A Bayesian Co-clustering of Anopheles Gene Expression Time Series Responses to Multiple Immune Challenges

Nicholas A. Heard[*], Christopher C. Holmes[†], David A. Stephens[*], David J. Hand[*], and George Dimopoulos[‡]

[*]Department of Mathematics, Imperial College London, Huxley Building, 180 Queens Gate, London SW7 2AZ, UK; [†]Oxford Centre for Gene Function, Department of Statistics, University of Oxford, Oxford, UK and MRC Mammalian Genetics Unit, Harwell, Oxford, OX11 0RD, U.K; and [‡]Department of Molecular Microbiology and Immunology, Johns Hopkins School of Public Health, 615 N. Wolfe Street, Baltimore, MD21205, USA.

ABSTRACT

We present a method for Bayesian model-based hierarchical co-clustering of gene expression data and use it to study the temporal transcription responses of an *Anopheles gambiae* cell line upon challenge with multiple microbial elicitors. The method fits statistical regressions to the gene expression time series for each experiment, and performs co-clustering on the genes by optimizing a joint probability model, characterizing gene co-regulation between multiple experiments. We compute the model using a two-stage EM-type algorithm, first fixing the cross-experiment covariance structure and using efficient Bayesian hierarchical clustering to obtain a locally optimal clustering of the gene expression profiles, and then conditional on that clustering carrying out Bayesian inference on the cross-experiment covariance using Markov chain Monte Carlo simulation to obtain an expectation. For the problem of model choice we use a cross-validatory approach to decide between individual experiment modeling and varying levels of co-clustering. Our method successfully generates tightly co-regulated clusters of genes that are implicated in related processes, and can therefore be used for analysis of global transcript responses to various stimuli, and prediction of gene functions.

INTRODUCTION

The *A. gambiae* mosquito is the major vector of human malaria and its innate immune system, which is capable of killing *Plasmodium* parasites, is a key area of research; see (9; 10; 11; 12). Here we analyse cDNA microarray data measuring gene transcription of an *A. gambiae* immune competent cell line Sua 1b in response to challenge with the microbial elicitors *Escherichia coli*, *Salmonella typhimurium*, *Micrococcus luteus*, *Listeria monocytogenes* and the yeast wall extract Zymosan. The microarray slides used represented approximately 2,400 unique genes and expression was assayed at 1, 4, 8, 12, 18 and 24 hours after challenge as previously described in (8). Fig. 1 shows the ranked values of the data plotted as a heat-map.

Cluster analyses of microarray data are used to identify potentially functionally related groups of genes that are transcriptionally co-regulated, that is, respond similarly at the transcriptional level to external stimuli and are most likely to be controlled by the same transcription factors and pathways. These analyses thereby allow assignment of putative functions to novel genes based on their shared cluster membership with genes of known functions. Due to the similarity of the different treatments here it is anticipated that we should observe some correlation in the gene transcription patterns across the different experiments. So for a cluster analysis of the genes in this setting, it is natural to consider jointly clustering ('co-clustering') the genes according to their response profiles across all of the experiments. Genes displaying similar expression profiles across the treatments, and hence being part of the same clusters, can be predicted to share similar functions or act in the same processes.

Cluster analysis in gene expression has mainly relied on Pearson's correlation or Euclidean distance based methods for clustering expression measurements and profiles. Typically, hierarchical clustering is used to obtain a sequence of partitions of N data observations, ranging from a single group containing all observations to N groups each containing just one observation. The partitioning occurs in either a *divisive* fashion (one group to N groups) or by *agglomeration* (N groups to one group), with the hierarchy represented by a tree, or *dendrogram*. For the analysis of time course expression data, such as the data shown in Fig. 1, a multivariate clustering algorithm is required whereby the time dependency is respected.

In (1), we developed a Bayesian model-based agglomerative scheme for clustering time course

microarray data, using non-linear regression splines to capture temporal variation within each cluster. The use of a Bayesian procedure allows us to compute measures of uncertainty for quantities of interest, such as the number of clusters in the data, and to report posterior probabilities that are comparable across all models, experiments, and computational methods. The use of non-linear regression splines allows us to accommodate non-stationary time-dependence in the data as well as unequal sampling intervals and yet affords analytic calculation of marginal probabilities. See (1) for further details. Model-based clustering of gene expression time series data has also been considered by amongst others, (2; 3; 4; 5; 6). Superior performance of non-stationary model-based clustering of time series data against the more standard clustering algorithms was established in (1).

The key extension we present here involves the co-clustering of multiple gene expression profiles obtained under related experimental treatments; to our knowledge, no explicit methodology has previously been developed for multiple experiment co-clustering of time series; though of course people have considered two way clustering outside of time series application, such as (7). We extend the models of (1) by allowing the co-expression of genes, not only *within* experiments at consecutive time points but now also *between* experiments. The benefits of joint modelling across parallel experiments are two-fold. Firstly, we obtain a more robust clustering of the genes, with the borrowing of strength between experiments working to stabilize the low signal to noise ratio inherent in current cDNA microarray study. Secondly, we can characterize the dependence structure in co-expressed profiles across treatments.

Figure 1 here

METHODS

The data are composed of the expression levels of N = 2,392 putative genes that are represented by DNA spots (probes) on a glass slide microarray, (see (8) for details). Of these 2,392 genes, 332 had putatively known function based on sequence similarity. The expression profiles shown in Fig. 1 relate to log2 transformed normalized ratios of intensities of hybridized samples (RNA from challenged versus non-challenged cells) to the spotted gene probes as previously described in (8). Assays were done at 6 time points, at 1, 4, 8, 12, 18 and 24 hours, after each

of the microbial elicitor challenges Full experimental details are given in (8).

Inspection of Fig. 1 indicates that many of the genes display similar expression profiles across the different microbial elicitor challenges, suggesting that it may be possible to co-cluster the genes using all of the data collected across the different treatments; this is made more explicit in Fig. 2, where for each pair of experiments we show the distribution across the genes of cross-correlations against the global mean of zero of the relative expression time series. That is, letting y_{ge} be the T-vector time series of expression levels for gene g in experiment e, for each pair of experiments e_i, e_j, $i \neq j$, we calculate

$$\frac{y_{ge_i}{}'y_{ge_j}}{\sqrt{y_{ge_i}{}'y_{ge_i}}\sqrt{y_{ge_j}{}'y_{ge_j}}}, \quad g = 1, \ldots, N. \tag{1}$$

Note that this is the gene correlation measure used for clustering of a single experiment by (7). Under the scenario of no underlying experimental correlation (easily simulated by permuting the gene indices for one of the pair of experiments) the distribution of Eq. 1 is symmetric about zero. The median values of the distributions for each of our experiment pairs are also shown in the plots in Fig. 2, and are all strongly positive.

Figure 2 here

A single experiment cluster analysis of the data from the first of the bacterial agents, S. typhimurium, appeared in (1). There, Bayesian model-based clustering procedures were developed to accommodate the large sample size and non-stationary time dependence structure of the data. Here we extend that methodology to the case of multiple experimental conditions for joint clustering across the four microbial elicitors.

Bayesian Modelling of Gene Expression Profiles Model-based clustering requires the specification of a probability distribution for the data residing within a group. We choose to model the gene expression profiles in a regression context via linear models and non-linear basis functions. This takes the form

$$y_{ge} = X\beta_{ge} + \sigma_g\varepsilon_{ge}, \quad g = 1, \ldots, N; e = 1, \ldots, E \tag{2}$$

where X is a $T \times T$ design matrix made up of non-linear basis functions evaluated at the experimental time points, β_{ge} is a T-vector of basis coefficient parameters, σ_g^2 is an error variance and ε_{ge} are independent standard Gaussian errors.

This approach readily accommodates non-uniform sampling times and any non-stationarity in the data (which can be seen for our mosquito data most clearly on examination of the clustered data in Fig. 3). In (1) we demonstrated that the use of fixed basis functions (X) with random coefficients and error variance induces a non-stationary stochastic process model for the underlying variation in expression for which we can analytically evaluate the marginal likelihood. The marginal likelihood forms the basis of our potential function of our agglomerative clustering scheme.

A fully Bayesian approach to clustering gene expression profiles was described by (1). There, for any clustering of the genes, defined by a partition $\mathcal{C}$ of the integers from 1 to N, genes in the same cluster of $\mathcal{C}$ share common values for the regression coefficients and error variance. In particular, the use of regression models, and a computationally efficient agglomerative algorithm for hierarchical clustering was established. In this paper we demonstrate that the models and algorithms can be extended to the more general case where multiple parallel time course profiles are available. Technical details of the extensions can be found in the Supplementary Material.

The key extension proposed in this paper is the modelling of cross-experiment correlation. This is achieved via the covariance matrix V of the regression coefficients β_{ge}, which models the dependence between these parameters across the experiments for gene profiles in each cluster. Specifically, for E experiments we take

$$V = \Sigma \otimes I_T$$

where Σ is an $E \times E$ symmetric positive definite matrix acting as a between experiment covariance matrix and the symbol $\otimes$ represents the *direct* (or *Kronecker*) *product* of two matrices.

To recognize that strong experimental correlation may not exist for some of the gene clusters, we propose a two-component mixture prior distribution for V. Defining $D_{\Sigma} = \text{diag}\{\Sigma_{11}, \ldots, \Sigma_{EE}\}$ to be the decorrelated experimental covariance matrix containing the diagonal elements of Σ,

we use

$$p(V) = \begin{cases} \Sigma \otimes I_T, & \text{with probability } q \\ D_\Sigma \otimes I_T, & \text{with probability } 1-q. \end{cases} \qquad [3]$$

It transpires that, even under this extended model, the basic approach of (1) can still be implemented. Full details are given in the Supplementary Material.

Bayesian Hierarchical Clustering As in (1), the method proposed uses the Bayesian posterior distribution on the unknown partition $\mathcal{C}$ given the expression data (and now conditional upon the experimental covariance parameters (Σ, q)) as a potential function for agglomerative clustering. The improved clustering performance under this approach and its speed of computation were outlined for the single experiment case in (1).

The full algorithm used to implement the Bayesian hierarchical clustering here incorporates a novel Markov chain Monte Carlo (MCMC) based approximation to the Expectation-Maximization (EM) algorithm to enable us to jointly learn about $\mathcal{C}$ and (Σ, q), and is described in detail in the Supplementary Material.

It should be noted here that for fully-Bayesian inference on the joint parameter space of $(\mathcal{C}, \Sigma, q)$ we should, for example, calculate expectations or find maxima over the whole joint posterior distribution of these parameters. However, with such large, high dimensional data as encountered here, standard MCMC methods for exploring the resulting vast joint space prove impractical and in any case severely increase the computation time. For further discussion again see the Supplementary Material.

Model choice We now consider the decision problem of choosing when to perform co-clustering. That is, we wish to ascertain whether it is appropriate to cluster the genes using the data from all of the experiments together, treat each experiment separately or perhaps co-cluster across some but not all of the experiments. Co-clustering cannot always be assumed to be beneficial, as there could be underlying differences in the function groupings of the genes if the treatments are sufficiently different in their action.

There were four microbial elicitor challenges in the study analysed here, and so there are

15 ways of partitioning these experiments into non-empty groups. For any given partition of the experiments {*S. typhimurium*, *L. monocytogenes*, *M. luteus*, Zymosan}, each set within the partition can represent a group of experiments we wish to cluster jointly, though independently from the remaining experiments in the other sets of the partition. The 15 partitions thus represent every possible level of joint modelling of the four experiments, including the special cases of modelling all the experiments individually or all jointly.

A further advantage of using a model based clustering technique is that we are able to perform a cross-validation (CV) study to identify an appropriate joint modelling structure by comparing predictive power. Taking any partition of the experiments, we sequentially leave out one of the interior time points in turn from the experimental data; for each group in the partition we then run our MCMC/EM algorithm to find an optimal clustering and the corresponding predictive distribution of the missing data points. As experiments in different groups are independent of one another we can take products of these predictive densities to find the predictive distribution of all the observations at the deleted time point. This procedure is repeated for each time point and then for each of the possible partitions of the experiments.

Note that we only have four experiments to consider here and so there is no difficulty in looking exhaustively at all of the possible partitions. For larger numbers of experiments this would become prohibitive, and instead one could resort once more to an agglomerative clustering procedure, though now on the experiments, to find a locally optimal joint modelling structure.

RESULTS

We present the results of our cluster analysis of the *A. gambiae* cell line transcription responses to microbial challenge described above. It is common practice in a cluster analysis of microarray data to filter out all genes whose observed increase or decrease in expression relative to unchallenged cells is never greater than some significance threshold, usually a minimum of two-fold up or two-fold down regulation (see, for example, 8). For our model-based method this pre-processing is not necessary, with these non-regulated genes naturally grouping together to form low variance clusters. Thus we included all the data in our analysis.

Firstly, to find the appropriate level of joint modelling a cross validation study as described

above was performed. The best experimental clustering structure found was to model all of the experiments together, and the worst was to model each experiment separately. So for this study, the borrowing of strength through co-clustering achieves more robust clustering. The CV scores for these two models are shown in Table 1, indicating overwhelming support for our co-clustering approach.

t left out	Individual	Joint
4	-44.146	799.275
8	389.805	1056.590
12	-1366.335	-184.610
18	-1892.495	-543.0239
Ave.	-706.220	282.058

Table 1: Log predictive densities. The first column gives the particular time point left out in the cross-validation. The second and third give the log predictive probabilities through modelling the experiments separately and using co-clustering.

In fact, the best two and three cluster models form a hierarchy, meaning the optimal clusterings at any desired number of clusters could have been found here by agglomerative clustering of the experiments; the merger sequence would have been:

1. $\{S.\ typhimurium\}, \{M.\ luteus\}, \{L.\ monocytogenes\}, \{Zymosan\}$ (-706.220)

2. $\{S.\ typhimurium\}, \{M.\ luteus\}, \{L.\ monocytogenes,\ Zymosan\}$ (-188.080)

3. $\{S.\ typhimurium\}, \{M.\ luteus,\ L.\ monocytogenes,\ Zymosan\}$ (79.2775)

4. $\{S.\ typhimurium,\ M.\ luteus,\ L.\ monocytogenes,\ Zymosan\}$ (282.058)

where the numbers in brackets are the CV scores for that experimental clustering.

For a simple comparison, we also tried standard hierarchical agglomerative clustering of the four experiments by simply concatenating the time series of all the genes into one long vector for each experiment. Dendrograms under different distance metrics and linkage choices are given in the Supplementary Material. In particular, under the Euclidean distance metric we obtained the same optimal experimental merger hierarchy as above using single or average-link clustering.

Having identified the optimal experimental clustering, our Bayesian hierarchical clustering method was used to learn jointly about the experimental covariance matrix and the gene clustering; see the Supplementary Material for full details. The reordered gene expressions after the final hierarchical clustering are shown in Fig. 3, and the expected covariance matrix giving rise to this clustering is given in Table 2. Note that strong, positive correlations have been found between all of the experiments, a finding supported by our earlier exploratory plots in Fig. 2. This correlation is much more apparent in Fig. 3 than it was for the unordered data plot, with genes generally up (down) regulated across the four treatments on the left (right) hand side of the plot. The two dendrograms plotted on Fig. 3 indicate the optimal hierarchy of mergers for the genes (horizontal axis) and challenges (vertical axis).

Figure 3 Here

	S. typhimurium	L. monocytogenes	M. luteus	Zymosan
S. typhimurium	1	0.782	0.624	0.614
L. monocytogenes	0.782	1	0.702	0.876
M. luteus	0.624	0.702	1	0.842
Zymosan	0.614	0.876	0.842	1

Table 2: Estimated correlation between the experiments. Final conditional posterior mean correlation matrix after 20 EM steps each with 10,000 MCMC iterations for full joint model correlating all four experiments.

The optimal clustering found had 159 clusters, many more than we get when clustering on the expressions of a single experiment (1). This increase reflects the extra information contained in the combined gene expression time series, so that genes which may have appeared similar by chance under a smaller number of sampling points are revealed as quite different. However, note that if a smaller number of clusters is sought, then as the method is hierarchical such a clustering can be read off from the dendrogram at any desired level. In fact the partitions for up to say ± 50 clusters from the optimal clustering had only slightly lower posterior probability and thus offer many plausible alternatives.

DISCUSSION

We have tested our clustering method on a dataset of mosquito immune responses to microbial challenge. Of the 2,392 assayed genes, 332 had predicted functions and 22 were related to the mosquito immune system, based on DNA sequence similarity. Three neighboring (in the sense of the dendrogram) clusters in the optimal clustering, clusters 1,4 and 5, contained 16 genes with predicted function, of which 11 belong to the immunity functional class. This compares favorably with the single bacterial experiment analysis of (1), where a single immune defense cluster was identified, with 9 of its 27 genes of known function were immunity related.

Figure 4 here

The three clusters are therefore of particular interest due to the tight co-regulation of various immune and other genes that are likely to be involved in common defense mechanisms. Indeed, cluster 1 comprises two pattern recognition receptors, GNBPB1 and PGRPLB, belonging to the gram-negative bacteria-binding gene family and the peptidoglycan recognition protein gene family, respectively. GNBPs and PGRPs have been shown to function in the same mechanism implicated in the activation of the Toll immune signaling pathway in *Drosophila melanogaster* (13).

Figure 5 here

Cluster 1 also comprises a leucine rich repeat (LRR) domain containing protein, which belongs to a gene family comprising several putative Toll receptor genes. This LRR-domain transcript may also be implicated in the pathway activated by GNBP and PGRP receptors. The Toll pathway control activation of anti-microbial effectors such as cecropin which is found in cluster 1 (13). Hence, cluster 1 comprises putative pattern recognition receptors, signaling factors and an effector gene that may be part of the same immune response process (Fig. 5). Cluster 4 comprises among other immune genes, a thioester containing protein TEP4 and a LRR domain protein that may function as pattern recognition receptors. Other genes in cluster 4 are two serine proteases, CLIPD1 and ENSANGG00000013355, and a prophenoloxidase gene PPO5. Prophenoloxidases are implicated in melanization defense reactions and are activated

10

by serine protease cascades that, in turn, are triggered by recognition of pathogens by pattern recognition receptors (14). The tight co-regulation of these components may be indicative for functional relations in the activation of melanization reactions (Fig. 5). Cluster 5 comprises an antimicrobial peptide gene gambicin and a thioester containing protein gene TEPIV. Gambicin and TEP1 have been shown to possess anti-*Plasmodium* activity in addition to antimicrobial action in *A. gambiae* (15; 16).

We have demonstrated a successful implementation of Bayesian hierarchical co-clustering on *A. gambiae* immune responsive genes, grouping them into putative functionally-related clusters. Functional relations between the identified genes can now be tested and validated through other experimental approaches. By examining predictive performance through cross-validation studies we have seen how clustering can be improved by joint modelling across different but related experimental conditions.

Sophisticated computational methods (MCMC and approximate EM) enabled us to implement a model that learned about the degree of correlation between the experiments. The algorithm we have proposed also yields a parameter, q in Eq. **3**, that measures the probability that a given gene has correlated response across experiments; of the 159 clusters, 109 (68% and including the immune defense clusters) attributed probability greater than $\frac{1}{2}$ to a model where there was correlation between the experiments.

The method is readily implementable, with the clustering M-step and Monte Carlo E-step of 10,000 Markov chain iterations each taking just over three and four minutes respectively for the data analysed here on a 2GHz processor PC. The 20 iterations of the overall EM algorithm performed proved more than sufficient, with the expected experimental covariance matrix stabilizing fairly well after as few as 5 iterations. C++ code of the full EM algorithm is available on request from the first author.

ACKNOWLEDGMENTS

References

1. Heard, N. A, Holmes, C. C, & Stephens, D. A. (2005, to appear) *J. Amer. Statist. Assoc.*

2. Wakefield, J, Zhou, C, & Self, S. (2003) in *Bayesian Statistics 7*, eds. Bernardo, J. M, Bayarri, M. J, O, B. J, Dawid, A. P, Heckerman, D, Smith, A. F. M, & West, M. (Oxford: Clarendon Press).

3. Ramoni, M, Sebastiani, P, & Kohane, P. R. (2002) *Proc. Nat. Acad. Sci. USA* **99**, 9121–9126.

4. Yeung, K. Y, Fraley, C, Murua, A, Raftery, A. E, & Ruzzo, W. L. (2001) *Bioinformatics* **17**, 977–987.

5. Luan, Y & Li, H. (2003) *Bioinformatics* **19**, 474–482.

6. Lu, X, Zhang, W, Qin, Z. S, Kwast, K. E, & Liu, J. S. (2004) *Nucleic Acids Research* **32**, 447–455.

7. Eisen, M. B, Spellman, P. T, Brown, P. O, & Botstein, D. (1998) *Proc. Nat. Acad. Sci. USA* **95**, 14863–14868.

8. Dimopoulos, G, Christophides, G. K, Meister, S, Schultz, J, White, K. P, & Barillas-Mury, C andKafatos, F. C. (2002) *Proc. Nat. Acad. Sci. USA* **99**, 8814–8819.

9. Dimopoulos, G, Seeley, D, Wolf, A, & Kafatos, F. C. (1998) *The EMBO Journal* **17**, 6115–6123.

10. Christophides, G. K, Zdobnov, E, Barillas-Mury, C, Birney, E, Blandin, S, Blass, C, Brey, P. T, Collins, F. H, Danielli, A, Dimopoulos, G, Hetru, C, Hoa, N. T, Hoffmann, J. A, Kanzok, S. M, Letunic, I, Levashina, E. A, Loukeris, T. G, Lycett, G, Meister, S, Michel, K, Moita, L. F, Muller, H.-M, Osta, M. A, Paskewitz, S. M, Reichhart, J.-M, Rzhetsky, A, Troxler, L, Vernick, K. D, Vlachou, D, Volz, J, von Mering, C, Xu, J, Zheng, L, Bork, P, & Kafatos, F. C. (2002) *Science* **298**, 159–165.

11. Alphey, L, Beard, C. B, Billingsley, P, Coetzee, M, Crisanti, A, Curtis, C, Eggleston, P, Godfray, C, Hemingway, J, Jacobs-Lorena, M, James, A. A, Kafatos, F. C, Mukwaya, L. G,

Paton, M, Powell, J. R, Schneider, W, Scott, T. W, Sina, B, Sinden, R, Sinkins, S, Spielman, A, Toure, Y, & Collins, F. H. (2002) *Science* **298**, 119–121.

12. Kumar, S, Christophides, G, K, Cantera, R, Charles, Band Han, Y. S, Meister, S, Dimopoulos, G, Kafatos, F. C, & Barillas-Mury, C. (2003) *Proc. Nat. Acad. Sci. USA* **100**, 14139–14144.

13. Leclerc, V & Reichart, J. M. (2004) *Immunological Reviews* **198**, 59–71.

14. Cerenius, L & Soderhall, K. (2004) *Immunological Reviews* **198**, 116–126.

15. Blandin, S, Shiao, S, Moita, L, Janse, C, Waters, A, Kafatos, F, & Levashina, E. (2004 Mar 5) *Cell* **116**, 661–70.

16. Vizioli, J, Bulet, P, Hoffmann, J, Kafatos, F, Mller, H, & Dimopoulos, G. (2001 Oct 23) *Proc Natl Acad Sci U S A* **98**, 12630–5.

17. Dimopoulos, G. (2003) *Cell Microbiology* **5**, 3–14.

FIGURE AND TABLE LEGENDS

Fig. 1: *A. gambiae* cell line gene expression profiles at 6 consecutive time points post microbial challenge. Vertically, each block corresponds to one the four bacterial or chemical challenges, in order from bottom to top *Salmonella typhimurium*, *Micrococcus luteus*, *Listeria monocytogenes* and the yeast wall extract Zymosan. Colour scheme has red (green) corresponding to gene up- (down-) regulation. Genes are presented in random order. The width of each assay has been adjusted to the actual time interval between assayed time points, to provide a more realistic temporal dimension to the cluster matrix.

Fig. 2: Histograms of empirical cross correlations, computed by taking the set of gene-by-gene (Pearson) correlation coefficients across all possible pairs of experiments, via equation (1). The bias towards a positive correlation evident in this picture indicates that there evidence for dependence between experiments.

Fig. 3: Hierarchically clustered expression profiles of challenged cell line experimental data. From bottom upwards: *S. typhimurium*, *L. monocytogenes*, *M. luteus*, Zymosan. Hierarchy of cluster of genes and experiments plotted as horizontal and vertical dendrograms respectively. Colour codes are the same as in Fig. 1.

Fig. 4: Three co-regulated gene clusters from Bayesian agglomerative clustering (clusters 1, 4 and 5) that are highly enriched with putative immune genes. Black lines are individual profiles in the cluster, red lines are Bayesian posterior mean estimates of response profiles, blue lines are approximate 95 % pointwise credible intervals for response. Each vertical block of plots corresponds to the gene expression profiles under the four challenges. Genes contained in each cluster are presented in Supplementary Data Table I. $\log_2$ transformed fold regulation is indicated on the vertical axis and assayed time-points are indicated on the horizontal axis.

Fig. 5: The model illustrates the activation of the Toll pathway and melanization melanization reactions, predicted from studies in *D. melanogaster* and other invertebrates (14; 17). The GNBP and PGRP receptors are believed to associate upon recognition of pathogens and activate serine protease carcades that can convert the cytokine-like factor spaetzle to its active form. Spaetzle will activate the Toll pathway which control transcription of effector genes such

as antimicrobial peptides. Cluster 1 comprises the GNBP and PGRP pattern recognition receptors and a Toll-like LRR-domain protein gene, and an anti-microbial peptide gene cecropin. Melanization reactions are catalyzed by phenoloxidases (PO) which are activated by clip-domain serine protease cascades that, in turn, are triggered by pattern recognition receptors upon binding to pathogen surfaces. Cluster 4 comprises two putative pattern recognition receptors (PRR), TEP3 and a LRR-domain protein gene, two clip-domain serine proteases and a prophenoloxidase gene PPO3.

Table 1 Log predictive densities. The first column gives the particular time point left out in the cross-validation. The second and third give the log predictive probabilities through modelling the experiments separately and using co-clustering.

Table 2 Estimated correlation between the experiments. Final conditional posterior mean correlation matrix after 20 EM steps each with 10,000 MCMC iterations for full joint model correlating all four experiments.

SUPPLEMENTARY MATERIAL

SUPPLEMENTARY MATERIAL I: The Regression Modelling Approach

Considering the data in Fig. 1, we wish to capture the behaviour of the relative gene expression y as a function of time t and measurement error. We propose a basis function regression model for the time series. In particular, we model the time series of expression levels for an individual gene g in experiment e as

$$y_{ge} = X_{ge}\beta_{ge} + \sigma_g\varepsilon_{ge}, \quad g = 1,\ldots,N; e = 1,\ldots,E \tag{4}$$

where y_{ge} is a vector of length T_e, X_{ge} is in general a $T_e \times p$ time-dependent design matrix with ith row a p-vector of specified basis functions of time evaluated at the ith of the T_e sampling time points of experiment e; β_{ge} is a p-vector of basis coefficient parameters, σ_g^2 is an error variance and $\{\varepsilon_{ge}\}$ is some standardized error process that we shall model as independent and Gaussian.

Concatenating the gene expression time series across the E experiments for each gene g, $y_g = (y_{g1}',\ldots,y_{gE}')'$, we have

$$y_g = X\beta_g + \sigma_g\varepsilon_g \tag{5}$$

for $X = \oplus_{e=1}^{E}X_{ge}$, where the symbol $\oplus$ represents the *direct sum* of a set of matrices. Thus X is of dimension $T. \times Ep$, with $T. = \sum_{e=1}^{E} T_e$, and $\beta_g = (\beta_{g1}',\ldots,\beta_{gE}')'$ and $\varepsilon_g = (\varepsilon_{g1}',\ldots,\varepsilon_{gE}')'$ are (Ep)-vectors. The model is simply a linear regression with respect to the basis functions of time defining X. The precise form of the basis functions and resulting design matrix X will be considered later.

Now let $\mathcal{C}$ be a partition of the genes dividing them into C groups of sizes $\{N_1,\ldots,N_C\}$, so $\sum_{k=1}^{C} N_k = N$. For the kth set of genes in this partition, let the vector $y^{(k)} = \left(y_{g1}^{(k)'},\ldots,y_{gN_k}^{(k)'}\right)'$ be the concatenated expression profiles from that set. The key assumption underlying our clustering method is that within each set k of the partition, the gene expression levels $y^{(k)}$ follow the regression model in Eq. 5 with a parameter vector β_k and error variance σ_k^2 specific to that group. Since we are assuming that the random error terms $\{\varepsilon\}$ form an i.i.d standard Gaussian sequence, the conditional distribution of the random variable $Y^{(k)}$ is multivariate normal

$$Y^{(k)}|X_k,\beta_k,\sigma_k^2 \sim \mathrm{N}\left(X_k\beta_k, \sigma_k^2 I_{N_kT.}\right) \tag{6}$$

where now X_k, the design matrix of the group, is of size $N_k T. \times p$ and is simply N_k copies of the individual gene design matrix X stacked on top of one another. $I_{N_k T.}$ is the appropriate dimension identity matrix. We can define the scalar $s_k = \sum_{i=1}^{N_k} y_{g_i}^{(k)'} y_{g_i}^{(k)}$ to be the sum of squared responses for the genes in cluster k, and the vector $z_k = X' \sum_{i=1}^{N_k} y_{g_i}^{(k)}$ to be a corresponding sum of the linearly transformed gene responses under X'; we shall see that the pair (s_k, z_k) are sufficient statistics for $y^{(k)}$ under this model. This sufficiency leads to computational savings in the clustering algorithm we present.

Also for each gene g define the linearly transformed response $z_g = y_g$.

The form of X (or X_k) in Eq. **5** and Eq. **6** relates to the specific *basis function representation* used. The class of suitable basis function models for time series is wide and includes Fourier representations, splines, wavelets, and radial basis functions. Basis function representations form some of the most flexible and convenient approaches to nonlinear modelling as, conditional on the basis functions, the model is simply a linear regression in a non-linear design space.

The conditional linear structure allows for many of the standard computational and methodological techniques surrounding linear models to be employed when making inference. This is an essential feature for our application where the dimensionality of the data is large, and computationally efficient procedures are required in order to make inference in reasonable time.

For general discussions of basis function representations and the so-called *Extended Linear Model*, see, for example, (1), (2), (3) and (4).

Bayesian Regression Modelling A Bayesian analysis of the model in Eq. **6** requires specification of a joint prior distribution for (β_k, σ_k^2). A convenient conjugate prior specification $p(\beta_k | \sigma_k^2, V) p(\sigma_k^2)$ is given by

$$
\begin{aligned}
p\left(\beta_k | \sigma_k^2, V\right) &\equiv \mathrm{N}\left(0, \sigma_k^2 V\right) \\
p\left(\sigma_k^2\right) &\equiv \mathrm{IG}\left(\frac{\alpha}{2}, \frac{\gamma}{2}\right)
\end{aligned}
\qquad [7]
$$

The matrix V is $Ep \times Ep$, positive definite and symmetric, α and γ are positive scalars and $\mathrm{IG}(\cdot)$ denotes the inverse gamma distribution. If we use this prior independently for each group of genes k in the partition, standard calculations (see, for example, (3)) show that conditional

17

on the observed data and the matrix V

$$p\left(\beta_k|y^{(k)}, \sigma_k^2, V\right) \equiv \mathrm{N}\left(m_k^*, \sigma_k^2 V_k^*\right)$$

$$p\left(\sigma_k^2|y^{(k)}\right) \equiv \mathrm{IG}\left(\frac{N_k T. + \alpha}{2}, \frac{d_k + \gamma}{2}\right)$$

where

$$V_k^* = \left(X_k'X_k + V^{-1}\right)^{-1}$$

$$m_k^* = V_k^* X_k' y^{(k)}$$

$$d_k = y^{(k)'} y^{(k)} - y^{(k)'} X_k V_k^* X_k' y^{(k)}$$

Noting that $X_k'X_k = N_k X'X$ and $X_k'y^{(k)} = z_k$, we can simplify these expressions to

$$V_k^* = \left(N_k X'X + V^{-1}\right)^{-1}$$

$$m_k^* = V_k^* z_k$$

$$d_k = s_k - z_k' m_k^* \tag{8}$$

Marginal Likelihood for a Single Cluster The critical quantity in our clustering procedure will be the marginal likelihood or prior predictive distribution for each cluster k,

$$p\left(y^{(k)}|V\right) = \int\int p\left(y^{(k)}|\beta_k, \sigma_k^2\right) p\left(\beta_k|\sigma_k^2, V\right) p\left(\sigma_k^2\right) d\beta_k d\sigma_k^2.$$

Combining Eq. **6** and Eq. **7** firstly gives

$$p(y^{(k)}|\sigma_k^2, V) \equiv \mathrm{N}(0, \sigma_k^2(X_k V X_k' + I_{N_k T.})) \tag{9}$$

which, after marginalizing over σ_k, leads to

$$p\left(y^{(k)}|V\right) = \frac{|V_k^*|^{1/2} g\left(N_k T., \alpha, \gamma\right)}{|V|^{1/2} \{d_k + \gamma\}^{(N_k T. + \alpha)/2}} \tag{10}$$

where

$$g\left(N_k T., \alpha, \gamma\right) = \sqrt{\frac{\gamma^\alpha}{\pi^{N_k T.}}} \frac{\Gamma\left(\dfrac{N_k T. + \alpha}{2}\right)}{\Gamma\left(\dfrac{\alpha}{2}\right)}$$

is a normalizing constant independent of the data.

Choice of Design Matrix For the data presented here, the sampling time points are the same across each of the bacterial experiments, so $T_e \equiv T$ across the experiments. We thus for simplicity of exposition we need only consider the design matrix, X_{ge}, for a single gene expression curve in a single experiment, y_{ge}, that appears in Eq. **4**. Note that all the methodology presented here readily extends to the general case of different sampling points for different experiments.

Following (1) we use linear spline basis functions and thus set the $T \times p$ matrix X_{ge} to have row s, $s = 1, \ldots, T$, given by

$$[1, (t_s - t_1)_+, \ldots, (t_s - t_{T-1})_+]$$

where $(\cdot)_+ = \max\{0, \cdot\}$. For further discussion of other possible choices of the basis functions for the design matrix, see (1).

Prior Covariance Modelling We extend the model in (1) by defining a joint prior covariance matrix V to accommodate co-expression between treatments. In (1) V for a single experiment was taken as some multiple of the identity matrix I_p to reflect prior ignorance of the correlations between the spline coefficients within an experiment. In contrast, however, in the case of joint clustering (co-clustering) of multiple parallel experiments, we would expect there to be correlation in the corresponding spline coefficients between experiments in some of the clusters. Therefore to allow for correlation between the experiments we consider a covariance matrix specification of the form

$$V = \Sigma \otimes I_p$$

where Σ is an $E \times E$ symmetric positive definite matrix acting as a between experiment covariance matrix and the symbol $\otimes$ represents the *direct* (or *Kronecker*) *product* of two matrices.

The matrix Σ is unknown and its inverse is considered *a priori* to follow a Wishart distribution (6) with m degrees of freedom,

$$p(\Sigma^{-1}) = \text{Wi}(m, I_E), \tag{11}$$

Typically we choose $m = E + 1$. This formulation for Σ allows for both positive and negative correlations between the experiments. As a consequence of this, up regulation of genes under one experimental condition can be matched with down regulation under another condition.

To recognize that strong experimental correlation may not exist for some of the gene clusters, we propose a two-component mixture prior distribution for V. Defining $D_\Sigma = \text{diag}\{\Sigma_{11}, \ldots, \Sigma_{EE}\}$ to be the decorrelated experimental covariance matrix containing the diagonal elements of Σ, we use

$$p(V) = \begin{cases} \Sigma \otimes I_p, & \text{with probability } q \\ D_\Sigma \otimes I_p, & \text{with probability } 1 - q. \end{cases}$$

It follows that the mixture likelihood for the kth cluster is given by

$$\begin{aligned} p\left(y^{(k)}\big|\Sigma, q\right) &= q\, p\left(y^{(k)}\big|\Sigma \otimes I_p\right) \\ &+ (1 - q)\, p\left(y^{(k)}\big|D_\Sigma \otimes I_p\right) \end{aligned} \quad [\mathbf{12}]$$

with the likelihoods on the right hand side given by Eq. **10**. The mixture weight $q \in (0, 1)$ can also be treated as unknown and, in the absence of any prior information on the expected proportion of clusters exhibiting correlation in expression across the experiments, assigned a default uniform prior distribution, so

$$p(q) = 1, \qquad q \in (0, 1). \quad [\mathbf{13}]$$

It should be emphasized that even under the uncorrelated experiment formulation (D_Σ) we are still co-clustering with genes belonging to common sets across the treatments; only in this case we are not making any assumptions about the similarity of gene regulation patterns across the different treatments. Also note that this model can be easily extended to a more general mixture model with an arbitrary number of components when we anticipate multiple modes of correlation between the experiments, characterized by a sequence of Σ matrices and a Dirichlet distributed vector of weights.

Prior and marginal likelihood on clustering Assuming independence between clusters, the marginal likelihood for a clustering of the data is

$$p(y|\mathcal{C}, \Sigma, q) = \prod_{k=1}^{C} p\left(y^{(k)}|\Sigma, q\right) \quad [\mathbf{14}]$$

where the likelihood terms on the right hand side are given by Eq. **12**.

A Bayesian specification of a clustering regression model also requires a prior model for the clustering. Assuming exchangeability in the assignment of genes to clusters, it is sufficient to specify prior distributions for the number of clusters C and the cluster sizes $N_1, \ldots, N_C$ to satisfy this requirement. We use a default specification placing a uniform prior on C over the range $\{1, 2, \ldots, N\}$ and for the allocations to those C clusters, a Multinomial-Dirichlet prior with uniform settings. This leads to the prior model

$$p(\mathcal{C}) = \frac{(C-1)!N_1!N_2!\ldots N_C!}{N(N+C-1)!} \qquad [\mathbf{15}]$$

Note that a uniform prior distribution on the number of clusters is unrealistic since, for example, we are unlikely to believe that there are as many clusters as we have genes on the arrays. However, the resulting posterior analysis is robust to the choice of prior on C, with the expression data controlling the number of clusters.

Posterior distribution We have specified a likelihood model for the data given the parameter triple $(\mathcal{C}, \Sigma, q)$ (Eq. **14**) and independent priors for those parameters through $p(\Sigma)$, $p(q)$ and $p(\mathcal{C})$ (Eq. **11**, Eq. **13**, Eq. **15**). These elements can be combined through Bayes Theorem to give the posterior distribution of the parameters

$$p(\mathcal{C}, \Sigma, q|y) = \frac{p(\mathcal{C})p(\Sigma)p(q)p(y|\mathcal{C}, \Sigma, q)}{\sum_{\mathcal{C}} p(\mathcal{C}) \int_{\Sigma, q} p(\Sigma)p(q)p(y|\mathcal{C}, \Sigma, q)d\Sigma dq} \qquad [\mathbf{16}]$$

For all but the most trivially small clustering problems the summation part of the denominator alone prohibits analytic computation of Eq. **16** due to the large number of possible partitions. However, since this denominator is a normalizing constant of the posterior invariant to $(\mathcal{C}, \Sigma, q)$, then both for our potential function for agglomerative clustering and for MCMC exploration of the covariance parameter space we can simply work with the numerator

$$\pi(\mathcal{C}, \Sigma, q) = p(\mathcal{C})p(\Sigma)p(q)p(y|\Sigma, q, \mathcal{C}) \qquad [\mathbf{17}]$$

which is the joint distribution of the parameters and all the data, and equal to the posterior distribution of $(\mathcal{C}, \Sigma, q)$ up to proportionality.

SUPPLEMENTARY MATERIAL II: EM and Clustering Algorithms

Ideally we would like to perform full Bayesian inference on the joint parameter space $(\mathcal{C}, \Sigma, q)$. However, for the large data sets arising from microarray experiments standard reversible jump Markov Chain Monte Carlo (MCMC) methods for moving between different numbers of clusters in $\mathcal{C}$ are rendered ineffective. Methods we have developed to work around the low acceptance problems of such schemes are computationally intensive and will appear separately in another paper. Furthermore, it should be noted that here our aim is to identify a small group of candidate optimal clustering configurations to report to the experimental biologists rather than a more full exploration of the space. This paper is more concerned with learning about between experiment covariances, and thus MCMC will be performed only on (Σ, q) conditional on a clustering $\mathcal{C}$.

However, the distributions of (Σ, q) and $\mathcal{C}$ are very much interrelated and thus we will want to be able to continue to optimize over $\mathcal{C}$ in some way whilst we continue to learn about (Σ, q). The solution we present is to use an approximation to the Expectation-Maximization (EM) algorithm (5), where the expectation step is performed on (Σ, q) and the maximization step on $\mathcal{C}$.

So in each iteration of the algorithm, if our current optimal clustering is $\mathcal{C}^*$ we would wish to maximize

$$\mathrm{E}_{p(\Sigma, q | \mathcal{C}^*, y)}[\log\{p(\Sigma, q, y | \mathcal{C})\}] + \log\{p(\mathcal{C})\} \qquad [\mathbf{18}]$$

with respect to $\mathcal{C}$. As mentioned earlier, for large N full exploration over the space of all clusterings for a maximum is not possible. For the maximization step we thus perform agglomerative clustering as in (1), with the revised algorithm given below, to find a local optimum.

For a given covariance specification (Σ, q), a simple agglomerative clustering algorithm can proceed as follows:

Step 1: Start with $C = N$ clusters, each cluster containing the expression levels for one gene. Calculate the marginal posterior unnormalized probability kernel π_N in Eq. **17**.

Step 2: Let $\mathcal{C}$ be the current clustering and for each pair of clusters k, l, let $\mathcal{C}^{(kl)}$ represent the hypothetical clustering that would result from their merger and $y^{(kl)}$ the corresponding

concatenated gene expression profile vectors. We calculate the multiplicative increase in marginal posterior that would be gained by merging the two clusters to obtain an inter-cluster closeness

$$c_{kl} = c_{lk} = \frac{p\left(\mathcal{C}^{(kl)}\right) p\left(y^{(kl)}|\Sigma, q\right)}{p(\mathcal{C}) p\left(y^{(k)}|\Sigma, q\right) p\left(y^{(l)}|\Sigma, q\right)} \qquad [19]$$

which follows from the prior on $\mathcal{C}$ (Eq. **15**) and where $p\left(y^{(\cdot)}|\Sigma, q\right)$ is given by Eq. **12** ($N(N-1)/2$ calculations).

Step 3: For each cluster k, identify the closest other cluster according to the metric of Eq. **19** and the corresponding maximum closeness value

$$k' = \arg\max_l c_{kl}, \quad c_k = c_{kk'}.$$

Step 4: Find the cluster $\hat{k}$ with largest c_k value, and merge with cluster $\hat{k}'$ to form a new cluster $\hat{k}$. Set $C = C - 1$ and relabel the other remaining clusters accordingly. Calculate the revised marginal unnormalized posterior kernel value

$$\pi_C = c_{\hat{k}\hat{k}'}\pi_{C+1}.$$

Step 5: For each cluster $l \neq \hat{k}$, calculate the closeness to cluster $\hat{k}$, $c_{\hat{k}l}$ (C calculations), and identify the new nearest cluster $\hat{k}'$.

Step 6: For each cluster $l \neq \hat{k}$, update the stored nearest cluster l'; unless the stored cluster l' was just merged, we only need to check the value of c_l against $c_{\hat{k}l}$.

Step 7: Repeat *Steps 4-6* until $C = 1$.

Step 8: Looking back over the clusterings visited, find the number of clusters C in the hierarchy maximizing the posterior distribution, $\arg\max_C \pi_C$. This is our optimal clustering $\mathcal{C}^*_{\Sigma,q}$.

Computational Efficiency A principal feature of data from microarray-based gene expression studies is the dimensionality, as the technology enables thousands of gene expression measurements to be taken simultaneously and in the case of gene expression profiling this is repeated

at a series of time points. Therefore when constructing statistical methods to analyze these types of data, it is crucial to examine the implications of this dimensionality on the feasibility of implementation.

Performing the calculations in Step 2 when implementing the above algorithm can be made very efficient. By always storing the current marginal likelihoods of each cluster in the current configuration, the only work to be done in obtaining Eq. **19** is calculating $p\left(y^{(kl)}|V\right)$ for each possible V in the mixture prior. Evaluating this quantity via Eq. **10** depends upon calculating the new posterior quantities $\{V_{kl}^*, m_{kl}^*, d_{kl}\}$ of Eq. **8**, which we now consider each in turn: Firstly, for each V in the mixture, V_k^* can only take one of N values, and hence these matrices and their determinants only need to be calculated once, stored and then looked up when needed; m_{kl}^* relies on V_{kl}^* and $z^{(kl)}$, the latter of which is simply $z^{(k)} + z^{(l)}$; and d_{kl} additionally requires the quantity s_{kl}, which again is just $s_k + s_l$. So to minimize the amount of computation required we simply need to store the quantities $\left\{N_k, s_k, z^{(k)}, p\left(y^{(k)}|\Sigma\right)\right\}$ for each cluster k in the current configuration, and the V^* matrices and their determinants when we calculate them (and usually we will require much fewer than all N possibilities for each choice of V).

MCMC and the EM Algorithm

Analytic evaluation of the expectation Eq. **18** for a particular clustering $\mathcal{C}$ is also not possible. Thus we resort to MCMC sampling from the distribution $p(\Sigma, q|\mathcal{C}^*, y)$, obtaining a sample $(\Sigma, q)^{(1)}, \ldots, (\Sigma, q)^{(M)}$ for a relatively large integer M. More specifically, we perform a Metropolis Hastings random walk ($M = 10,000$ iterations seemed to suffice for the data presented here) on both q and the continuous valued square root matrix W of Σ to obtain an approximate sample from the conditional posterior distribution of (Σ, q)

The standard Monte Carlo estimate of the expectation term of Eq. **18** would be given by

$$\mathrm{E}_{p(\Sigma,q|\mathcal{C}^*,y)}[\log\{p(\Sigma, q, y|\mathcal{C})\}] \approx \frac{1}{M}\sum_{i=1}^{M}\log\{p(\Sigma^{(i)}, q^{(i)}, y|\mathcal{C})\} \qquad [20]$$

and in theory this could indeed be used as the potential function for evaluating a cluster merger in the agglomerative clustering algorithm above. However, in practice such a procedure would increase the computational time by a factor of M. So instead we approximate Eq. **20** by plugging

in the sample mean values of the parameters,

$$\overline{\Sigma} = \frac{1}{M} \sum_{i=1}^{M} \Sigma^{(i)}, \qquad \overline{q} = \frac{1}{M} \sum_{i=1}^{M} q^{(i)}$$

(note that $\overline{\Sigma}$ is of course also symmetric and positive definite) giving

$$\mathrm{E}_{p(\Sigma, q | \mathcal{C}^*, y)}[\log\{p(\Sigma, q, y | \mathcal{C})\}] \approx \log\left\{p\left(\overline{\Sigma}, \overline{q}, y | \mathcal{C}\right)\right\}. \qquad [\mathbf{21}]$$

Hence we seek to maximize

$$\log\left\{p\left(\overline{\Sigma}, \overline{q}, y | \mathcal{C}\right)\right\} + \log\{p(\mathcal{C})\} = \log\{\pi(\mathcal{C}, \overline{\Sigma}, \overline{q})\}$$

with respect to $\mathcal{C}$. So we simply use the agglomerative clustering algorithm above with the pair $(\overline{\Sigma}, \overline{q})$ acting as our new experimental covariance matrix Σ and mixture component weight q.

The full EM algorithm can be summarised as followed:

Step 1: Initialise $\Sigma = E[\Sigma] = I_E$ so the experiments are initially assumed uncorrelated, and $q = E[q] = \frac{1}{2}$.

Step 2: Run the agglomerative clustering algorithm to find the optimal hierarchical clustering $\mathcal{C}^*_{\Sigma, q}$.

Step 3: Given the clustering $\mathcal{C}^*_{\Sigma, q}$, perform M iterations of a Metropolis Hastings random walk on W and q to obtain an approximate sample $(\Sigma, q)^{(1)}, \ldots, (\Sigma, q)^{(M)}$ from the conditional posterior of (Σ, q) given $\mathcal{C}$ (given up to proportionality by Eq. $\mathbf{17}$).

Step 4: Set $\Sigma = \frac{1}{M} \sum_{i=1}^{M} \Sigma^{(i)} \approx E[\Sigma | \mathcal{C}^*_{\Sigma, q}, y]$ and $q = \frac{1}{M} \sum_{i=1}^{M} q^{(i)} \approx E[q | \mathcal{C}^*_{\Sigma, q}, y]$.

Step 5: Repeat *Steps 2-4* until $E[\Sigma | y]$ and $E[q | y]$ are deemed to have sufficiently stabilized, then finally repeat step 2 again and take this covariance matrix, component weight and clustering triple as our optimal clustering model.

References

1. Vidakovic, B. (1999) *Statistical Modelling by Wavelets.* (John Wiley).

2. Schimek, M. G, ed. (2000) *Smoothing and Regression: Approaches, Computation and Application.* (John Wiley).

3. Denison, D. G. T, Holmes, C. C, Mallick, B. K, & Smith, A. F. M. (2002) *Bayesian Methods for Nonlinear Classification and Regression.* (Chichester: Wiley).

4. Hansen, M & Kooperberg, C. (2002) *Statistical Science* **17**, 2–51.

5. Dempster, A. P, Laird, N. M, & Rubin, D. B. (1977) *J. Roy. Statist. Soc. B* **39**, 1–38.

6. Press, S. J. (1972)*Applied multivariate analysis* (New York: Holt, Rinehart and Winston).

FIGURES
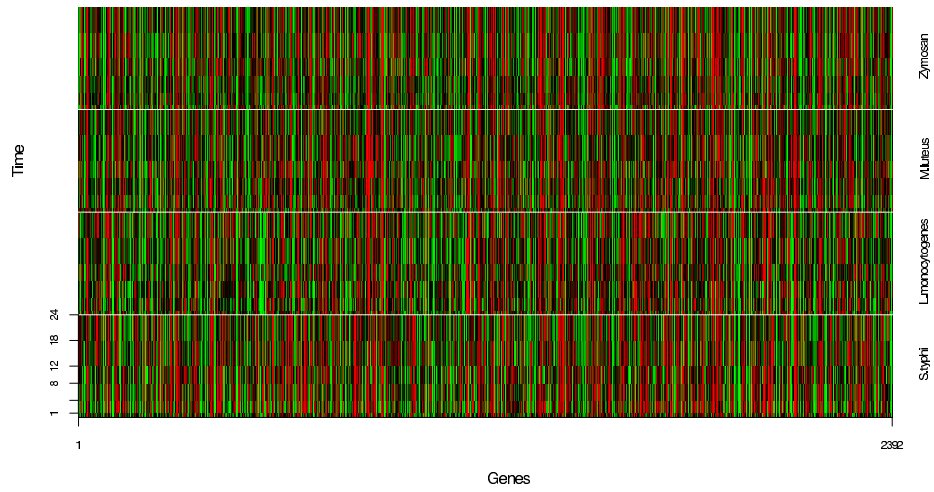
Figure 1:
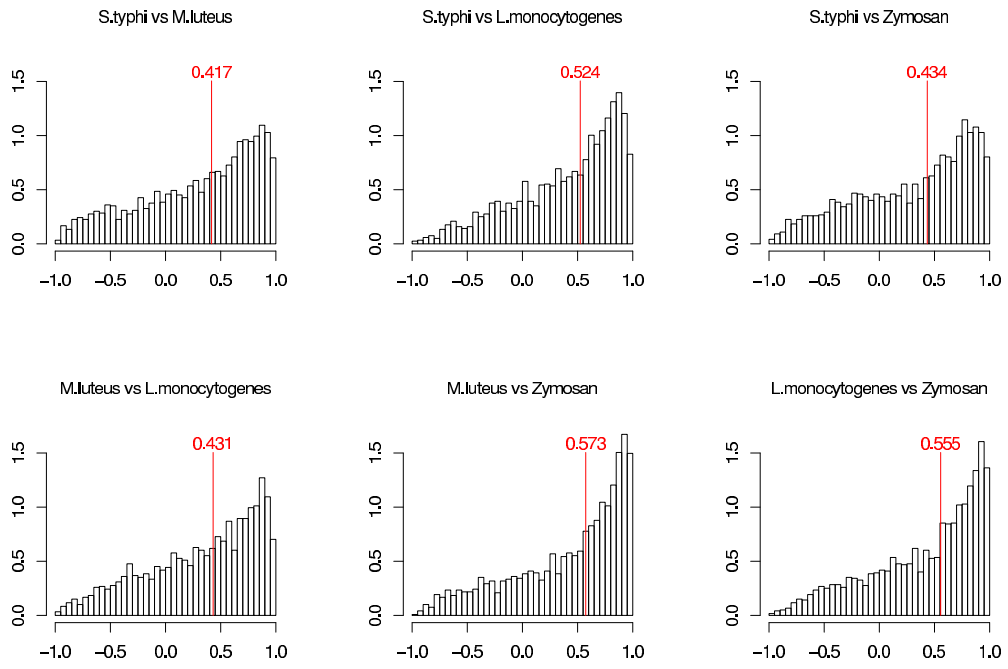
Figure 2:
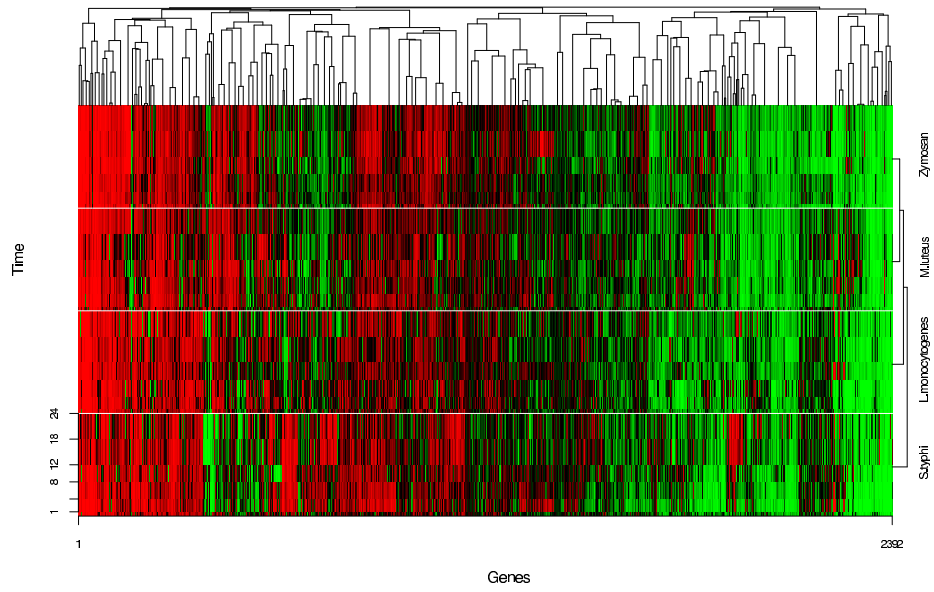
Figure 3:

Figure 4:

Figure 5: