AN EMPIRICAL BAYES APPROACH TO ESTIMATING DYNAMIC MODELS OF CO-REGULATED GENE EXPRESSION

SARA VENKATRAMAN[1], SUMANTA BASU[1], ANDREW G. CLARK[2], SOFIE DELBARE[1,2], MYUNG HEE LEE[3], MARTIN T. WELLS[1]

[1]Department of Statistics and Data Science, Cornell University
[2]Department of Molecular Biology and Genetics, Cornell University
[3]Center for Global Health, Department of Medicine, Weill Cornell Medical College

Abstract. Time-course gene expression datasets provide insight into the dynamics of complex biological processes, such as immune response and organ development. It is of interest to identify genes with similar temporal expression patterns because such genes are often biologically related. However, this task is challenging due to the high dimensionality of such datasets and the nonlinearity of gene expression time dynamics. We propose an empirical Bayes approach to estimating ordinary differential equation (ODE) models of gene expression, from which we derive similarity metrics that can be used to identify groups of genes with co-moving or time-delayed expression patterns. These metrics, which we call the Bayesian lead-lag R^2 values, can be used to construct clusters or networks of functionally-related genes. A key feature of this method is that it leverages biological databases that document known interactions between genes. This information is automatically used to define informative prior distributions on the ODE model's parameters. We then derive data-driven shrinkage parameters from Stein's unbiased risk estimate that optimally balance the ODE model's fit to both the data and external biological information. Using real gene expression data, we demonstrate that our biologically-informed similarity metrics allow us to recover sparse, interpretable gene networks. These networks reveal new insights about the dynamics of biological systems.

Keywords: Bayesian regression, ordinary differential equations, dimension reduction, normal-inverse gamma model, Zellner's g-prior, shrinkage estimation, biological dynamical systems, biological database

1 Introduction

The time dynamics of complex biological processes, such as immune response, organ development, and disease progression, provide insight into how the systems regulating these processes interact with one another. Experimental evidence has revealed that groups of genes exhibiting similar patterns in their expressions over time often share several biological characteristics [Eisen et al., 1998]. For instance, such genes may be co-regulated by the same *transcription factors* [Tavazoie et al., 1999]: proteins that directly control the expression of genes, which ultimately contributes to changes in cellular function. Clustering genes based on their time dynamics can therefore reveal networks of genes that comprise interconnected biological systems, as well as generate testable hypotheses about the functional roles of orphan genes that exhibit similar expression patterns to ones that are better understood.

Technological advances such as whole-genome RNA sequencing have made it possible to study biological processes through the generation of time-course gene expression datasets, which typically consist of gene expression measurements for thousands of genes at a few (typically 5-20) time points. However, identifying genes with similar time dynamics in such datasets is challenging due to three main factors. First, the time dynamics of gene expression are highly nonlinear in nature. Second, these datasets are high dimensional, consisting of several thousands of genes and small sample sizes (number of time points). Finally, groups of interacting genes may exhibit lead-lag or inverted expression patterns [Qian et al., 2001] due to delayed activation of underlying transcription factors. Measuring the "similarity" between the overall shapes of the expression patterns of two genes is therefore a non-trivial task.

Ordinary differential equations (ODEs) or discrete time difference equations have been successfully used to model the nonlinear expression patterns of a small number of genes [D'haeseleer et al., 1999, Chen et al., 1999, De Jong, 2002, Bansal et al., 2006, Polynikis et al., 2009]. It is possible to derive similarity metrics for the time dynamics of two genes from such ODEs, thus enabling putative identification of co-regulated genes and the reconstruction of gene regulatory networks [Farina et al., 2007, Farina et al., 2008, Wu et al., 2019]. The approach proposed in [Farina et al., 2007] in particular allows explicit modeling of lead-lag as well as contemporaneous associations between gene expression trajectories, so we use it as a starting point in our analysis of the time dynamics of gene expression.

Because gene expression datasets contain thousands of genes, the potentially large number of gene pairs receiving high similarity scores may be prohibitive to further analysis of novel gene-gene interactions. In this study, we consider addressing this challenge by leveraging external sources of biological information that document known associations amongst genes. Incorporating such information into similarity calculations can encourage gene pairs with known biological associations to receive higher similarity scores, while filtering away those known to be unrelated. Possible sources include the vast public and commercial curated databases that assign genes to functional groups. For instance, Gene Ontology (GO) terms are keywords that describe a gene's molecular function, role in a biological process, or cellular localization [Ashburner et al., 2000]. Other curated databases include KEGG [Kanehisa & Goto, 2000], Reactome [Fabregat et al., 2018], BioCyc [Karp et al., 2019], and STRING [Szklarczyk et al., 2019].

Several previous studies have considered integrating biological knowledge into statistical measures of genetic association. Some works take a Bayesian approach to this task by using external data sources to determine prior distributions over genes or proteins that influence a biological response [Li & Zhang, 2010, Stingo et al., 2011, Hill et al., 2012, Lo et al., 2012, Peng et al., 2013]. Other works develop biologically-informed regularization terms in graph-regularized methods for reconstructing gene networks [Zhang et al., 2013, Li & Jackson, 2015]. [Nepomuceno et al., 2015] propose an algorithm for clustering genes based on both gene expression data and gene ontology annotations. However, less attention has been given to using both data and prior biological knowledge to model the complex time dynamics of gene expression, e.g. with ODEs.

In this work, our objective is to identify biologically-meaningful similarities in the time dynamics of a large number of genes. Our methodological contribution is a Bayesian approach to estimating ODE models of gene expression that leverages external databases of genetic informa-

tion. Specifically, previously-curated biological networks are used to define prior distributions on the parameters of an ODE that describes a gene's temporal expression as a function of another potentially co-regulated gene's expression. We propose data-driven estimators for hyperparameters that balance the posterior model's fit to the data and to prior information, making our approach an *empirical* Bayes method. From the posterior model, we derive "Bayesian lead-lag R^2" values, which are measures of similarity between the temporal expression patterns of two genes. As indicated by the name, these metrics are based on the familiar R^2 statistic (the coefficient of determination) and are simple and fast to compute for all $\binom{N}{2}$ gene pairs, where N is the number of genes under study. We validate our approach in detail on real time-course gene expression data collected by [Schlamp et al., 2020]. Our similarity computations yield clusters and networks of genes with clear biological interpretations. These clusters include less well-understood genes that exhibit strikingly similar expression patterns to more well-studied ones, which allows us to conjecture new relationships between them.

The remainder of this paper is organized as follows. Section 2 describes the ODE model of gene expression we adopt and introduces the similarity measures we use as indicators of co-regulation. Section 3 details our empirical Bayes method for fitting the ODE model, including how a prior network of genes can be translated into prior distributions for this model's parameters. Section 4 discusses the application of our method to real gene expression data, and the clusters and networks of genes that we can construct as a result. We conclude with a brief discussion in Section 5.

2 Dynamic models of gene expression

2.1 Model derivation

We consider an ODE model of gene expression proposed by [Farina et al., 2007]. Let $m_A(t)$ denote the expression of some gene A at time t, measured for instance as the $\log_2$-fold change in mRNA levels relative to time 0. The model assumes the rate of change in gene A's expression is given by some regulatory signal $p(t)$:

$$\frac{\mathrm{d}m_A(t)}{\mathrm{d}t} = p(t) - \kappa_A m_A(t) + \eta_A,$$

where κ_A denotes the mRNA decay rate for gene A, and η_A denotes natural and experimental noise. This model can be naturally extended to consider two genes A and B that might be associated with one another, i.e. that are governed by the same underlying signal $p(t)$, yielding a pair of coupled differential equations:

$$\frac{\mathrm{d}m_A(t)}{\mathrm{d}t} = \alpha_A p(t) + \beta_A - \kappa_A m_A(t) + \eta_A, \tag{1}$$

$$\frac{\mathrm{d}m_B(t)}{\mathrm{d}t} = \alpha_B p(t) + \beta_B - \kappa_B m_B(t) + \eta_B. \tag{2}$$

The common signal $p(t)$ accounts for the effect of one or more transcription factors that potentially regulate both genes A and B. The coefficients α_A and α_B measure the strength of $p(t)$ in the expression patterns of genes A and B, respectively. β_A and β_B are affine coefficients allowing $m_A(t)$ and $m_B(t)$ to exhibit linear time trends.

We can now obtain a model of gene A's expression in terms of gene B's expression by first rearranging (2) to isolate the common regulatory signal $p(t)$:

$$p(t) = \frac{1}{\alpha_B}\left(\frac{\mathrm{d}m_B(t)}{\mathrm{d}t} - \beta_B + \kappa_B m_B(t) - \eta_B\right). \tag{3}$$

Substituting (3) into (1) yields

$$\frac{\mathrm{d}m_A(t)}{\mathrm{d}t} = \frac{\alpha_A}{\alpha_B}\frac{\mathrm{d}m_B(t)}{\mathrm{d}t} + \frac{\alpha_A\kappa_B}{\alpha_B}m_B(t) - \kappa_A m_A(t) + \beta_A - \frac{\alpha_A\beta_B}{\alpha_B} + \eta_A - \frac{\alpha_A\eta_B}{\alpha_B}.$$

Integrating from 0 to t, we obtain:

$$\begin{aligned} m_A(t) &= \frac{\alpha_A}{\alpha_B}m_B(t) + \frac{\alpha_A\kappa_B}{\alpha_B}\int_0^t m_B(s)\,\mathrm{d}s - \kappa_A\int_0^t m_A(s)\,\mathrm{d}s + \left(\beta_A - \frac{\alpha_A\beta_B}{\alpha_B}\right)t \\ &\quad + \int_0^t\left(\eta_A - \frac{\alpha_A\eta_B}{\alpha_B}\right)\mathrm{d}s + c \\ &= c_1 m_B(t) + c_2\int_0^t m_B(s)\,\mathrm{d}s + c_3\int_0^t m_A(s)\,\mathrm{d}s + c_4 t + c_5, \end{aligned} \tag{4}$$

where the parameters c_k are defined as

$$c_1 = \frac{\alpha_A}{\alpha_B}, \qquad c_2 = \frac{\alpha_A\kappa_B}{\alpha_B}, \qquad c_3 = -\kappa_A,$$
$$c_4 = \beta_A - \frac{\alpha_A\beta_B}{\alpha_B}, \qquad c_5 = \int_0^t\left(\eta_A - \frac{\alpha_A\eta_B}{\alpha_B}\right)\mathrm{d}s + c.$$

Observe that (4) is linear in the parameters c_k. Given measurements $\{m_A(t_1), ..., m_A(t_n)\}$ and $\{m_B(t_1), ..., m_B(t_n)\}$ of the expression levels of genes A and B at time points $t_1, ..., t_n$, we can then express (4) as the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\mathbf{Y} = \begin{bmatrix} m_A(t_1) \\ m_A(t_2) \\ \cdots \\ m_A(t_n) \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} m_B(t_1) & \int_0^{t_1} m_B(s)\,\mathrm{d}s & \int_0^{t_1} m_A(s)\,\mathrm{d}s & t_1 & 1 \\ m_B(t_2) & \int_0^{t_2} m_B(s)\,\mathrm{d}s & \int_0^{t_2} m_A(s)\,\mathrm{d}s & t_2 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ m_B(t_n) & \int_0^{t_n} m_B(s)\,\mathrm{d}s & \int_0^{t_n} m_A(s)\,\mathrm{d}s & t_n & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix}, \tag{5}$$

with the standard assumption that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. Although only samples from the functions $m_A(t)$ and $m_B(t)$ are given, we can estimate the integral entries of the second and third columns of $\mathbf{X}$ by numerically integrating spline or polynomial interpolants fitted to these samples.

$\mathbf{X}$ and $\mathbf{Y}$ can be constructed from the data $\{(m_A(t_i), m_B(t_i))\}_{i=1}^n$, so fitting the model (4) to this data amounts to estimating the unknown vector $\boldsymbol{\beta}$. The ordinary least-squares (OLS) estimator for $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Then, the amount of variation in gene A's expression that is captured by the estimated linear model is expressed as the familiar R^2 value

$$R^2 = \frac{\|\mathbf{X}\hat{\boldsymbol{\beta}} - \overline{Y}\mathbb{1}_n\|^2}{\|\mathbf{Y} - \overline{Y}\mathbb{1}_n\|^2}, \tag{6}$$

where $\overline{Y} = (1/n)\mathbf{Y}^T\mathbb{1}_n$ and $\mathbb{1}_n$ denotes the $n \times 1$ vector of ones. Adopting the terminology in [Farina et al., 2008], we refer to the R^2 in (6) as the *lead-lag R^2 between genes A and B*. A high lead-lag R^2 between genes A and B may indicate that the two genes are co-regulated in time by common transcription factors, or are at least associated with one another in some way. The term "lead-lag" comes from the lead-lag compensator in control theory. In this context, a "lead-lag relationship" between genes refers to the presence of a common regulatory signal (input) that, in conjunction with the process of mRNA decay, modulates the expression of genes with the same biological function (output).

2.2 Our contributions towards measuring the similarity of gene time dynamics

Our primary contribution in this work is a Bayesian method for incorporating prior biological information into similarity measures between the temporal expressions of two genes, such as the lead-lag R^2. We can motivate this via Figure 1, which shows genes from two functional groups, known as pathways, namely circadian rhythms and immune response. Within each group, we would expect to see high pairwise lead-lag R^2 values. The data comes from an experiment on fruit flies by [Schlamp et al., 2020] that aimed to profile the dynamics of genes involved in or affected by immune response immediately following an infection; more details on this dataset can be found in Section 4.1. Incorporating biological information, such as pathway membership or protein-protein interaction networks, into the lead-lag R^2 enables us to encourage genes in the same pathways to receive higher pairwise similarity scores. Importantly, we can also suggest possible pathway membership for previously uncharacterized genes.

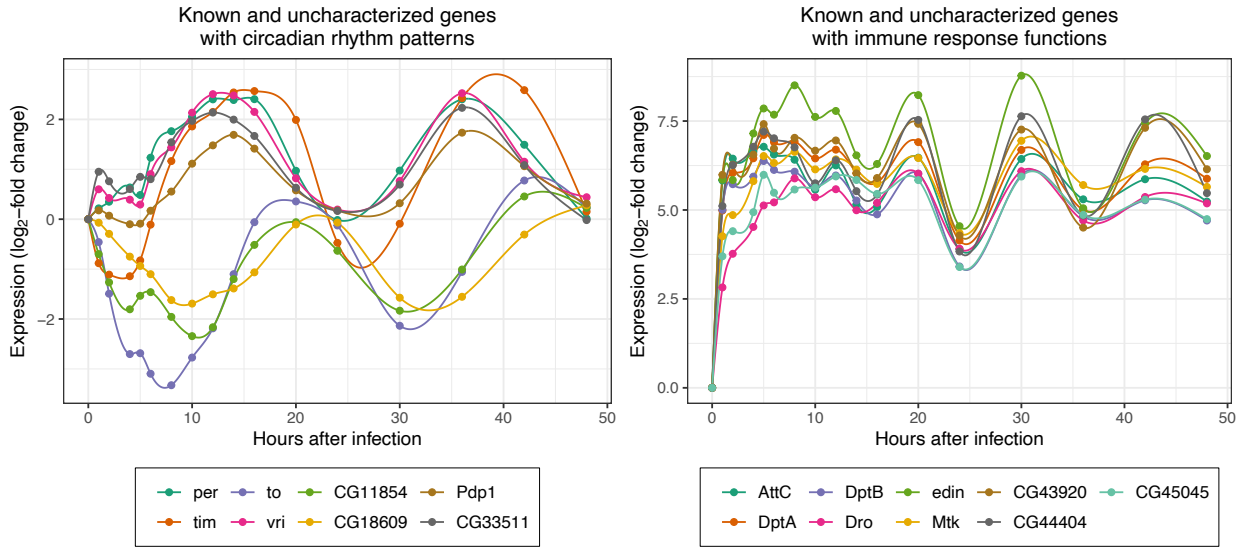

Figure 1: *Left: The uncharacterized gene CG33511 exhibits similar time dynamics to known circadian rhythm genes with 24-hour cyclic temporal expressions. Right: Uncharacterized genes CG43920 and CG45045 exhibit similar temporal expression patterns to known immune response genes, which are up-regulated in response to infection.*

In our Bayesian method, external biological information is used to determine the locations of normal prior distributions on the parameters $c_1, ..., c_5$ in the model (4). Upon obtaining the posterior estimates of these parameters, we can recompute the lead-lag R^2 to obtain what we call

the *Bayesian lead-lag* R^2, denoted LLR^2. In doing so, we observe a desirable shrinkage effect in the distribution of the Bayesian lead-lag R^2 values that pares down the number of significant associations. The hierarchical model and its hyperparameters are detailed in Section 3.

Our second contribution addresses the fact that it is quite possible for two genes to have a spuriously high lead-lag R^2. For example, if gene A's expression levels increase monotonically with time, the response vector $\mathbf{Y}$ in (5) will be highly correlated with the third and fourth columns of $\mathbf{X}$, which consist of gene A's estimated time integrals and the n time points. The lead-lag R^2 between genes A and B will be high in this case, but not because the genes have similar time dynamics or a biological relationship. Such gene pairs are referred to as "false positives". Figure 2 displays the temporal expressions of a few of these false positive pairs.

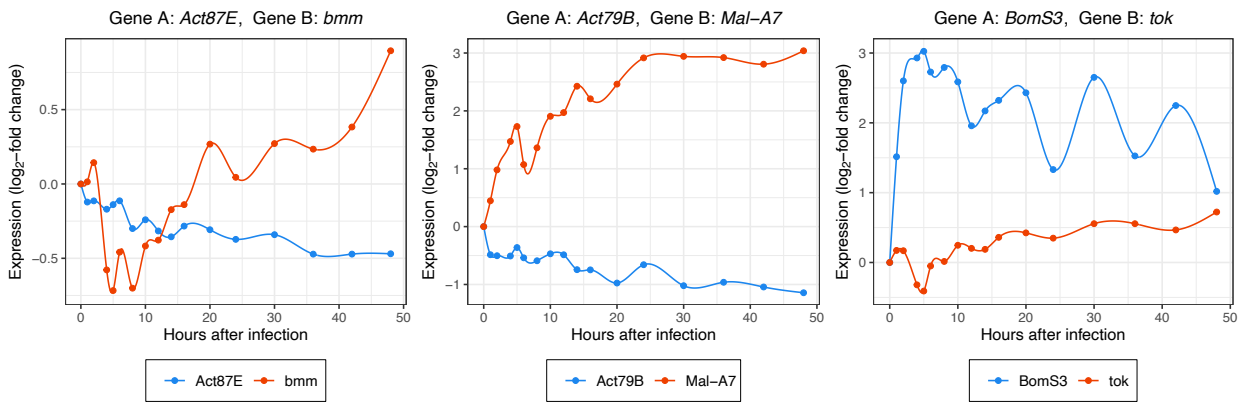

Figure 2: *Left: Lead-lag* $R^2 = 0.978$. *Middle: Lead-lag* $R^2 = 0.973$. *Right: Lead-lag* $R^2 = 0.956$. *For all three gene pairs, the first listed gene (blue) is treated as gene A and is regressed on gene B (red) when fitting the model (4). Spline interpolants were fit through the observed points, which are indicated by the solid dots. Functional annotations for these genes in Flybase [Larkin et al., 2020] do not suggest a clear link within each pair.*

To separate false positives from true positives, meaning high lead-lag R^2 values that are biologically-meaningful, we consider the model (4) in addition to two sub-models that capture different sources of variation. The first sub-model describes the temporal expression of gene A as a function only of a second, potentially co-regulated gene B. The second sub-model is "autoregressive" in the sense that it describes gene A's expression only in terms of its own past and linear time trends. Using the notation of the model (4), these two sub-models are specifically of the form

$$m_A(t) = c_1 m_B(t) + c_2 \int_0^t m_B(s)\,\mathrm{d}s + c_5, \qquad \text{(Sub-model 1)} \tag{7}$$

$$m_A(t) = c_3 \int_0^t m_A(s)\,\mathrm{d}s + c_4 t + c_5. \qquad \text{(Sub-model 2)} \tag{8}$$

We again apply our Bayesian approach to fitting these two sub-models and compute new variants of the Bayesian lead-lag R^2 from each, denoted LLR^2_{other} and LLR^2_{own} respectively. The LLR^2_{other} value indicates the amount of variation in gene A's temporal expression that can be explained by the dynamics of an*other* gene B. LLR^2_{own} indicates the amount of variation in gene A that is explained by its *own* past, via its time integrals, and linear time trends. Recalling that LLR^2 denotes the Bayesian lead-lag R^2 derived from (4), we can view LLR^2 − LLR^2_{own} as the

amount of *additional* variation in gene A's temporal dynamics that we can explain by considering gene B, on top of the variation captured by the second sub-model. Intuitively, if $\text{LL}R^2$ is large, then a large value of $\text{LL}R^2 - \text{LL}R^2_{\text{own}}$ suggests that the $\text{LL}R^2$ value is unlikely to mean a false positive relationship between the two genes. In Section 4.4, we demonstrate empirically that using $\text{LL}R^2_{\text{other}}$ and $\text{LL}R^2 - \text{LL}R^2_{\text{own}}$ together can help identify gene pairs with similar time dynamics.

3 Empirical Bayes estimation of dynamic gene expression models

In this section, we propose our empirical Bayes approach to deriving biologically-informed similarity metrics between genes for clustering and network analysis. The components of our method are: 1) encoding external biological information into a prior adjacency matrix, 2) defining a normal-inverse gamma prior, specifically Zellner's g-prior, on the parameters of the ODE model of gene expression, 3) optimally selecting the hyperparameter g in Zellner's g-prior, and 4) calculating the Bayesian lead-lag R^2. Note that parts 2-4 lead to the computation of the Bayesian lead-lag R^2 for a single gene pair; a summary of the full algorithm for all pairs is can be found in Appendix A.2.

3.1 Part 1: Leveraging biological priors

A key component of our method is placing biologically-informed prior distributions on the parameters of an ODE model of gene expression, from which we derive the Bayesian lead-lag R^2. There exist numerous databases that extensively document known or predicted interactions between genes as well as their functional roles. For instance, Gene Ontology (GO) terms are keywords that describe a gene's molecular function, role in a biological process (e.g., "lipid metabolism"), or cellular localization (e.g., "nucleus") [Ashburner et al., 2000]. Semantic similarity methods, such as the R package `GOSemSim` [Yu et al., 2010], have been developed to determine how similar genes are based on their associated GO terms. Other curated databases that similarly assign genes to pathways include KEGG [Kanehisa & Goto, 2000], Reactome [Fabregat et al., 2018], and BioCyc [Karp et al., 2019]. Alternatively, the STRING database [Szklarczyk et al., 2019] aggregates multiple sources of information to generate a more holistic measurement of the association between two genes. For each pair of genes in an organism, STRING provides a score between 0 and 1 indicating how likely the two genes' encoded proteins are to interact physically based on experimental evidence, belong to the same pathways according to the aforementioned databases, be conserved across species, or be mentioned in the same studies.

Regardless of which sources of external biological information one chooses to employ, the first step of our method is to encode this information into a matrix $\mathbf{W}$ of size $N \times N$, where N is the total number of genes under study. The entries of $\mathbf{W}$ are:

$$\mathbf{W}_{ij} = \begin{cases} 1 & \text{if prior evidence suggests that the } i^{\text{th}} \text{ and } j^{\text{th}} \text{ genes are associated} \\ \text{NA} & \text{if the relationship between the } i^{\text{th}} \text{ and } j^{\text{th}} \text{ genes is unknown} \\ 0 & \text{if the } i^{\text{th}} \text{ and } j^{\text{th}} \text{ genes are unlikely to be associated.} \end{cases}$$

Intuitively, $\mathbf{W}$ can be viewed as an adjacency matrix for a network that reflects the known relationships amongst the genes. The proposed method uses this network as an informed basis for

what the network constructed from the time-course gene expression data itself should look like. Importantly, the data can indicate what the "NA" entries of $\mathbf{W}$ ought to be, as well as confirm (or refute) the "0" or "1" entries.

In the event that $\mathbf{W}$ consists largely of "NA" entries, one can also use time-course gene expression data from previous studies to fill in some of these unknown entries. Such datasets often consist of multiple replicates of the expression measurements, possibly gathered under the same experimental conditions to account for sampling variation, or different conditions to assess the effect of a treatment.

3.2 Part 2: The normal-inverse gamma model and Zellner's g-prior

Recall from Section 2.1 that given measurements $\{(m_A(t_i), m_B(t_i))\}_{i=1}^n$ of the expressions of two genes A and B at times $t_1, ..., t_n$, the model we aim to fit is

$$m_A(t) = c_1 m_B(t) + c_2 \int_0^t m_b(s)\,\mathrm{d}s + c_3 \int_0^t m_B(s)\,\mathrm{d}s + c_4 t + c_5,$$

where $c_1, ..., c_5$ are unknown parameters, and the integrals $\int_0^t m_A(s)\,\mathrm{d}s$ and $\int_0^t m_B(s)\,\mathrm{d}s$ are estimated by numerically integrating spline interpolants of the given data. Recall from (5) that this model can be represented in matrix form as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

Since c_1 and c_2 link the expressions of genes A and B, we intuitively ought to place priors of non-zero mean on these two parameters if $\mathbf{W}$ indicates that the genes are associated. To do this, we adopt the normal-inverse gamma model for $\boldsymbol{\beta}$ and σ^2, which is used frequently in Bayesian regression and allows for flexible modeling of the prior mean and covariance matrix of $\boldsymbol{\beta}$. The normal-inverse gamma model is specified as follows:

$$\boldsymbol{\beta}|\sigma^2 \sim N\left(\boldsymbol{\beta}_0, \sigma^2 \mathbf{V}_0\right) \quad \text{and} \quad \sigma^2 \sim \Gamma^{-1}(a, b),$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^{p\times 1}$, $\mathbf{V}_0 \in \mathbb{R}^{p\times p}$ is a positive semidefinite matrix, and $a, b > 0$. It is then said that $(\boldsymbol{\beta}, \sigma^2)$ jointly follow a normal-inverse gamma distribution with parameters $(\boldsymbol{\beta}_0, \mathbf{V}_0, a, b)$. This is a conjugate prior, so the posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$ is also normal-inverse gamma with parameters $(\boldsymbol{\beta}_*, \mathbf{V}_*, a_*, b_*)$ defined as

$$\boldsymbol{\beta}_* = \left(\mathbf{V}_0^{-1} + \mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{V}_0^{-1}\boldsymbol{\beta}_0 + \mathbf{X}^T\mathbf{Y}\right), \tag{9}$$

$$\mathbf{V}_* = \left(\mathbf{V}_0^{-1} + \mathbf{X}^T\mathbf{X}\right)^{-1}, \tag{10}$$

$$a_* = a + \frac{n}{2}, \tag{11}$$

$$b_* = b + \frac{1}{2}\left(\boldsymbol{\beta}_0^T\mathbf{V}_0^{-1}\boldsymbol{\beta}_0 + \mathbf{Y}^T\mathbf{Y} - \boldsymbol{\beta}_*^T\mathbf{V}_*^{-1}\boldsymbol{\beta}_*\right). \tag{12}$$

That is, the conditional posterior of $\boldsymbol{\beta}$ given σ^2 and the posterior of σ^2 are:

$$\boldsymbol{\beta}|\sigma^2, \mathbf{Y} \sim N(\boldsymbol{\beta}_*, \sigma^2\mathbf{V}_*) \;\text{ and }\; \sigma^2|\mathbf{Y} \sim \Gamma^{-1}(a_*, b_*).$$

The posterior mean $\boldsymbol{\beta}_*$ can be used as the vector of estimated regression parameters. This estimator has the desirable properties of consistency and asymptotic efficiency in large samples and admissibility in finite samples [Giles & Rayner, 1979].

The hyperparameters $\boldsymbol{\beta}_0$ and $\mathbf{V}_0$ are of particular interest as they allow us to incorporate biological information into our modeling framework. In defining $\boldsymbol{\beta}_0$, recall that we wish to place priors with non-zero mean on the parameters c_1 and c_2 when external sources suggest that genes A and B are co-regulated or at least associated. It was stated in Section 2.1 that c_1 represents the ratio α_A/α_B, where α_A and α_B denote the strength of a common regulatory signal in the first-order dynamics of the two genes. If the genes are known to be associated, it is reasonable to believe *a priori* that $\alpha_A = \alpha_B$, implying $c_1 = 1$. Then we could also say *a priori* that $c_2 = 1$, since c_2 represents a parameter that is proportional to c_1. If prior information about genes A and B is lacking or suggests that they are unrelated, a prior mean of zero for c_1 and c_2 is appropriate. Supposing genes A and B are the ith and j^{th} genes in the dataset, we can thus set the prior mean $\boldsymbol{\beta}_0$ of $\boldsymbol{\beta}$ as follows:

$$\boldsymbol{\beta}_0 = \begin{cases} [1,1,0,0,0]^T & \text{if } \mathbf{W}_{ij} = 1 \\ [0,0,0,0,0]^T & \text{if } \mathbf{W}_{ij} = 0 \text{ or NA.} \end{cases}$$

As for the prior covariance matrix $\sigma^2\mathbf{V}_0$ of $\boldsymbol{\beta}$, we first note that for the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}_n)$, the least-squares estimator $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ has the covariance matrix $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. A popular choice for $\mathbf{V}_0$ is therefore $g(\mathbf{X}^T\mathbf{X})^{-1}$, where $g > 0$. This choice of $\mathbf{V}_0$ yields a particularly tractable special case of the normal-inverse gamma model known as Zellner's g-prior [Zellner, 1986]. Substituting this $\mathbf{V}_0$ into the expressions for $\boldsymbol{\beta}_*$ in (9) and $\mathbf{V}_*$ in (10), which are the posterior mean and covariance matrix of $\boldsymbol{\beta}$, we obtain

$$\boldsymbol{\beta}_* = \frac{1}{1+g}\boldsymbol{\beta}_0 + \frac{g}{1+g}\hat{\boldsymbol{\beta}}_{\text{OLS}}, \tag{13}$$

$$\mathbf{V}_* = \frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}. \tag{14}$$

(13) reveals that the posterior mean $\boldsymbol{\beta}_*$ is a convex combination of the prior mean $\boldsymbol{\beta}_0$ and the least-squares estimator $\hat{\boldsymbol{\beta}}_{\text{OLS}}$. The parameter g therefore balances the weight placed on external information encoded by $\boldsymbol{\beta}_0$ and the data used to compute $\hat{\boldsymbol{\beta}}_{\text{OLS}}$, and so the selection of g is an important component of our modeling strategy. We next describe our data-driven approach to choosing it.

3.3 Part 3: Optimal selection of g in Zellner's g-prior

Several methods for choosing g in Zellner's g-prior have been proposed previously. For instance, [George & Foster, 2000] discuss an empirical Bayes method in which one selects the value of g that maximizes the marginal likelihood of $\mathbf{Y}$. [Liang et al., 2008] provide a closed-form expression for this maximizing value of g that is nearly identical to the F-statistic for testing the hypothesis that $\boldsymbol{\beta} = \mathbf{0}$. They show that this maximum marginal likelihood approach has the desirable property that the Bayes factor for comparing the full model to the null (intercept-only) model diverges as the R^2 approaches 1. For our application, one concern is that the F-statistic defining this particular estimate of g is likely to be overwhelmingly large for many gene pairs. If g is large, then $\boldsymbol{\beta}_*$ will be very close to $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ according to (13). As a result, any biological evidence of association captured by $\boldsymbol{\beta}_0$ will have a negligible impact on the model.

A fully Bayesian approach to selecting g involves placing a prior distribution on g that is then integrated out when defining the prior on $\boldsymbol{\beta}$. This method is motivated by [Zellner & Siow, 1980], who propose placing a Cauchy prior on $\boldsymbol{\beta}$ to derive a closed-form posterior odds ratio for hypothesis testing purposes. These priors can be represented as a mixture of g-priors with an inverse gamma prior on g, although a closed-form posterior estimate of g is unavailable. [Cui & George, 2008] and [Liang et al., 2008] alternatively consider a class of priors of the form $\pi(g) = \frac{a-2}{2}(1+g)^{a/2}$ for $a > 2$, known as the hyper-g priors. Under these priors, the posterior mean of g can be expressed in closed-form in terms of the Gaussian hypergeometric function.

Because our application involves fitting regression models for potentially thousands of gene pairs, the computational cost of fully Bayesian methods for selecting g requires us to consider alternative approaches. One idea is to select the value of g that minimizes the sum of squared residuals $\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$, where $\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}_*$ is the vector of fitted values obtained with $\boldsymbol{\beta}_*$ in (13):

$$\hat{\mathbf{Y}} \;=\; \mathbf{X}\boldsymbol{\beta}_* \;=\; \mathbf{X}\left(\frac{1}{1+g}\boldsymbol{\beta}_0 + \frac{g}{1+g}\hat{\boldsymbol{\beta}}_{\text{OLS}}\right) \;=\; \frac{1}{1+g}\mathbf{Y}_0 + \frac{g}{1+g}\hat{\mathbf{Y}}_{\text{OLS}}, \tag{15}$$

where $\mathbf{Y}_0 = \mathbf{X}\boldsymbol{\beta}_0$ and $\hat{\mathbf{Y}}_{\text{OLS}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}}$. However, we found that there are no analytical solutions to $g_* = \text{argmin}_{g>0}\,\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = 0$. Instead, we can minimize Stein's unbiased risk estimate (SURE), which is an unbiased estimate of $\|\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|^2$. Below is a rephrased version of Theorem 8.7 in [Fourdrinier et al., 2018], which defines SURE for the general problem of estimating $\mathbb{E}(\mathbf{Y})$ using a linear estimator of the form $\hat{\mathbf{Y}} = \mathbf{a} + \mathbf{S}\mathbf{Y}$. This theorem statement differs from its original version in that we have rewritten the divergence of $\mathbf{X}\hat{\boldsymbol{\beta}}$ with respect to $\mathbf{Y}$ using the notion of generalized degrees of freedom developed in [Efron, 2004].

Theorem 3.1 (SURE for linear models)**.** Let $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, where the dimensions of $\mathbf{X}$ are $n \times p$, and let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{Y})$ be a weakly differentiable function of the least squares estimator $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ such that $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ can be written in the form $\hat{\mathbf{Y}} = \mathbf{a} + \mathbf{S}\mathbf{Y}$ for some vector $\mathbf{a}$ and matrix $\mathbf{S}$. Let $\hat{\sigma}^2 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}}\|^2/(n-p)$. Then,

$$\delta_0(\mathbf{Y}) = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + (2\,\text{Tr}(\mathbf{S}) - n)\,\hat{\sigma}^2 \tag{16}$$

is an unbiased estimator of $\|\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|^2$.

From (15), observe that we can write $\hat{\mathbf{Y}}$ as

$$\hat{\mathbf{Y}} = \frac{1}{1+g}\mathbf{Y}_0 + \frac{g}{1+g}\mathbf{H}\mathbf{Y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Therefore, the matrix $\mathbf{S}$ in Theorem 3.1 is $g\mathbf{H}/(1+g)$, whose trace is $gp/(1+g)$. SURE in (16) then becomes

$$\delta_0(\mathbf{Y}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_*\|^2 + \left(\frac{2gp}{1+g} - n\right)\hat{\sigma}^2, \tag{17}$$

where we have also substituted the posterior mean $\boldsymbol{\beta}_*$ in (13) for $\hat{\boldsymbol{\beta}}$.

We next present the value of g that minimizes SURE.

Theorem 3.2 (SURE minimization with respect to g)**.** The value of g that minimizes SURE in (17) is

$$g_* = \frac{\|\hat{\mathbf{Y}}_{\text{OLS}} - \mathbf{Y}_0\|^2 - p\hat{\sigma}^2}{p\hat{\sigma}^2}.$$

The proof of Theorem 3.2 is provided in Appendix A.1.

It is quite possible that $p\hat{\sigma}^2$ is small, resulting in a large value of g_* (i.e., $g_* \gg 1$) in Theorem 3.2. In this case, $\boldsymbol{\beta}_*$ in (13) will be largely influenced by the data via $\hat{\boldsymbol{\beta}}_{\text{OLS}}$, rather than by prior information via $\boldsymbol{\beta}_0$. This is desirable when the relationship between the two genes is unknown (i.e. $\mathbf{W}_{ij} = \text{NA}$), but not if the relationship is known to be unlikely (i.e. $\mathbf{W}_{ij} = 0$). In the latter case, we prefer to shrink the regression coefficients towards the prior mean $\boldsymbol{\beta}_0 = \mathbf{0}$ so as to yield a smaller lead-lag R^2 value. To address this, we set g conditionally on $\mathbf{W}_{ij}$ as

$$g = \begin{cases} g_* & \text{if } \mathbf{W}_{ij} = \text{NA or } 1 \\ 1 & \text{if } \mathbf{W}_{ij} = 0, \end{cases} \tag{18}$$

where g_* is defined according to Theorem 3.2.

3.4 Part 4: Computing the R^2 for Bayesian regression models

Once a posterior estimate of the model coefficients (13) has been obtained, with the parameter g selected optimally, we can compute the Bayesian lead-lag R^2 between genes A and B.

Recall that for a linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\beta}$ is estimated by the least-squares estimator $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, the R^2 (coefficient of determination) is defined as

$$R^2 = \frac{\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}} - \overline{Y}\mathbb{1}_n\|^2}{\|\mathbf{Y} - \overline{Y}\mathbb{1}_n\|^2}, \tag{19}$$

where $\overline{Y} = (1/n)\mathbf{Y}^T\mathbb{1}_n$. In Bayesian regression, however, the standard decomposition of total sum of squares into residual and regression sums of squares no longer holds. Thus, when we replace $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ with an estimator such as the posterior mean of $\boldsymbol{\beta}$, the formula (19) can potentially yield $R^2 > 1$. As a remedy to this issue, we compute the R^2 as the ratio of the variance of the model fit to the variance of the model fit plus the variance of the residuals. This ratio is within $[0, 1]$ by construction, and is given by

$$R^2 = \frac{\widehat{\text{Var}}(\mathbf{X}\boldsymbol{\beta}_*)}{\widehat{\text{Var}}(\mathbf{X}\boldsymbol{\beta}_*) + \widehat{\text{Var}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_*)}, \tag{20}$$

where, for a vector $\mathbf{Z} = [z_1 \ \dots \ z_n]^T$ with mean $\bar{z}$, we define $\widehat{\text{Var}}(\mathbf{Z}) = \frac{1}{n-1}\sum_{i=1}^n (z_i - \bar{z})^2$. This calculation is based on the approach to computing R^2 for Bayesian regression models proposed in [Gelman et al., 2018]. For our application, we refer to (20) as the *Bayesian lead-lag* R^2.

One may further desire a notion of statistical significance for the Bayesian lead-lag R^2. One option is to simulate the posterior distribution of this R^2 using draws from the posterior distribution of $\boldsymbol{\beta}$, as described in [Gelman et al., 2018]. Recall from Section 3.2 that the posterior distribution

of $(\boldsymbol{\beta}, \sigma^2)$ is normal-inverse gamma with parameters $(\boldsymbol{\beta}_*, \mathbf{V}_*, a_*, b_*)$, defined respectively in (13), (14), (11), and (12). To draw samples from this posterior distribution, we can first sample σ^2 from the $\Gamma^{-1}(a_*, b_*)$ distribution, and then sample $\boldsymbol{\beta}$ from the $N(\boldsymbol{\beta}_*, g\sigma^2\mathbf{V}_*)$ distribution. In particular, if $\mathbf{W}_{ij} = 1$, we may wish to simulate the posterior distribution of the R^2 in (20) under a null hypothesis of no relationship between genes A and B. This can be reflected in the sampling procedure by calculating $\boldsymbol{\beta}_*$ with $\boldsymbol{\beta}_0$ set to $\mathbf{0}$.

Alternatively, one could use the Bayes factors presented in [Liang et al., 2008] to corroborate the Bayesian lead-lag R^2. Let $\mathcal{M}_F$ denote the model (4) of gene expression, which has an intercept and $p = 4$ "covariates" with coefficients c_1 through c_4. Let $\mathcal{M}_N$ denote the null (intercept-only) model. Then the Bayes factor for comparing $\mathcal{M}_F$ to $\mathcal{M}_N$ is

$$\mathrm{BF}(\mathcal{M}_F : \mathcal{M}_N) = (1+g)^{(n-p-1)/2} \frac{1}{(1+g(1-R^2))^{(n-1)/2}} ,$$

where R^2 is the usual coefficient of determination in (19). [Kass & Raftery, 1995] interpret a $\log_{10}(\mathrm{BF}(\mathcal{M}_F : \mathcal{M}_N))$ value between 1 and 2 as "strong" evidence in favor of $\mathcal{M}_F$, or above 2 as "decisive" evidence.

In Appendix A.2, we give a summary of our methodology in the form of a generic algorithm that can be run on any time-course gene expression dataset.

3.5 Possible modifications to prior hyperparameters

In Section 3.2, we set the prior mean $\boldsymbol{\beta}_0$ of the parameters of the ODE model to $[1, 1, 0, 0, 0]^T$ when $\mathbf{W}_{ij} = 1$, i.e. there is prior evidence suggesting genes A and B are associated. To make the method even more data-driven, one could alternatively set $\boldsymbol{\beta}_0 = [\xi, \xi, 0, 0, 0]^T$ in the $\mathbf{W}_{ij} = 1$ case, where $\xi \neq 0$ is chosen adaptively from the data. The following theorem presents the values of ξ and g that simultaneously minimize SURE in (17) in this setting.

Theorem 3.3 (SURE minimization with respect to ξ and g)**.** Assume the entries of the least-squares estimator $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ are all distinct and the expression of gene B is non-zero for at least one time point, i.e. $m_B(t_i) \neq 0$ for at least one i. Let $\boldsymbol{\beta}_0 = [\xi, \xi, 0, 0, 0]^T$ in the case that $\mathbf{W}_{ij} = 1$. Then the values of ξ and g that minimize SURE in (17) are

$$\xi_* = \frac{\mathbf{Y}^T\mathbf{X}_{12}}{\|\mathbf{X}_{12}\|^2} \quad \text{and} \quad g_* = \frac{\|\hat{\mathbf{Y}}_{\mathrm{OLS}}\|^2\|\mathbf{X}_{12}\|^2 - (\mathbf{Y}^T\mathbf{X}_{12})^2}{\|\mathbf{X}_{12}\|^2 p\hat{\sigma}^2} - 1,$$

where $\mathbf{X}_{12} \in \mathbb{R}^{n\times 1}$ is the element-wise sum of the first two columns of $\mathbf{X}$.

The proof of Theorem 3.3 is provided in Appendix A.1.

4 Results for the *Drosophila melanogaster* dataset

In this section, we apply our method to a real time-course gene expression dataset. That is, given a dataset of N genes whose expressions are measured at n time points, we compute a $N \times N$ similarity matrix $\mathbf{S}$ where $\mathbf{S}_{i,j}$ stores the Bayesian lead-lag R^2 (LLR^2) in (20) between the ith

and j^{th} genes. Note that this quantity is asymmetric: $\text{LL}R^2(i,j) \neq \text{LL}R^2(j,i)$. Here, $\text{LL}R^2(i,j)$ denotes the Bayesian lead-lag R^2 where we treat the ith gene as the response (gene A) in the model (4). In Section 4.2 we consider what this asymmetry suggests about the directionality of the genes' relationship, but we thereafter symmetrize $\mathbf{S}$ for exploratory purposes by setting $\mathbf{S}_{i,j} = \mathbf{S}_{j,i} = \text{LL}R^2(i,j)$.

4.1 Generating a time-course gene expression dataset

The expression of a gene is typically measured via RNA sequencing (RNA-seq) as the count of a particular type of messenger RNA found in a tissue sample. These counts can be normalized to library size and transformed using the limma-voom transformation [Law et al., 2014]. This transformation produces near-normally distributed gene expression measurements, making them more amenable to analysis with linear models such as those described in Section 2.1.

Our primary time-course gene expression dataset, introduced in [Schlamp et al., 2020], profiles the expression dynamics for 12,657 genes in *Drosophila melanogaster* (fruit fly) in response to an immune challenge. An immune response was triggered in flies by injecting them with commercial lipopolysaccharide, which contains peptidoglycan (cell wall molecules) derived from the bacterium *E. coli*. Following injection, the flies were sampled via RNA-seq at 20 time points over five days. The data was normalized by the aforementioned limma-voom transformation and expressed as the $\log_2$-fold change with respect to the first time point, which was sampled pre-injection as a control. We focus on the first 17 time points, ranging from zero to 48 hours post-injection, as this is when most of the variation in expression occurs.

Differential expression analysis is typically used to identify genes exhibiting significant expression changes, and thus reduce a set of thousands of genes into a smaller set meriting further study. We first reduce the set of 12,657 genes into a set of 951 "differentially-expressed" (DE) genes identified in [Schlamp et al., 2020], defined as genes satisfying either of the following criteria: 1) there is at least one time point at which the gene's expression undergoes a $\log_2$-fold change of at least two, or 2) a spline-based method for differential expression analysis returns a significant result. The latter method involves fitting a cubic spline to the temporal expression measurements of a gene under treatment and control settings, and testing whether the difference in the resulting two sets of coefficients is significant. We then add back a set of 784 genes that are not DE by these criteria, but that are "neighbors" of at least one DE gene. We define a neighbor of a DE gene as a non-DE gene that has a STRING score of at least 0.95 with the DE gene. The purpose of adding such neighbors back into the dataset, now consisting of $951 + 784 = 1735$ genes, is to enable more complete biological pathways to be reconstructed from our cluster analysis.

In Section 3.1, we describe several sources of biological information that can be encoded into a prior adjacency matrix $\mathbf{W}$. We choose to use the STRING database, and we mark two genes as "associated" (i.e., $\mathbf{W}_{ij} = 1$) if their STRING score is greater than 0.5. We additionally use replicate information from the time-course dataset to fill in some of the unknown STRING scores. Specifically, if two genes have entries in the STRING database but have an unknown STRING score, we set $\mathbf{W}_{ij} = 1$ if the correlation between their replicated temporal expressions is greater than 0.8. We keep $\mathbf{W}_{ij} = \text{NA}$ for the gene pairs that do not have entries in STRING.

4.2 Small-scale case study: immunity and metabolism

We first validate our methodology on a small set of genes whose behavior exhibits a known interplay between immunity and metabolism. [Schlamp et al., 2020] observe that exposure to bacterial peptidoglycan has an effect not only on the time dynamics of immune response, but also on the expression of genes involved in metabolism. In particular, some genes involved in immune response are up-regulated immediately following peptidoglycan injection, while other genes associated with metabolic processes are down-regulated more slowly. Interestingly, the metabolic genes return to their pre-infection levels well before the immune response has subsided. This phenomenon can be observed in Figure 3, which shows the temporal expression patterns of two immune response genes (*IM1, IM2*) and four metabolic process genes (*FASN1, UGP, mino, fbp*).

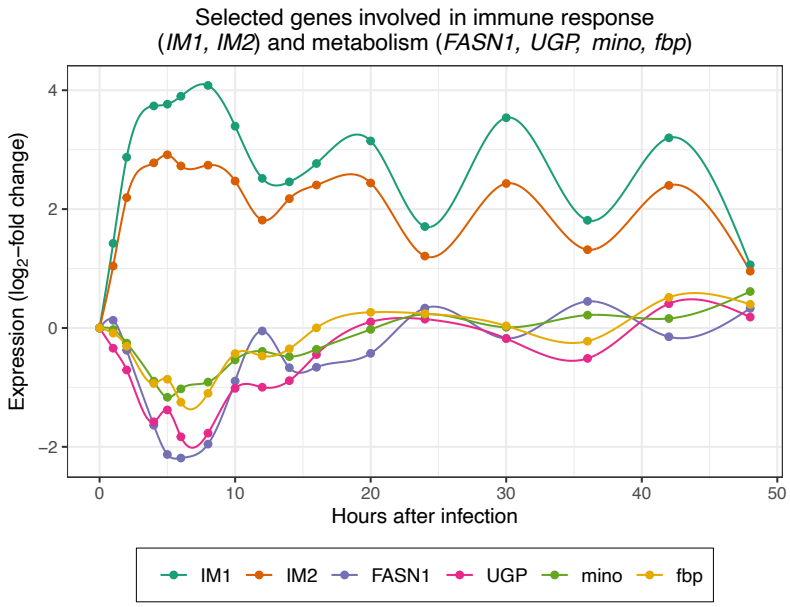

Figure 3: *Temporal expression patterns of selected genes involved in immune response (IM1, IM2) and metabolic processes (FASN1, UGP, mino, fbp). Upon infection, immune response genes undergo immediate up-regulation. At the same time, metabolic genes undergo slower down-regulation and return to pre-infection expression levels after 12 to 24 hours, which is before the immune response is resolved.*

Tables 1-3 show the prior adjacency matrix $\mathbf{W}$, the Bayesian LLR^2 values, and the values of the quantity LLR^2 − LLR^2_{own} (introduced in Section 2.2) corresponding to these six genes. Recall that the Bayesian LLR^2 between genes A and B, denoted LL$R^2(A, B)$, comes from the model

$$m_A(t) = c_1 m_B(t) + c_2 \int_0^t m_B(s)\,\mathrm{d}s + c_3 \int_0^t m_A(s)\,\mathrm{d}s + c_4 t + c_5, \tag{21}$$

while the Bayesian LL$R^2_{\text{own}}(A, B)$ comes from the model

$$m_A(t) = c_3 \int_0^t m_A(s)\,\mathrm{d}s + c_4 t + c_5. \tag{22}$$

	IM1	IM2	FASN1	UGP	mino	fbp
IM1	-	1	NA	0	0	NA
IM2	1	-	NA	0	0	NA
FASN1	NA	NA	-	NA	NA	NA
UGP	0	0	NA	-	1	NA
mino	0	0	NA	1	-	NA
fbp	NA	NA	NA	NA	NA	-

Table 1: *Portion of the prior adjacency matrix* $\mathbf{W}$ *corresponding to the six genes in Figure 3. Red cells indicate that there is prior evidence of association between the two corresponding genes. "NA" entries indicate that the association between the two genes is unavailable in the STRING database.*

	IM1	IM2	FASN1	UGP	mino	fbp
IM1	-	0.99	0.76	0.21	0.33	0.52
IM2	0.98	-	0.71	0.18	0.31	0.46
FASN1	0.82	0.80	-	0.77	0.97	0.78
UGP	0.30	0.30	0.83	-	0.88	0.99
mino	0.40	0.39	0.98	0.91	-	0.90
fbp	0.68	0.66	0.82	0.99	0.86	-

Table 2: *Bayesian* LLR^2 *values corresponding to each gene pair in Figure 3. The* LLR^2 *value in each cell comes from treating the "row gene" as gene A and the "column gene" as gene B in the model (21), hence the asymmetry of this table. Colored cells mark Bayesian* LLR^2 *values above 0.72, the* 95th *percentile of the empirical distribution of the Bayesian* LLR^2 *for this dataset. Red cells are consistent with prior evidence of association in* $\mathbf{W}$*. Blue cells point to potential associations that were uncharacterized in Table 1.*

	IM1	IM2	FASN1	UGP	mino	fbp
IM1	-	0.64	0.42	0.10	0.22	0.17
IM2	0.70	-	0.43	0.09	0.22	0.17
FASN1	0.27	0.26	-	0.22	0.42	0.23
UGP	0.05	0.04	0.20	-	0.26	0.36
mino	0.09	0.09	0.27	0.20	-	0.19
fbp	0.09	0.06	0.23	0.39	0.26	-

Table 3: *Values of* $LLR^2 - LLR^2_{own}$ *corresponding to each gene pair in Figure 3. In each cell, the "row gene" is treated as gene A and the "column gene" is treated as gene B in the model (21). Colored cells mark values of* $LLR^2 - LLR^2_{own}$ *above 0.22, the* 95th *percentile of the empirical distribution of this metric, and have the same meaning as in Table 2. Most of the colored cells in Table 2 are colored here as well, indicating that the associations drawn from Table 2 are unlikely to be false positives.*

Within this set of six genes, Table 1 states that there is prior evidence of association between the immune response genes *IM1* and *IM2* and between the metabolic genes *mino* and *UGP*. The "NA" entries in the off-diagonal blocks of Table 1 signify that the relationships between these groups of immune and metabolic genes are uncharacterized in the STRING database. Table 2 shows that we can recover this immune-metabolic relationship through the high values of the Bayesian $\mathrm{LL}R^2$ observed between the metabolic gene *FASN1* and the two immune response genes. Indeed, Figure 3 shows that the temporal expression pattern of *FASN1* is quite similar in overall shape to those of *IM1* and *IM2*, but is reflected along the horizontal axis with more damped oscillations.

Observe that Table 2 is not symmetric: for any two genes A and B, we compute both $\mathrm{LL}R^2(A, B)$ and $\mathrm{LL}R^2(B, A)$ with our Bayesian method. Doing so can allow us to infer the directionality of the relationship between the two genes. In particular, $\mathrm{LL}R^2(A, B) > \mathrm{LL}R^2(B, A)$ suggests that the temporal expression of gene B explains more of the expression of gene A than vice versa, pointing to a possible "lead-lag" relationship[1]. In Table 2, we see that $\mathrm{LL}R^2(\mathit{FASN1}, \mathit{IM1}) > \mathrm{LL}R^2(\mathit{IM1}, \mathit{FASN1})$, and similarly for *IM2*. This suggests that changes in immune response genes may be more likely to lead to changes in metabolic genes, rather than the other way around.

Lastly, Table 3 shows the values of $\mathrm{LL}R^2 - \mathrm{LL}R^2_{\mathrm{own}}$ for each gene pair. This metric captures

[1]It is possible to develop tests of significance for this based on the beta or F distributions.

the amount of additional variation in one gene that can be explained by modeling it as a function of another potentially co-regulated gene. Most values in Table 2 that are highlighted as "large", i.e. greater than the 95th percentile of the empirical distribution of the Bayesian LLR^2, are also highlighted in Table 3. This suggests that the high Bayesian LLR^2 values observed in Table 2 are indeed due primarily to an association between two genes rather than to artifacts of their individual expression patterns. Thus, the metric LLR^2 − LLR^2_{own} helps verify that associations suggested by the LLR^2 value alone are not false positives.

4.3 Biologically-meaningful hierarchical clustering with the Bayesian LLR^2

We now apply hierarchical clustering using Ward's method [Ward Jr, 1963] to the $N \times N$ distance matrix $\mathbf{J} - \mathbf{S}$, where $\mathbf{J}$ is a matrix of ones and $\mathbf{S}$ is the symmetrized similarity matrix containing the Bayesian LLR^2 values between all gene pairs. These genes come from the dataset described in Section 4.1, so $N = 1735$ here. Cutting the resulting dendrogram at a height of nine yields 15 clusters, which we display in Figure 4.

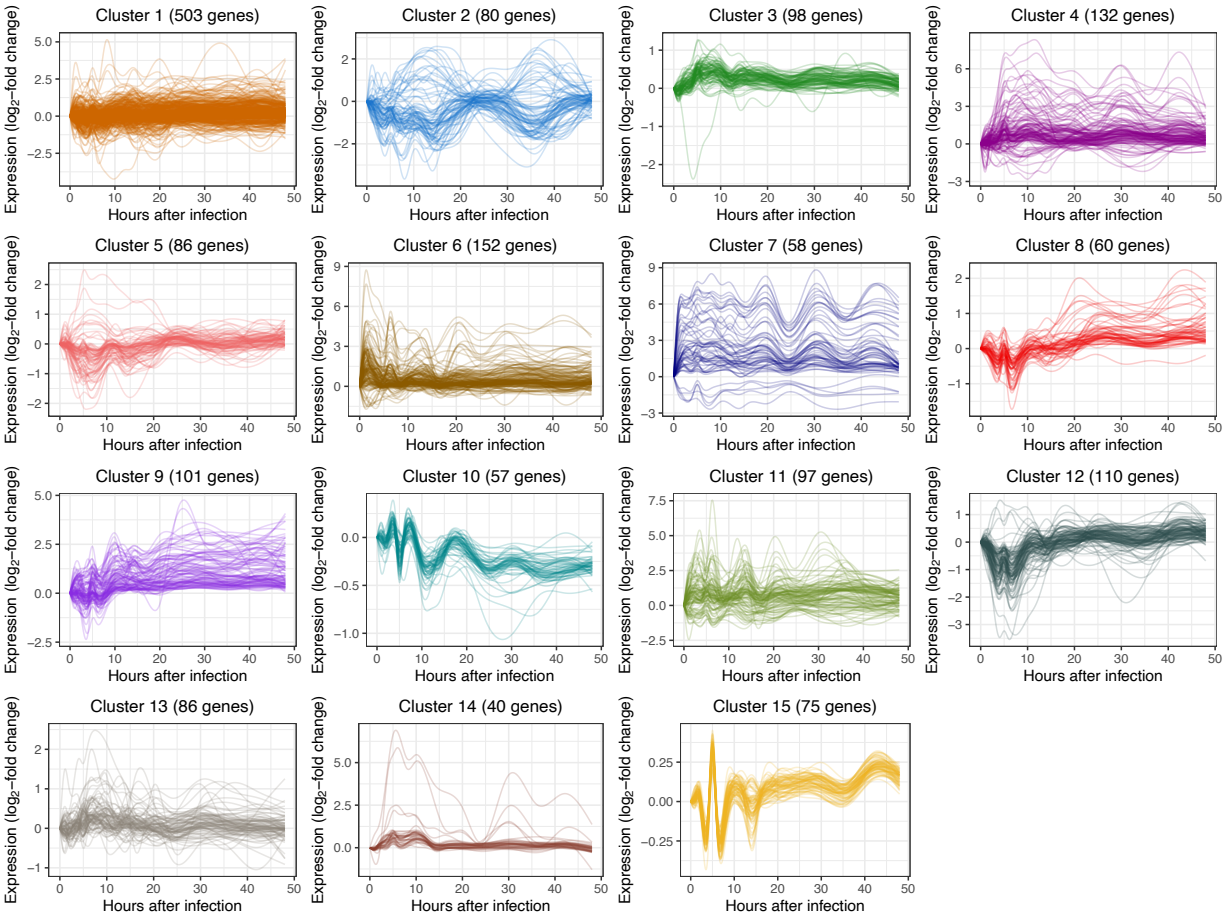

Figure 4: *Plots of the temporal expression patterns of genes in each cluster obtained from applying hierarchical clustering to the distance matrix* $\mathbf{J} - \mathbf{S}$*, where* $\mathbf{S}$ *contains the Bayesian LLR*^{2} *for all gene pairs. Cluster sizes range from 40 to 503 genes, with a mean of 86 and a median of 116. These plots visually demonstrate that the Bayesian LLR*^{2} *is capable of capturing similarities in the overall shapes of two temporal expression patterns, even if one gene is a time-delayed copy of the other or is reflected across a horizontal axis, for instance.*

From each cluster, we can also construct a network in which an edge is defined between two genes if their corresponding Bayesian LLR^2 exceeds a certain threshold. In our analysis, we choose this threshold to be 0.9.

To understand the biological processes that are represented by these clusters, we made use of the Gene Ontology (GO) resource [Ashburner et al., 2000, Carbon et al., 2021]. GO links genes with standardized terms that describe their functions. To determine whether a GO term is significantly enriched within a cluster, i.e. whether the cluster contains significantly more genes associated with that term than expected by chance, we performed a hypergeometric test using the R package `ClusterProfiler` [Yu et al., 2012].

Our analysis shows that each cluster is significantly enriched for specific biological functions. Recall that our dataset profiles the dynamics of immune response, which is an energetically costly process that is also associated with metabolic changes [DiAngelo et al., 2009]. The interplay between immunity and metabolism, which we briefly explored in Section 4.2, is represented particularly well in these clusters. Clusters 6, 7, and 11 are significantly enriched for GO terms related to immune response; clusters 5, 8, 10, and 12 are significantly enriched for metabolic processes; and clusters 4 and 9 are significantly enriched for both metabolic processes and immune and stress responses. Here, "significantly enriched" refers to Benjamini-Hochberg (B-H) corrected p-values below 0.05 [Benjamini & Hochberg, 1995]. Below, we highlight biologically relevant findings from two clusters, and we discuss three additional clusters in Appendix A.3. These examples show that clustering with the Bayesian LLR^2 allows genes with both similar and lagged expression patterns to be grouped together, even in the absence of known prior information. Finally, the Bayesian LLR^2 is not influenced by the direction of gene expression changes (i.e., positive or negative changes), making it easier to detect tradeoffs or negative regulatory interactions between genes.

4.3.1 Analysis of cluster 7

In cluster 7, which consists of 58 genes, the most significantly enriched GO biological process term is "defense response" (B-H corrected p-value of 7.4×10^{-25}). This GO term is supported by 24 genes in the cluster, within which there are two distinct groups that we display in Figure 5: eight genes that are known to respond to "Imd" signaling and eight genes that respond to "Toll" signaling. The Imd and Toll signaling pathways represent well-studied molecular responses to infection in flies. The Imd pathway is tailored to fight off infections from gram-negative bacteria [Kaneko et al., 2004, Zaidman-Rémy et al., 2011, Hanson & Lemaitre, 2020], while the Toll pathway fights off infections from gram-positive bacteria and fungi [Gobert et al., 2003, Hanson & Lemaitre, 2020].

Since the flies profiled in this gene expression dataset were injected with peptidoglycan derived from *E. coli*, a gram-negative bacterium, we expect to see an activation of Imd-regulated genes. Indeed, as seen in Figure 5, the eight Imd-regulated genes in this cluster underwent an immediate and strong up-regulation and reached their highest expression one to two hours after peptidoglycan injection. By contrast, Toll-regulated genes underwent a delayed up-regulation of smaller magnitude, and reached their highest expression two to four hours after injection. This cluster also contained genes that were down-regulated after peptidoglycan injection, including *Acp1* and *CG7214*, which encode components of the cuticle. Genes that encode cuticle components have

been found to respond to infections [Irving et al., 2001], but their functional role after an immune response is not known. Overall, the Bayesian LLR^2 method successfully grouped together functionally related genes with distinct activation kinetics in this cluster.

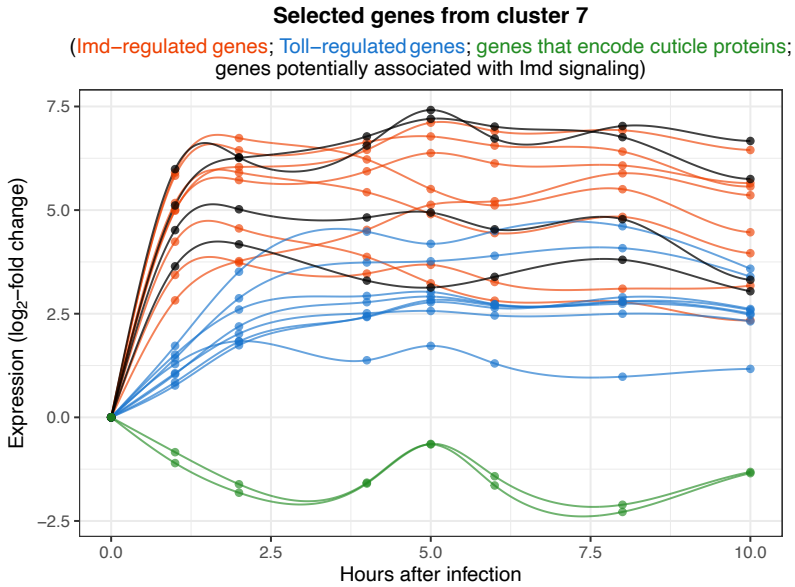

Figure 5: *Temporal expression patterns of selected genes in cluster 7 during the first ten hours following peptidoglycan injection. The eight red lines correspond to Imd-regulated genes: AttA, AttB, AttC, Dro, CecA2, DptA, DptB, PGRP-SB1. Imd-regulated genes show immediate up-regulation that peaks one to two hours after infection. The eight blue lines correspond to Toll-regulated genes: PGRP-SA, IMPPP, IM1/BomS1, IM2/BomS2, IM4/Dso1, IM14/Dso2, IM23/BomBc1, IM3/BomS3. Toll-regulated genes show a smaller and delayed up-regulation that peaks two to four hours after infection. The two green lines correspond to genes Acp1 and CG7214, which encode cuticle proteins and are down-regulated following injection. The four black lines correspond to genes CG44404/IBIN, CG43236/Mtk-like, CG43202/BomT1, and CG43920, which exhibited high Bayesian LLR^2 values (> 0.9) with Imd-regulated genes; while there is no prior information in STRING linking them with the Imd pathway, their co-clustering with Imd-regulated genes was also observed in [Schlamp et al., 2020].*

In addition to recovering known dynamics of immune response pathways in cluster 7, the Bayesian LLR^2 metric identified several new relationships between genes. In this cluster, four genes (*CG44404*, also known as *IBIN*; *CG43236*, also known as *Mtk-like*; *CG43202*, also known as *BomT1*; *CG43920*) had no prior information available in the STRING database to link them with Imd-regulated genes. These four genes exhibit similar expression patterns to Imd-regulated genes, as seen in Figure 5, although previous studies examine *IBIN* and *BomT1* expression downstream of Toll signaling [Clemmons et al., 2015, Valanne et al., 2019]. This suggests that these four genes are not exclusively controlled by Toll signaling, but that they can also respond to Imd signaling after a gram-negative immune challenge. Furthermore, while Imd-regulation of *CG43236/Mtk-like* and *CG43920* has not been experimentally validated, their co-clustering pattern with Imd-regulated genes was also observed by [Schlamp et al., 2020].

In Figure 6, we show a network consisting of the four aforementioned genes (*CG44404/IBIN, CG43236/Mtk-like, CG43202/BomT1, CG43920*) and their neighbors, i.e. the genes with which they have a Bayesian LLR^2 of at least 0.9. Red edges in Figure 6 connect genes that were known

to be associated according to prior information, i.e. $\mathbf{W}_{ij} = 1$ for these pairs. Blue edges connect genes with an uncharacterized relationship in the STRING database, i.e $\mathbf{W}_{ij} = \text{NA}$ for these pairs.

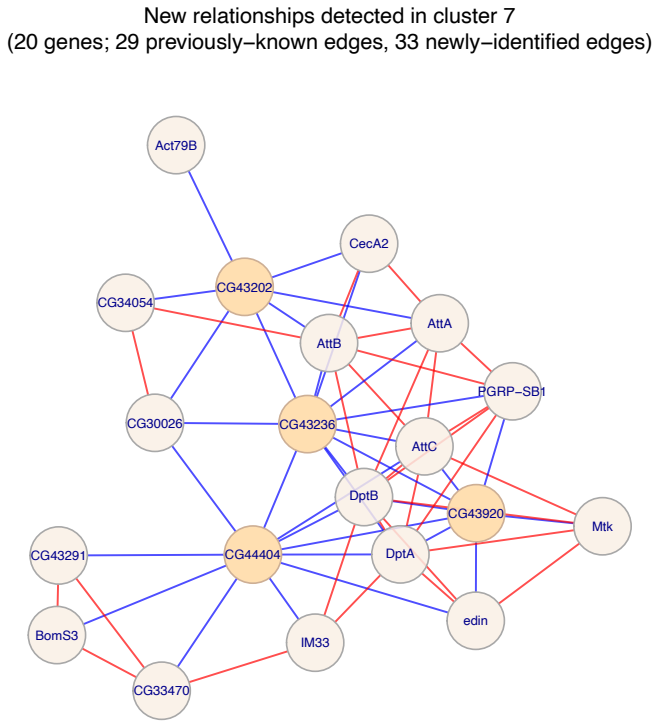

Figure 6: *Network of genes formed from CG44404/IBIN, CG43236/Mtk-like, CG43202/BomT1, CG43920, and their neighbors. Two genes are connected by an edge if they have a Bayesian LLR^2 value above 0.9. Red edges connect genes with known relationships, while blue edges connect genes with uncharacterized relationships according to the STRING database. Blue edges connect the four genes of interest with genes known to be regulated by the Imd signaling pathway, suggesting a possible role for them in fighting gram-negative infections; previously, these four genes were studied after gram-positive or fungal infections.*

4.3.2 Analysis of cluster 12

Genes in cluster 12, some of which are displayed in Figure 7, are characterized by a transient decrease in expression during the first 24 hours after peptidoglycan injection. This cluster was significantly enriched for GO terms related to metabolic processes, one of which is "carbohydrate metabolic process" (B-H corrected p-value of 1.2×10^{-9}). A highly-connected gene involved in carbohydrate metabolism is *fbp*, which has a degree of 87 in our reconstructed network shown in Figure 8. *fbp* plays an important role in gluconeogenesis, the storage of energy as glycogen. Some of the genes connected to *fbp* include *Galk*, *AGBE*, and *Gba1b* (Figure 7), which are also known to be involved in carbohydrate metabolism but are not linked to *fbp* in the STRING database. Twenty-five other genes connected to *fbp* in our network are not well-studied. These connections suggest roles for these uncharacterized genes in carbohydrate metabolism or energy homeostasis.

Another gene connected to *fbp* is *takeout*, which has been implicated in feeding behavior and is regulated by the circadian clock [Sarov-Blat et al., 2000]. In this dataset, *takeout* follows a circadian rhythm, but was interestingly not assigned to cluster 2 (Appendix A.3, Figure 12),

which includes many genes with known circadian rhythms. Its inclusion with metabolic genes in cluster 12 instead suggests an interplay between feeding behavior and metabolism after infection. [Ponton et al., 2020] also showed that feeding can be altered by infection.

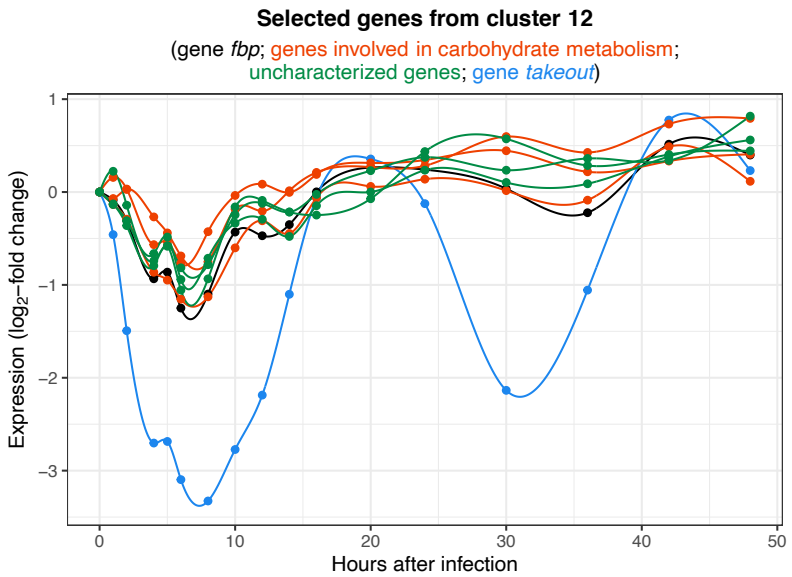

Figure 7: *Temporal expression patterns of selected genes in cluster 12 that have Bayesian LLR^2 values above 0.9 with gene fbp, shown in black. The three red lines correspond to genes AGBE, Galk, and Gba1b, which are known to be involved in carbohydrate metabolism, but are not linked to fbp in the STRING database. The three green lines correspond to genes CG11594, CG10469, and CG13315, which have uncharacterized roles. The blue line corresponds to the gene takeout. takeout exhibits a circadian pattern that is more strongly down-regulated in the first 24 hours, suggesting that its expression can be influenced by both the circadian clock and the immune response.*

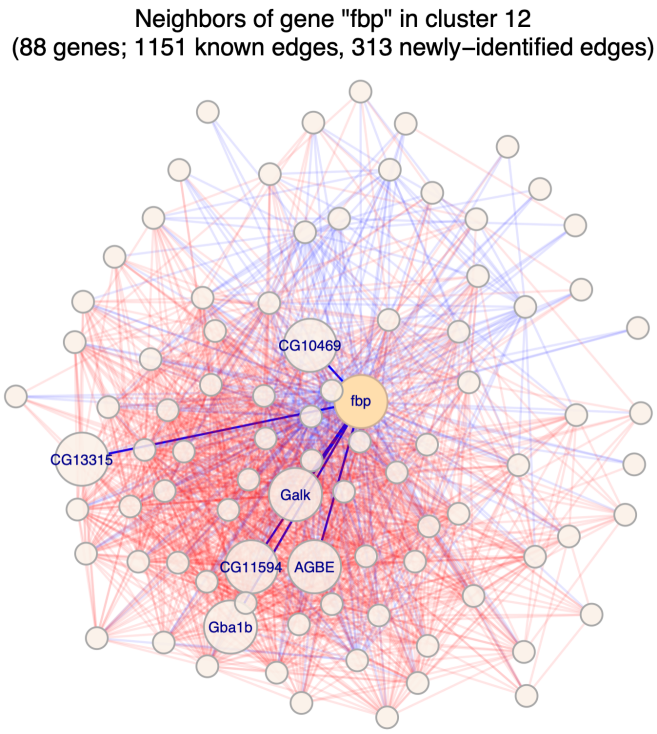

Figure 8: *Network of genes formed from gene fbp and its 87 neighbors, i.e. genes with which fbp has a Bayesian LLR^2 value above 0.9. Red edges connect genes with known relationships, while blue edges connect genes with uncharacterized relationships according to STRING. Selected genes with uncharacterized relationships to fbp (AGBE, Galk, Gba1b, CG11594, CG10469, CG13315) are highlighted.*

4.4 The Bayesian LLR^2 produces a sparse set of significant associations

The quantities $\text{LL}R^2_{\text{other}}$ and $\text{LL}R^2 - \text{LL}R^2_{\text{own}}$ introduced in Section 2.2 can be computed without biological information via ordinary least-squares regression, or with biological information via our proposed Bayesian methodology. We now examine how the Bayesian approach changes the distributions of these quantities in a way that is conducive to identifying pairs or groups of genes with highly similar time dynamics.

In Figure 9, we randomly select 150 genes and place all $\binom{150}{2} = 11{,}175$ pairs on two scatterplots whose x-axes display the $\text{LL}R^2_{\text{other}}$ values and y-axes display the $\text{LL}R^2 - \text{LL}R^2_{\text{own}}$ values. These R^2 values are computed via ordinary least-squares regression in one scatterplot and via our Bayesian method in the other. When two genes are indeed associated and exhibit similar temporal expression patterns, we expect both the $\text{LL}R^2_{\text{other}}$ and the $\text{LL}R^2 - \text{LL}R^2_{\text{own}}$ values between them to be large. Such gene pairs fall into the upper-right quadrant of the scatterplot.

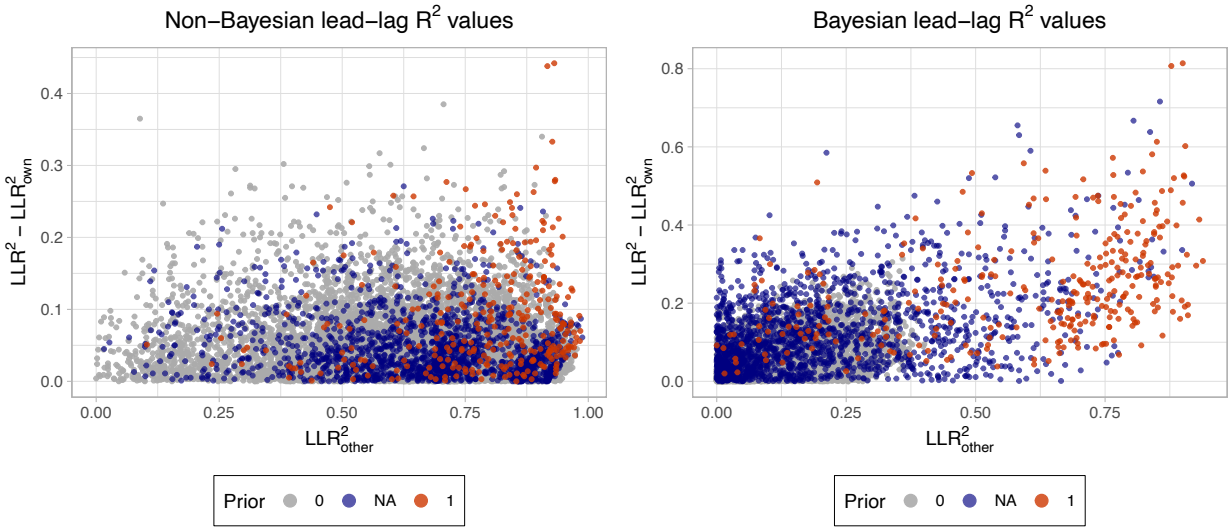

Figure 9: *Scatterplots of LLR$^2_{other}$ (x-axis) and LLR^2 − LLR$^2_{own}$ (y-axis) for a random selection of 150 genes, resulting in 11,175 gene pairs. Each point corresponds to one gene pair. Left: R^2 values are computed via ordinary least-squares, without leveraging external biological information. Right: R^2 values are computed with the proposed Bayesian approach. Points are colored according to how prior information characterizes the relationship between the corresponding two genes: unlikely to be associated, i.e. $\mathbf{W}_{ij} = 0$ (gray); uncharacterized association, i.e. $\mathbf{W}_{ij} = NA$ (blue); or known association, i.e. $\mathbf{W}_{ij} = 1$ (red).*

Figure 9 shows that when we use the ordinary least-squares approach, i.e. without incorporating external biological information, we obtain a very large number of gene pairs with high $\text{LL}R^2_{\text{other}}$ scores. Many of these gene pairs are those that are unlikely to be associated, according to our chosen sources of prior information. By contrast, the Bayesian approach leverages this information

to shift the distribution of the lead-lag R^2 values noticeably, yielding a smaller set of gene pairs that are worth examining further. This distributional shift is due to both the estimator $\boldsymbol{\beta}_*$ for the coefficients in the model (21) and the method in (18) for setting the parameter g in the prior distribution of these coefficients. In particular, g controls how much $\boldsymbol{\beta}_*$ is influenced by either the data or prior information.

Importantly, Figure 9 shows that gene pairs with previously uncharacterized relationships but highly similar time dynamics are more easily identified with the Bayesian versions of the lead-lag R^2 values. Figure 10 shows three examples of small gene groups for which prior evidence of pairwise associations is unavailable in the STRING database, yet the corresponding Bayesian $\mathrm{LL}R^2_{\mathrm{other}}$ and $\mathrm{LL}R^2 - \mathrm{LL}R^2_{\mathrm{own}}$ values place the pairs in the upper-right region of the scatterplot. Figure 11 shows three more examples of gene groups that fall pairwise in the upper-right region of the Bayesian $\mathrm{LL}R^2$ scatterplot in Figure 9. For these genes, the pairwise associations are already well-studied, further demonstrating that the proposed Bayesian methodology successfully recovers known relationships amongst genes.

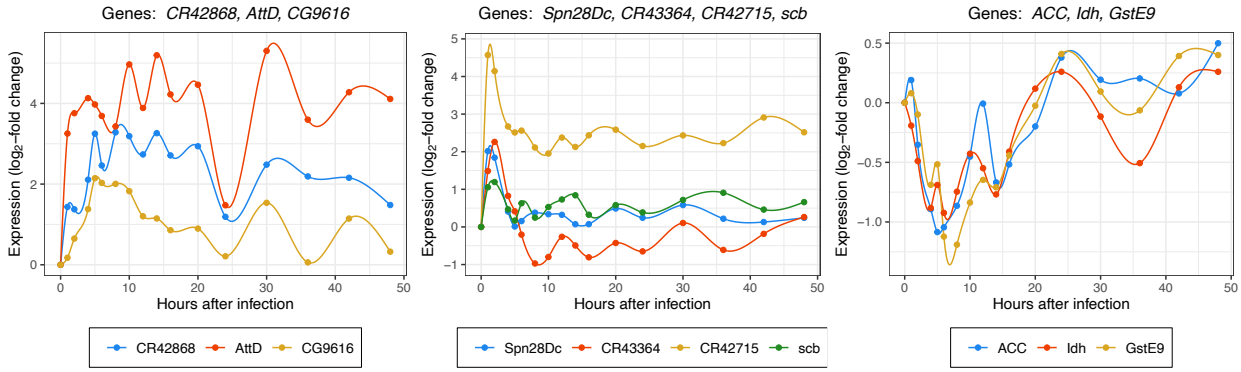

Figure 10: *Sets of genes in the upper-right region of the Bayesian LLR^2 scatterplot in Figure 9, several of which have uncharacterized pairwise associations. Left: The gene AttD is involved in immune response against Gram-negative bacteria. AttD exhibits similar patterns to CR42868 and CG9616, neither of which have known molecular functions according to FlyBase [Larkin et al., 2020]. Middle: The gene Spn28Dc is involved in response to stimuli and protein metabolism, and the gene scb is involved in cell death and organ development. These two genes display similar, vertically shifted expression patterns that are also similar to the less well-understood RNAs CR43364 and CR42715. Right: The genes ACC, Idh, and GstE9 are involved in a variety of metabolic processes, but not all of their pairwise interactions are known in the STRING database.*

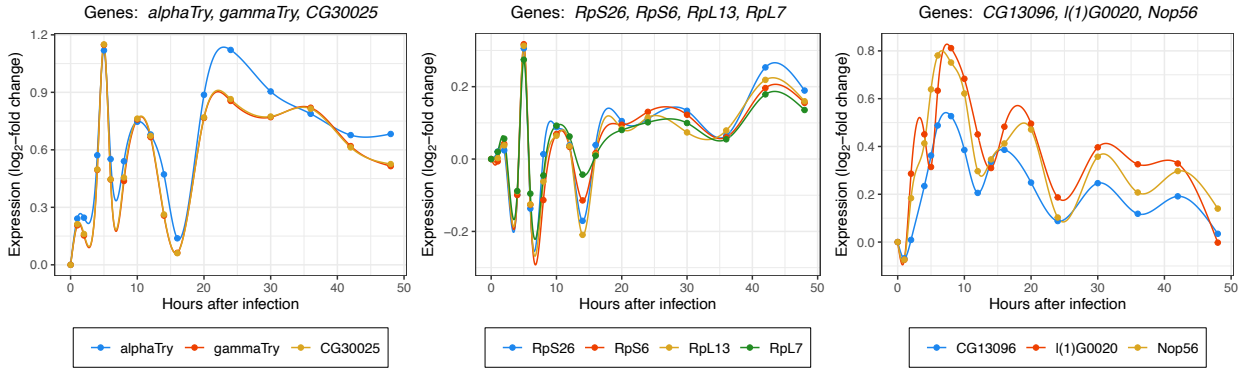

Figure 11: *Sets of genes appearing in the upper-right region of the Bayesian LLR^2 scatterplot in Figure 9, all of which have known pairwise associations. Left: The genes alphaTry, gammaTry, and CG30025 are known to be involved in proteolysis. Middle: RpS26, RpS6, RpL13, and RpL7 encode ribosomal proteins. Right: The genes CG13096, l(1)G0020, and Nop56 are involved in RNA binding.*

5 Discussion

Time-course gene expression datasets are a valuable resource for understanding the complex dynamics of interconnected biological systems. An important statistical task in the analysis of such data is to identify clusters or networks of genes that exhibit similar temporal expression patterns. These patterns yield systems-level insight into the roles of biological pathways and have important implications for the study of processes such as disease progression and recovery.

The main statistical challenges in studying time-course gene expression datasets stem from their high dimensionality and small sample sizes, combined with the nonlinearity of gene expression time dynamics. To overcome these difficulties, we proposed a method for identifying potentially associated genes that treats temporal gene expression as a dynamical system governed by ODEs, whose parameters are determined in a Bayesian way using gene networks curated *a priori* from biological databases. The ODE model and relevant sub-models are fit via Bayesian regression, from which we derived the biologically-informed Bayesian lead-lag R^2 similarity measure. The Bayesian regression procedure leverages Zellner's g-prior and ideas from shrinkage estimation, namely minimization of Stein's unbiased risk estimate (SURE), to balance the posterior ODE model's fit to the data and to prior information. As a result, we automatically encourage gene pairs with known associations to receive higher Bayesian lead-lag R^2 scores, while reducing the scores of gene pairs that are unlikely to be related and allowing new relationships to be discovered.

In Section 4, we explored groups of genes that were identified by our method as having similar temporal dynamics. These groups comprise clusters and networks of genes with clear interpretations in genetics. In particular, the clusters highlighted the known interplay between immune response and metabolism, and suggested roles for uncharacterized genes displaying remarkably similar temporal patterns to more well-studied ones.

Acknowledgements

SKV, SB, AGC, SYND, MHL and MTW were supported by the NIH award R01GM135926. In addition, SB was supported by an NSF award DMS-1812128, and an NIH award R21NS120227. MW was supported by NIH awards U19AI111143-07 and 1P01-AI159402. AGC was supported by R01HD059060.

References

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool

for the unification of biology. *Nature Genetics*, 25(1), 25–29.

[Bansal et al., 2006] Bansal, M., Gatta, G. D., & di Bernardo, D. (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7), 815–822.

[Bausek & Zeidler, 2014] Bausek, N. & Zeidler, M. P. (2014). Gα73b is a downstream effector of jak/stat signalling and a regulator of rho1 in drosophila haematopoiesis. *Journal of Cell Science*, 127(1), 101–110.

[Benjamini & Hochberg, 1995] 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.

[Buch et al., 2008] Buch, S., Melcher, C., Bauer, M., Katzenberger, J., & Pankratz, M. J. (2008). Opposing effects of dietary protein and sugar regulate a transcriptional target of drosophila insulin-like peptide signaling. *Cell Metabolism*, 7(4), 321–332.

[Carbon et al., 2021] Carbon, S., Douglass, E., Good, B. M., Unni, D. R., Harris, N. L., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., et al. (2021). The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49(D1), D325–D334.

[Ceriani et al., 2002] Ceriani, M. F., Hogenesch, J. B., Yanovsky, M., Panda, S., Straume, M., & Kay, S. A. (2002). Genome-wide expression analysis in drosophilareveals genes controlling circadian behavior. *Journal of Neuroscience*, 22(21), 9305–9319.

[Chen et al., 1999] Chen, T., He, H. L., & Church, G. M. (1999). Modeling gene expression with differential equations. In *Biocomputing'99* (pp. 29–40). World Scientific.

[Clemmons et al., 2015] Clemmons, A. W., Lindsay, S. A., & Wasserman, S. A. (2015). An effector peptide family required for drosophila toll-mediated immunity. *PLoS Pathogens*, 11(4), e1004876.

[Cui & George, 2008] Cui, W. & George, E. I. (2008). Empirical bayes vs. fully bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4), 888–900.

[Cyran et al., 2003] Cyran, S. A., Buchsbaum, A. M., Reddy, K. L., Lin, M.-C., Glossop, N. R., Hardin, P. E., Young, M. W., Storti, R. V., & Blau, J. (2003). vrille, pdp1, and dclock form a second feedback loop in the drosophila circadian clock. *Cell*, 112(3), 329–341.

[De Jong, 2002] De Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*, 9(1), 67–103.

[D'haeseleer et al., 1999] D'haeseleer, P., Wen, X., Fuhrman, S., & Somogyi, R. (1999). Linear modeling of mrna expression levels during cns development and injury. In *Biocomputing'99* (pp. 41–52). World Scientific.

[DiAngelo et al., 2009] DiAngelo, J. R., Bland, M. L., Bambina, S., Cherry, S., & Birnbaum, M. J. (2009). The immune response attenuates growth and nutrient storage in drosophila by reducing insulin signaling. *Proceedings of the National Academy of Sciences*, 106(49), 20853–20858.

[Efron, 2004] Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467), 619–632.

[Eisen et al., 1998] Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25), 14863–14868.

[Fabregat et al., 2018] Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The reactome pathway knowledgebase. *Nucleic Acids Research*, 46(D1), D649–D655.

[Farina et al., 2007] Farina, L., De Santis, A., Morelli, G., & Ruberti, I. (2007). Dynamic measure of gene co-regulation. *IET Systems Biology*, 1(1), 10–17.

[Farina et al., 2008] Farina, L., De Santis, A., Salvucci, S., Morelli, G., & Ruberti, I. (2008). Embedding mRNA stability in correlation analysis of time-series gene expression data. *PLoS Computational Biology*, 4(8).

[Fourdrinier et al., 2018] Fourdrinier, D., Strawderman, W. E., & Wells, M. T. (2018). *Shrinkage Estimation*. Springer.

[Gaudet et al., 2011] Gaudet, P., Livstone, M. S., Lewis, S. E., & Thomas, P. D. (2011). Phylogenetic-based propagation of functional annotations within the gene ontology consortium. *Briefings in bioinformatics*, 12(5), 449–462.

[Gelman et al., 2018] Gelman, A., Goodrich, B., Gabry, J., & Vethari, A. (2018). R-squared for Bayesian regression models. *American Statistician*.

[George & Foster, 2000] George, E. I. & Foster, D. P. (2000). Calibration and empirical bayes variable selection. *Biometrika*, 87(4), 731–747.

[Giles & Rayner, 1979] Giles, D. & Rayner, A. (1979). The mean squared errors of the maximum likelihood and natural-conjugate bayes regression estimators. *Journal of Econometrics*, 11(2-3), 319–334.

[Gobert et al., 2003] Gobert, V., Gottar, M., Matskevich, A. A., Rutschmann, S., Royet, J., Belvin, M., Hoffmann, J. A., & Ferrandon, D. (2003). Dual activation of the drosophila toll pathway by two pattern recognition receptors. *Science*, 302(5653), 2126–2130.

[Goto et al., 2003] Goto, A., Kadowaki, T., & Kitagawa, Y. (2003). Drosophila hemolectin gene is expressed in embryonic and larval hemocytes and its knock down causes bleeding defects. *Developmental Biology*, 264(2), 582–591.

[Hanson & Lemaitre, 2020] Hanson, M. A. & Lemaitre, B. (2020). New insights on drosophila antimicrobial peptide function in host defense and beyond. *Current Opinion in Immunology*, 62, 22–30.

[Hill et al., 2012] Hill, S. M., Neve, R. M., Bayani, N., Kuo, W. L., Ziyad, S., Spellman, P. T., Gray, J. W., & Mukherjee, S. (2012). Integrating biological knowledge into variable selection: an empirical Bayes approach with an application in cancer biology. *BMC Bioinformatics*, 13(1), 1–15.

[Irving et al., 2001] Irving, P., Troxler, L., Heuer, T. S., Belvin, M., Kopczynski, C., Reichhart, J.-M., Hoffmann, J. A., & Hetru, C. (2001). A genome-wide analysis of immune responses in drosophila. *Proceedings of the National Academy of Sciences*, 98(26), 15119–15124.

[Kanehisa & Goto, 2000] Kanehisa, M. & Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30.

[Kaneko et al., 2004] Kaneko, T., Goldman, W. E., Mellroth, P., Steiner, H., Fukase, K., Kusumoto, S., Harley, W., Fox, A., Golenbock, D., & Silverman, N. (2004). Monomeric and polymeric gram-negative peptidoglycan but not purified lps stimulate the drosophila imd pathway. *Immunity*, 20(5), 637–649.

[Karp et al., 2019] Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., et al. (2019). The biocyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*, 20(4), 1085–1093.

[Kass & Raftery, 1995] Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.

[Krejčová et al., 2019] Krejčová, G., Danielová, A., Nedbalová, P., Kazek, M., Strych, L., Chawla, G., Tennessen, J. M., Lieskovská, J., Jindra, M., Doležal, T., et al. (2019). Drosophila macrophages switch to aerobic glycolysis to mount effective antibacterial defense. *Elife*, 8, e50414.

[Larkin et al., 2020] Larkin, A., Marygold, S. J., Antonazzo, G., Attrill, H., Dos Santos, G., Garapati, P. V., Goodman, J. L., Gramates, L. S., Millburn, G., Strelets, V. B., et al. (2020). Flybase: updates to the drosophila melanogaster knowledge base. *Nucleic Acids Research*.

[Law et al., 2014] Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, 15(2), R29.

[Li & Zhang, 2010] Li, F. & Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105(491), 1202–1214.

[Li & Jackson, 2015] Li, Y. & Jackson, S. A. (2015). Gene network reconstruction by integration of prior biological knowledge. *G3: Genes, Genomes, Genetics*, 5(6), 1075–1079.

[Liang et al., 2008] Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481), 410–423.

[Lo et al., 2012] Lo, K., Raftery, A. E., Dombek, K. M., Zhu, J., Schadt, E. E., Bumgarner, R. E., & Yeung, K. Y. (2012). Integrating external biological knowledge in the construction of regulatory networks from time-series expression data. *BMC Systems Biology*, 6.

[Melcarne, 2020] Melcarne, C. (2020). Role of nimrod receptors in the drosophila melanogaster cellular immune response. (pp. 127).

[Nepomuceno et al., 2015] Nepomuceno, J. A., Troncoso, A., Nepomuceno-Chamorro, I. A., & Aguilar-Ruiz, J. S. (2015). Integrating biological knowledge based on functional annotations for biclustering of gene expression data. *Computer Methods and Programs in Biomedicine*, 119(3), 163–180.

[Peng et al., 2013] Peng, B., Zhu, D., Ander, B. P., Zhang, X., Xue, F., Sharp, F. R., & Yang, X. (2013). An integrative framework for bayesian variable selection with informative priors for identifying genes and pathways. *PloS One*, 8(7), e67672.

[Polynikis et al., 2009] Polynikis, A., Hogan, S., & di Bernardo, M. (2009). Comparing different ode modelling approaches for gene regulatory networks. *Journal of Theoretical Biology*, 261(4), 511–530.

[Ponton et al., 2020] Ponton, F., Morimoto, J., Robinson, K., Kumar, S. S., Cotter, S. C., Wilson, K., & Simpson, S. J. (2020). Macronutrients modulate survival to infection and immunity in drosophila. *Journal of Animal Ecology*, 89(2), 460–470.

[Qian et al., 2001] Qian, J., Dolled-Filhart, M., Lin, J., Yu, H., & Gerstein, M. (2001). Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *Journal of Molecular Biology*, 314(5), 1053–1066.

[Sarov-Blat et al., 2000] Sarov-Blat, L., So, W. V., Liu, L., & Rosbash, M. (2000). The drosophila takeout gene is a novel molecular link between circadian rhythms and feeding behavior. *Cell*, 101(6), 647–656.

[Schlamp et al., 2020] Schlamp, F., Delbare, S., Early, A. M., Wells, M. T., Basu, S., & Clark, A. G. (2020). Dense time-course gene expression profiling of the drosophila melanogaster innate immune response.

[Stingo et al., 2011] Stingo, F. C., Chen, Y. A., Tadesse, M. G., & Vannucci, M. (2011). Incorporating biological information into linear models: A bayesian approach to the selection of pathways and genes. *The Annals of Applied Statistics*, 5(3).

[Szklarczyk et al., 2019] Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., et al. (2019). String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1), D607–D613.

[Tavazoie et al., 1999] Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., & Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics*, 22(3), 281–285.

[Valanne et al., 2019] Valanne, S., Salminen, T. S., Järvelä-Stölting, M., Vesala, L., & Rämet, M. (2019). Immune-inducible non-coding rna molecule lincrna-ibin connects immunity and metabolism in drosophila melanogaster. *PLoS Pathogens*, 15(1), e1007504.

[Walter et al., 1996] Walter, M. F., Zeineh, L. L., Black, B. C., McIvor, W. E., Wright, T. R., & Biessmann, H. (1996). Catecholamine metabolism and in vitro induction of premature cuticle melanization in wild type and pigmentation mutants of drosophila melanogaster. *Archives of Insect Biochemistry and Physiology: Published in Collaboration with the Entomological Society of America*, 31(2), 219–233.

[Ward Jr, 1963] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.

[Wolfram Research, 2019] Wolfram Research (2019). Mathematica 12.0.

[Wu et al., 2019] Wu, L., Qiu, X., xiang Yuan, Y., & Wu, H. (2019). Parameter estimation and variable selection for big systems of linear ordinary differential equations: A matrix-based approach. *Journal of the American Statistical Association*, 114(526), 657–667.

[Yu et al., 2010] Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., & Wang, S. (2010). Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, 26(7), 976–978.

[Yu et al., 2012] Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: A Journal of Integrative Biology*, 16(5), 284–287.

[Zaidman-Rémy et al., 2011] Zaidman-Rémy, A., Poidevin, M., Hervé, M., Welchman, D. P., Paredes, J. C., Fahlander, C., Steiner, H., Mengin-Lecreulx, D., & Lemaitre, B. (2011). Drosophila immunity: analysis of pgrp-sb1 expression, enzymatic activity and function. *PLoS One*, 6(2), e17231.

[Zellner, 1986] Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques*.

[Zellner & Siow, 1980] Zellner, A. & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadistica Y de Investigacion Operativa*, 31(1), 585–603.

[Zhang et al., 2013] Zhang, W., Wan, Y.-w., Allen, G. I., Pang, K., Anderson, M. L., & Liu, Z. (2013). Molecular pathway identification using biological network-regularized logistic models. *BMC Genomics*, 14(8), 1–8.

A Appendices

A.1 Proofs

In this section, we provide proofs of Theorems 3.2 and 3.3.

Proof of Theorem 3.2. We write $\delta_0(\mathbf{Y})$ as $\delta_0(g)$, treating $\mathbf{Y}$ as fixed. Expanding the expression for $\delta_0(\mathbf{g})$, we obtain

$$\begin{aligned}
\delta_0(g) &= \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \frac{2gp\hat{\sigma}^2}{1+g} - n\hat{\sigma}^2 \\
&= \left(\mathbf{Y} - \frac{1}{1+g}\mathbf{Y}_0 - \frac{g}{1+g}\hat{\mathbf{Y}}_{\text{OLS}}\right)^T \left(\mathbf{Y} - \frac{1}{1+g}\mathbf{Y}_0 - \frac{g}{1+g}\hat{\mathbf{Y}}_{\text{OLS}}\right) + \frac{2gp\hat{\sigma}^2}{1+g} - n\hat{\sigma}^2 \\
&= \|\mathbf{Y}\|^2 - \frac{2}{1+g}\mathbf{Y}^T\mathbf{Y}_0 - \frac{2g}{1+g}\mathbf{Y}^T\hat{\mathbf{Y}}_{\text{OLS}} + \frac{1}{(1+g)^2}\|\mathbf{Y}_0\|^2 + \frac{2g}{(1+g)^2}\hat{\mathbf{Y}}_{\text{OLS}}^T\mathbf{Y}_0 \\
&\quad + \frac{g^2}{(1+g)^2}\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 + \frac{2gp\hat{\sigma}^2}{1+g} - n\hat{\sigma}^2. \qquad (23)
\end{aligned}$$

Next, observe that $\mathbf{Y}^T\hat{\mathbf{Y}}_{\text{OLS}} = \|\hat{\mathbf{Y}}_{\text{OLS}}\|^2$, because

$$\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 = (\mathbf{HY})^T(\mathbf{HY}) = \mathbf{Y}^T\mathbf{HY} = \mathbf{Y}^T\hat{\mathbf{Y}}_{\text{OLS}}. \qquad (24)$$

Furthermore, observe that $\hat{\mathbf{Y}}_{\text{OLS}}^T\mathbf{Y}_0 = \mathbf{Y}^T\mathbf{Y}_0$, because

$$\hat{\mathbf{Y}}_{\text{OLS}}^T\mathbf{Y}_0 = (\mathbf{HY})^T\mathbf{Y}_0 = \mathbf{Y}^T\mathbf{HY}_0 = \mathbf{Y}^T\mathbf{Y}_0, \qquad (25)$$

where the last equality follows from the fact that $\mathbf{Y}_0 = \mathbf{X}\boldsymbol{\beta}_0$, i.e. $\mathbf{Y}_0$ is in the column span of $\mathbf{X}$, which is the space onto which $\mathbf{H}$ projects. The identities (24) and (25) can now be used to write $\delta_0(g)$ in (23) as

$$\begin{aligned}
\delta_0(g) &= \|\mathbf{Y}\|^2 - \frac{2}{1+g}\mathbf{Y}^T\mathbf{Y}_0 - \frac{2g}{1+g}\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 + \frac{1}{(1+g)^2}\|\mathbf{Y}_0\|^2 + \frac{2g}{(1+g)^2}\mathbf{Y}^T\mathbf{Y}_0 \\
&\quad + \frac{g^2}{(1+g)^2}\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 + \frac{2gp\hat{\sigma}^2}{1+g} - n\hat{\sigma}^2 \\
&= a - \frac{2b}{1+g} - \frac{2gc}{1+g} + \frac{d}{(1+g)^2} + \frac{2gb}{(1+g)^2} + \frac{g^2c}{(1+g)^2} + \frac{2gp\hat{\sigma}^2}{1+g} - n\hat{\sigma}^2,
\end{aligned}$$

where $a = \|\mathbf{Y}\|^2$, $b = \mathbf{Y}^T\mathbf{Y}_0$, $c = \|\hat{\mathbf{Y}}_{\text{OLS}}\|^2$, and $d = \|\mathbf{Y}_0\|^2$. Differentiating $\delta_0(g)$ with respect to g, we obtain

$$\begin{aligned}
\frac{\mathrm{d}\delta_0(g)}{\mathrm{d}g} &= \frac{2b}{(1+g)^2} - \frac{2c}{1+g} + \frac{2gc}{(1+g)^2} - \frac{2d}{(1+g)^3} + \frac{2b}{(1+g)^2} - \frac{4gb}{(1+g)^3} \\
&\quad + \frac{2gc}{(1+g)^2} - \frac{2g^2c}{(1+g)^3} + \frac{2p\hat{\sigma}^2}{1+g} - \frac{2gp\hat{\sigma}^2}{(1+g)^2} \\
&= \frac{2p\hat{\sigma}^2 - 2c}{1+g} + \frac{4b + 4gc - 2gp\hat{\sigma}^2}{(1+g)^2} - \frac{2d + 4gb + 2g^2c}{(1+g)^3} \\
&= \frac{2p\hat{\sigma}^2 - 2c + 2gp\hat{\sigma}^2 - 2gc}{(1+g)^2} + \frac{4b + 4gc - 2gp\hat{\sigma}^2}{(1+g)^2} - \frac{2d + 4gb + 2g^2c}{(1+g)^3}
\end{aligned}$$

$$= \frac{4b + 2gc + 2p\hat{\sigma}^2 - 2c}{(1+g)^2} - \frac{2d + 4gb + 2g^2c}{(1+g)^3}.$$

Setting this derivative to zero and rearranging yields:

$$\begin{aligned}
\frac{2d + 4gb + 2g^2c}{(1+g)^3} &= \frac{4b + 2gc + 2p\hat{\sigma}^2 - 2c}{(1+g)^2} \\
\Rightarrow \quad 2d + 4gb + 2g^2c &= (1+g)\left(4b + 2gc + 2p\hat{\sigma}^2 - 2c\right) \\
&= 4b + 2gc + 2p\hat{\sigma}^2 - 2c + 4gb + 2g^2c + 2gp\hat{\sigma}^2 - 2gc \\
\Rightarrow \quad 2d &= 4b + 2p\hat{\sigma}^2 - 2c + 2gp\hat{\sigma}^2 \\
\Rightarrow \quad g_* &= \frac{c + d - 2b - p\hat{\sigma}^2}{p\hat{\sigma}^2}.
\end{aligned}$$

We now substitute the definitions of b, c, and d back into this expression to obtain

$$g_* = \frac{\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 + \|\mathbf{Y}_0\|^2 - 2\mathbf{Y}^T\mathbf{Y}_0 - p\hat{\sigma}^2}{p\hat{\sigma}^2}. \tag{26}$$

The numerator of (26) can be simplified by observing that

$$\begin{aligned}
\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 + \|\mathbf{Y}_0\|^2 - 2\mathbf{Y}^T\mathbf{Y}_0 &= \|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 + \|\mathbf{Y}_0\|^2 - 2\hat{\mathbf{Y}}_{\text{OLS}}^T\mathbf{Y}_0 \\
&= (\hat{\mathbf{Y}}_{\text{OLS}} - \mathbf{Y}_0)^T(\hat{\mathbf{Y}}_{\text{OLS}} - \mathbf{Y}_0)^T \\
&= \|\hat{\mathbf{Y}}_{\text{OLS}} - \mathbf{Y}_0\|^2.
\end{aligned}$$

Therefore, (26) becomes

$$g_* = \frac{\|\hat{\mathbf{Y}}_{\text{OLS}} - \mathbf{Y}_0\|^2 - p\hat{\sigma}^2}{p\hat{\sigma}^2}. \tag{27}$$

When $\boldsymbol{\beta}_0 = \mathbf{0}$, we have $\mathbf{Y}_0 = \mathbf{X}\boldsymbol{\beta}_0 = \mathbf{0}$. In this case, (27) becomes

$$g_* = \frac{\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 - p\hat{\sigma}^2}{p\hat{\sigma}^2}.$$

Finally, the second derivative of $\delta_0(g)$ evaluated at $g = g_*$ in (27) is equal to

$$\left.\frac{\mathrm{d}^2\,\delta_0(g)}{\mathrm{d}g^2}\right|_{g=g_*} = \frac{2p^4\hat{\sigma}^8}{\|\hat{\mathbf{Y}}_{\text{OLS}} - \mathbf{Y}_0\|^6},$$

which is positive, thus confirming that $\delta_0(g)$ is indeed minimized at $g = g_*$. This second derivative calculation was verified in *Mathematica* [Wolfram Research, 2019]. □

Proof of Theorem 3.3. We write $\delta_0(\mathbf{Y})$ as $\delta_0(g, \xi)$, treating $\mathbf{Y}$ as fixed. First, if $\boldsymbol{\beta}_0 = [\xi, \xi, 0, 0, 0]^T$, then $\mathbf{X}\boldsymbol{\beta}_0 = \xi\mathbf{X}_{12}$, where $\mathbf{X}_{12}$ is the element-wise sum of the first two columns of $\mathbf{X}$. We now proceed similarly to the proof of Theorem 3.2 by expanding $\delta_0(g, \xi)$:

$$\begin{aligned}
\delta_0(g, \xi) &= \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \frac{2gp\hat{\sigma}^2}{1+g} - n\hat{\sigma}^2 \\
&= \left\|\mathbf{Y} - \mathbf{X}\left(\frac{1}{1+g}\boldsymbol{\beta}_0 + \frac{g}{1+g}\hat{\boldsymbol{\beta}}_{\text{OLS}}\right)\right\|^2 + \frac{2gp\hat{\sigma}^2}{1+g} - n\hat{\sigma}^2
\end{aligned}$$

$$
\begin{aligned}
&= \left\| \mathbf{Y} - \frac{\xi}{1+g}\mathbf{X}_{12} - \frac{g}{1+g}\hat{\mathbf{Y}}_{\text{OLS}} \right\|^2 + \frac{2gp\hat{\sigma}^2}{1+g} - n\hat{\sigma}^2 \\
&= \left(\mathbf{Y} - \frac{\xi}{1+g}\mathbf{X}_{12} - \frac{g}{1+g}\hat{\mathbf{Y}}_{\text{OLS}} \right)^T \left(\mathbf{Y} - \frac{\xi}{1+g}\mathbf{X}_{12} - \frac{g}{1+g}\hat{\mathbf{Y}}_{\text{OLS}} \right) + \frac{2gp\hat{\sigma}^2}{1+g} - n\hat{\sigma}^2 \\
&= \|\mathbf{Y}\|^2 - \frac{2\xi}{1+g}\mathbf{Y}^T\mathbf{X}_{12} - \frac{2g}{1+g}\mathbf{Y}^T\hat{\mathbf{Y}}_{\text{OLS}} + \frac{\xi^2}{(1+g)^2}\|\mathbf{X}_{12}\|^2 + \frac{2\xi g}{(1+g)^2}\hat{\mathbf{Y}}_{\text{OLS}}^T\mathbf{X}_{12} \\
&\quad + \frac{g^2}{(1+g)^2}\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 + \frac{2gp\hat{\sigma}^2}{1+g} - n\hat{\sigma}^2. \qquad (28)
\end{aligned}
$$

Next, observe that $\hat{\mathbf{Y}}_{\text{OLS}}^T\mathbf{X}_{12} = \mathbf{Y}^T\mathbf{X}_{12}$, because

$$
\hat{\mathbf{Y}}_{\text{OLS}}^T\mathbf{X}_{12} = (\mathbf{H}\mathbf{Y})^T\mathbf{X}_{12} = \mathbf{Y}^T\mathbf{H}\mathbf{X}_{12} = \mathbf{Y}^T\mathbf{X}_{12}, \qquad (29)
$$

where the last equality follows from the fact that $\mathbf{X}_{12}$ is in the column span of $\mathbf{X}$, which is the space onto which $\mathbf{H}$ projects. The identities (24) and (29) can now be used to write $\delta_0(g,\xi)$ in (28) as

$$
\begin{aligned}
\delta_0(g,\xi) &= \|\mathbf{Y}\|^2 - \frac{2\xi}{1+g}\mathbf{Y}^T\mathbf{X}_{12} - \frac{2g}{1+g}\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 + \frac{\xi^2}{(1+g)^2}\|\mathbf{X}_{12}\|^2 + \frac{2\xi g}{(1+g)^2}\mathbf{Y}^T\mathbf{X}_{12} \\
&\quad + \frac{g^2}{(1+g)^2}\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 + \frac{2gp\hat{\sigma}^2}{1+g} - n\hat{\sigma}^2 \\
&= a - \frac{2\xi b}{1+g} - \frac{2gc}{1+g} + \frac{\xi^2 d}{(1+g)^2} + \frac{2\xi gb}{(1+g)^2} + \frac{g^2 c}{(1+g)^2} + \frac{2gp\hat{\sigma}^2}{1+g} - n\hat{\sigma}^2,
\end{aligned}
$$

where $a = \|\mathbf{Y}\|^2$, $b = \mathbf{Y}^T\mathbf{X}_{12}$, $c = \|\hat{\mathbf{Y}}_{\text{OLS}}\|^2$, and $d = \|\mathbf{X}_{12}\|^2$. Differentiating $\delta_0(g,\xi)$ with respect to ξ, we obtain

$$
\frac{\partial \delta_0(g,\xi)}{\partial \xi} = -\frac{2b}{1+g} + \frac{2\xi d}{(1+g)^2} + \frac{2gb}{(1+g)^2}.
$$

Setting this derivative to zero and rearranging yields:

$$
\begin{aligned}
\frac{2b}{1+g} &= \frac{2\xi d + 2gb}{(1+g)^2} \\
\Rightarrow \quad b(1+g) &= \xi d + gb \\
\Rightarrow \quad b &= \xi d \\
\Rightarrow \quad \xi_* &= \frac{b}{d}.
\end{aligned}
$$

We now substitute the definitions of b and d back into this expression to obtain

$$
\xi_* = \frac{\mathbf{Y}^T\mathbf{X}_{12}}{\|\mathbf{X}_{12}\|^2}. \qquad (30)
$$

Next, we differentiate $\delta_0(g,\xi)$ with respect to g:

$$
\begin{aligned}
\frac{\partial \delta_0(\mathbf{Y})}{\partial g} &= \frac{2\xi b}{(1+g)^2} - \frac{2c}{1+g} + \frac{2gc}{(1+g)^2} - \frac{2\xi^2 d}{(1+g)^3} + \frac{2\xi b}{(1+g)^2} - \frac{4\xi gb}{(1+g)^3} \\
&\quad + \frac{2gc}{(1+g)^2} - \frac{2g^2 c}{(1+g)^3} + \frac{2p\hat{\sigma}^2}{1+g} - \frac{2gp\hat{\sigma}^2}{(1+g)^2}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{2p\hat{\sigma}^2 - 2c}{1+g} + \frac{4\xi b + 4gc - 2gp\hat{\sigma}^2}{(1+g)^2} - \frac{2\xi^2 d + 4\xi gb + 2g^2 c}{(1+g)^3} \\
&= \frac{2p\hat{\sigma}^2 - 2c + 2gp\hat{\sigma}^2 - 2gc}{(1+g)^2} + \frac{4\xi b + 4gc - 2gp\hat{\sigma}^2}{(1+g)^2} - \frac{2\xi^2 d + 4\xi gb + 2g^2 c}{(1+g)^3} \\
&= \frac{4\xi b + 2gc + 2p\hat{\sigma}^2 - 2c}{(1+g)^2} - \frac{2\xi^2 d + 4\xi gb + 2g^2 c}{(1+g)^3}.
\end{aligned}
$$

Setting this derivative to zero and rearranging yields:

$$
\begin{aligned}
\frac{2\xi^2 d + 2\xi gb + 2g^2 c}{(1+g)^3} &= \frac{4\xi b + 2gc + 2p\hat{\sigma}^2 - 2c}{(1+g)^2} \\
\Rightarrow \quad 2\xi^2 d + 4\xi gb + 2g^2 c &= (1+g)(4\xi b + 2gc + 2p\hat{\sigma}^2 - 2c) \\
&= 2\xi b + 2gc + 2p\hat{\sigma}^2 - 2c + 4\xi gb + 2g^2 c + 2gp\hat{\sigma}^2 - 2gc \\
\Rightarrow \quad 2\xi^2 d &= 4\xi b + 2p\hat{\sigma}^2 - 2c + 2gp\hat{\sigma}^2 \\
\Rightarrow \quad g_* &= \frac{c + \xi^2 d - 2\xi b - p\hat{\sigma}^2}{p\hat{\sigma}^2}.
\end{aligned}
$$

We now substitute the definitions of b, c, and d back into this expression to obtain

$$
g_* = \frac{\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 + \xi^2 \|\mathbf{X}_{12}\|^2 - 2\xi \mathbf{Y}^T \mathbf{X}_{12} - p\hat{\sigma}^2}{p\hat{\sigma}^2}.
$$

Substituting ξ in (30) into this expression for g yields

$$
\begin{aligned}
g_* &= \frac{\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 + \left(\frac{\mathbf{Y}^T\mathbf{X}_{12}}{\|\mathbf{X}_{12}\|^2}\right)^2 \|\mathbf{X}_{12}\|^2 - 2\left(\frac{\mathbf{Y}^T\mathbf{X}_{12}}{\|\mathbf{X}_{12}\|^2}\right)\mathbf{Y}^T\mathbf{X}_{12} - p\hat{\sigma}^2}{p\hat{\sigma}^2} \\
&= \frac{\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 + \frac{(\mathbf{Y}^T\mathbf{X}_{12})^2}{\|\mathbf{X}_{12}\|^2} - 2\frac{(\mathbf{Y}^T\mathbf{X}_{12})^2}{\|\mathbf{X}_{12}\|^2} - p\hat{\sigma}^2}{p\hat{\sigma}^2} \\
&= \frac{\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 \|\mathbf{X}_{12}\|^2 - (\mathbf{Y}^T\mathbf{X}_{12})^2}{\|\mathbf{X}_{12}\|^2 p\hat{\sigma}^2} - 1. \qquad (31)
\end{aligned}
$$

Finally, the determinant of the Hessian matrix of $\delta_0(g,\xi)$, denoted $\nabla^2\delta_0$, evaluated at $g = g_*$ in (31) and $\xi = \xi_*$ in (30) is

$$
\begin{aligned}
\det \nabla^2 \delta_0|_{(g_*,\xi_*)} &= \left.\frac{\partial^2 \delta_0}{\partial g^2}\right|_{(g_*,\xi_*)} \left.\frac{\partial^2 \delta_0}{\partial \xi^2}\right|_{(g_*,\xi_*)} - \left(\left.\frac{\partial^2 \delta_0}{\partial g \partial \xi}\right|_{(g_*,\xi_*)}\right)^2 \\
&= \frac{-4\|\mathbf{X}_{12}\|^2\, p^6\, \hat{\sigma}^{12}}{\left((\mathbf{Y}^T\mathbf{X}_{12})^2 - \|\hat{\mathbf{Y}}_{\text{OLS}}\|^2\|\mathbf{X}_{12}\|^2\right)^5}. \qquad (32)
\end{aligned}
$$

We must now verify that $\det \nabla^2\delta_0|_{(g_*,\xi_*)} > 0$, i.e. that (g_*, ξ_*) is indeed an extremum of δ_0. First, we observe that

$$
\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2 \|\mathbf{X}_{12}\|^2 > (\mathbf{Y}^T\mathbf{X}_{12})^2, \qquad (33)
$$

by direct application of the Cauchy-Schwarz inequality (recall that $\mathbf{Y}^T\mathbf{X}_{12} = \hat{\mathbf{Y}}_{\text{OLS}}^T\mathbf{X}_{12}$ from (29)). The inequality is strict because of the assumption that the entries of $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ are distinct, which

prevents $\hat{\mathbf{Y}}_{\text{OLS}}$ and $\mathbf{X}_{12}$ from being linearly dependent. Thus, (33) implies that the denominator in (32) is strictly negative. The numerator is also strictly negative, because the assumption that gene B is not zero everywhere ensures that $\|\mathbf{X}_{12}\|^2 \neq 0$ (recall that $\mathbf{X}$ is defined in (5), and its first two columns consist of gene B's expression measurements over time and its time integrals). Therefore, (32) is strictly positive. To verify that (g_*, ξ_*) is indeed a minimizer of δ_0, we now check that $\frac{\partial^2 \delta_0}{\partial \xi^2}|_{(g_*,\xi_*)} > 0$ as well. We have:

$$\left.\frac{\partial^2 \delta_0}{\partial \xi^2}\right|_{(g_*,\xi_*)} = \frac{2\|\mathbf{X}_{12}\|^2 p^2 \hat{\sigma}^4}{\left((\mathbf{Y}^T\mathbf{X}_{12})^2 - \|\hat{\mathbf{Y}}_{\text{OLS}}\|^2\|\mathbf{X}_{12}\|^2\right)^2}$$

which is strictly positive. This second derivative calculation, as well as those in (32), were verified in *Mathematica*. □

A.2 Bayesian lead-lag R^2 algorithm

Following is an algorithm for computing the Bayesian lead-lag R^2 for all gene pairs in a time-course gene expression dataset, consisting of N genes measured at n time points $t_1, ..., t_n$.

Algorithm 1: Bayesian lead-lag R^2 calculations for all gene pairs

Input: $\mathbf{Z}$, a gene expression dataset of size $N \times n$ (N genes observed at n time points);
$\mathbf{W}$, a prior biological information matrix of size $N \times N$;
$\mathbf{t}$, a vector of n time points $t_1, ..., t_n$.

Output: $\mathbf{S}$, a Bayesian lead-lag R^2 similarity matrix of size $N \times N$.

Initialize $\mathbf{S}$;

for gene $i = 1, ..., N-1$ **do**

  for gene $j = 2, ..., N$ **do**

    `// Get expression measurements for genes` i `and` j

    Let $\mathbf{x}_i^T = \mathbf{Z}[i, :]$ and $\mathbf{x}_j^T = \mathbf{Z}[j, :]$;

    `// Get spline interpolants of given expression measurements`

    Let $s_i(t) = \texttt{spline}(\mathbf{x}_i)$ and $s_j(t) = \texttt{spline}(\mathbf{x}_j)$;

    `// Numerically integrate spline interpolants up to each time point`

    Let $\tilde{\mathbf{s}}_i = [\int_0^{t_1} s_i(t)\,\mathrm{d}t,\ ...,\ \int_0^{t_n} s_i(t)\,\mathrm{d}t]^T$ and $\tilde{\mathbf{s}}_j = [\int_0^{t_1} s_j(t)\,\mathrm{d}t,\ ...,\ \int_0^{t_n} s_j(t)\,\mathrm{d}t]^T$;

    `// Define` $n \times 5$ `matrix` $\mathbf{X}$ `and` $n \times 1$ `vector` $\mathbf{Y}$ `according to (5)`

    Let $\mathbf{X} = [\mathbf{x}_j \quad \tilde{\mathbf{s}}_j \quad \tilde{\mathbf{s}}_i \quad \mathbf{t} \quad \mathbb{1}_n]$ and $\mathbf{Y} = \mathbf{x}_i$;

    `// Define prior mean of regression coefficients`

    if $\mathbf{W}_{ij} = 1$ **then**

      Let $\boldsymbol{\beta}_0 = [1, 1, 0, 0, 0]^T$;

    else

      Let $\boldsymbol{\beta}_0 = [0, 0, 0, 0, 0]^T$;

    end

    `// Compute least-squares estimates`

    Let $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$;

    `// Compute` g `for Zellner's` g`-prior according to Theorem 3.2`

    Let $\mathbf{Y}_0 = \mathbf{X}\boldsymbol{\beta}_0$ and $\hat{\mathbf{Y}}_{\mathrm{OLS}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$;
    Let $\hat{\sigma}^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}_{\mathrm{OLS}}\|^2/(n-p)$;
    Let $g = (\|\hat{\mathbf{Y}}_{\mathrm{OLS}} - \mathbf{Y}_0\|^2 - p\hat{\sigma}^2)/(p\hat{\sigma}^2)$;

    `// Compute posterior mean of regression coefficients according to (13)`

    Let $\boldsymbol{\beta}_* = (1/(1+g))\boldsymbol{\beta}_0 + (g/(1+g))\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$;

    `// Compute Bayesian lead-lag` R^2

    Let $\mathbf{S}_{ij} = \widehat{\mathrm{Var}}(\mathbf{X}\boldsymbol{\beta}_*)\,/(\widehat{\mathrm{Var}}(\mathbf{X}\boldsymbol{\beta}_*) + \widehat{\mathrm{Var}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_*))$;

  end

end

To instead compute the Bayesian lead-lag R^2 from sub-model 1 in (7), it suffices to change the definition of $\mathbf{X}$ to $[\mathbf{x}_j \;\; \tilde{\mathbf{s}}_j \;\; \mathbb{1}_n]$, and to define $\boldsymbol{\beta}_0$ as $[1,1,0]^T$ if $\mathbf{W}_{ij} = 1$ and $[0,0,0]^T$ otherwise. To compute the Bayesian lead-lag R^2 from sub-model 2 in (8), we change $\mathbf{X}$ to $[\tilde{\mathbf{s}}_i \;\; \mathbf{t} \;\; \mathbb{1}_n]$, and $\boldsymbol{\beta}_0 = [0,0,0]$ regardless of the value of $\mathbf{W}_{ij}$. An optimized version of this algorithm runs in 21.8 minutes for $N = 1735$ genes on a 2017 3.1 GHz Intel Core i5 MacBook Pro.

A.3 Additional results

We continue our analysis in Section 4.3 of the results of hierarchical clustering with the Bayesian lead-lag R^2 (LLR^2).

A.3.1 Analysis of cluster 2

Cluster 2 contains both up- and down-regulated genes with circadian rhythms, according to GO term enrichment. Several of these genes are displayed in Figure 12. Among the up-regulated genes are four regulators of the circadian clock (*tim, per, vri, Pdp1*). A fifth regulator of the circadian clock, *Clk*, is down-regulated. *Pdp1* has been reported to reach its peak expression three to six hours after *vri*'s peak expression [Cyran et al., 2003], a pattern that is visible in this cluster. Cluster 2 further contains genes that are regulated downstream of the circadian clock. These include six genes that are down-regulated during the day: four genes that encode cuticle proteins (*Cpr49Ab, Cpr49Ae, Cpr62Ba, Cpr72Ec*) [Gaudet et al., 2011], and two genes involved in dopamine synthesis (*e, ple*), which contributes to cuticle pigmentation [Walter et al., 1996]. *ple* was found to be under circadian control by [Ceriani et al., 2002].

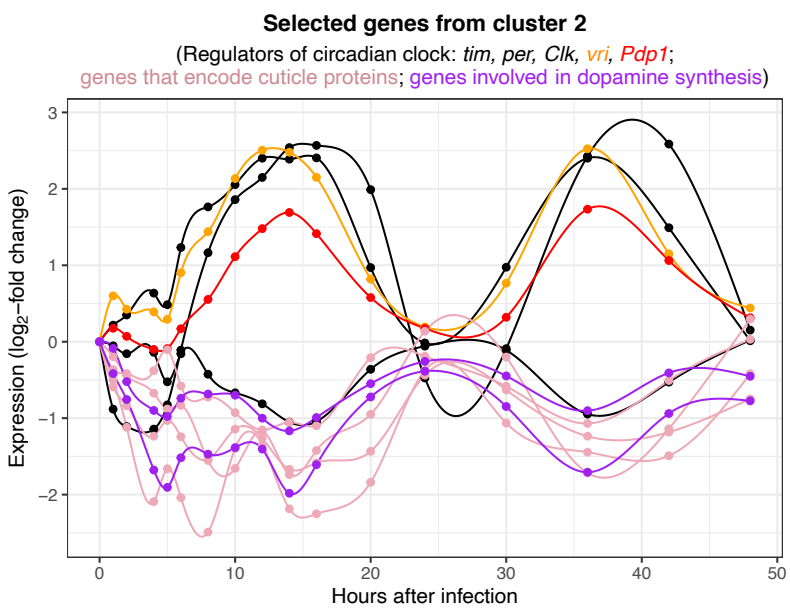

Figure 12: *Temporal expression patterns of selected genes in cluster 2. Black, red, and orange lines correspond to regulators of the circadian clock (Black: tim, per, Clk; orange: vri; red: Pdp1). Pdp1 is known to reach its peak expression after vri does. The four pink lines correspond to genes that encode cuticle proteins (Cpr49Ab, Cpr49Ae, Cpr62Ba, Cpr72Ec). The two purple lines correspond to genes involved in dopamine synthesis (e, ple).*

A.3.2 Analysis of cluster 5

Cluster 5 includes numerous genes with divergent expression patterns, displayed in Figure 13. In this cluster, GO enrichment analysis identified six genes involved in ribosome biogenesis (B-H corrected p-value of 0.01), which are transiently up-regulated starting one hour after peptidoglycan injection. GO analysis also identified eight genes involved in cellular lipid metabolism (B-H corrected p-value of 0.006). This set of eight genes included both genes implicated in lipid catabolism (*dob, Lsd-1, Lip2*) and genes involved in fatty acid biosynthesis (*FASN1, ACC, AcCoAS, mino, ATPCL*). These genes are all transiently down-regulated, apart from the small up-regulation observed for *dob* and *Lsd-1* one hour after injection. These patterns suggest that an immune challenge elicits an increased need for protein synthesis, in a tradeoff with alterations to the fly's lipid homeostasis.

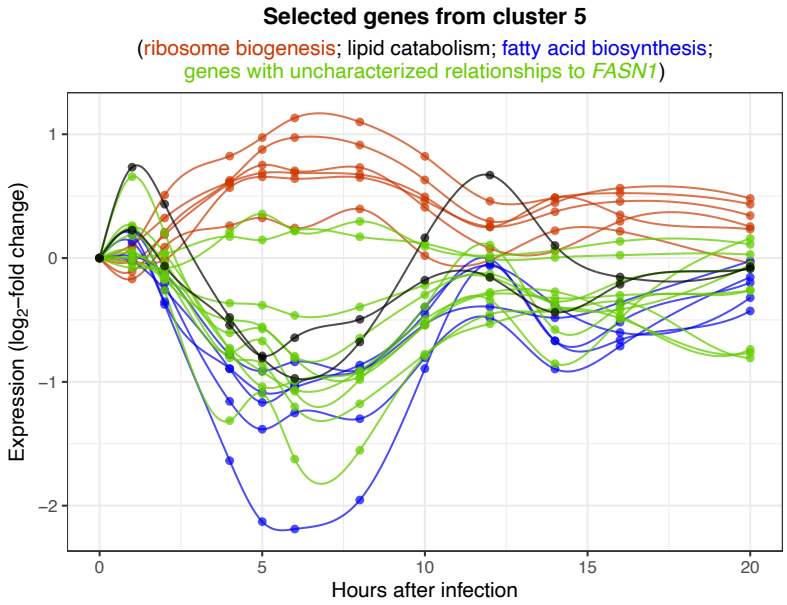

Figure 13: *Temporal expression profiles of selected genes in cluster 5 during the first 20 hours following peptidoglycan injection. The six red lines correspond to genes involved in ribosome biogenesis (nop5, Fib, Nop60B, CG12301, CG32409, U3-55K). The three black lines correspond to genes involved in lipid catabolism (dob, Lip2, Lsd-1). The five blue lines correspond to genes involved in fatty acid biosynthesis (FASN1, ACC, AcCoAS, ATPCL, mino). The ten green lines correspond to genes with uncharacterized relationships to FASN1, according to the STRING database, but that had Bayesian LLR^2 values with FASN1 above 0.9 (CG3756, CG3940, CG8036, eRF1, aralar1, CG32409, CG31904, CG15120, CG1640, CG16926).*

Of the genes described above, *FASN1* undergoes the strongest down-regulation, with more than a four-fold change relative to time zero. Because of this, we were interested in determining whether there were any genes that were not detected in the GO enrichment analysis and had uncharacterized prior associations with *FASN1*, but still had a high Bayesian LLR^2 value (> 0.9) with *FASN1*. We identified ten such genes (*CG3756, CG3940, CG8036, eRF1, aralar1, CG32904, CG31904, CG15120, CG1640, CG16926*). Two of these, *aralar1* and *eRF1*, have an opposite expression pattern to *FASN1* and align with the genes involved in ribosome biosynthesis. Both *aralar1* and *eRF1* have known functions related to protein synthesis: *eRF1* is involved in translation, and

aralar1 is predicted to be involved in amino acid transmembrane transport [Gaudet et al., 2011]. The eight remaining genes have expression patterns that largely align with *FASN1*'s expression profile, suggesting a potential role for these genes in lipid homeostasis.

A.3.3 Analysis of cluster 9

Cluster 9 is significantly enriched for GO terms related to metabolic processes, particularly the term "carbohydrate metabolic process" (B-H corrected p-value of 7.5×10^{-9}). An intriguing feature of this cluster is that it also contains genes involved in immune response, specifically genes known to be expressed in hemocytes, which are *Drosophila* blood cells. Several of these metabolic genes and genes expressed in hemocytes are displayed in Figure 14.

Among the genes in cluster 9 involved in carbohydrate metabolism are five mannosidases (*LManI, LManIII, LManIV, LManV, LManVI*) and six maltases (*Mal-A1, Mal-A3, Mal-A4, Mal-A6, Mal-A7, Mal-A8*). These genes encode enzymes that break down complex sugars into simple sugars such as glucose. Also of interest is the gene *tobi* ("target of brain insulin"), which is regulated by *Drosophila* insulin and dietary sugar [Buch et al., 2008].

Five genes in cluster 9 are expressed in hemocytes. These include three genes that belong to the Nimrod gene family (*NimC1, NimB4, eater*), which have also been shown to mediate Gram-positive bacteria phagocytosis [Melcarne, 2020]; *Hml*, which is involved in the clotting reaction in larvae [Goto et al., 2003]; and *Galphaf*, which in larvae has been implicated in hemocyte proliferation and filopodia formation [Bausek & Zeidler, 2014].

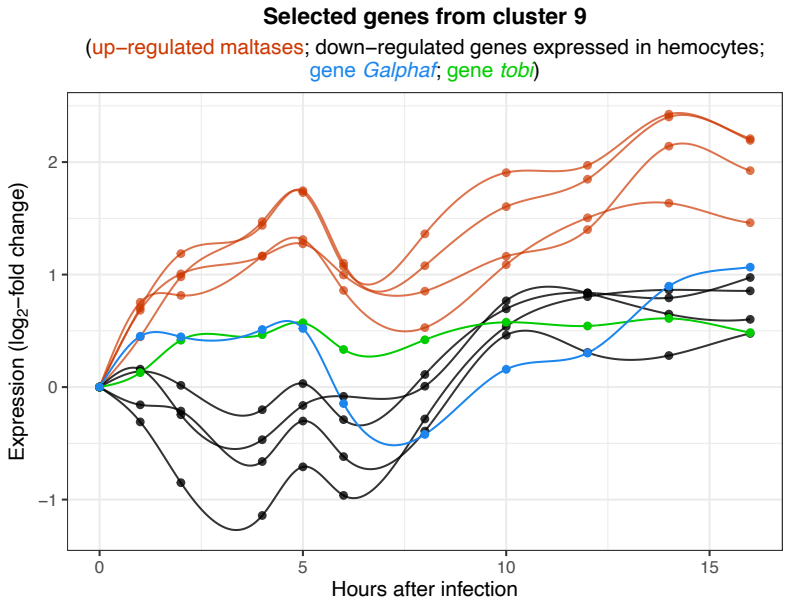

Figure 14: *Temporal expression patterns of selected genes in cluster 9 during the first 16 hours after peptidoglycan injection. The four red lines correspond to up-regulated maltases (Mal-A1, Mal-A6, Mal-A7, Mal-A8). The four black lines correspond to down-regulated genes expressed in hemocytes (NimB4, NimC1, eater, Hml). The blue line corresponds to Galphaf, and the magenta line corresponds to tobi. Genes expressed in hemocytes are initially down-regulated, and then experience a delayed up-regulation relative to maltases; this behavior is consistent with known metabolic changes in how energy production is sustained in hemocytes following an infection.*

Figure 14 shows that these metabolic genes and genes expressed in hemocytes exhibit distinct expression patterns. Most of the maltases, as well as *tobi*, are up-regulated throughout the timeframe. All mannosidases and genes expressed in hemocytes, apart from *Galphaf*, are initially down-regulated but are eventually up-regulated with a lag relative to the maltases and *tobi*. These patterns are sensible in the context of known hemocyte biology. After an infection, hemocytes undergo a metabolic switch, whereby their energy production is sustained mostly by aerobic glycolysis rather than oxidative phosphorylation [Krejčová et al., 2019]. Since aerobic glycolysis is dependent on glucose, the observed up-regulation of glucose-producing enzymes prior to the up-regulation of genes expressed in hemocytes is aligned with our expectations. This lead-lag relationship would be worth following up on using wet lab experiments, in order to determine whether there truly is a dependency between hemocyte function and maltase or *Galphaf* activity.

It is also worth noting that the fold changes of genes expressed in hemocytes are generally small, e.g., less than two-fold, which is often used as a minimal cutoff in RNA-seq analyses. However, the coordinated expression changes detected by the Bayesian LLR^2 suggest that these are biologically relevant patterns.