Bioinformatics, 32(19), 2016, 2956–2964
doi: 10.1093/bioinformatics/btw329
Advance Access Publication Date: 10 June 2016
Original Paper

OXFORD

Gene expression

Inferring the perturbation time from biological time course data

Jing Yang[1,*], Christopher A. Penfold[2], Murray R. Grant[3] and Magnus Rattray[1,*]

[1]Faculty of Life Sciences, University of Manchester, Manchester, UK, [2]Warwick Systems Biology Centre, University of Warwick, Coventry, UK and [3]School of Biosciences, University of Exeter, Exeter, UK

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on April 11, 2015; revised on March 16, 2016; accepted on May 23, 2016

Abstract

Motivation: Time course data are often used to study the changes to a biological process after perturbation. Statistical methods have been developed to determine whether such a perturbation induces changes over time, e.g. comparing a perturbed and unperturbed time course dataset to uncover differences. However, existing methods do not provide a principled statistical approach to identify the specific time when the two time course datasets first begin to diverge after a perturbation; we call this the perturbation time. Estimation of the perturbation time for different variables in a biological process allows us to identify the sequence of events following a perturbation and therefore provides valuable insights into likely causal relationships.

Results: We propose a Bayesian method to infer the perturbation time given time course data from a wild-type and perturbed system. We use a non-parametric approach based on Gaussian Process regression. We derive a probabilistic model of noise-corrupted and replicated time course data coming from the same profile before the perturbation time and diverging after the perturbation time. The likelihood function can be worked out exactly for this model and the posterior distribution of the perturbation time is obtained by a simple histogram approach, without recourse to complex approximate inference algorithms. We validate the method on simulated data and apply it to study the transcriptional change occurring in Arabidopsis following inoculation with *Pseudomonas syringae* pv. tomato DC3000 versus the disarmed strain DC3000*hrpA*.

Availability and Implementation: An R package, DEtime, implementing the method is available at https://github.com/ManchesterBioinference/DEtime along with the data and code required to reproduce all the results.

Contact: Jing.Yang@manchester.ac.uk or Magnus.Rattray@manchester.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene expression time profiles can reveal important information about cellular function and gene regulation (see, e.g. Bar-Joseph, 2004). A common experimental design is to perturb a biological system either before or during a time course experiment. In this case, a fundamental problem is to identify the precise *perturbation time* when a gene's time profile is first altered. In this paper we present an exactly tractable Bayesian inference procedure to infer the perturbation time by comparing perturbed and wild-type gene expression profiles. Ordering genes by their perturbation time gives valuable insight into the likely causal sequence of events following a perturbation. We demonstrate the applicability of our method by studying the timing of transcriptional changes in *Arabidopsis thaliana* leaves following inoculation with the

hemibiotrophic bacteria *Pseudomonas syringae* pv. tomato DC3000 versus the disarmed strain DC3000*hrpA*.

Most methods for the analysis of differentially expressed genes are based upon snapshots of gene expression (Dudoit *et al.*, 2002; Kerr *et al.*, 2000) and there are many well-established software packages for that purpose targeted at microarray and RNA-Seq data (Anders and Huber, 2010; Hardcastle and Kelly, 2010; Robinson *et al.*, 2010). However, most of these methods cannot easily be extended to time course gene expression data and ignoring the temporal nature of the data is statistically inefficient. Methods have therefore been developed specifically for time-series applications. In the case of gene expression profiles under a single condition, one-sample methods have been developed to discriminate differentially expressed genes from constitutively expressed genes. For example, probabilistic models have been designed for this purpose which use a likelihood-ratio test to rank genes based on a comparison between a dynamic and a constant profile (Angelini *et al.*, 2008; Kalaitzis and Lawrence, 2011).

When expression profiles are available from two or more conditions then a two-sample test is more appropriate (Conesa *et al.*, 2006; Kim *et al.*, 2013; Stegle *et al.*, 2010; Storey *et al.*, 2005). Storey *et al.* (2005) apply a polynomial regression model to simulate the temporal behaviour of genes and a statistical test to identify differentially expressed genes. Conesa *et al.* (2006) adopt a two-step regression model in analyzing temporal profiles of genes with time treated as an extra experimental factor. Kim *et al.* (2013) apply Fourier analysis to time course gene expression data and identify differentially expressed genes in the Fourier domain. Stegle *et al.* (2010) apply a model based on Gaussian Process (GP) regression which is closely related to our proposed approach. In this model, when two time series are the same they are represented by a shared GP function but where they differ they are better represented by two independent GP functions. Binary latent variables are used to model whether a particular time interval is better represented by two independent GPs or one combined GP. More recently, the GP regression framework has been refined through use of a non-stationary covariance function and a simplified scoring approach to detect time periods of differential gene expression (Heinonen *et al.*, 2014). Similar to the work of Stegle *et al.* (2010), a log-likelihood ratio is used to identify time periods of differential expression. In order to better adapt to the case where unevenly or sparsely distributed times are used, they introduce a non-stationary covariance function and proposed two novel likelihood ratio tests to evaluate the likelihood at arbitrary time points. All these approaches can be used to find differentially expressed genes and some can be used to identify temporal domains where there is support for profiles being different. However, these methods do not directly score the probability of the perturbation time where two profiles first diverge, which is the aim of our approach. Although the methods of Stegle *et al.* (2010) and Heinonen *et al.* (2014) can be adapted to provide an estimate of the perturbation time, e.g. by applying a thresholding procedure to their differential expression scores, we show here that direct inference of the perturbation time is a more powerful approach when that is the object of interest.

In this paper, we propose a method to identify the perturbation point given data from two time course experiments. We use a non-parametric GP to describe the joint posterior distribution of two time profiles which are equal up to a proposed perturbation time. The perturbation time is then a model parameter which can be inferred. We derive the covariance function of the GP model and show that the likelihood function is exactly tractable. The posterior distribution of the perturbation time can be computed through a simple one-dimensional histogram approach, with no assumptions over the shape of the posterior distribution and no need to resort to complex approximate inference schemes. This differs from Stegle *et al.* (2010) and Heinonen *et al.* (2014) in that we focus specifically on inferring the perturbation time and derive an exact approach to this problem. Stegle *et al.* (2010) creates a mixed model in pre-specified time intervals with the transition between independent GPs and shared GPs. The likelihood in that case must be approximated using Expectation Propagation (EP) due to its non-Gaussian nature. Heinonen *et al.* (2014) provide a simpler approach by adopting the expected marginal log-likelihood ratio or the noisy posterior concentration ratio to construct a smooth curve indicating time periods of differential expression. However, their approach does not allow direct inference of the perturbation time.

The paper is organized as follows. In Section 2, we present background on GP regression and derive the covariance function, likelihood function and posterior inference procedure for our new model. In Section 3, the algorithm is demonstrated on simulated data and subsequently applied to identify the perturbation times for Arabidopsis genes in a microarray time series dataset detailing the transcriptional changes that occur in Arabidopsis following inoculation with DC3000 versus the disarmed strain DC3000*hrpA* (Lewis *et al.*, 2015) and with a brief conclusion presented in Section 4.

2 Methods

2.1 Gaussian process regression

Gaussian Processes (GPs) (Rasmussen and Williams, 2006) extend multivariate Gaussian distributions to infinite dimensionality and can be used as probabilistic models that specify a distribution over functions (Lawrence, 2005). GPs have been used in a range of gene expression applications, e.g. to model the dynamics of transcriptional regulation (Gao *et al.*, 2008; Honkela *et al.*, 2010) and in temporal differential expression scoring (Heinonen *et al.*, 2014; Kalaitzis and Lawrence, 2011; Stegle *et al.*, 2010; Yuan, 2006).

We have a dataset $\mathcal{D}$ with N inputs $\mathbf{X} = \{x_n\}_{n=1}^N$ and corresponding real valued targets $\mathbf{Y} = \{y_n\}_{n=1}^N$. In the case of time course data the data are ordered such that $x_n \geq x_{n-1}$ but there is no restriction on the spacing since GPs operate over a continuous domain. We allow the case $x_n = x_{n-1}$ since that provides a simple way to incorporate replicates. We assume that measurement noise in $\mathbf{Y}$, denoted by ϵ, is i.i.d Gaussian distributed $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and the underlying model for $\mathbf{Y}$ as a function of $\mathbf{X}$ is $f(\cdot)$, so that

$$\mathbf{Y} = f(\mathbf{X}) + \epsilon,$$

and $f(\mathbf{X})$ represents the mean of the data generating process. Our prior modelling assumption is that the function f is drawn from a GP prior with mean function $\mu(\mathbf{X})$, covariance function $K(\mathbf{X}, \mathbf{X})$ and hyperparameters θ. We write,

$$f(\mathbf{X}) \sim \mathcal{GP}(\mu(\mathbf{X}), K(\mathbf{X}, \mathbf{X})),$$

and the likelihood of $\mathbf{Y}$ becomes

$$p(\mathbf{Y}|\mathbf{X}, \theta) \sim \mathcal{N}(\mu(\mathbf{X}), K(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I}),$$

where $K(\mathbf{X}, \mathbf{X})$ is the $N \times N$ covariance matrix with elements $K(x_n, x_m)$. The covariance function describes typical properties of the function f, e.g. whether it is rough or smooth, stationary or non-stationary etc. We choose the squared exponential function,

$$K(x_n, x_m) = \alpha \exp\left(-\frac{(x_n - x_m)^2}{2l^2}\right), \quad (1)$$

with hyper-parameters $\theta = (\alpha, l)$ specifying the amplitude and length-scale of samples drawn from the prior. This choice corresponds to a prior assumption of smooth and stationary functions. However, our model can be applied with any other choice of covariance function, e.g. the non-stationary covariance introduced by Heinonen *et al.* (2014). The hyper-parameters can be estimated from the data by maximum likelihood or through a Bayesian procedure (Rasmussen and Williams, 2006). We can also consider the noise variance, σ^2, as an additional hyper-parameter to be estimated similarly.

A typical regression analysis will be focused on a new input x_* and its prediction f_*. Based upon Gaussian properties (Rasmussen and Williams, 2006) the posterior distribution of f_* given data $\mathbf{Y}$ is $p(f_*|\mathbf{Y}) \sim \mathcal{N}(\mu_*, C_*)$ with

$$\mu_* = K(\mathbf{X}, x_*)^\top (K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I})^{-1} \mathbf{Y},$$

$$C_* = K(x_*, x_*) - K(\mathbf{X}, x_*)^\top (K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I})^{-1} K(\mathbf{X}, x_*).$$

We see then that the posterior distribution is also a GP but it is adapted to the data. The mean prediction is a weighted sum over data with weights larger for nearby points in a manner determined by the covariance function. The posterior covariance captures our uncertainty in the inference of $f*$ and will typically be reduced as we incorporate more data.

A special case of GP regression, which is useful in deriving our model below, is the case where $(\mathbf{X}, \mathbf{Y})$ is a single point (x_p, u) measured with zero noise. In this case the GP regression of all new points $\overline{X}$ given (x_p, u) is then

$$p(f(\overline{\mathbf{X}})|\mathbf{Y}) \sim \mathcal{N}(\mu(\overline{\mathbf{X}}), C(\overline{\mathbf{X}}, \overline{\mathbf{X}})), \tag{2}$$

with

$$\mu(\overline{\mathbf{X}}) = \frac{K(\overline{\mathbf{X}}, x_p) u}{K(x_p, x_p)}, \tag{3}$$

$$C(\overline{\mathbf{X}}, \overline{\mathbf{X}}) = K(\overline{\mathbf{X}}, \overline{\mathbf{X}}) - \frac{K(\overline{\mathbf{X}}, x_p) K(\overline{\mathbf{X}}, x_p)^\top}{K(x_p, x_p)}. \tag{4}$$

2.2 Joint distribution of two functions constrained to cross at one point

Consider the case where two time profiles, $f(\mathbf{X})$ and $g(\mathbf{Z})$, evaluated at specified sets of time points $\mathbf{X}$ and $\mathbf{Z}$, respectively, cross at the point x_p with $f(x_\mathrm{p}) = g(x_\mathrm{p}) = u$ at the crossing point. Before considering the constraint we use the same GP prior for each function with hyperparameters θ,

$$f(\mathbf{X}) \sim \mathcal{GP}(\mu(\mathbf{X}), K(\mathbf{X}, \mathbf{X})), \quad g(\mathbf{Z}) \sim \mathcal{GP}(\mu(\mathbf{Z}), K(\mathbf{Z}, \mathbf{Z})).$$

Imposing the constraint that the functions cross at x_p is equivalent to observing a data point (x_p, u) with zero noise. Then $p(f|\mathbf{X}, u)$ and $p(g|\mathbf{Z}, u)$ are as in Eq. (2),

$$p(f(\mathbf{X})|u) \sim \mathcal{N}(\mu_\mathbf{X}, C_\mathbf{X}), \quad p(g(\mathbf{Z})|u) \sim \mathcal{N}(\mu_\mathbf{Z}, C_\mathbf{Z}),$$

with

$$\mu_\mathbf{X} = \frac{K(\mathbf{X}, x_\mathrm{p}) u}{K(x_\mathrm{p}, x_\mathrm{p})}, \quad C_\mathbf{X} = K(\mathbf{X}, \mathbf{X}) - \frac{K(\mathbf{X}, x_\mathrm{p}) K(\mathbf{X}, x_\mathrm{p})^\top}{K(x_\mathrm{p}, x_\mathrm{p})},$$

$$\mu_\mathbf{Z} = \frac{K(\mathbf{Z}, x_\mathrm{p}) u}{K(x_\mathrm{p}, x_\mathrm{p})}, \quad C_\mathbf{Z} = K(\mathbf{Z}, \mathbf{Z}) - \frac{K(\mathbf{Z}, x_\mathrm{p}) K(\mathbf{Z}, x_\mathrm{p})^\top}{K(x_\mathrm{p}, x_\mathrm{p})},$$

In practice, the time profiles $f(\mathbf{X})$ and $g(\mathbf{Z})$ are typically measured at the same time points, so that $\mathbf{Z}$ can be replaced by $\mathbf{X}$. The value of the functions at the crossing point, u, is not known and we marginalize it out using the prior Gaussian distribution $u \sim \mathcal{N}(0, K(x_p, x_p))$. The joint probably distribution of f and g is then given by Eq. (5) below,

$$\begin{aligned} p(f(\mathbf{X}), g(\mathbf{X})) &= \int p(f|\mathbf{X}, u) p(g|\mathbf{X}, u) p(u) du, \\ &\propto \exp\left(-\frac{1}{2} (f \quad g) \Sigma^{-1} (f \quad g)^\top\right), \end{aligned} \tag{5}$$

so that the two functions are jointly Gaussian distributed as $\mathcal{N}(0, \Sigma)$ with covariance given by,

$$\Sigma = \begin{pmatrix} K_{ff} & K_{fg} \\ K_{gf} & K_{gg} \end{pmatrix} = \begin{pmatrix} K_\mathbf{X} & \frac{k_\mathbf{X} k_\mathbf{X}^\top}{k_{x_p}} \\ \frac{k_\mathbf{X} k_\mathbf{X}^\top}{k_{x_p}} & K_\mathbf{X}, \end{pmatrix}, \tag{6}$$

where $K_\mathbf{X}$, k_{x_p} and $k_\mathbf{X}$ are abbreviations for $K(\mathbf{X}, X)$, $K(x_p, x_p)$ and $K(\mathbf{X}, x_p)$, respectively. We show an example of this covariance function in Figure 1 (upper panel) for $\mathbf{X}$ in the range [0,100] and $x_p = 40$. The detailed derivations of Eqs. (5) and (6) are illustrated in the Supplementary.

2.3 The data likelihood under the model

We define the perturbation time x_p as the point where two time profiles first begin to diverge. If the time profiles are measured without noise then it would be trivial to identify this point. However, biological time course data from high-throughput experiments are often corrupted by significant biological and technical sources of noise and our task is to *infer* the perturbation time given noisy time course data. In order to do that we must first derive the likelihood function under the new model.

Let two sets of gene expression time course data, $y^c(\mathbf{X})$ and $y^p(\mathbf{X})$, represent noisy measurements with i.i.d Gaussian measurement noise, $\mathcal{N}(0, \sigma^2 \mathbf{I})$, from the control condition and perturbed condition, respectively. A GP prior is placed on the mean functions

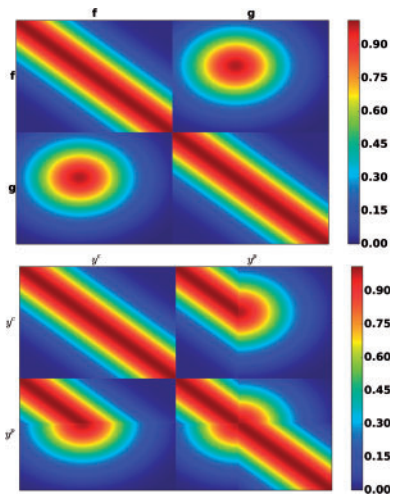

Fig. 1. Illustration of the covariance matrix, Σ, for two functions *f* and *g* evaluated at points evenly distributed in [0,100] and crossing at $x_p = 40$ (upper) and the resulting data covariance matrix, $\widehat{\Sigma}$, for time course data y^c and y^p from a wild-type and perturbed system respectively (lower) (Color version of this figure is available at *Bioinformatics* online.)

underlying y^c and y^p and a time point x_p is defined as the perturbation time point. The data model is defined as:

1. The two datasets y^c and y^p before x_p are noise-corrupted versions of the same underlying mean function f which has a GP prior,

$$y^c(x_n) \sim \mathcal{N}(f(x_n), \sigma^2),$$

$$y^p(x_n) \sim \mathcal{N}(f(x_n), \sigma^2) \quad \text{for } x_n \leq x_p\,.$$

2. The mean function for y^c stays intact after x_p while the mean function for y^p changes to follow g,

$$y^c(x_n) \sim \mathcal{N}(f(x_n), \sigma^2),$$

$$y^p(x_n) \sim \mathcal{N}(g(x_n), \sigma^2) \quad \text{for } x_n > x_p,$$

where f and g are constrained to cross at x_p and follow the GP described in Eq. (5).

The joint distribution of y^c and y^p is then

$$p(y^c(\mathbf{X}), y^p(\mathbf{X})|x_p) = \exp\left(-\frac{1}{2}\begin{pmatrix} y^c \\ y^p \end{pmatrix}^\top \widehat{\Sigma}^{-1} \begin{pmatrix} y^c \\ y^p \end{pmatrix}\right), \tag{7}$$

where the covariance matrix $\widehat{\Sigma}$ can be worked out in terms of the covariance matrix Σ for the joint distribution of f and g defined by Eq. (6),

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{K}_{y^c y^c} & \widehat{K}_{y^c y^p} \\ \widehat{K}_{y^p y^c} & \widehat{K}_{y^p y^p} \end{pmatrix}, \tag{8}$$

with

$$\widehat{K}_{y^c(\mathbf{X}_1)y^c(\mathbf{X}_2)} = K_{f(\mathbf{X}_1)f(\mathbf{X}_2)} + \sigma^2\mathbf{I} \qquad \mathbf{X}_1 \in \mathbf{X}, \mathbf{X}_2 \in \mathbf{X}$$

$$\widehat{K}_{y^c(\mathbf{X}_1)y^p(\mathbf{X}_2)} = \begin{cases} K_{f(\mathbf{X}_1)f(\mathbf{X}_2)} & \mathbf{X}_1 \in \mathbf{X}, \mathbf{X}_2 \leq x_p \\ K_{f(\mathbf{X}_1)g(\mathbf{X}_2)} & \mathbf{X}_1 \in \mathbf{X}, \mathbf{X}_2 > x_p \end{cases}$$

$$\widehat{K}_{y^p(\mathbf{X}_1)y^c(\mathbf{X}_2)} = \begin{cases} K_{f(\mathbf{X}_1)f(\mathbf{X}_2)} & \mathbf{X}_1 \leq x_p, \mathbf{X}_2 \in \mathbf{X} \\ K_{g(\mathbf{X}_1)f(\mathbf{X}_2)} & \mathbf{X}_1 > x_p, \mathbf{X}_2 \in \mathbf{X} \end{cases}$$

$$\widehat{K}_{y^p(\mathbf{X}_1)y^p(\mathbf{X}_2)} = \begin{cases} K_{f(\mathbf{X}_1)f(\mathbf{X}_2)} + \sigma^2\mathbf{I} & \mathbf{X}_1 \leq x_p, \mathbf{X}_2 \leq x_p \\ K_{g(\mathbf{X}_1)f(\mathbf{X}_2)} & \mathbf{X}_1 > x_p, \mathbf{X}_2 \leq x_p \\ K_{f(\mathbf{X}_1)g(\mathbf{X}_2)} & \mathbf{X}_2 > x_p, \mathbf{X}_1 \leq x_p \\ K_{g(\mathbf{X}_1)g(\mathbf{X}_2)} + \sigma^2\mathbf{I} & \mathbf{X}_1 > x_p, \mathbf{X}_2 > x_p \end{cases}$$

The lower panel in Figure 1 shows the data covariance matrix $\widehat{\Sigma}$ for $\mathbf{X}$ evenly spread in the range $[0, 100]$ and with a perturbation occurring at $x_p = 40$.

2.4 Posterior distribution of the perturbation point

According to Bayes' rule the posterior distribution of x_p is,

$$p(x_p|y^c(\mathbf{X}), y^p(\mathbf{X})) = \frac{p(y^c(\mathbf{X}), y^p(\mathbf{X})|x_p)p(x_p)}{\int p(y^c(\mathbf{X}), y^p(\mathbf{X})|x_p)p(x_p)dx_p}.$$

We assume a uniform prior on x_p within the range $[x_{\min}, x_{\max}]$ of the observed data. We use a simple discretization $x_p \in [x_{\min}, x_{\min} + \delta, x_{\min} + 2\delta, \ldots, x_{\max}]$ in this range. Then the posterior can be approximated as a simple summation over this grid,

$$p(x_p|y^c(\mathbf{X}), y^p(\mathbf{X})) \simeq \frac{p(y^c(\mathbf{X}), y^p(\mathbf{X})|x_p)}{\sum_{x=x_{\min}}^{x=x_{\max}} p(y^c(\mathbf{X}), y^p(\mathbf{X})|x)},$$

only requiring that we evaluate the likelihood at each grid point. There are hyper-parameters θ also involved in the posterior distribution of x_p which would potentially complicate matters. We choose to estimate these hyper-parameters prior to inferring x_p. To do this we use maximum likelihood optimization for the case where x_p approaches $-\infty$ which corresponds to the two GPs for the control and perturbed conditions being independent,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left(\lim_{x_p \to -\infty} p_\theta(y^c(\mathbf{X}), y^p(\mathbf{X})|x_p, \theta) \right).$$

Since we have a simple histogram representation for the posterior distribution of the perturbation time point x_p then we can easily estimate the mean, median or mode (MAP) of the posterior distribution to provide a point estimate.

2.5 Pre-filtering to remove non-DE genes

In many applications a large number of genes will show no strong evidence for DE at any time or will have a low signal-to-noise due to being weakly expressed. We therefore filter genes prior to using our model. A DE gene will be better represented by two independent GPs rather than a shared GP under control and perturbed conditions. We therefore filter genes using the log-likelihood ratio r between the independent GP model (equivalent to x_p approaching $-\infty$ in the perturbation model) and the integrated GP (with x_p approaching $+\infty$):

$$r = \log\mathcal{L}(y_c(\mathbf{X}), y_p(\mathbf{X})|x_p \to -\infty) - \log\mathcal{L}(y_c(\mathbf{X}), y_p(\mathbf{X})|x_p \to +\infty)$$

We note that it is difficult to distinguish genes with a late perturbation time from those that are non-DE and our filtering approach may remove some genuine late perturbation genes. In many applications we are primarily interested on relatively early perturbations (e.g. in the application considered here) in which case this will not significantly impact the results. In the Supplementary we consider an alternative filtering approach which is based on detecting genes with time-varying profile in either the control or perturbed condition and is therefore less likely to filter out late x_p genes.

The method has been implemented in the DEtime R-package (github.com/ManchesterBioinference/DEtime) and also as the DEtime kernel in the GPy Python package (github.com/SheffieldML/GPy). The running time for the whole genome (32 578 genes) for the example in Section 3.3 on a Intel(R) Core(TM) i7-3770 CPU of 3.40 GHz is around 11 h using the DEtime R-package.

3 Results and discussion

3.1 Generating simulated data

We generated data under a range of different scenarios to explore performance and robustness to deviations from the model. We generated expression profiles from three different covariance models, one matching the one used for inference and the other two generating rougher profiles. We then add noise using three different noise models, one matching the Gaussian model used for inference and two from heavier-tailed distributions.

1. *profile*$_1$: simulated noise-free profile generated from the model $\mathcal{G}\mathcal{P}_\theta(\mathbf{0}, \widehat{\Sigma}_\theta)$ with $\widehat{\Sigma}_\theta$ given in Eq. (8) assuming a squared exponential covariance function (recall Eq. (1)) with the hyperparameters $\theta = \{\alpha = 30.0, l = 2.0\}$.
2. *profile*$_2$: simulated noise-free profile generated from above model with the covariance function in the form of a matern32

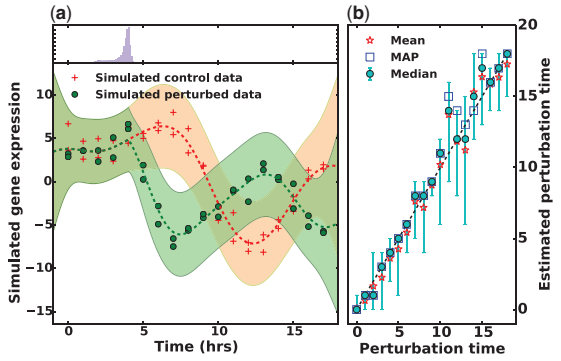

Fig. 2. (a) The shaded area in the lower panel represents the 95% credible region of the GP regression result. In the top panel we show the inferred posterior distribution for the perturbation time x_p. (b) The mean, mode and median of the posterior distribution of x_p with the 5-95 percentile coverage of the posterior distribution for 19 simulated dataset at different perturbation time points (dashed line shows the ground truth) (Color version of this figure is available at *Bioinformatics* online.)

covariance function (see Rasmussen and Williams, 2006) with the same hyperparameters as above.

3. *profile*$_3$: simulated noise-free profile generated from above model with the covariance function in the form of a matern12 covariance function (an Ornstein-Uhlenbeck (OU) process) with the same hyperparameters as above.

Nine simulated dataset are induced with different kinds of i.i.d noise on top of *profile*$_1$, *profile*$_2$ and *profile*$_3$, respectively: Gaussian $\mathcal{N}(1.5)$, Student-t distributed with 3 ($T(3)$) and 6 ($T(6)$) degrees of freedom. The simulated data are sampled every hour from 0 h until 18 h. We simulate data with a range of perturbation times $\mathbf{x}_p \in \{0, 1, \ldots, 17, 18\}$ and 100 different sets of data are produced for each x_p value.

Figure 2(a) shows an example of simulated data (using the *profile*$_1$+$\mathcal{N}(1.5)$ scenario) with two replicates and a perturbation at 4 h. The estimated posterior distribution of x_p is shown in the upper panel and in the lower panel we show the GP regression function after fixing x_p at the MAP value. In this case the MAP estimate for x_p is very close to the ground truth. The mean, mode and median of the posterior distribution of x_p for 19 simulated datasets are illustrated in Figure 2(b) together with the 5–95 percentile coverage of the posterior distribution. It is clear that the posterior distributions of the perturbation time cover the actual perturbation time to a great extent and that the three different point estimates are typically close to the ground truth values.

3.2 Comparison with a thresholding approach

Related methods have been introduced to identify regions of differential expression from time course data (Heinonen *et al.*, 2014; Stegle *et al.*, 2010). Such methods can in principle also be used to identify the perturbation time by locating the first time point where the DE score passes some threshold value. Here we compare our approach to the most recently published package of this type, developed by Heinonen *et al.* (2014) implemented in the nsgp R-package. The nsgp package infers the differentially expressed time periods and uses four likelihood ratios: marginal log-likelihood ratio (MLL), expected marginal log-likelihood ratio (EMLL), the posterior concentration (PC) and the noisy posterior concentration (NPC) to quantify these regions. We adopt thresholds of 0.5 and 1.0 to define the initial perturbation points, respectively. The mean, median and mode of the posterior distribution of the inferred perturbation points from our method are also computed. The performance of ranking x_p using each method is measured by Spearman's rank correlation coefficient with the known ground truth and the mean and standard deviation of the rank correlation coefficients across 100 dataset are illustrated in Table 1.

From the table, it is clear that the mean, median and MAP estimates from the DEtime package provide better ranking performance. The results from the nsgp package vary significantly through different ratios and thresholds, among which, EMLL with threshold 1.0 performs the best in this task, giving rank correlation coefficient of 0.67 ± 0.16 when tested on the simulated *profile*$_1$ contaminated with Gaussian noise $\mathcal{N}(1.5)$, which is still considerably lower than the rank correlation coefficients from mean, median or mode of the DEtime package. In order to compare the performance of the algorithm on data with varied signal-to-noise ratios, we adjusted the signal amplitude hyperparameter α and compared the results from DEtime and nsgp with $\alpha = 1.5, 10.0, 20.0, 30.0$. Supplementary Table S1 illustrates the results which shows the robustness of the proposed model. Supplementary Figure S1 shows the errorbar of the mean, median, mode from DEtime package and EMLL with thresholds of 0.5 and 1.0 from nsgp package across 100 replicates along all perturbation times for all simulated datasets. We observe that the DEtime package provides reasonable estimation of the initial perturbation time under various noise distributions whereas the performance of the EMLL ratio from nsgp package varies substantially and its performance seems to be deteriorating with later initial perturbations.

We note that methods in the nsgp package are not designed specifically for the task of inferring the initial perturbation point as they were proposed for the more general problem of identifying DE regions. Nevertheless, a common application of time-series DE studies is to distinguish early and late DE events. We have demonstrated that one can obtain greater accuracy by focusing on this specific task rather than adapting a more general DE method.

3.3 Bacterial infection response in *A. thaliana*

To determine the biological utility of estimating perturbation times, we re-examined a large dataset recently published by Lewis *et al.* (2015) that captures the transcriptional reprogramming associated with defence and disease development in *A. thaliana* leaves inoculated with *P. syringae* pv. tomato DC3000 and the non-pathogenic DC3000hrpA mutant strain. The differences in gene expression between these two challenges is a result of the action of virulence factors delivered by the DC3000 strain into the plant cell, in this case predominately the collaborative activities of 28 bacterial effector proteins. Figure 3 shows examples of an early and late perturbed gene identified by our method. A preliminary investigation of the perturbation times of differentially expressed genes revealed two peak times (Supplementary Fig. S2), allowing genes to be assigned to one of three groups: early, intermediate and late perturbed genes. This initial characterization was consistent with major phase changes in the infection process, and the onset of effector mediated transcriptional reprogramming: effectors are not delivered into plant cells until 90–120 min post inoculation (Grant *et al.*, 2000), and do not promote bacterial growth until ~8 hpi, when they have effectively disabled host defence processes. This general progression is reflected in GO and pathway analysis outlined in Supplementary Section 4.

The recent study by Lewis *et al.* (2015) provided a comprehensive overview of the transition from defence to disease. Thus we investigated if the calculation of perturbation times provided

Table 1. Comparison of the means and stds of the Spearman's rank correlation coefficients of the mean, median and MAP estimation of the perturbation times from our DEtime package and different likelihood ratios with various thresholds from the nsgp package. M_n represents the M ratio with a threshold of n.

R Package	DEtime			nsgp						
Data	Mean	Median	MAP	$MLL_{0.5}$	$EMLL_{0.5}$	$PC_{0.5}$	$NPC_{0.5}$	$MLL_{1.0}$	$EMLL_{1.0}$	$PC_{1.0}$	$NPC_{1.0}$
$profile_1$+$\mathcal{N}(1.5)$	0.94 ± 0.04	0.94 ± 0.04	0.93 ± 0.05	−0.02 ± 0.23	0.26 ± 0.22	0.36 ± 0.22	0.26 ± 0.23	−0.02 ± 0.23	0.67 ± 0.16	0.29 ± 0.22	0.48 ± 0.21
$profile_2$+$\mathcal{N}(1.5)$	0.90 ± 0.06	0.90 ± 0.06	0.88 ± 0.08	−0.05 ± 0.23	0.17 ± 0.23	0.21 ± 0.25	0.24 ± 0.22	−0.06 ± 0.23	0.57 ± 0.20	0.18 ± 0.24	0.42 ± 0.21
$profile_3$+$\mathcal{N}(1.5)$	0.93 ± 0.04	0.93 ± 0.04	0.92 ± 0.05	−0.04 ± 0.26	0.31 ± 0.22	0.21 ± 0.25	0.28 ± 0.25	−0.05 ± 0.26	0.75 ± 0.16	0.20 ± 0.23	0.47 ± 0.23
$profile_1$+$T(6)$	0.93 ± 0.05	0.92 ± 0.06	0.91 ± 0.07	0.01 ± 0.24	0.22 ± 0.24	0.24 ± 0.27	0.23 ± 0.24	0.02 ± 0.23	0.59 ± 0.18	0.17 ± 0.23	0.40 ± 0.23
$profile_2$+$T(6)$	0.87 ± 0.07	0.86 ± 0.07	0.83 ± 0.09	0.04 ± 0.23	0.27 ± 0.25	0.08 ± 0.24	0.15 ± 0.22	0.03 ± 0.23	0.42 ± 0.24	0.02 ± 0.24	0.28 ± 0.24
$profile_3$+$T(6)$	0.89 ± 0.06	0.89 ± 0.06	0.87 ± 0.08	0.01 ± 0.26	0.24 ± 0.24	0.13 ± 0.25	0.26 ± 0.26	−0.00 ± 0.27	0.64 ± 0.20	0.07 ± 0.25	0.41 ± 0.23
$profile_1$+$T(3)$	0.91 ± 0.05	0.90 ± 0.06	0.89 ± 0.06	−0.02 ± 0.24	0.14 ± 0.22	0.19 ± 0.22	0.20 ± 0.21	−0.03 ± 0.24	0.48 ± 0.21	0.15 ± 0.23	0.32 ± 0.21
$profile_2$+$T(3)$	0.83 ± 0.09	0.83 ± 0.10	0.80 ± 0.11	−0.03 ± 0.26	0.08 ± 0.23	0.12 ± 0.23	0.16 ± 0.21	−0.04 ± 0.26	0.36 ± 0.23	0.05 ± 0.22	0.24 ± 0.24
$profile_3$+$T(3)$	0.87 ± 0.07	0.87 ± 0.07	0.84 ± 0.09	−0.01 ± 0.25	0.20 ± 0.21	0.09 ± 0.23	0.20 ± 0.18	0.00 ± 0.25	0.54 ± 0.22	0.09 ± 0.23	0.33 ± 0.23

supporting evidence and additional novel insights not highlighted by Lewis *et al.* (2015). To do so, genes were first grouped according to their GO or AraCyc Pathway annotation, and the cumulative perturbation time for each term calculated. The time at which more than 50% of the genes associated with a particular term were perturbed could then be used to rank terms, allowing a high resolution understanding of the infection process. Heat maps showing the cumulative density function (CDF) of perturbation times for each term are shown in Supplementary Figures S4–S10. For clarity, we chose to focus predominately on the earliest processes perturbed by bacterial effectors as these are predicted to be processes integral to the suppression of innate immunity. As an initial proof of concept we focussed on the perturbation of hormone pathways, as modulation of these pathways are well known to be integral to pathogen virulence strategies (Fig. 4).

First we looked at abscisic acid (ABA) pathways, as it has previously been shown that DC3000 rapidly induces de novo ABA biosynthesis and hijacks ABA signalling pathways to promote virulence (de Torres-Zabala *et al.*, 2007; de Torres Zabala *et al.*, 2009). Figure 4A shows a strong link between various GOs associated with ABA processes and early perturbation, which is what is predicted in the literature and demonstrated by Lewis *et al.* (2015). Amongst these early ABA signalling components induced were the classic ABA responsive TFs, RD26 and both ATAIB and AFP2 were induced around 2 hpi. This prediction suggests that effectors are targeting ABA signaling very early in the infection process. Furthermore > 50% of genes annotated with 'regulation of abscisic acid biosynthetic process' were perturbed by 2.3 hpi, consistent with measurable increased in de novo ABA biosynthesis 6 hpi, (de Torres-Zabala *et al.*, 2007), with subsequent perturbation of 'cellular response to abscisic acid stimulus' occurring by 3.5 hpi. Two genes showing perturbation at 4.1 hpi and annotated as ABA responsive, BLHL1 and TCP14, are predicted to be targeted by the DC3000 effector AvrPto in yeast two hybrid protein–protein interaction studies (Mukhtar *et al.*, 2011). Moreover a knockout of TCP14 results in enhanced disease resistance to DC3000, consistent with TCP14 being a virulence target of effectors (Weßling *et al.*, 2014). Subsequently, a number of ABA related pathways appear to be further targeted later in the infection. Interestingly 'negative regulation of abscisic acid-activated signaling pathway' was perturbed at 4.4 hpi suggesting this is an example of a failed host response Lewis *et al.* (2015). Other notable perturbed ABA related ontologies included 'abscisic acid transport' (4.9 hpi), 'abscisic acid catabolic process' (5.1 hpi), 'abscisic acid binding' (5.1 hpi), 'abscisic acid-activated signaling pathway' (6.3 hpi), 'abscisic acid biosynthetic process' (7.2 hpi) and 'positive regulation of abscisic acid-activated signaling pathway' (7.2 hpi). Thus we can validate the importance of ABA in the infection process but, moreover, using our estimation of perturbation process we can see fine resolution of the increased impact of ABA biosynthesis and signaling on the infection process not evidenced by the previous analyses (Lewis *et al.*, 2015) as illustrated in Figure 4A.

As expected, we also identified strong early perturbations in salicylic acid (Fig. 4B) related ontologies, as these are key targets for effector mediated suppression (DebRoy *et al.*, 2004). For further validation, we looked at ontologies associated with the hormone jasmonic acid (Fig. 4C). The JA ontologies show more delayed perturbation than ABA, particularly notably the ontologies associated with 'response to jasmonic acid' (2.3 hpi), 'jasmonic biosynthetic processes' (3.7 hpi) and 'regulation of jasmonic acid mediate signaling pathways' (3.8 hpi). This is consistent with the recent study by de Torres *et al.* (2015) using a specific targeted analysis of the same

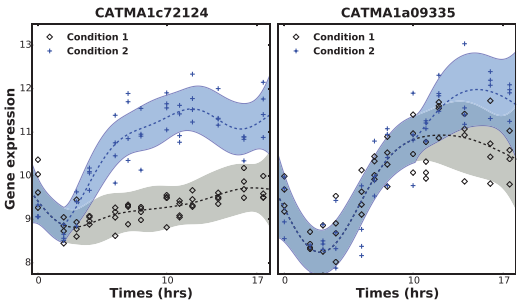

Fig. 3. Examples of fitting the DEtime model to an early-perturbed (left) and late-perturbed gene from an experiment comparing arabadopsis leaves collected from plants infected with DC3000 (condition 1) and the mutant DC3000*hrpA* (condition 2). The shaded area represents the 95% credible region of the GP and the dashed line is the estimated mean of the model (Color version of this figure is available at *Bioinformatics* online.)

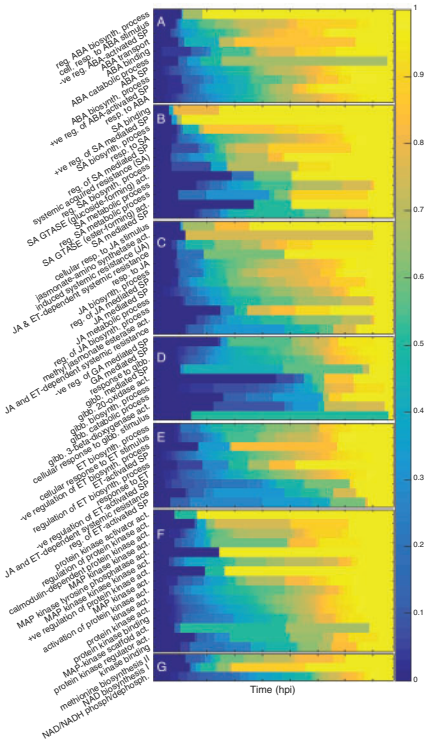

Fig. 4. The CDF of inferred perturbation times of gene sets associated with hormone, signalling and metabolism related gene ontology terms. Abbreviations: reg., regulation; biosynth., biosynthesis; SP, signalling pathway; resp., response; act., activity; gibb., gibberellin; phos., phosphorylation; dephosph., dephosphorylation; GTASE, glucosyltransferase (Color version of this figure is available at *Bioinformatics* online.)

dataset which demonstrated that the JA contribution to DC3000 pathogenesis was preceded by a stronger ABA component. Thus both the ABA and JA analyses provide two examples that validate the utility of the perturbation estimation approach. Two other hormone signalling pathways, gibberellic acid (Fig. 4D) and ethylene (Fig. 4E), are predicted to play a minor role in establishment of virulence, with their contributions only occurring late in the infection process.

We next identified two signalling and two primary metabolism pathways that are predicted to be important in the early conflict between plant defence and pathogen virulence: MAP kinase kinase (MAPKK) activity, regulation of protein kinase activity, NAD biosynthesis process and methionine biosynthesis (Fig. 4F/G).

MAMP signaling activates an early kinase phosphorylation cascade that initiates transcriptional activation (Zipfel, 2014), however little is known about the transcriptional activation or kinases. Remarkably, 8 out of the 10 MAPKKs encoded by the Arabidopsis genome were perturbed early. Given that these MAPKKs are responsible for phosphorylation of the 20 downstream MAPKs their respective roles are naturally extensive. However, MAPKKs are strongly implicated in biotic stress. Most notably, the DC3000 effector HopF2 can interact with Arabidopsis MKK5 and most likely other MAPKKs to inhibit MAPKs and PAMP-triggered immunity. This is probably through MAPKK inhibition via ADP-ribosylation as HopF2 delivery inhibited PAMP-induced MPK phosphorylation (Wang *et al.*, 2010). Functional evidence for a positive role of MKKs in defence comes from work in tobacco, where transient expression of AtMKK7/AtMKK9 and AtMKK4/AtMKK5 caused a hypersensitive response (Zhang *et al.*, 2008). However, the roles of MKKs are likely to be multifunctional and may be manipulated by effectors to promote virulence. The MAPKK, MKK1 was shown to negatively regulate immunity (Kong *et al.*, 2012). This may be through a dual role in activating ABA signalling as AtMKK1 as well as AtMKK2 and AtMKK3, could activate the ABA responsive RD29A promoter and MKK8 could activate the RD29B promoter (HUA *et al.*, 2006). Concomitant with perturbation of the MKK pathway was a significant early perturbation of a sets of genes associated with regulation of protein kinase activity. Strikingly, these genes belong to a class of evolutionarily conserved kinases functioning as metabolic sensors and are activated in response to declining energy levels. Their co-regulation is probably because they typically function as a heterotrimeric complex comprising two regulatory subunits, β and γ and an α-catalytic subunit. Intriguingly, a recent study predicted that the two clade A type 2C protein phosphatases that are negative regulators of ABA signalling, ABI1 and PP2CA, negatively regulate the Snf1-related protein kinase1 and that PP2C inhibition by ABA results in SnRK1 activation (Rodrigues *et al.*, 2013). Moreover, SnRK1 and ABA were shown to induce largely overlapping transcriptional responses, thus these data reveal a previously unknown link between ABA and energy signalling during DC3000 infection.

A pathway intimately linked to energy signalling and redox reactions is NAD biosynthesis, one of the most significantly perturbed pathways following effector delivery (Fig. 4G). Although powdery mildew infection of barley leaves was reported to be associated with increased NAD content more than 40 years ago (Ryrie and Scott, 1969) and recently the identification of the fin4 (flagellin insensitive 4) mutant as aspartate oxidase (Macho *et al.*, 2012), a precursor of the NANP biosynthetic pathway, the role of pyridines in plant defence has received little attention. NAD and NADP play crucial roles in pro-oxidant and antioxidant metabolism and have been linked to biotic stress responses, including production of nitric oxide and metabolism of reactive lipid derivatives (Crawford and Guo, 2005; Mano *et al.*, 2005). We highlight two possible, and contrasting, roles for rapid induction of NAD biosynthesis components by effectors. First, it has recently been shown that chloroplast ROS production is influenced by NADP:NADPH ratios and bacteria effector delivery rapidly suppresses a MAMP triggered chloroplast burst of hydrogen peroxide in an ABA dependent manner (de Torres Zabala *et al.*, 2015). Second, poly(ADP-Ribose) polymerases (PARPs) is emerging as a key regulator of defence responses. PARPs are important NAD+ consuming enzymes induced by biotic stress, polymerizing long poly(ADP-ribose) chains on target proteins including histones. Adams-Phillips *et al.* (2010) reported a 40–50% decrease in NAD+ 12 hpi of DC3000 challenged leaves compared to a mock

control and ~50% increase in total cellular and nuclear poly(ADP-Rib) polymers (Adams-Phillips *et al.*, 2010). Consistent with these results, a knockout of PARP2, which is induced by MAMPs, restricts DC3000 growth (Song *et al.*, 2015) demonstrating that loss of poly(ADP-ribosyl)ation activity affects the capacity of Arabidopsis to limit DC3000 growth.

The second primary metabolism example we choose to highlight is the very rapid induction methionine biosynthesis pathway (Fig. 4G). Methionine is a sulphur amino acid involved in multiple cellular processes from being a protein constituent, to initiation of mRNA translation as well as functioning as a regulatory molecule in the form of S-adenosylmethionine (SAM). There are 13 unique genes associated with this ontology, and while it is outside the scope of this manuscript to explore these in detail it is worth noting that this includes DMR1 (Downy Mildew Resistance 1) (van Damme *et al.*, 2009), encoding homoserine kinase, which produces O-phospho-L-homoserine, a compound at the branching point of methionine and threonine biosynthesis. Mutations in dmr1 lead to elevated foliar homoserine and resistance to the biotrophic pathogens *Hyaloperonospora arabidopsidis*, *Oidium neolycopersici*, *F. culmorum* and *F. graminearum*, although the mechanism has yet to be identified (Brewer *et al.*, 2014; Huibers *et al.*, 2013; van Damme *et al.*, 2009).

Thus in summary, we have validated perturbation times against previous analyses, and provide four new examples derived from examining early perturbation times of biological pathways to identify novel signalling and, particularly, primary metabolic pathways that are implicated in the transition from defence to disease following infection with DC3000. These examples provide compelling leads for further investigation.

4 Conclusion

We have introduced a fully Bayesian approach to infer the initial point where two gene expression time profiles diverge using a novel GP regression approach. We model the data as noise-corrupted samples coming from a shared function prior to some 'perturbation time' after which it splits into two conditionally independent functions. The full posterior distribution of the perturbation point is obtained through a simple histogram approach, providing a straightforward method to infer the divergence time between two gene expression time profiles under different conditions. The proposed method is applied to a study of the timing of transcriptional changes in *A. thaliana* under a bacterial challenge with a wild-type and disarmed strain. Analysis of differences in the gene expression profiles between strains is shown to be informative about the immune response.

Many transcriptional perturbation experiments are focused on a single perturbation. However, multiple perturbations occurring at different times or a single perturbation targeting many conditions will be needed to unmask complex gene regulatory strategies. An interesting future line of research would be the development of GP covariance structures to uncover the ordering of events under these more general scenarios.

Funding

M.R. and J.Y. were supported by the EU FP7 project RADIANT (Grant 305626 to M.R.) and MRC award MR/N00017X/1, M.R.G. and C.A.P. by the BBSRC [BB/F005806/1 to M.R.G. and C.A.P.] and EPSRC [EP/I036575/1 to C.A.P.] UK research councils.

Conflict of Interest: none declared.

References

Adams-Phillips,L. *et al.* (2010) Disruption of poly (adp-ribosyl) ation mechanisms alters responses of arabidopsis to biotic stress. *Plant Physiol.*, **152**, 267–280.

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Angelini,C. *et al.* (2008) Bats: a Bayesian user-friendly software for analyzing time series microarray experiments. *BMC Bioinformatics*, **9**, 415.

Bar-Joseph,Z. (2004) Analyzing time series gene expression data. *Bioinformatics*, **20**, 2493–2503.

Brewer,H.C. *et al.* (2014) Mutations in the arabidopsis homoserine kinase gene dmr1 confer enhanced resistance to *F. culmorum* and *F. graminearum*. *BMC Plant Biol.*, **14**, 317.

Conesa,A. *et al.* (2006) masigpro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, **22**, 1096–1102.

Crawford,N.M. and Guo,F.Q. (2005) New insights into nitric oxide metabolism and regulatory functions. *Trends Plant Sci.*, **10**, 195–200.

de Torres,Z.M. *et al.* (2015) Novel JAZ co-operativity and unexpected ja dynamics underpin Arabidopsis defence responses to *Pseudomonas syringae* infection. *New Phytol*, **209**, 1120–1134.

de Torres-Zabala,M. *et al.* (2007) *Pseudomonas syringae* pv. tomato hijacks the *Arabidopsis abscisic* acid signalling pathway to cause disease. *EMBO J.*, **26**, 1434–1443.

de Torres Zabala,M. *et al.* (2009) Antagonism between salicylic and abscisic acid reflects early host–pathogen conflict and moulds plant defence responses. *Plant J.*, **59**, 375–386.

de Torres Zabala,M. *et al.* (2015) Chloroplasts play a central role in plant defence and are targeted by pathogen effectors. *Nat. Plants*, **1**, 15074.

DebRoy,S. *et al.* (2004) A family of conserved bacterial effectors inhibits salicylic acid-mediated basal immunity and promotes disease necrosis in plants. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 9927–9932.

Dudoit,S. *et al.* (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.*, **12**, 111–140.

Gao,P. *et al.* (2008) Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, **24**, i70–i75.

Grant,M. *et al.* (2000) The RPM1 plant disease resistance gene facilitates a rapid and sustained increase in cytosolic calcium that is necessary for the oxidative burst and hypersensitive cell death. *Plant J.*, **23**, 441–450.

Hardcastle,T.J. and Kelly,K.A. (2010) bayseq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.

Heinonen,M. *et al.* (2014) Detecting time periods of differential gene expression using Gaussian processes: an application to endothelial cells exposed to radiotherapy dose fraction. *Bioinformatics*.

Honkela,A. *et al.* (2010) Model-based method for transcription factor target identification with limited data. *Proc. Natl. Acad. Sci.*, **107**, 7793–7798.

Hua,Z.M. *et al.* (2006) Activation of the nacl-and drought-induced RD29A and RD29B promoters by constitutively active Arabidopsis MAPKK or MAPK proteins. *Plant, Cell Environ.*, **29**, 1761–1770.

Huibers,R.P. *et al.* (2013) Powdery mildew resistance in tomato by impairment of SlPMR4 and SlDMR1. *PLoS ONE*, **8**, e67467.

Kalaitzis,A.A. and Lawrence,N.D. (2011) A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinformatics*, **12**, 180.

Kerr,M.K. *et al.* (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Kim,J. *et al.* (2013) A method to identify differential expression profiles of time-course gene data with Fourier transformation. *BMC Bioinformatics*, **14**, 310.

Kong,Q. *et al.* (2012) The MEKK1-MKK1/MKK2-MPK4 kinase cascade negatively regulates immunity mediated by a mitogen-activated protein kinase kinase kinase in Arabidopsis. *Plant Cell*, **24**, 2225–2236.

Lawrence,N. (2005) Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *J. Mach. Learn. Res.*, **6**, 1783–1816.

Lewis,L. A. *et al*. (2015) Transcriptional dynamics driving MAMP-triggered immunity and pathogen effector-mediated immunosuppression in Arabidopsis leaves following infection with *Pseudomonas syringae* pv tomato dc3000. *Plant Cell*, **27**, 3038–3064.

Macho,A. P. *et al*. (2012) Aspartate oxidase plays an important role in Arabidopsis stomatal immunity. *Plant Physiol*., **159**, 1845–1856.

Mano,J. *et al*. (2005) Protection against photooxidative injury of tobacco leaves by 2-alkenal reductase. Detoxication of lipid peroxide-derived reactive carbonyls. *Plant Physiol*., **139**, 1773–1783.

Mukhtar,M. S. *et al*. (2011) Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science*, **333**, 596–601.

Rasmussen,C.E. and Williams,C.K. (2006) *Gaussian Processes for Machine Learning*, vol. **2**. The MIT Press.

Robinson,M. D. *et al*. (2010) edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Rodrigues,A. *et al*. (2013) ABI1 and PP2CA phosphatases are negative regulators of snf1-related protein kinase1 signaling in Arabidopsis. *Plant Cell*, **25**, 3871–3884.

Ryrie,I. and Scott,K. (1969) Nicotinate, quinolinate and nicotinamide as precursors in the biosynthesis of nicotinamide-adenine dinucleotide in barley. *Biochem. J*., **115**, 679–685.

Song,J. *et al*. (2015) PARP2 is the predominant poly (ADP-ribose) polymerase in Arabidopsis DNA damage and immune responses. *PLoS Genet*., **11**, e1005200.

Stegle,O. *et al*. (2010) A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *J. Comput. Biol*., **17**, 355–367.

Storey,J. D. *et al*. (2005) Significance analysis of time course microarray experiments. *Proc. Natl. Acad. Sci. U. S. A*., **102**, 12837–12842.

van Damme,M. *et al*. (2009) Downy mildew resistance in arabidopsis by mutation of homoserine kinase. *Plant Cell*, **21**, 2179–2189.

Wang,Y. *et al*. (2010) A *Pseudomonas syringae* ADP-ribosyltransferase inhibits Arabidopsis mitogen-activated protein kinase kinases. *Plant Cell*, **22**, 2033–2044.

Weßling,R. *et al*. (2014) Convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life. *Cell Host Microbe*, **16**, 364–375.

Yuan,M. (2006) Flexible temporal expression profile modelling using the Gaussian process. *Comput. Stat. Data Anal*., **51**, 1754–1764.

Zhang,X. *et al*. (2008) The arabidopsis map kinase kinase 7: a crosstalk point between auxin signaling and defense responses? *Plant Signal. Behav*., **3**, 272–274.

Zipfel,C. (2014) Plant pattern-recognition receptors. *Trends Immunol*., **35**, 345–351.