Inferring networks from microarray data

Latent position models for microarray experiment treatments

Nick Heard, Dave Stephens, Chris Holmes[†]

Department of Mathematics, Imperial College

[†] OCGF, Department of Statistics, University of Oxford

9th September, 2005

Outline

- Functional Genomics of *Anopheles Gambiae* and *Plasmodium Falciparum.*
- Bayesian Hierarchical Model-Based Clustering
- Dynamic/Network Models
- Co-clustering, Covariance and Latent Space Models
- A Latent Position Model for Multiple Treatments

I: Mosquito and Parasite Genomics

cDNA array data from Imperial College CMMI

- 2001-2003 :
  - mosquito with parasite infestation
  - tissue specific, cell line, male/female
  - mosquito only under artificial experimental challenge
  - bacterial, chemical, heat shock
- 2002: Mosquito and parasite genomes published.
- 2005 : Affymetrix data, genome wide, transcriptome

Typically multiple time points, sometimes replicated.

Major data fusion issues.

Objectives

- to understand regulatory mechanisms within each organism
- to produce plausible models for the patterns of regulation
- to extract subsets of genes that have similar patterns
- to classify genes to functional classes of interest (i.e. immune defense clusters)
  - biologists often reluctant to use advanced statistical methods
  - computational feasibility is an important factor

“Sophistication buys you nothing"

Regression Models for the Profiles

For gene i at time t

$$y_{i_t} = X(t)\beta + \varepsilon_t$$

where $X(t)$ is a p-vector of specified basis functions of t, β is a p-vector of basis coefficient parameters

Note: $\{\varepsilon_t\}$ is an independent and Gaussian error process; assumption can be relaxed.

Design matrix based on truncated polynomial spline.

Bayesian Clustering

In a Bayesian analysis a conjugate prior specification is used for (β, σ^2)

$$p\left(\beta|\sigma^2\right) \equiv N\left(m, \sigma^2 V\right) \qquad p\left(\sigma^2\right) \equiv IGamma\left(\frac{\alpha}{2}, \frac{\gamma}{2}\right)$$

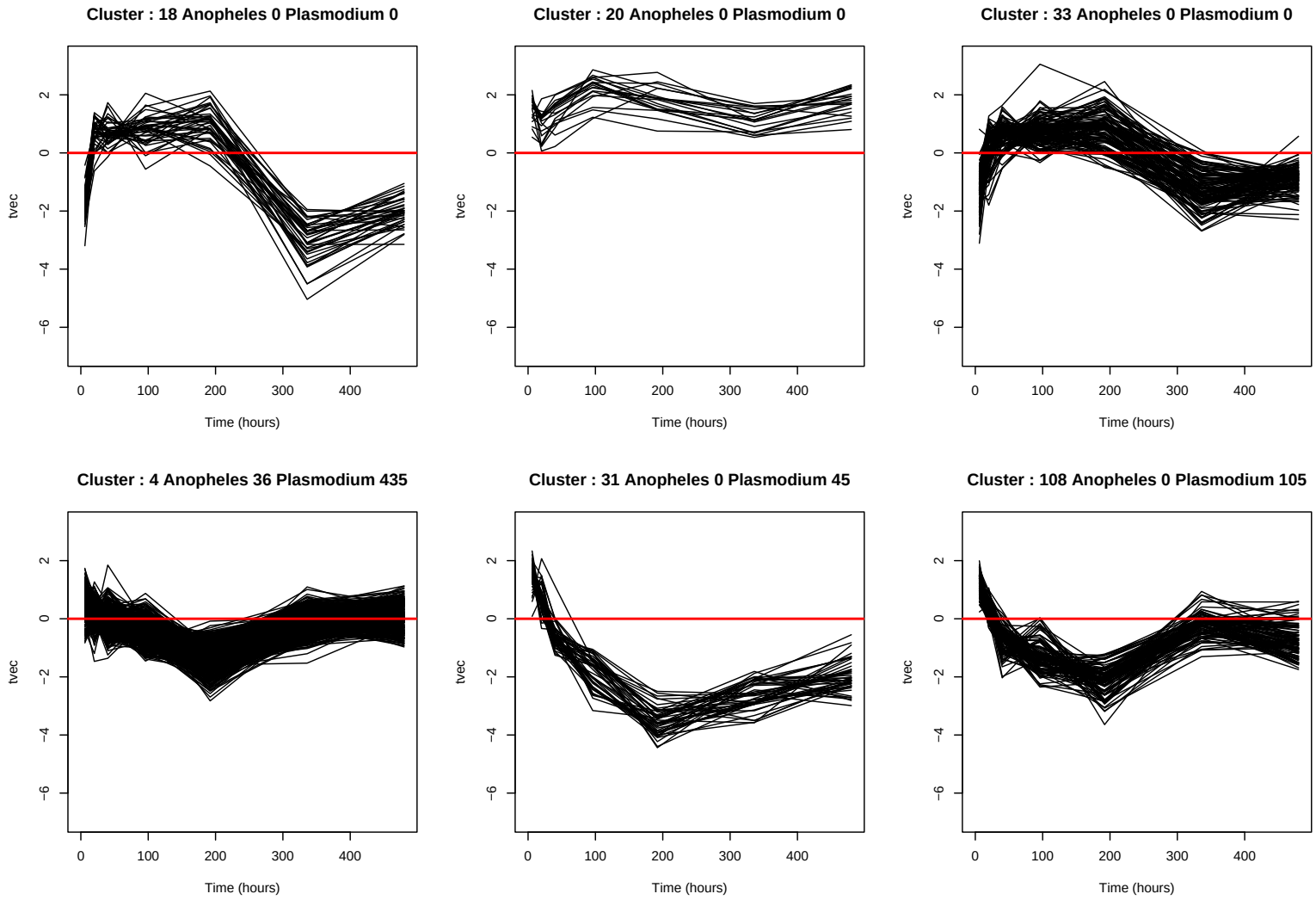
m is $p \times 1$, V is $p \times p$ positive definite and symmetric, all other parameters are scalars. Then

$$p\left(\beta|y, \sigma^2\right) \equiv N\left(m^*, \sigma^2 V^*\right) \; p\left(\sigma^2|y\right) \equiv IGamma\left(\frac{T+\alpha}{2}, \frac{c+\gamma}{2}\right)$$

Clustering approach based on the marginal likelihood which is analytically available, and a hierarchical clustering procedure

Two species data

2800 Genes: 1400 Mosquito/1400 Parasite.



Advantages of the approach

- respects the time ordering of the data
- choice of covariance matrix (induced by the basis function representation) heavily influences resulting clustering.
- outperforms Euclidean clustering, `mclust`.

Full Bayesian Analysis

Agglomerative clustering optimizes only over a very small portion of the space of possible clusterings.

Size of model space implies exhaustive search not possible.

MCMC methods for clustering (mixtures)

Conventional approaches to MCMC in high- or variable-dimension problems (e.g. naive split/merge steps)

DO NOT WORK ! ... maybe ...

- using naive MCMC, the marginal likelihood (or unnormalized posterior) decreases and wanders from the hierarchical optimal clustering
- the vast number of possible cluster configurations cannot be explored fully
- identifying meaningful cluster profiles not negatively affected by this; good enough for future classification purposes

MCMC methods for clustering (mixtures)

Need more advanced MCMC approaches

- simulated tempering
- evolutionary/population MCMC
- stratification
- multivariate DPM models look promising

see e.g. Jasra (2005), Jasra et al. (2005a, b).

II. Regulatory Models Derived From ODEs

Development of a dynamic system of interacting objects governed (modelled) by a set of ordinary differential equations (ODEs).

- many examples in mathematical biology
- viral dynamics
- drug therapy for controlling viral load
- dynamic models of signalling pathways
- pharmacokinetics (compartmental models)
- immune defense

Developing “predator-prey" non-linear ODE models for mosquito/parasite interaction.

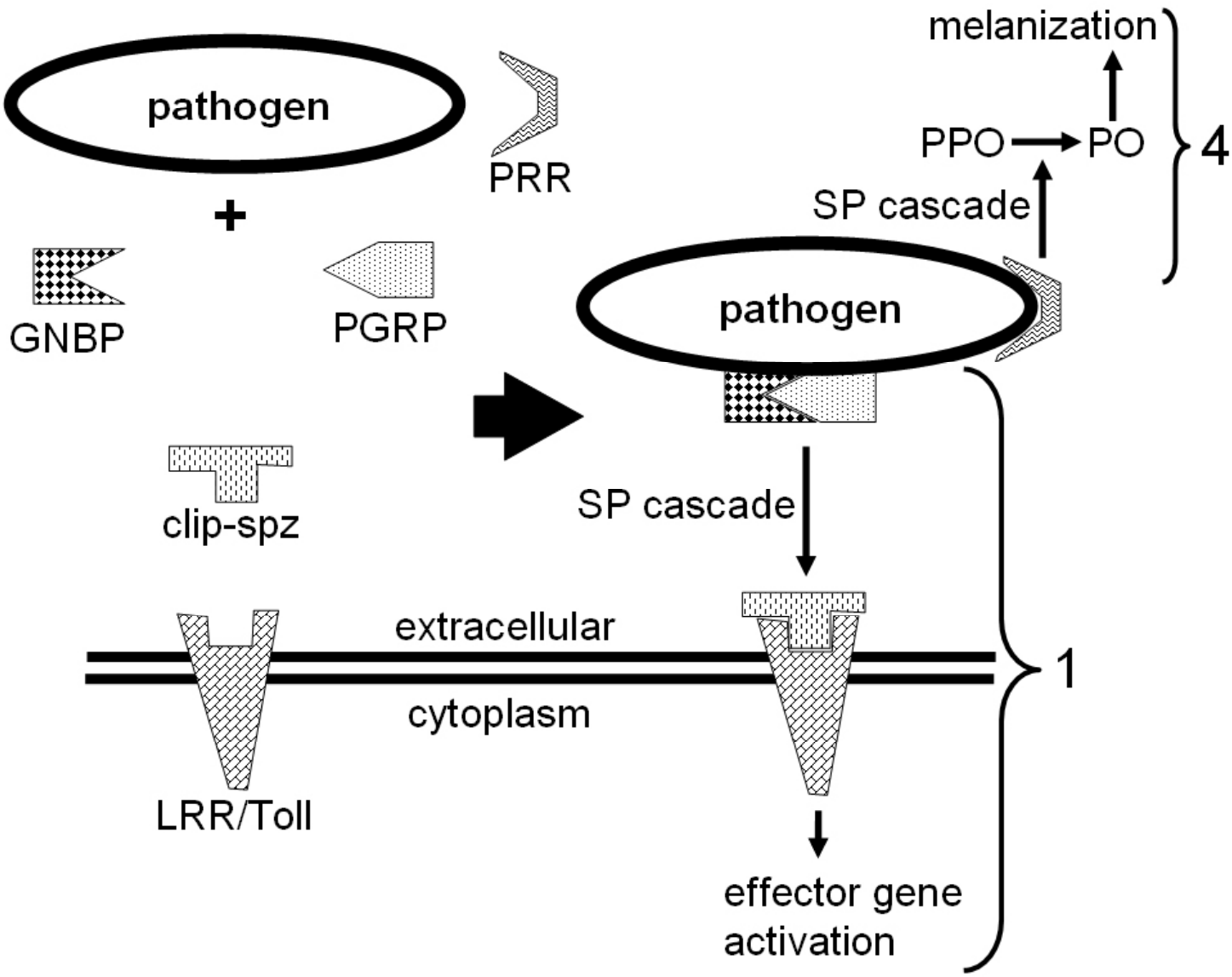
III. Pathway and Network Discovery

Clustering can generate biologically plausible pathways.

e.g. Multiple challenge mosquito experiments

> *Cluster 1 comprises two pattern recognition receptors, GNBPB1 and PGRPLB, belonging to the gram-negative bacteria-binding gene family and the peptidoglycan recognition protein gene family, respectively. GNBPs and PGRPs have been shown to function in the same mechanism implicated in the activation of the Toll immune signaling pathway in Drosophila melanogaster*

See Heard et al. (2005).


melanization
pathogen
PRR
PPO
PO
4
SP cascade
+
GNBP
PGRP
pathogen
clip-spz
SP cascade
extracellular
cytoplasm
1
LRR/Toll
effector gene
activation

Network Modelling: Inference strategies

Two probabilistic/statistical approaches to inference for graphical models

- covariance selection modelling (Jones et al. 2004).
  - identification of key covariances in multivariate (Gaussian) model
  - related to classical “relevance" discovery methods ?
- latent position models (Hoff et al. 2002)
  - takes binary “interaction" matrix
  - uses low dimensional (latent) mixture model to represent (functional) subgroupings of objects.
  - objects close in latent space more likely to be in same functional unit.

Latent Position Models

$Y_{ij} \in \{0, 1\}$ represents absence/presence of “interaction" between two objects. Binary regression model given by

$$\text{logit}\{P[Y_{ij} = 1]\} = X_{ij}\beta_0 + \beta_1|z_i - z_j|$$

where *a priori*

$$f(z) = \sum_{k=1}^{K} \pi_k f_k(z|\theta_k)$$

Obvious applications in protein-protein interaction modelling.

Model for patterns of coregulation in microarray experiments ?

IV. Multiple Microarray Treatments

General data structure:

- N genes of an organism represented on each microarray.

IV. Multiple Microarray Treatments

General data structure:

- N genes of an organism represented on each microarray.
- E experimental treatments applied to different cohorts of the organism.

IV. Multiple Microarray Treatments

General data structure:

- N genes of an organism represented on each microarray.
- E experimental treatments applied to different cohorts of the organism.
- gene expression of organism under treatment $i \in \{1, \ldots, E\}$ measured at T_i time points $\mathbf{t_i} = (t_{i1}, \ldots, t_{iT_i})$.
  $\rightarrow$ The $\{\mathbf{t_i}\}$ are of different lengths and non-nested.

IV. Multiple Microarray Treatments

General data structure:

- N genes of an organism represented on each microarray.
- E experimental treatments applied to different cohorts of the organism.
- gene expression of organism under treatment
  $i \in \{1, \ldots, E\}$ measured at T_i time points
  $\mathbf{t_i} = (t_{i1}, \ldots, t_{iT_i})$.
  $\rightarrow$ The $\{\mathbf{t_i}\}$ are of different lengths and non-nested.

  The model described extends automatically to incomplete data sets or when genes measured for only a $\subset$ of the treatments (at high computational expense).

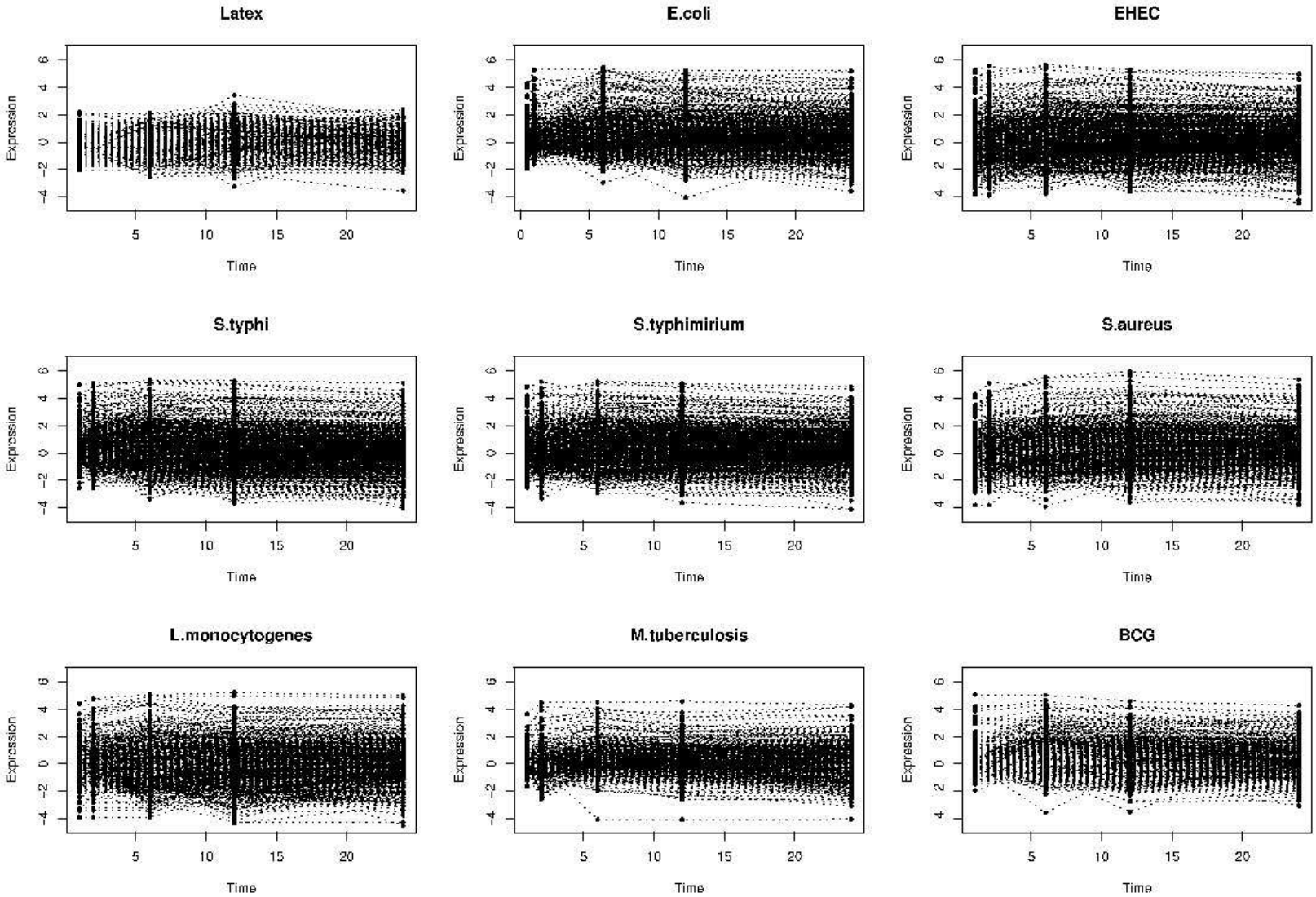
Example

Nau *et al* (PNAS, 2002) investigated human macrophage activation induced by different bacterial pathogens.

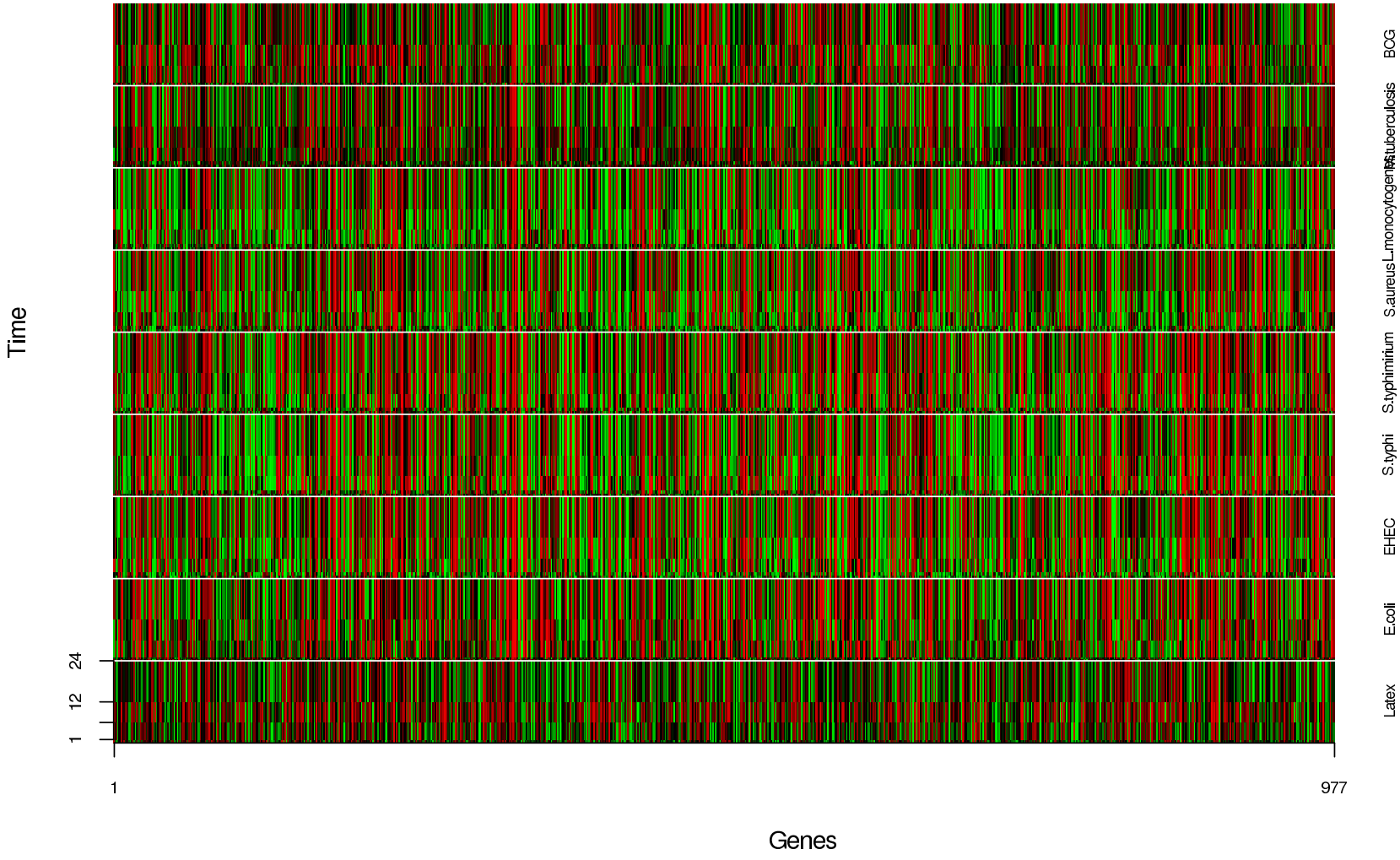
Learning about the gene response of innate immune cells to these pathogens may provide insight into host defenses and tactics used by pathogens to circumvent these defenses.

Sequences of microarray experiments were performed over a duration of 24 hours for 8 different bacterial pathogens plus a control.

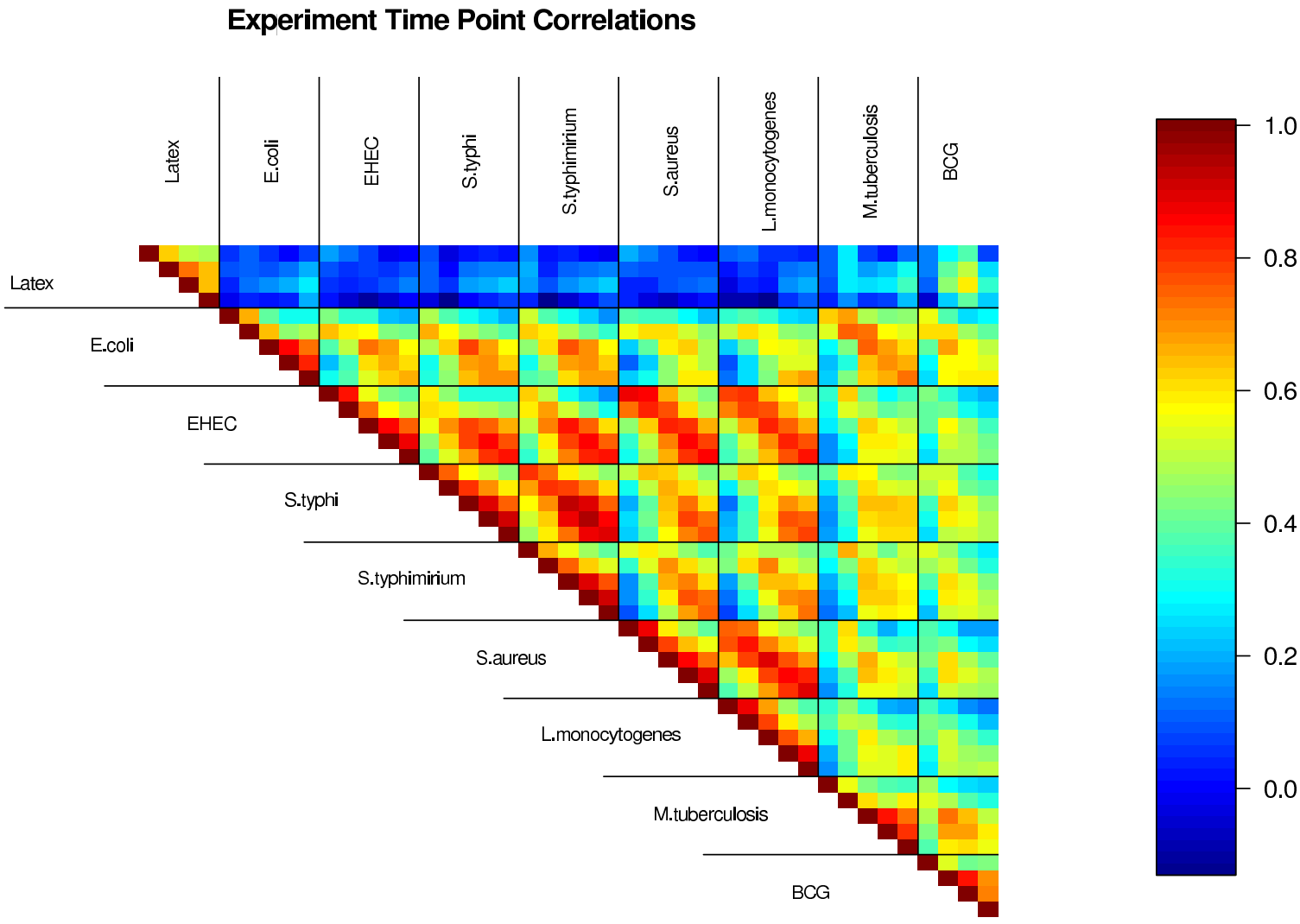
Nau *et al* data



Unclustered expression data



Nau *et al* data, time point correlations



Spline Regression Clustering Revisited

A partition of the genes $\mathcal{C} = (C, N_1, \ldots, N_C)$ divides them into C groups of sizes $N_1, \ldots, N_C$, $\sum_{k=1}^{C} N_k = N$.

For each set of genes k, we fit E spline regression models, one for each treatment.

Each spline curve characterised by the same set of p basis functions giving design matrix X_i $(T_i \times p)$, and a corresponding treatment coefficient p-vector $\beta_i^{(k)}$.

$\Rightarrow$ allows structured joint learning between treatment response profiles, despite the different $\{\mathbf{t_i}\}$.

Spline Regression Model

Fit a joint model for the combined expressions time series of the genes across the treatments, $y^{(k)}$.

$$\rightarrow \left.\begin{array}{ll} \text{spline coefficients} & \beta^{(k)} = \left(\beta_1^{(k)}, \ldots, \beta_E^{(k)}\right) \\ \text{error variance} & \sigma^{2(k)} \end{array}\right\} p\left(y^{(k)} | \beta^{(k)}, \sigma^{2(k)}\right)$$

$$p\left(\beta^{(k)} | \sigma^{2(k)}\right) \equiv \mathrm{N}\left(0, \sigma^{2(k)} V^{(k)}\right), \quad p\left(\sigma^{2(k)}\right) \equiv \mathrm{I\Gamma}\left(\frac{\alpha}{2}, \frac{\gamma}{2}\right)$$

lead to an explicit marginal likelihood $p(y|\mathcal{C}) = \prod_{k=1}^{C} p\left(y^{(k)}\right)$.

Coefficient covariance

We model $V^{(k)}$ as a two component mixture. First, we set

$$V = \Sigma \otimes I_p$$

where Σ is an $E \times E$ symmetric positive definite matrix.

- Σ acts as a between experiment covariance matrix.
- the symbol $\otimes$ represents the *direct* (or *Kronecker*) *product* of two matrices.

Coefficient covariance mixture prior

Let $D_\Sigma = \text{diag}\{\Sigma_{11}, \ldots, \Sigma_{EE}\}$ be the decorrelated experimental covariance matrix containing the diagonal elements of Σ, we use

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

$$q \sim \text{Unif}[0,1]$$

Easily extends to more general mixture model to allow multiple modes of correlation.

Finally, in the simplest case can use

$$\Sigma^{-1} \sim \text{Wishart}(cI_E)$$

Posterior distribution on $\mathcal{C}$

Uniform/Multinomial-Dirichlet distribution on $\mathcal{C}$.

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

The posterior p.m.f. of $\mathcal{C}$ given the expression profile data y is known up to proportionality and we can thus define a kernel of the p.m.f.

$$\pi(\mathcal{C}|y) = p(y,\mathcal{C}) = p(y|\mathcal{C})p(\mathcal{C}) \propto p(\mathcal{C}|y)$$

which we use as a potential function for clustering.

Agglomerative clustering algorithm

1. Start each gene observation vector in its own cluster. Calculate π for this partition.
2. Looking across all clusters, find the cluster pair whose merger would cause the biggest increase/smallest decrease in π.
3. Combine the two clusters and update π.
4. Repeat steps 2-3 until there is just one cluster containing all the gene observations.
5. Partition the genes according to the clustering visited with the highest (unnormalized) probability π.

Stochastic search strategy

Simulated annealing to find an *approximate* posterior mode

- MH moves on e.g. $\Sigma^{-\frac{1}{2}}$, q, etc are ok.
- Gibbs sampling of cluster allocations for fixed C (#Clusters) ok too.
- RJ moves to change C not ok
  - $-$ tempered Gibbs scans to find a launch state (Neal, 2000) seems best, but acceptance rates still low.
  - $+$ multiple restarts of agglomerative clustering algorithm using updated Σ improves performance.

Data Fusion

How do we incorporate extra sources of information into this framework?

Two possibilities

1. Gene information e.g. sequence similarity, position on genome, known pathways.
2. Treatment information e.g. chemical structure, data from other kinds of experiment.

Modelling extra data sources

Incorporating 1 is trivial - either

- just append some extra columns to X;
- or at worst, build $p(\text{other data}|y, \mathcal{C})$.

2 is less straightforward, requires a more flexible model than $\Sigma^{-1} \sim \text{Wishart}(\cdot)$.

Will consider 2 here.

Nau *et al* data treatment categories

The 8 bacterial treatments fell into 3 broad categories with "different cellular components and pathogenesis":

1. Gram-negative bacterial: *E. coli*, *EHEC*, *S. typhi*, *S. typhimurium*;
2. Gram-positive bacterial: *S.aureus*, *L. monocytogenes*;
3. Mycobacteria: *M. tuberculosis*, BCG.

These plus the control experiments give 4 groupings to the 9 experiment sets which we may wish to exploit.

Alternative model for Σ

For experiment variances, still use $\Sigma_{ii}^{-1} \sim \Gamma(\cdot)$, $i = 1, \ldots, E$.

To construct the correlations, introduce a *latent position* vector for each treatment $l_i \in \mathbb{R}^d$.

Then, letting $s_i \sim$ Bernoulli(0.5),

$$\Sigma_{ij} = (-1)^{s_i + s_j} \sqrt{\Sigma_{ii}\Sigma_{jj}} \exp(-||l_i - l_j||^{\nu})$$

$\Rightarrow$ Treatments closer together in latent space have greater correlation in gene expression.

Distributions for $\{l_i\}$

$l_i \sim \mathsf{N}_p(0, \tau I_p)$ $\rightarrow$ gives something very similar to Wishart model.

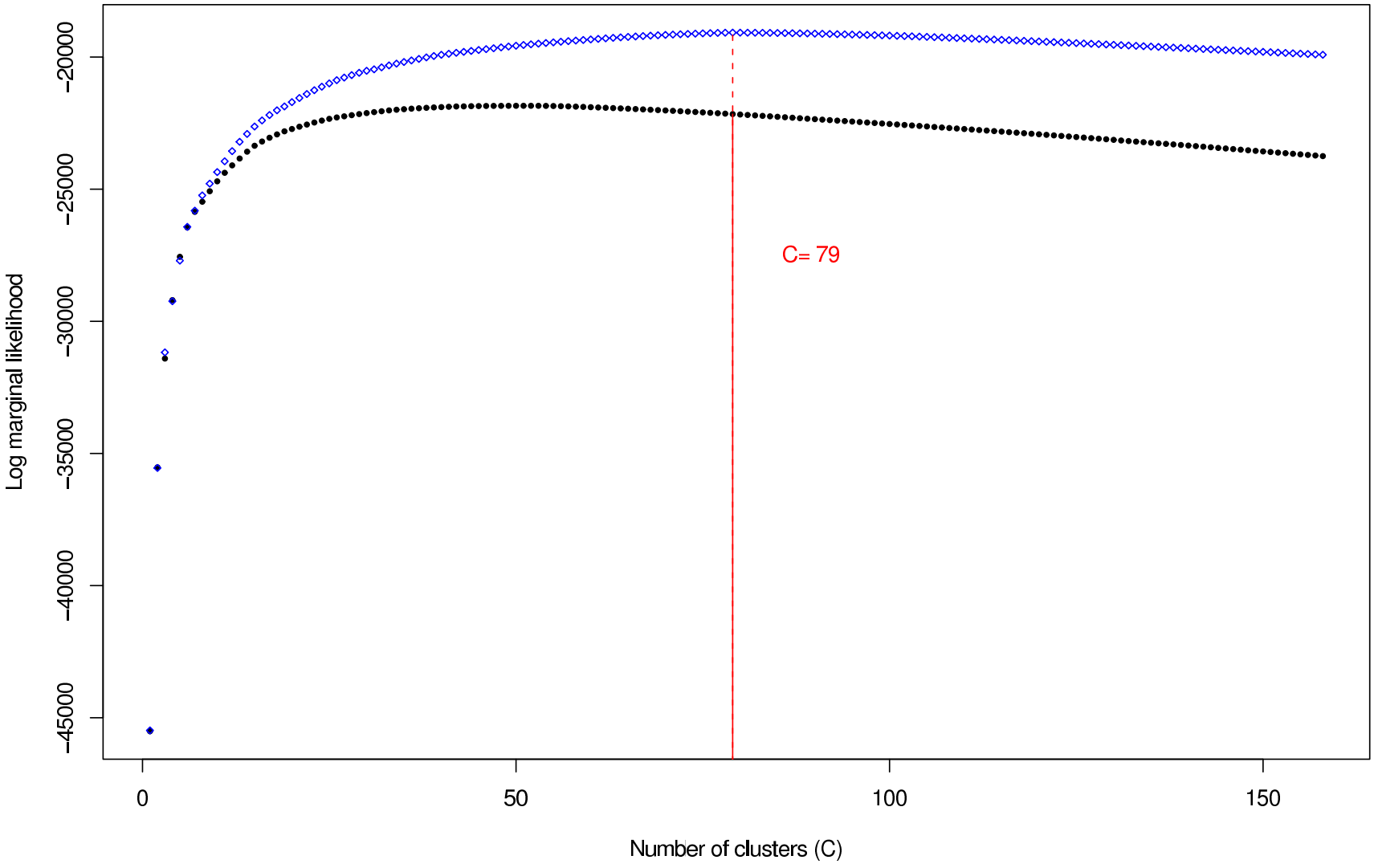
Instead, we build a hierarchy with parameters $\{\mu_1, \mu_2, \mu_3, \mu_4\}$ for the 4 treatment categories.

$$l_i \sim \mathsf{N}_p(\mu_{c(i)}, \lambda I_p)$$
$$\mu_j \sim \mathsf{N}_p(0, \tau I_p)$$

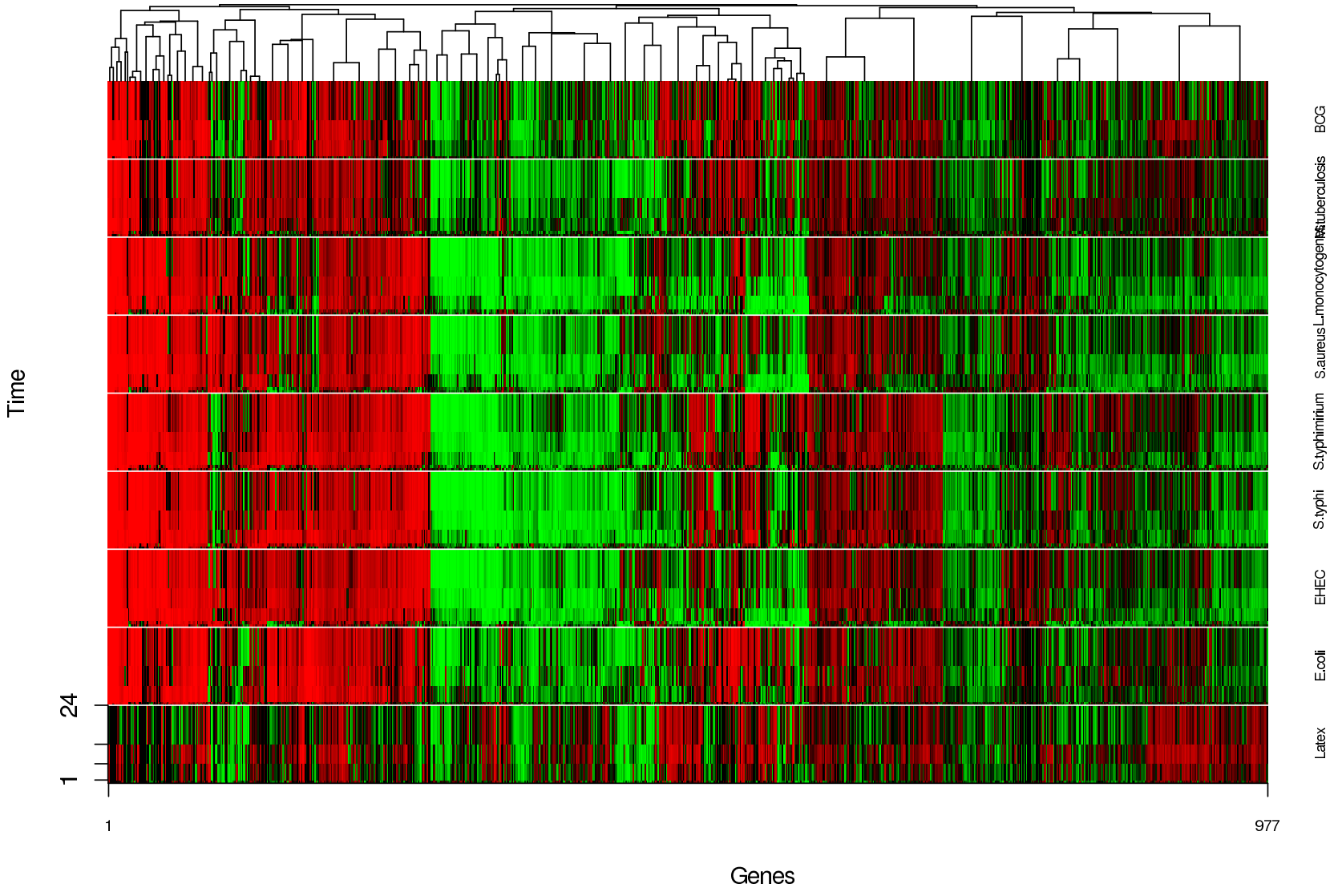
Maybe with $\lambda << \tau$.

Learn jointly about $\{l_i\}$ and $\{\mu_j\}$.

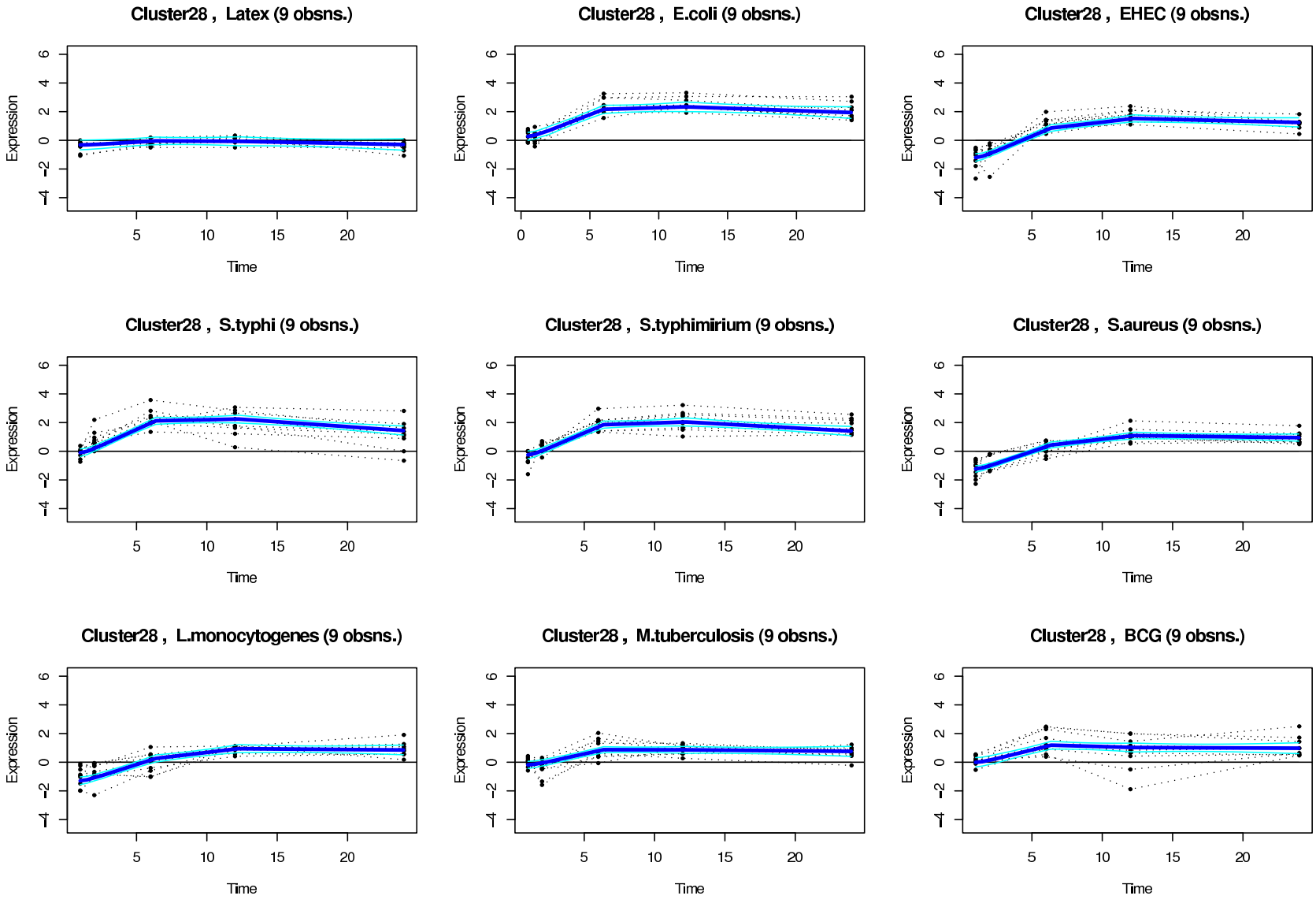
Agglomerative clustering posterior score



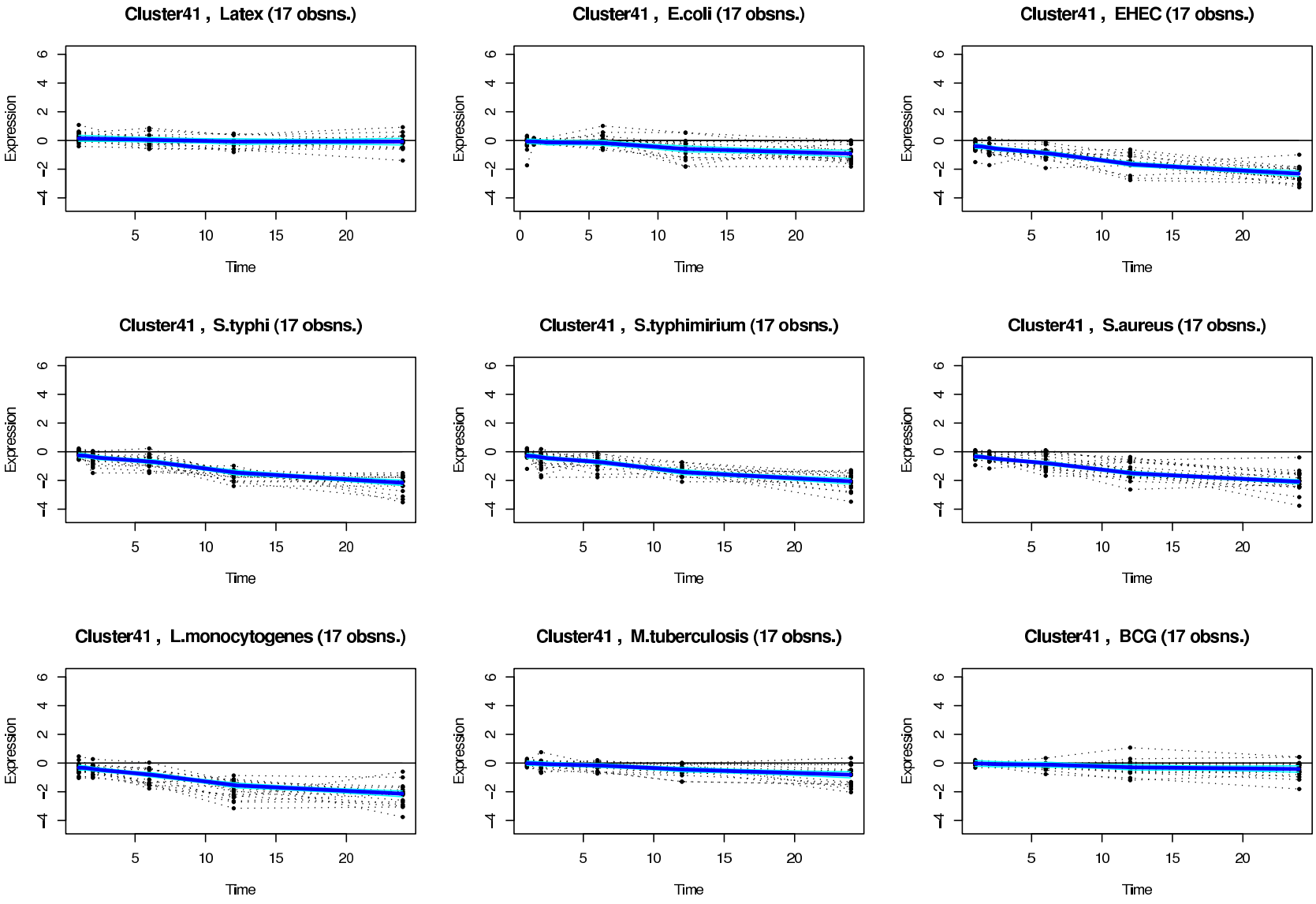
Clustered expression data



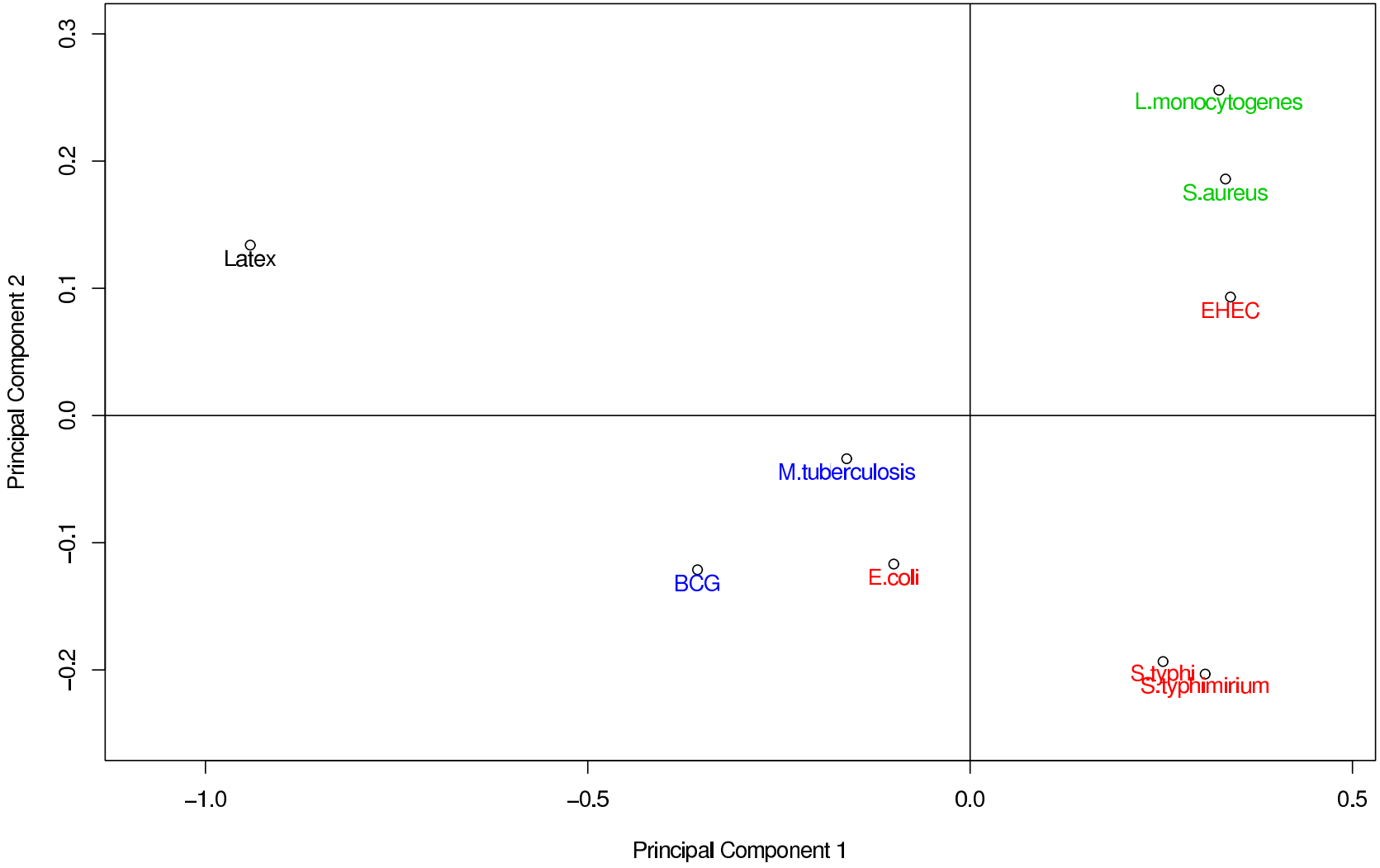
An up-regulated cluster



A down-regulated cluster



Treatment locations under hierarchical model



More details

- Related paper at
  http://stats.ma.imperial.ac.uk/~naheard/
- and code at
  http://stats.ma.imperial.ac.uk/~naheard/splinecluster