, traditional neurogenomic analyses have some well-known limitations; above all, the limited number of biological replicates compared to the number of genes tested – often referred to as "curse of dimensionality". Here we implemented a new Machine Learning (ML) approach that can be used as a complement to established methods of transcriptomic analyses. We tested three types of ML models for their performance in the identification of genes associated with honeybee waggle dance. We then intersected the results of these analyses with traditional outputs of differential gene expression analyses and identified two promising candidates for the neural regulation of the waggle dance: the G-protein coupled receptor *boss* and *hnRNP A1*, a gene involved in alternative splicing. Overall, our study demonstrates the application of Machine Learning to analyse transcriptomics data and identify genes underlying social behaviour. This approach has great potential for application to a wide range of different scenarios in evolutionary ecology, when investigating the genomic basis for complex phenotypic traits.

Introduction

The complex relationship between genes and behaviour has fuelled a large body of recent research and we now know that gene activity can influence brain function, which in turn may affect behaviour . Several studies have shown that behavioural states (distinct and well-characterised behaviours such as foraging or defensive behaviour) can be associated with distinct gene expression profiles in neural tissue, representing the basis for a neurogenomic approach: e.g., large gene networks have been associated with foraging and defence behaviour in honeybees , and numerous candidate neurological genes have been linked to aggression in a variety of organisms, including honeybees and zebrafish Nonetheless, most studies have focused on behavioural states that are long lasting or inherent to a species (Zayed & Robinson, 2012), whereas more plastic and transient social interactions among members of the same species (or colony) have been less characterised at the neurogenomic level . This is likely due to the challenges associated with combining accurate behavioural observations with complex experimental designs to obtain and analyse large sets of gene expression data .

The Western honeybee *Apis mellifera* has become a model organism for neurogenomics due to its fascinating sociobiology, the ecosystem services it provides as a pollinator and the availability of a fully annotated genome . Honeybees display perhaps one of the most iconic social behaviours in the animal world – the "waggle dance"– where foragers communicate the location of suitable food sources and possible nest locations to nestmates via stereotyped movements . This complex behaviour was described for the first time in in the last century by the Nobel Prize winner ethologist Karl Von Frisch and since then many details of its ecological, evolutionary, and physiological underpinnings have been characterized (reviewed in . Despite this, we still do not have a complete picture of how the waggle dance is regulated at the brain level. Pioneering studies have started to reveal some of the key players at the levels of molecules , cell types and genetic pathways associated with dance communication, but it is unclear what genes in the honeybee brain trigger the performance of dance behaviour once activated.

Traditionally the neurogenomic approach has consisted of using statistical methods to calculate differential gene expression , which requires robust data analysis techniques due to the large volumes of sequence reads generated per sample . An interesting development in the field to address the increased computational needs of these approaches has been the application of Machine Learning (ML) to genomics studies . ML methodologies have proved to be powerful resources for this purpose and have been the focus of extensive research recently to identify the possibilities of new applications to a wide range of fields in biology and medicine . Despite the abundance of studies applying ML frameworks to transcriptomic data, its use to characterise the molecular regulation of highly plastic and transient behaviours has not yet been properly explored.

In this study, we set out to identify the genes associated with the performance of dance behaviour in honeybee foragers using a ML approach. We obtained a transcriptomic dataset of brain tissues (mushroom bodies) from honeybee foragers that were sampled for another study designed to underpin the molecular basis for learning distance and direction through the waggle dance (Manfredini et al *in prep* .): mushroom bodies were targeted for this study as they are the best suited brain tissue to explore high cognitive functions in insects , including spatial tasks . We trained three types of ML models (classifiers) on the activation levels of 15,314 genes that correspond to the whole honeybee genome, with the direct goal of classifying honeybees according to whether or not they performed a waggle dance upon their return from a foraging trip (i.e., dancers vs non-dancers). Thereafter, we unified the information obtained from the different ML approaches to identify the genes associated with these complex behavioural states, and we compared these results with more traditional analyses of gene expression. Together, our study provides a deeper insight into the molecular regulations of the waggle dance, a plastic and transient behavioural state, and promotes incorporating ML in the analysis of transcriptomic data.

Methods

Experimental Setup and Initial Dataset

The transcriptomic data used for analysis was part of an experiment prepared to study the molecular basis for social learning of distance in honeybees through the waggle dance . In this experiment honeybees from 3 different colonies were trained to visit a feeder positioned at the end of a 6m long tunnel , which was used to alter the bee's perception of distance as follows: vertical stripes (with respect to the direction of flight) on the tunnel walls were used to increase the estimated flight distance, while horizontal stripes were used to decrease it (Figure 1). Honeybees were then marked at the feeder according to perceived distance (similarly to , yielding two groups: "honeybees perceiving long distance" and "honeybees perceiving short distance".

Honeybee colonies were housed in an observation hive, which allowed direct monitoring of the comb where honeybees normally performed the waggle dance after returning from a foraging trip, known as the "dance floor" . Bees were trained during the morning to visit the feeder at the end of the tunnel – it usually took approximately 5 hours to complete this part – and then in the afternoon (between 2pm and 4pm) foragers that regularly visited the feeder were monitored by an observer while flying from the dance floor to the feeder and vice-versa repeatedly. During this 2-hour time window, the dance floor was also recorded with a video camera, producing a recording of all waggle dance events that occurred in the focal colony. The footage was then carefully analysed to identify marked honeybees that performed waggle dances upon their return from the feeder (from now on "dancers") and separate them from those that instead were never seen performing any dance ("non-dancers") despite being visible on the dancefloor upon their return from a foraging trip. An analysis of the recorded dances confirmed that the manipulation was successful: bees exposed to vertical stripes advertised longer distances on average in their dances compared to bees exposed to horizontal stripes (Manfredini et al. *in prep*).

This resulted in the following 4 groups of honeybees: Dancer perceiving Long distance (DL), Dancer perceiving Short distance (DS), Non-dancer perceiving Long distance (NL), Non-Dancer perceiving Short distance (NS), with 8 replicate samples in each of the 4 groups (N = 32). Brain tissues from all these samples were processed individually for RNAseq analysis (see Supplemental Information). RNAseq read files were aligned to the most recent version of the *A. mellifera* genome (Amel_4.5) using the intron-aware STAR aligner version 2.6.1a . Read counts were extracted using the *featureCounts* function from the Bioconductor R package 'Subread' version 1.8.0 . The final dataset, which represent the starting material for this study, included the read counts for 15,314 genes, corresponding to the whole honeybee genome across 32 bees. As we noticed some variation in library size for some of the bee samples, we normalised read counts by the total library sizes to correct for the effect of possible outliers.

Model Hyperparameters and Data Pre-processing

We used the 'caret' package in the programming language R to train and assess the performance of the classifiers. To evaluate the models, 20% of the data (6 samples) were retained for testing and the remaining 80% (26 samples) were used to train each model. While training, we assessed the performance of each classifier on the validation set and selected the one with the highest area under the ROC curve which is available using the *twoClassSummary* as the summary function.

In order to avoid overfitting to the validation set, we performed cross-validation repeatedly 100 times We chose the number of folds to be 10, a standard practice in ML. The optimal hyperparameters were found by caret implicitly, by performing a grid search through the most likely values. Furthermore, as pre-processing, the data were centred, scaled, and freed from variables of (near) zero variance, as it is standard in ML.

Selected Machine Learning Algorithms

We used Principal Component Analysis (PCA) in order to explore the underlying structure of our dataset. As a result of this set of preliminary analyses, we carefully selected the classification algorithms shown in Table 1. For a brief description of these algorithms see Supplemental Information. We also made use of "Feature Selection" techniques (FS) in order to identify the most suitable features (genes) at predicting the correlation between gene expression data and dance behaviour.

We explored three approaches with implicit feature ranking procedures based on previous studies (see Table 1 and also Supplemental Information): Random Forests (RF), Lasso and Elastic net Regularized Generalized Linear Model (GLMNET), and Support Vector Machine (SVM). Due to the complexity of the data, we decided to use a radial kernel for SVM, as supported by previous research . These methods, also known as "embedded techniques", rank the features based on the already trained classifier, and as a result, the predictive power of the selected features is dependent on the performance of the model. The selected approaches proved to converge on the same final set of predictors even when subjected to repeated random starting conditions.

Whereas embedded methods obtain the importance of certain features from the trained model, wrapper methods, such as Recursive Feature Elimination (RFE), embed the model hypothesis search within the feature subset search . RFE uses backwards selection to assess the importance of each feature to the model. The ranking of the features is done by the underlying algorithm, which can be RF, SVM, or others . Considering the promising properties of RF for genomics studies , we decided to use RF as the underlying model for recursive feature elimination.

Characterisation of Focal Genes

The results of the described approaches were used comparatively to obtain a final set of predictors. First, we obtained the top 20 most important features according to each approach, which were then compiled into a single list of focal genes. To test the statistical significance of the overlaps, we calculated the Jaccard Index and Odds Ratio with the GeneOverlap R package . The annotations of overlapping genes were obtained using NCBI. Where NCBI could not provide any information on putative gene function, we used BLAST to retrieve such information from orthologs in other organisms. We then performed overlap analyses to detect candidate genes that were in common among the three algorithms. For comparison with standard analytical methods, we also analysed the same dataset of RNAseq read counts with a traditional transcriptomic approach to identify differentially expressed genes across groups (see . We used two different statistical analyses using the Bioconductor R package: we performed the Likelihood Ratio Test (LRT) using DESeq2, version 1.24 , and fit a Generalized Linear Model (GLM) using edgeR . We adopted a False-Discovery Rate (FDR) equal to 0.05 to invoke statistically significant difference in gene expression. Lastly, we compared the outputs of these analyses with the list of candidate genes from the machine learning approaches to identify common genes.

Results

Exploratory Analysis

PCA was unable to clearly separate the four groups of bees according to the combination of dance behaviour (dancer (D)/non-dancer (N)) and distance perceived (long (L) /short (S)) (Figure 2). However, when considering the dance factor alone, we obtained a low-dimensional representation/projection of the data using only few Principal Components (PCs), which produced easily distinguishable clusters of samples i.e., dancers and non-dancers (Figure 3). The representation was dominated by PC1, accounting for 51% of the variance in the data, while PC2 only accounted for 14.4% (Figure 3 and see also Supplemental Information).

In particular, dancers were clustered together towards the centre of the plot, showing lower variance than non-dancers – which is indicative of more consistent global patterns of gene expression in dancers vs. non-dancers. We also found 4 non-dancer (2 NL + 2 NS) which formed a separate cluster further along the first Principal Component (PC1). These samples showed the highest loadings for PC1, with levels around 200 that were much higher than dancers (centred around 0) and the other non-dancers (all below 0). We identified the 3 genes with maximal loadings for PC1: GB52651 (*diphthine-ammonia ligase*), GB49108 (*PDZ domain-containing protein 8*) and GB44753 (un uncharacterized gene). Moreover, dancers showed the highest levels of positive correlation between global patterns of gene expression and the first two principal components PC1 and PC2 (DL = 0.711 and DS = 0.574, Figure 2). Overall, the data show a clear underlying structure in the dataset with respect to the dance component (dancers vs non-dancers) while no evident structure appeared to be associated with the perceived distance (long vs short). Based on these findings, we proceeded in our ML analyses focusing on the “dance” factor alone.

Embedded Methods

We trained two classifiers with the underlying algorithms SVM and GLMNET (see Table 1). Both algorithms achieved 100% accuracy (ACC) on the test set, with a 95% likelihood that the true value lies between (54% - 100%); the wide range is due to the limited size of the test set. The No Information Rate (NIR) was 0.5, as we started from a balanced dataset, and the p-value for ACC > NIR was 0.01563. We concluded that both algorithms generalised successfully, as high performance was achieved on both the training and test datasets (see also Supplemental Information).

Recursive Feature Elimination

The model achieved high accuracy even when using only a small portion of the available features (genes in this case). In particular, with only 20 variables it achieved 0.9752 ROC, 0.906 Sens and 0.8715 Spec. Therefore, the model was able to represent the data using only limited information. The best subset however, proved to be marginally better using 5000 of the original features (see Supplemental Information), which is around 34% of the available data. We examined the first 20 of these genes and found significant overlap with the other methods (see Figure 4). Similar to the approaches above, RFE was able to generalise from the training data, and obtained 100% ACC on the test set. Below we present a characterization of the focal genes selected by each approach.

Genes Identified as Key Predictors

There were 18 genes (predictors) that were shared between at least two approaches (see Figure 4). The largest overlap was observed between RFE and SVM (16 genes) while the overlap between SVM and GLMNET was smaller (4 genes). No overlap was detected between RFE and GLMNET. The Jaccard Index between SVM and RFE was the most significant (0.739), while between SVM-GLMNET and GLMNET-RFE it was 0.111 and 0.052, respectively. Similarly, the odds ratio indicated strong association between SVM and RFE (10265.036, p < 0.001). Overall, elements in common corresponded mainly to protein coding genes, with the exception of "GB40714" which corresponds to non-coding RNA. We were able to retrieve functional information for most of the genes from annotations of the honeybee genome (see Table 2), with the exception of "GB50940" and" GB45448", where annotations were available only for closely related insects (*Apis dorsata* and *Apis cerana* , respectively), and "GB54617" that we could not find any information for.

Comparison with Standard Gene Expression Analyses

We characterized gene expression patterns in the same groups of honeybees as above with standard statistical approaches in order to identified possible elements in common with the ML approaches that we tested. The LRT approach identified 243 genes that were statistically different between any two groups of bees, while the GLM approach identified 373 genes that were specifically different between dancers vs. non-dancers (see Supplemental Information for the lists of these genes). We performed overlap analyses between these two gene sets and the list of 18 genes selected by the machine learning approaches. This resulted in 5 genes in common for the LRT approach, and 9 genes in common for the GLM approach: both overlaps are statistically significant (Representation Factors: 17.5 and 20.5; p-values < 0.001 in both comparisons) indicating more genes in common than expected by chance. Five genes were shared across all analyses that we performed (see Table 2). Interestingly, all focal genes identified by ML approaches (as well as elements in common with gene expression analyses) were expressed at higher levels in dancers, indicating their possible involvement with the regulation of dancing behaviour (see Figure 5).

Discussion

In the present study, we implemented a Machine Learning approach to investigate the transcriptomic signatures arising from a complex plastic phenotype. We explored the unique gene expression profiles of *Apis mellifera* associated with dance behaviour, in order to determine the set of focal genes that could be some of the key regulators for this complex behaviour. Training two embedded models (SVM & GLMNET) and one wrapper algorithm (RFE-RF), we were able to achieve perfect accuracy in assigning honeybees to the major behavioural response that we tested ("dancer" vs. "non-dancer") according to gene expression data. Using Feature Selection, we were able to obtain a set of key predictors for each classifier, which were then distilled into a list of genes. Our results show a set of genes that are promising candidates and could directly regulate dance behaviour.

While we were able to clearly separate dancers from non-dancers, our initial preliminary analyses (PCA) were unable to detect any major effects of distance perception, which was one of the research questions that we had initially pursued. Although we found that the impact of distance perception on gene expression was too subtle to be detected by our approaches, other studies have succeeded to identify the effect of distance perception alone on honeybee brain gene expression, using more traditional statistical tools of transcriptomic analyses . It is possible that with an increased sample size, we would have been able to investigate this behaviour further. Alternatively, the transcriptomic signature associated with distance perception might be more significant in honeybees experiencing real distance as opposed to the perceived distance that honeybees experienced through our tunnel manipulation setup. As a matter of fact, a larger set of genes was found to differ between foragers experiencing real long distance vs. short distance (Manfredini et al. *in prep.*) but we cannot exclude that a portion of these genes might have changed their patterns of expression from one group to the other according for example to different metabolic costs of flight.

A recurrent problem in transcriptomic analyses is that the number of genes (predictors) is far greater than the number of samples. The fact that we managed to obtain an overlapping subset of genes by different

ML and differential gene expression analyses in this setting indicates that the dance behaviour has a unique and well-defined transcriptomic signature. Moreover, our methods further prove that ML can be used as a complementary approach to provide further support to transcriptomic studies and help restricting the focus to a smaller group of genes that can be investigated for their association with a behaviour of interest. Note also that this approach can be expanded beyond honeybees and the waggle dance; a significant portion of animal behaviours are transient and plastic, making them very difficult to characterise, but they can be described by combining traditional gene expression analyses and ML approaches.

Interestingly the overlap between SVM and RFE-RF was greater than with GLMNET. In high-dimensional settings, such as ours, there can be more than one optimum for the learning objective, which is why ML models can end up in different final states on subsequent runs with random initialisation. In our case, it Is simply possible that GLMNET favoured reaching a different state than those reached by SVM and RFE-RF, hence selecting different features. It is also possible that since GLMNET has built in mechanisms to avoid overfitting, its learning procedure was slower than the other methods, and simply needed more training time.

Nevertheless, the extensive overlap between the SVM and RFE-RF approaches, and the fact that many of the identified genes were also in common with traditional methods to analyse gene expression shows great promise. We hypothesise that these genes are the best predictors for the dance behaviour, as they all appear to be expressed at higher levels in dancers vs. non-dancers. In particular, the two genes that are in common to all three ML approaches and were also identified by at least one approach of gene expression analysis deserve special attention.*Boss* (*bride of sevenless*) belongs to the group of G-protein coupled receptors; an important family of genes often associated to expression of behaviour in insects. In particular,*boss* has been associated to a set of different functions in*Drosophila* , including sight and eye development, energy homeostasis and response to glucose . *Boss* might have been co-opted in honeybees to regulate dance behaviour, an energetically expensive activity that is highly related to feeding behaviour (and therefore to sugar response) and relies on visual input for orientation purpose during flight to a foraging site. As for *heterogeneous nuclear ribonucleoprotein A1* , studies on the *Drosophila*orthologue *HRP59* have revealed a role for this gene in alternative splicing , a molecular process that consent the translation of a single mRNA molecule into multiple protein variants , significantly increasing the repertoire of responses to a stimulus. Even though the role of alternative splicing in the regulation of behaviour is largely unknown, this process has started to be characterized in multiple organisms, including honeybees , hinting at the possibility that the honeybee orthologue of *HRP59* might contribute to the high plasticity that is necessary to regulate a complex behaviour such as the waggle dance.

More functional approaches are needed to move beyond correlation and investigate whether a causal link exists between the expression levels of the genes that we identified and the performance of dance behaviour. If further research were to support this, the results could then be used to test the recruitment potential in a specific colony. By designing a diagnostic tool to directly measure the levels of expression of the focal genes and compare them against a reference, it would be possible to assess the overall ability of a colony at recruiting to a foraging site through dancing.

Acknowledgements

The authors would like to thank Dr Georgios Leontidis (The School of Natural and Computing Science, University of Aberdeen) for his valuable support during the selection and implementation of the ML models. The authors are also grateful to NERC (Natural Environment Research Council) for funding this project and supporting MV salary over 10 weeks through their Research Experience Placement (REP) program (DTG reference: NE/S007377/1). The honeybee work that was performed to obtain the sequencing data used in this study was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant number 638873 to EL). This funding also supported FM during the execution of the field and molecular work.

References

Data Accessibility and Benefit-Sharing Statement

All codes used in the analyses here reported are visible in a GitHub repository associated with this project https://github.com/Vejni/WaggleDance_MachineLearning. The raw sequencing data that represent the starting material for the analyses here described have been deposited on NCBI SRA (Bioproject PRJNA756776).

Author Contributions

Designed research: MV, JM, FM. Collected samples and prepared them for analysis: FM. Performed research: MV. Contributed new reagents and analytical tools: EL. Analysed data: MV, FM. Wrote the paper: MV with input from JM, EL and FM.

Figures and Tables

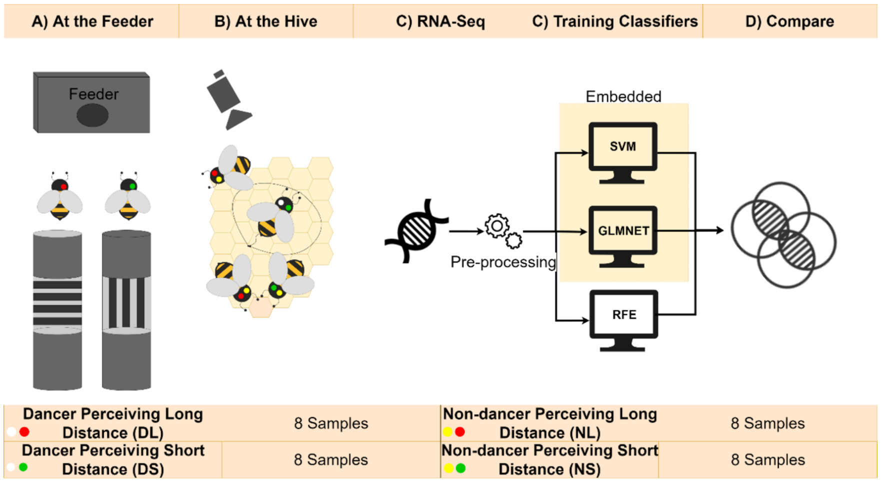

A) At the Feeder
B) At the Hive
C) RNA-Seq
C) Training Classifiers
D) Compare
Feeder
Embedded
SVM
GLMNET
RFE
Pre-processing
Dancer Perceiving Long Distance (DL)
8 Samples
Non-dancer Perceiving Long Distance (NL)
8 Samples
Dancer Perceiving Short Distance (DS)
8 Samples
Non-dancer Perceiving Short Distance (NS)
8 Samples

Figure 1 . Schematic Representation of the Experimental Design. A) Honeybees visited a feeder through a tunnel, used to alter their distance perception (vertical stripes with respect to the direction of flight increased, while horizontal stripes decreased the distance perceived), and were then marked accordingly. B) At the observation hive, honeybees were recorded while performing or not-performing dancing behaviour and were finally prepared for RNAseq analysis (C). D) Three machine learning algorithms (SVM, GLMNET and RFE) were trained on the pre-processed sequence reads (Training Classifiers). E) Key features from each model were compared to identify common elements (genes or predictors).

Table 1 . Benchmark Algorithms. We chose to test SVM, GLMNET, RF and RFE for our study, based on their use in previous research. The first three algorithms use embedded feature selection (FS) to obtain key predictors from the trained model (Embedded), while RFE requires an underlying embedded approach for the ranking (Wrapper). We report the studies that featured or reviewed these algorithms.

Algorithm	FS Method	Reviewed in	Featured in
Support Vector Machine (SVM)	Embedded		
Random Forest (RF)	Embedded		
Generalized Linear Model (GLMNET)	Embedded		
Recursive Feature Elimination (RFE)	Wrapper		

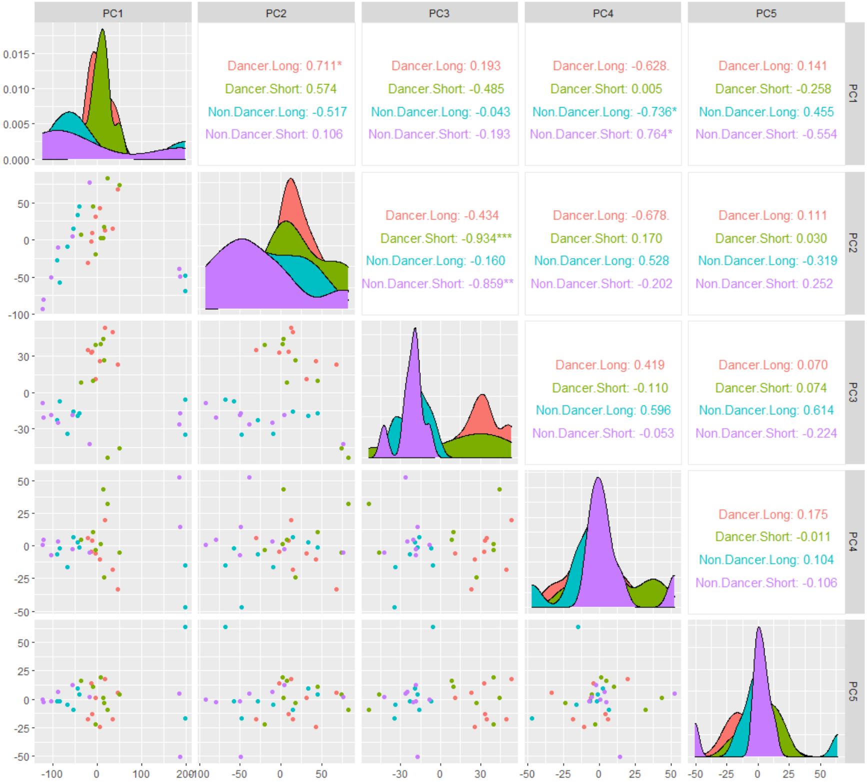


Figure 2 . Principal Component Analyses. Shown here are the two-dimensional comparisons of all Principal Components (PC) across 4 groups of bee foragers. The scatterplots (bottom-left hand side of the picture) show datapoints as they are represented with any 2 PCs. Each scatterplot corresponds to two PCs, indicated at the top of the figure and on the right: for example, the plot in the first column and fifth row corresponds to PC1 (top ID)) and PC5 (right-side ID). The diagonal shows the distributions for each PC over each group in the experiment (Dancers perceiving Long distance (red), Dancers perceiving Short distance (Green), Non-Dancers perceiving Long distance (Blue), and Non-Dancers perceiving Short distance (Purple)). The upper-right hand side of the figure shows the correlations between each of the 4 groups according to the corresponding PC. These values also indicate the direction (if any) of the groupwise trends in the scatter plots. Values with stars show statistical significance and *, **, *** correspond to p [?] 0.05, p [?] 0.01, and p [?] 0.001, respectively.

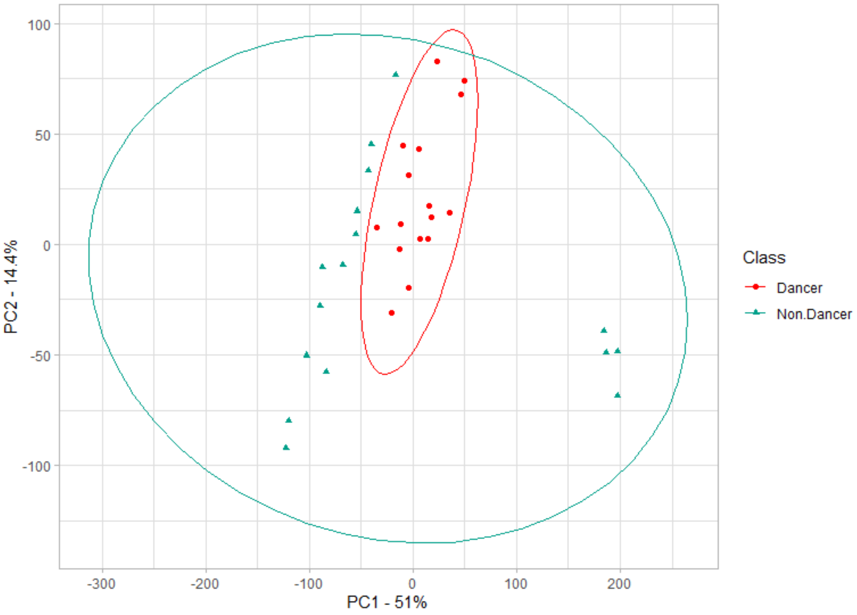


Figure 3 . 2D projection of the Dataset using the first two Principal Components. Datapoints are coloured (and shaped) by their group membership i.e., red circles correspond to Dancers and turquoise triangles correspond to Non-Dancers. Confidence ellipses of the appropriate colour also show the variance of the two groups.

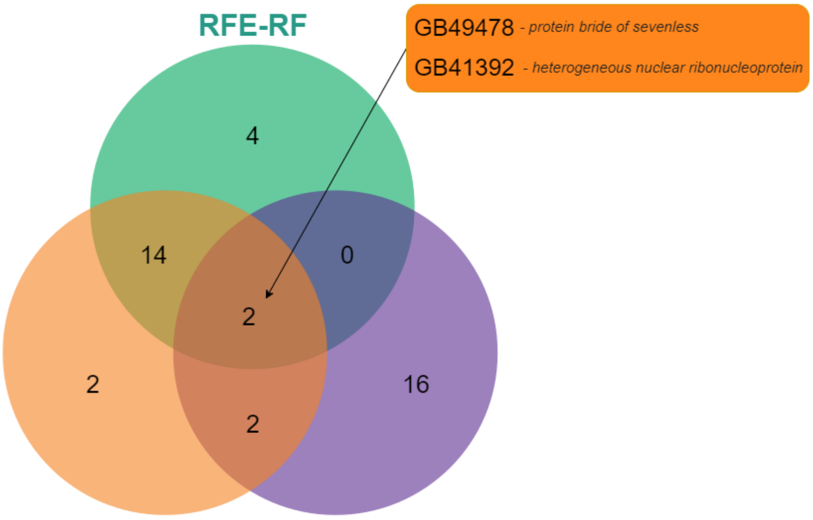

RFE-RF
GB49478 - protein bride of sevenless
GB41392 - heterogeneous nuclear ribonucleoprotein
4
14
0
2
2
16
2
SVM
GLMNET

Figure 4 . Overlap between Selected Features. We queried the 20 most important features in each trained classifier, SVM (Orange, bottom left), GLMNET (Purple, bottom right) and RFE-RF (Green, top), which were then compared for overlapping subsets of genes. There were 16 genes selected both by RFE-RF and SVM, with 2 genes (GB41392, GB49478) selected by all 3 approaches. GLMNET showed little overlap with SVM and RFE-RF.

Table 2 . Focal Genes. The annotations of overlapping genes (selected by at least 2 approaches) were obtained using NCBI or BLAST search (where needed). Additionally, Gene Expression Analysis (GLM and LRT) detected overlapping genes, which is indicated in the "Gene expr. Analysis" column. The "Reference" column indicates key studies that focused on the selected genes. The obtained list of focal genes included promising genes (e.g., GB49478, GB50290) that could play a key role in the dance behaviour observed in honeybees.

Hosted file

`image5.emf` available at https://authorea.com/users/437108/articles/538980-machine-learning-models-identify-gene-predictors-of-waggle-dance-behaviour-in-honeybees

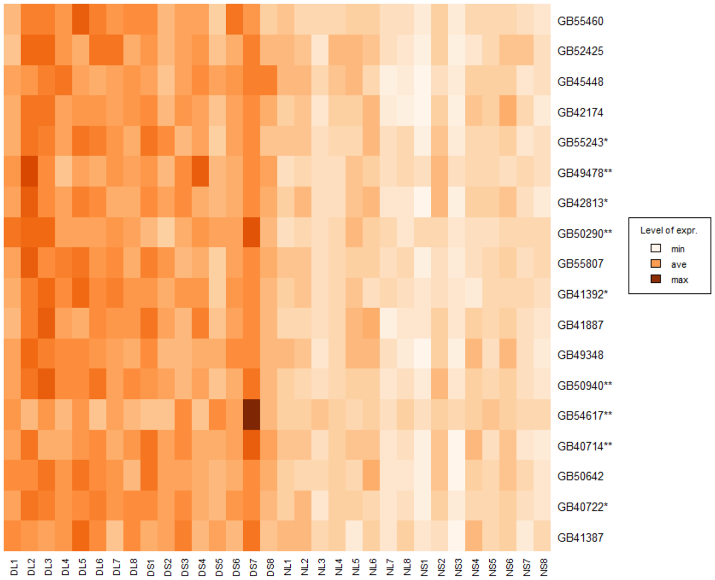
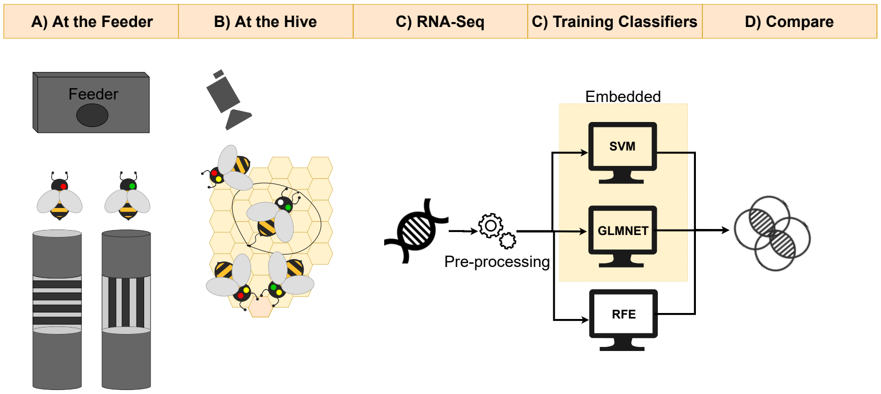
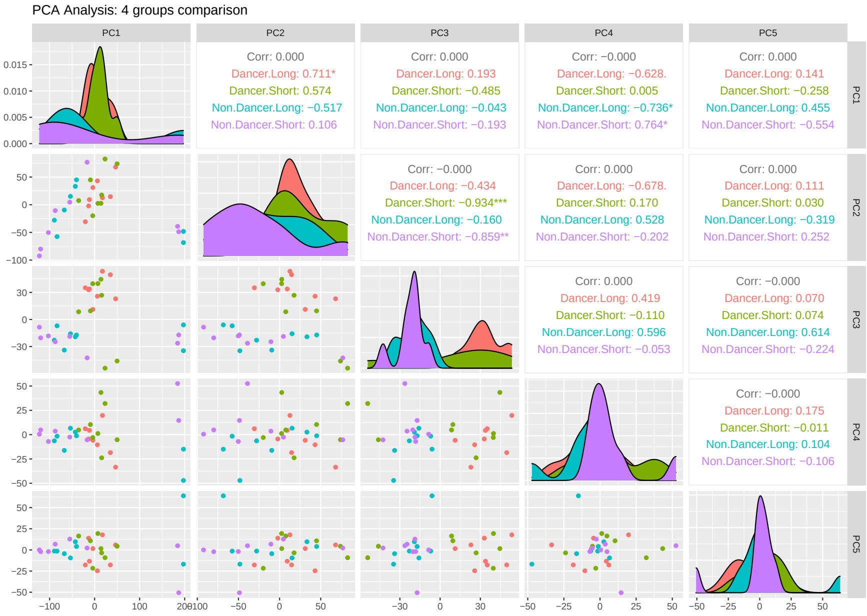
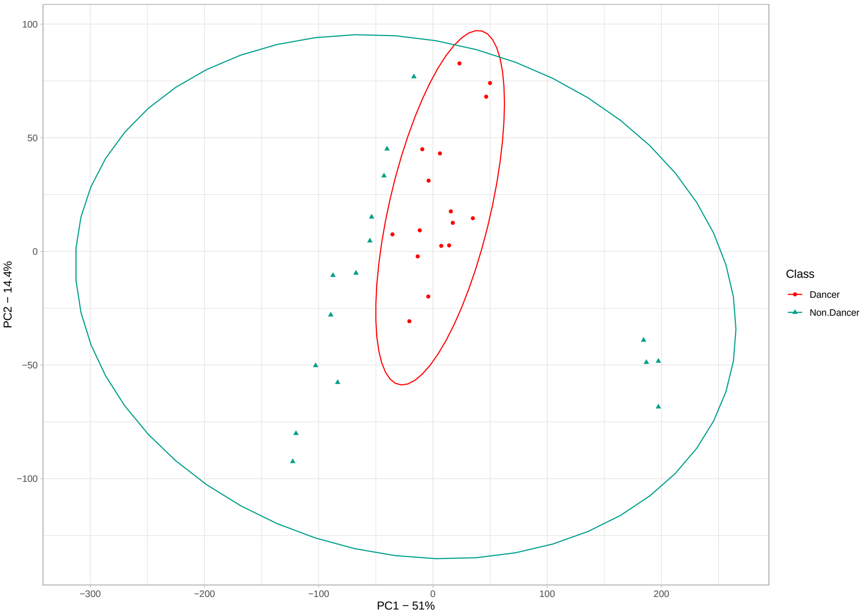


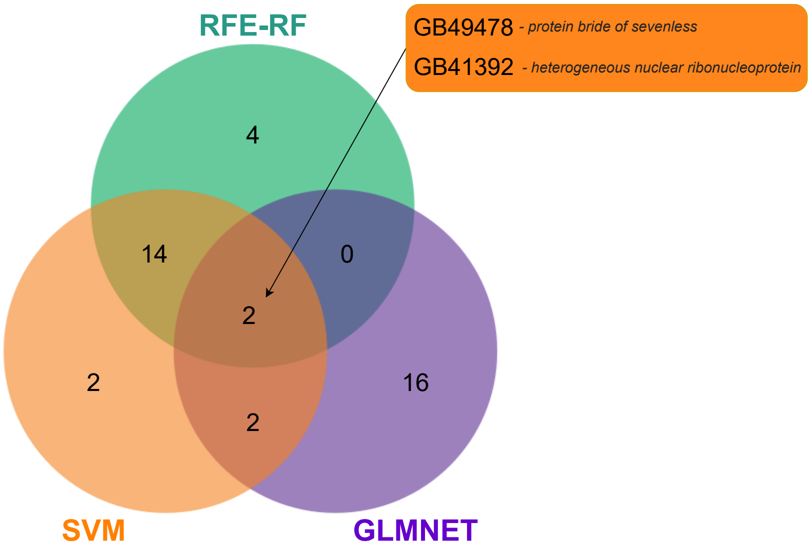
Figure 5. Focal Genes Heatmap. The heatmap shows the log counts per million (LogCPM) transformed expression levels of the 18 focal genes across all bee samples (Dancers grouped to the left, while Non-Dancers to the right). All focal genes showed higher levels of expression in Dancers (indicated by darker colours). The gene IDs marked with a single asterisk (*) were identified by gene expression analysis (GLM approach only), while IDs marked with two asterisks (**) were identified by both GLM and LRT.



○ ● Dancer Perceiving Long Distance (DL)	8 Samples	● ● Non-dancer Perceiving Long Distance (NL)	8 Samples
○ ● Dancer Perceiving Short Distance (DS)	8 Samples	● ● Non-dancer Perceiving Short Distance (NS)	8 Samples

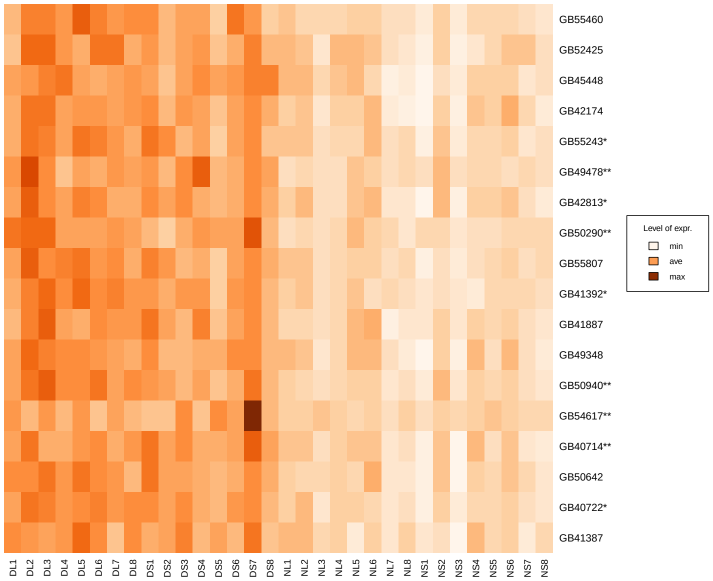






GB49478 - *protein bride of sevenless*

GB41392 - *heterogeneous nuclear ribonucleoprotein*


GB55460
GB52425
GB45448
GB42174
GB55243*
GB49478**
GB42813*
GB50290**
GB55807
GB41392*
GB41887
GB49348
GB50940**
GB54617**
GB40714**
GB50642
GB40722*
GB41387
DL1
DL2
DL3
DL4
DL5
DL6
DL7
DL8
DS1
DS2
DS3
DS4
DS5
DS6
DS7
DS8
NL1
NL2
NL3
NL4
NL5
NL6
NL7
NL8
NS1
NS2
NS3
NS4
NS5
NS6
NS7
NS8
Level of expr.
min
ave
max