Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Model-based Clustering of Gene Expression Profiles

David A Stephens

Department of Mathematics & Statistics
McGill University

d.stephens@math.mcgill.ca
www.math.mcgill.ca/∼dstephens/

October 25, 2006

Outline

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

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

I: Mosquito and Parasite Genomics

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

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 Plasmodium/Anopheles Genome Array: genome wide, 19200 transcripts.

Typically multiple time points, sometimes replicated.

Major data fusion issues.

Objectives

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- to understand regulatory mechanisms
- 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)
- computational feasibility is an important factor
  - biologists often reluctant to use advanced statistical methods

"Sophistication buys you nothing"

Professor of Statistical Bioinformatics, 2005.

Objectives

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- to understand regulatory mechanisms
- 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)
- computational feasibility is an important factor
  - biologists often reluctant to use advanced statistical methods

"Sophistication buys you nothing"

Professor of Statistical Bioinformatics, 2005.

Objectives

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- to understand regulatory mechanisms
- 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)
- computational feasibility is an important factor
  - biologists often reluctant to use advanced statistical methods

"Sophistication buys you nothing"

Professor of Statistical Bioinformatics, 2005.

Objectives

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- to understand regulatory mechanisms
- 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)
- computational feasibility is an important factor
  - biologists often reluctant to use advanced statistical methods

"Sophistication buys you nothing"

Professor of Statistical Bioinformatics, 2005.

EXCEPT SOMETIMES THE CORRECT ANSWER.

Malaria, Mosquitoes and Immune Defense.

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- Malaria is caused by the parasite *Plasmodium Falciparum* and is primarily spread by *Anopheles Gambiae* mosquito vectors.



Mosquito and Parasite Genomics

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- 2002: genomes of *P. Falciparum* and *A. Gambiae* released
- it has been demonstrated that the mosquito employs its own immune system against the parasite.
- the components operating mosquito immune system and their potential relevance to antimalarial responses are being investigated.
- emphasis has been placed on the study of anti-malarial responses involved in limiting the extent of infection.
- the mosquito has become the organism of choice for directly studying antiparasitic innate immune responses.

A. Gambaie Microarray studies

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Microarray experiments carried out at Imperial College have studied various aspects of genetic regulation.

- An immune defense response is mounted whenever the host mosquito is infected by the parasite
- Genes in the mosquito genome known to be involved in immune defense have been identified.
- Key task is to find genes with similar regulation patterns, as they too may be involved with immune defense activity.
- Clear correlation of immune responses with the passage of the parasite through the vector.

Expression data available for study

- mosquito with parasite infestation
- mosquito only under artificial experimental challenge

A. Gambaie Microarray studies

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Microarray experiments carried out at Imperial College have studied various aspects of genetic regulation.

- An immune defense response is mounted whenever the host mosquito is infected by the parasite
- Genes in the mosquito genome known to be involved in immune defense have been identified.
- Key task is to find genes with similar regulation patterns, as they too may be involved with immune defense activity.
- Clear correlation of immune responses with the passage of the parasite through the vector.

Expression data available for study

- mosquito with parasite infestation
- mosquito only under artificial experimental challenge

II. Microarrays: Key Statistical Problems

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- image analysis
- normalization/standardization
- analysis of differential expression
- longitudinal data analysis
- multiple testing and the control of FWER and FDR
- clustering/partitioning (identification of functionally distinct subsets)

"Most of the time microarrays measure nothing, badly"

Robert Gentleman, Fred Hutchinson Cancer Research Centre, Seattle.

Time Course Data: Example

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

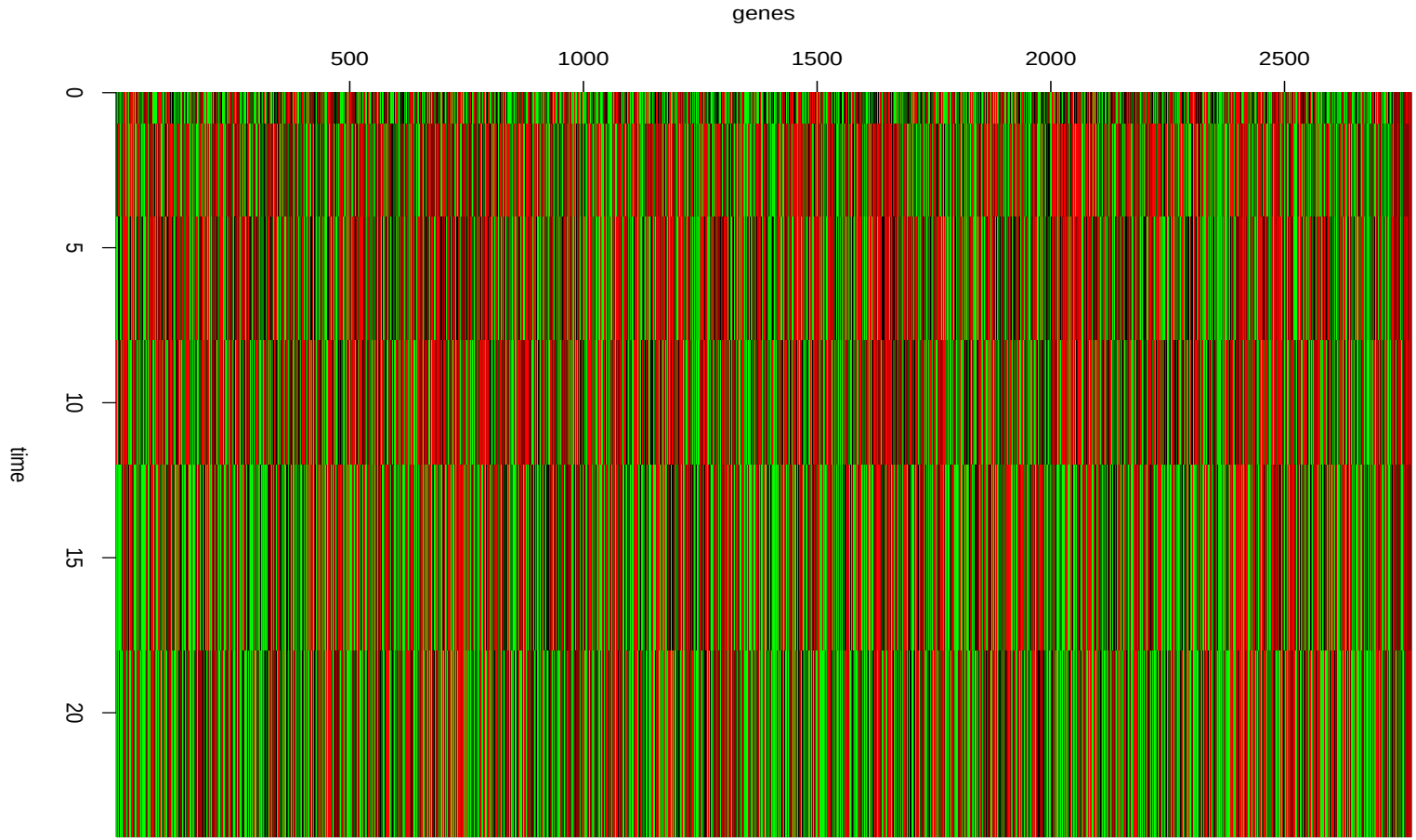
Analysis of gene expression profiles of mosquitos/cell-lines compromised by bacterial and chemical agents (challenges):

- relative expression of 2771 genes/sequence tags,
- probes selected from a specially constructed cDNA library.
- approximately 900 have associated/putative function.
- relative expression recorded at $T = 6$ time points, at 1, 4, 8, 12, 18 and 24 hours after the challenge.

Focus first on single bacterial challenge, *Salmonella typhi*.

Salmonella typhi challenge data.

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



III. Statistical Models for the Profiles

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

We model the gene expression profiles using a linear regression model and non-linear basis functions.

This induces a stochastic process model for the underlying variation in expression; our clustering approach utilizes the covariance structure of this process.

Aim is to use models that capture the characteristic behaviour of expression profiles corresponding to different forms of regulation.

Generically, we wish to capture the behaviour of the gene expression ratio y as a function of time t and measurement error.

Regression Representation

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

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;

- plausible for our experiments
- assumption can be relaxed
- $\{Y_{i_t}, t = 1, \ldots, T\}$ conditionally independent, unconditionally dependent.

Truncated Polynomial Spline Basis Model

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

$$X(t)\beta = \beta_0 + \sum_{j=1}^{p} \beta_j (t - \kappa_j)_+^q$$

for $q = 1, 2, ...,$ where $(\kappa_1, \ldots, \kappa_p)$ are knot positions spanning the range of t, and

$$(t - \kappa_j)_+^q = \max\{0, (t - \kappa_j)\}^q$$

This function is continuous at the knot points; here we presume that the knot positions are fixed at the data ordinates.

Any suitable basis function set may be used.

Bayesian Inference

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

For inference, under the assumption that the random error terms $\{\varepsilon_{i_t}\}$ form an i.i.d Gaussian sequence with variance σ^2, the conditional distribution of the concatenated response vectors of N genes Y is multivariate normal

$$Y|X, \beta, \sigma^2 \sim \mathrm{N}\left(X\beta, \sigma^2 I_{NT}\right)$$

where now X is $NT \times p$ and I_{NT} is the NT-dimensional identity matrix.

Prior and Posterior

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Under a conjugate prior specification: $\left(\beta, \sigma^2\right)$

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

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

$$p\left(\beta|y, \sigma^2\right) \equiv N\left(m^*, \sigma^2 V^*\right)$$

$$p\left(\sigma^2|y\right) \equiv \mathsf{IGamma}\left(\frac{T+\alpha}{2}, \frac{c+\gamma}{2}\right)$$

Marginal Likelihood

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Our clustering approach will be based on marginal likelihood considerations.

$$p(y) = \int \int p\left(y|\beta, \sigma^2\right) p\left(\beta|\sigma^2\right) p\left(\sigma^2\right) d\beta d\sigma^2.$$

so that, for our basis function representation of a single profile

$$p(y) = \left(\frac{1}{\pi}\right)^{T/2} \frac{\gamma^{\alpha/2}\Gamma\left(\frac{T+\alpha}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)} \frac{|V^*|^{1/2}}{|V|^{1/2}} \frac{1}{\{c+\gamma\}^{(T+\alpha)/2}}$$

In practice, V can be chosen to approximately maximize the marginal likelihood, whereas (α, γ) are chosen from genuine prior information.

Hierarchical Clustering

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- a method of organizing a collection of objects into disjoint sets
- uses a similarity/discrepancy measure/overall potential function
- *agglomerative clustering* places each of the N items in its own cluster, and then recursively (optimally) merges currently existing clusters until a single cluster remains (i.e. performs $N - 1$ merge operations in total)
- to find the ith optimal agglomeration requires $(N + 1 - i)(N - i)/2$ comparisons.
- at worst, an $O(N^3)$ procedure; here, $N \approx 2800$

Computational Efficiency

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

For both Euclidean distance clustering and our proposed method, the number of calculations is drastically reduced by noting the distance between any two clusters remains unchanged through successive iterations until one is agglomerated with another cluster.

In the Bayesian model-based approach, the distance between two clusters will be the change in marginal likelihood achieved by merging them.

The the number of marginal likelihood calculations actually required is

$$N^2 - N - 1$$

and so only $O\left(N^2\right)$.

Bayesian Model-based Clustering

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Model detects subgroups on the basis of the covariance structure induced by the underlying stochastic process rather than on Euclidean distances.

- hierarchical clustering approach assigns profiles to the same cluster if they are similar in **covariance** terms.
- covariance determined by X and V; marginally, vector response Y_i has covariance

$$XVX^{\mathsf{T}} + I$$

Completing the Bayesian Specification

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

The prior on

- the number of clusters, $\mathcal{C}$,
- the cluster sizes, $N_1, \ldots, N_C$,
- the cluster decomposition, $Z_1, \ldots, Z_n$

Natural default choice uses the specification

$$P(\mathcal{C} = C)P(N_1, \ldots, N_C|\mathcal{C} = C)P(Z_1, \ldots, Z_n|N_1, \ldots, N_C)$$

- $P(\mathcal{C} = C) = 1/N, C = 1, 2, \ldots, N$
- $P(N_1, \ldots, N_C|\mathcal{C} = C)$ Multinomial/Dirichlet
- $P(Z_1, \ldots, Z_n|N_1, \ldots, N_C)$ Uniform allocation

Advantages of the approach

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- respects the time ordering of the data
- leads to biologically appropriate covariance structures being discovered; can be used to incorporate knowledge of the dynamics of the underlying processes involved in the regulation of expression.
- clusters using the marginal probability that the objects came from the same generating process.
- choice of covariance matrix, induced by the basis function representation, leads to important simplifications in the calculation of the marginal likelihoods of each cluster, and for the hierarchical steps.
- allows coherent integration of different information sources: can input co-clustering prior probabilities.

Advantages of the approach

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- respects the time ordering of the data
- leads to biologically appropriate covariance structures being discovered; can be used to incorporate knowledge of the dynamics of the underlying processes involved in the regulation of expression.
- clusters using the marginal probability that the objects came from the same generating process.
- choice of covariance matrix, induced by the basis function representation, leads to important simplifications in the calculation of the marginal likelihoods of each cluster, and for the hierarchical steps.
- allows coherent integration of different information sources: can input co-clustering prior probabilities.

Advantages of the approach

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- respects the time ordering of the data
- leads to biologically appropriate covariance structures being discovered; can be used to incorporate knowledge of the dynamics of the underlying processes involved in the regulation of expression.
- clusters using the marginal probability that the objects came from the same generating process.
- choice of covariance matrix, induced by the basis function representation, leads to important simplifications in the calculation of the marginal likelihoods of each cluster, and for the hierarchical steps.
- allows coherent integration of different information sources: can input co-clustering prior probabilities.

Advantages of the approach

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- respects the time ordering of the data
- leads to biologically appropriate covariance structures being discovered; can be used to incorporate knowledge of the dynamics of the underlying processes involved in the regulation of expression.
- clusters using the marginal probability that the objects came from the same generating process.
- choice of covariance matrix, induced by the basis function representation, leads to important simplifications in the calculation of the marginal likelihoods of each cluster, and for the hierarchical steps.
- allows coherent integration of different information sources: can input co-clustering prior probabilities.

Advantages of the approach

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- respects the time ordering of the data
- leads to biologically appropriate covariance structures being discovered; can be used to incorporate knowledge of the dynamics of the underlying processes involved in the regulation of expression.
- clusters using the marginal probability that the objects came from the same generating process.
- choice of covariance matrix, induced by the basis function representation, leads to important simplifications in the calculation of the marginal likelihoods of each cluster, and for the hierarchical steps.
- allows coherent integration of different information sources: can input co-clustering prior probabilities.

The Algorithm

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- Starting with N clusters, each cluster containing the expression levels for one gene.
- At each step, merge the two clusters which cause the biggest increase (or smallest decrease) in the overall marginal likelihood/BIC/posterior probability.
- Continue until one cluster remains.
- Take the optimal number of clusters as that which maximized the marginal likelihood/BIC/posterior probability.

Full Bayesian Analysis

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

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

In theory, the full Bayesian posterior distribution can be computed using Markov chain Monte Carlo (MCMC)

However, there are B_N (Nth Bell Number) ways of partitioning a set of N data points into clusters.

$B_1 = 1, \ldots, B_{10} = 115,975, B_{11} = 678,570, B_{12} = 4,213,597, \ldots$

$\Longrightarrow$ exhaustive computation of the posterior distribution not possible.

Full Bayesian Analysis

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

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

In theory, the full Bayesian posterior distribution can be computed using Markov chain Monte Carlo (MCMC)

However, there are B_N (Nth Bell Number) ways of partitioning a set of N data points into clusters.

$B_1 = 1, \ldots, B_{10} = 115,975, B_{11} = 678,570, B_{12} = 4,213,597, \ldots$

$\Longrightarrow$ exhaustive computation of the posterior distribution not possible.

MCMC methods for clustering (mixtures)

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

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

DO NOT WORK !

- 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)

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Need more advanced MCMC approaches

- simulated annealing for optimization
- tempering/evolutionary/population MCMC
- multivariate DPM models look promising
- "jump-off" methods
- tree-based methods.

Current Strategy: Use the output from agglomerative clustering algorithm as seed of Markov chain Monte Carlo exploration of the space. Use simulated annealing (SA) to find optimal clustering.

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

Results for the challenge experiment

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- Multinomial-Dirichlet prior on the number and size of clusters.

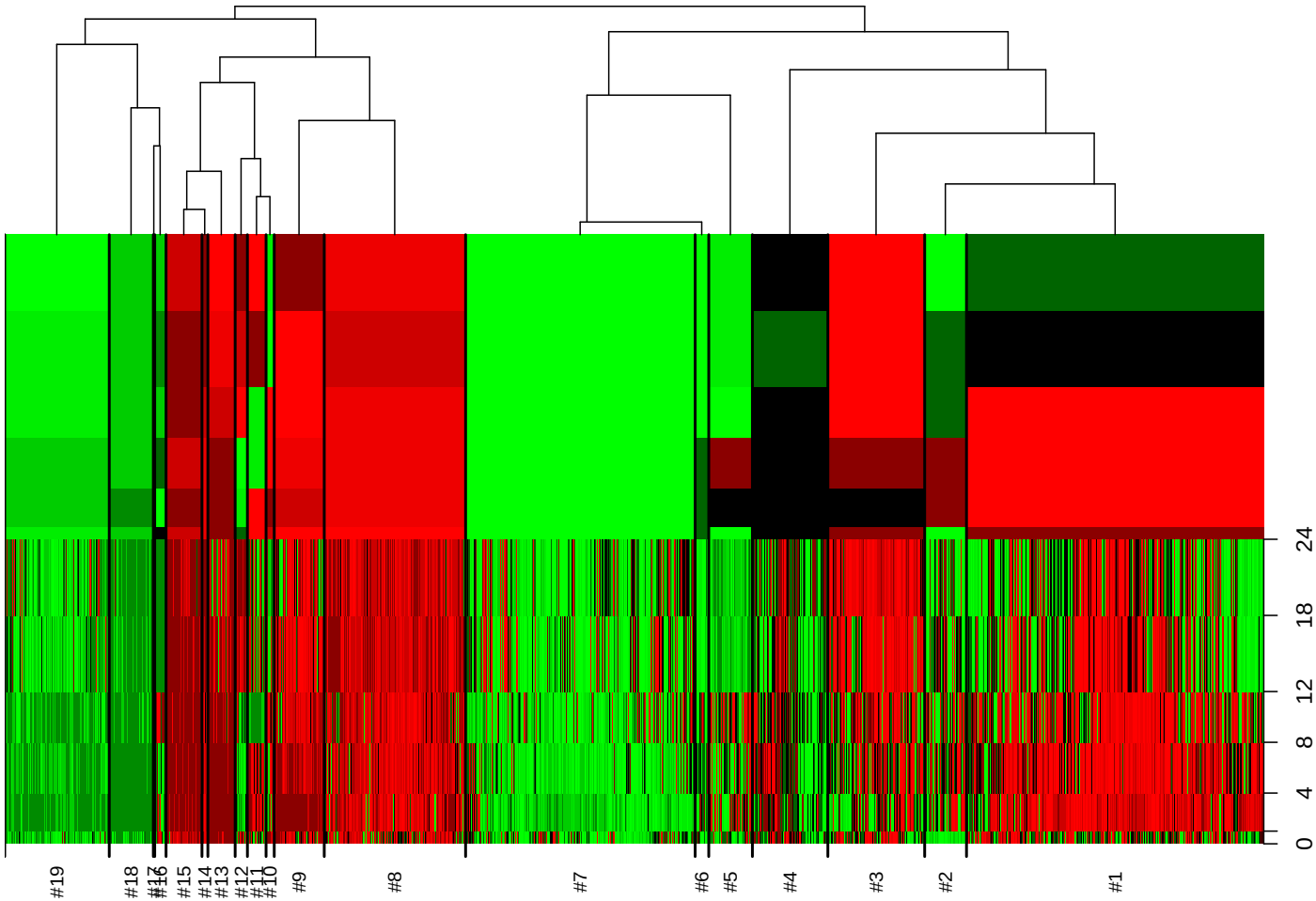
- Linear splines $q = 1$, knots at data points.

$$X(t)\beta = \beta_0 + \sum_{j=1}^{p} \beta_j (t - \kappa_j)_+$$

- Optimal number of clusters was 19 both before and after SA MCMC run.

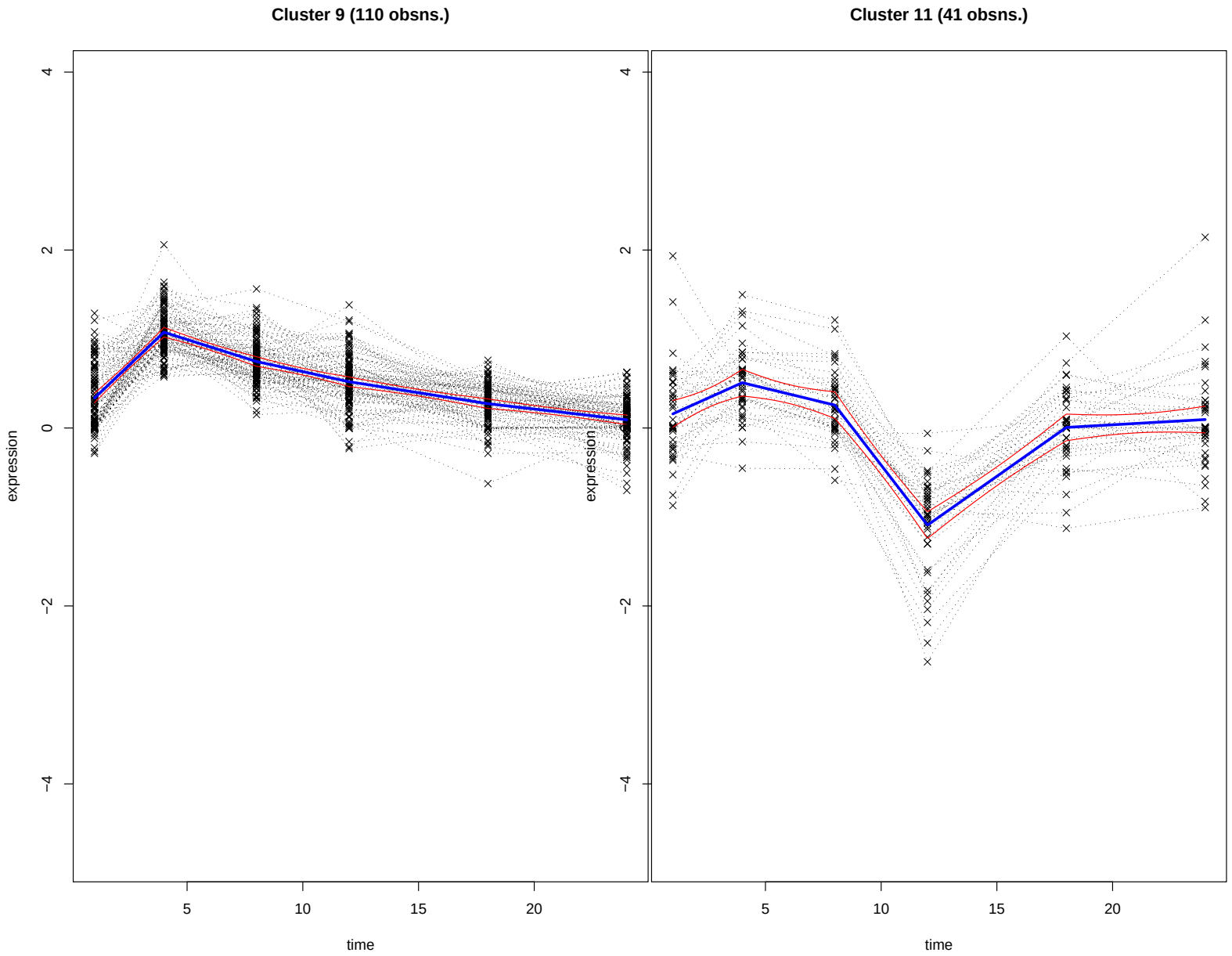
Final Dendrogram

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



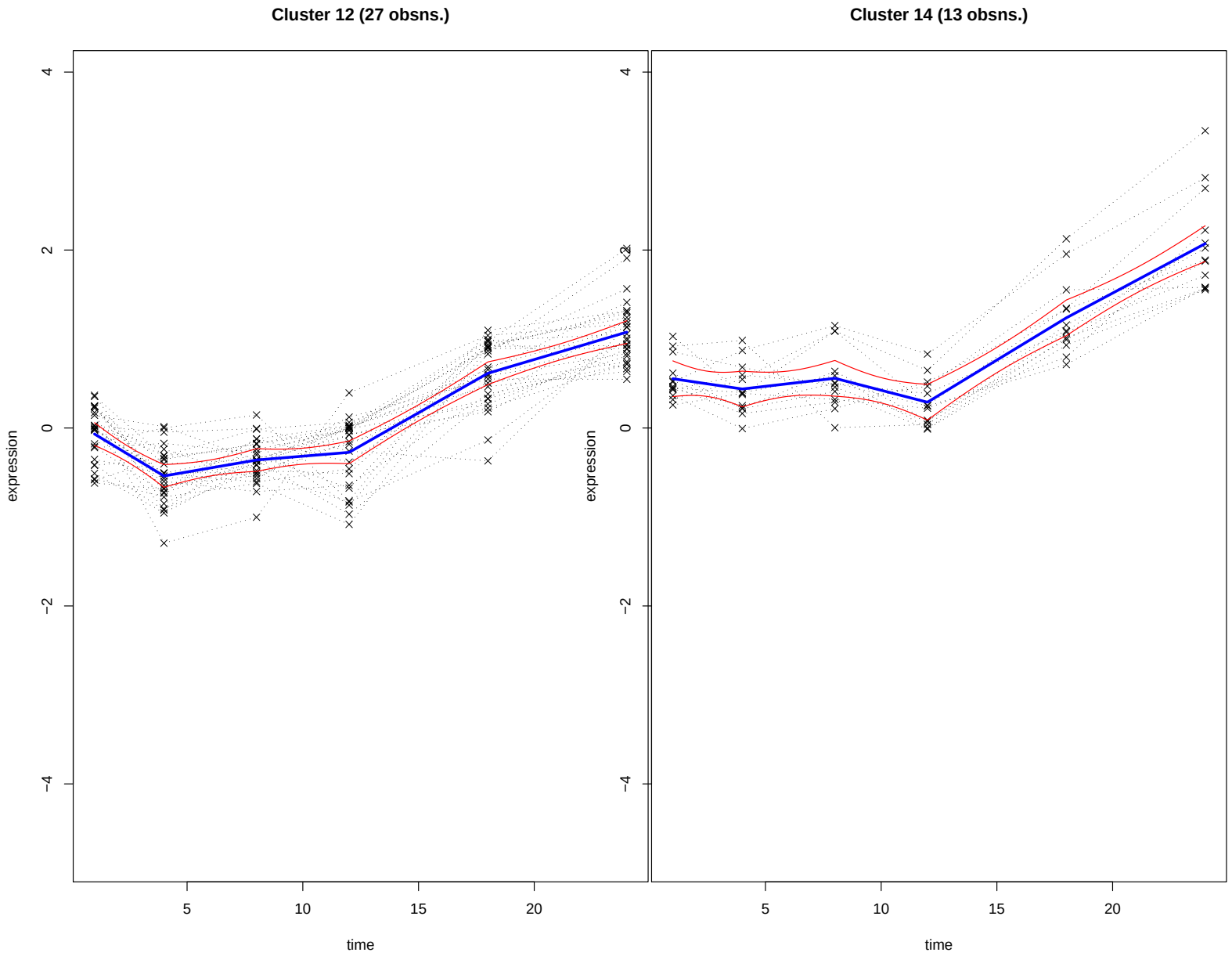
Clusters 9 and 11

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



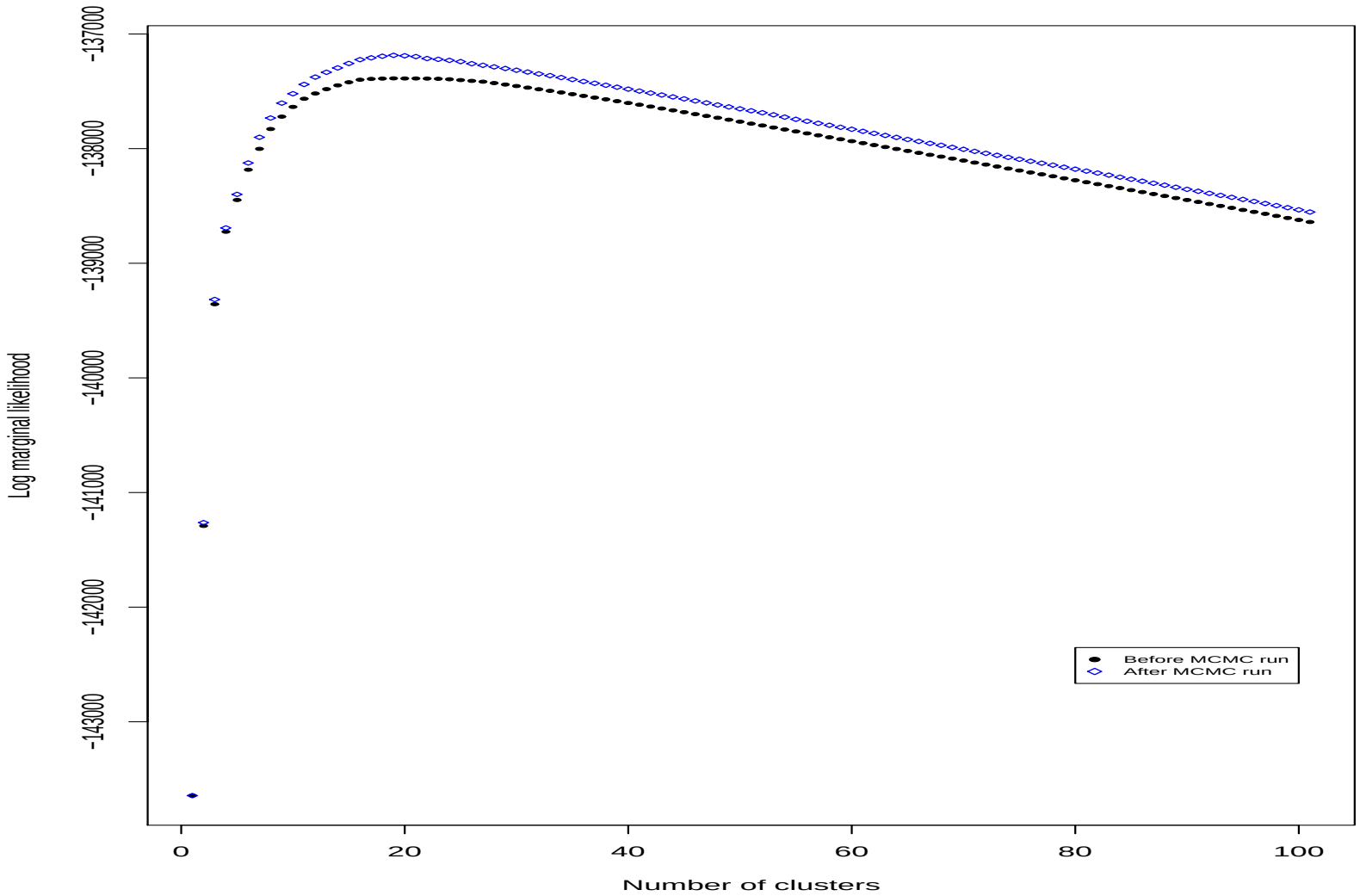
Clusters 12 and 14

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



Number of clusters

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



Results

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- Get 'tighter' clusters than Euclidean hierarchical clustering.
- Computationally feasible - agglomerative clustering takes ~ 2 minutes on a 2Gb processor PC.
- MCMC takes longer.
- Probabilistic model enables inference on the number of clusters, and provides a rigorous framework for classification of unknown genes.
- SA MCMC can find even higher probability regions.

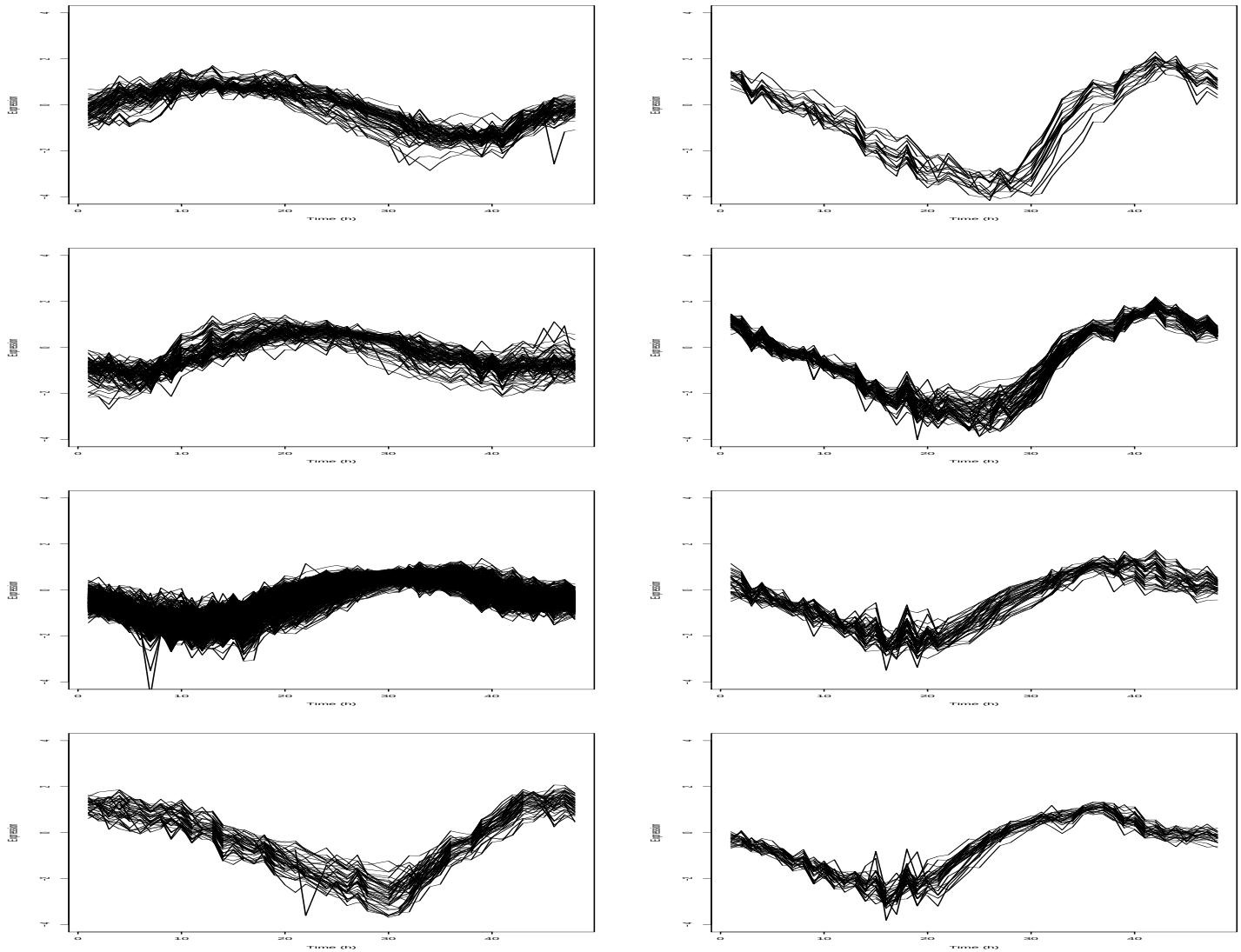
Gene Regulation in *P. Falciparum*

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- Patterns of gene regulation in the parasite also under study
- CAMDA challenge data
- 4221 genes, 46 time points over 48 hours studied here
- expression relative to 48 hour transcriptome
- using hierarchical clustering, results in ~ 30 clusters
- not strictly a clustering problem ?

CAMDA: Functionally related clusters

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



Modelling Options

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- retain basis function regression model
- in truncated polynomial basis function model, use
  - fewer knots than time points
  - cubic splines
- other bases may be more appropriate;

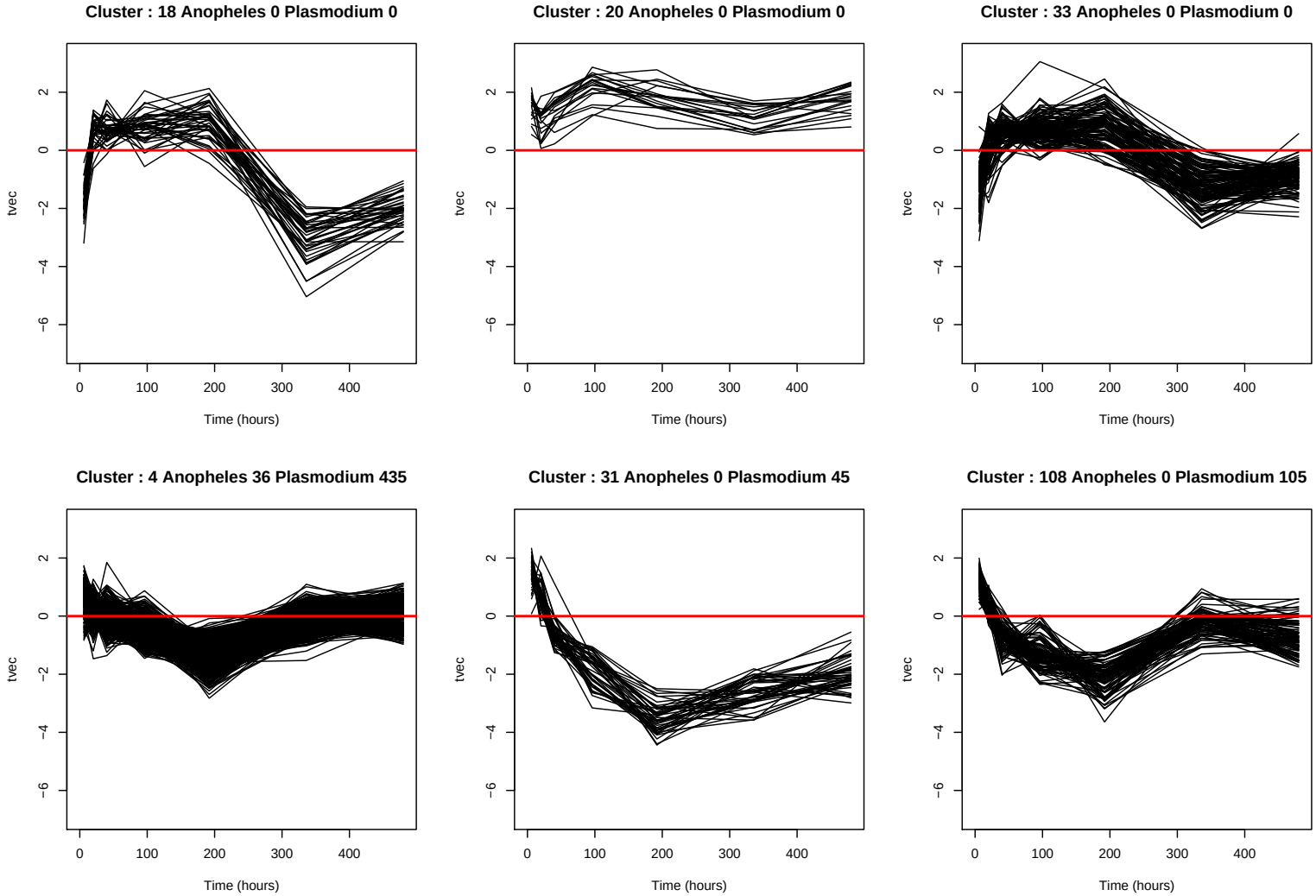
$$X(t)\beta = \sum_{j=1}^{p} \beta_j \cos(\omega_j t + \delta)$$

  This is still a linear model, for fixed $(\omega_1, \ldots, \omega_p, \delta)$, so can proceed as before.
- can fit this model, gene by gene, then rank via set of phase parameters, $(\delta_1, ..., \delta_n)$.

Two species data

2800 Genes: 1400 Mosquito/1400 Parasite.

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



IV. Multiple Microarray Treatments

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

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 $\mathbf{t_i} = t_{i1}, \ldots, t_{iT_i}$).
- $\longrightarrow$ 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).

Co-clustering for *A. Gambiae* Multiple Challenges

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

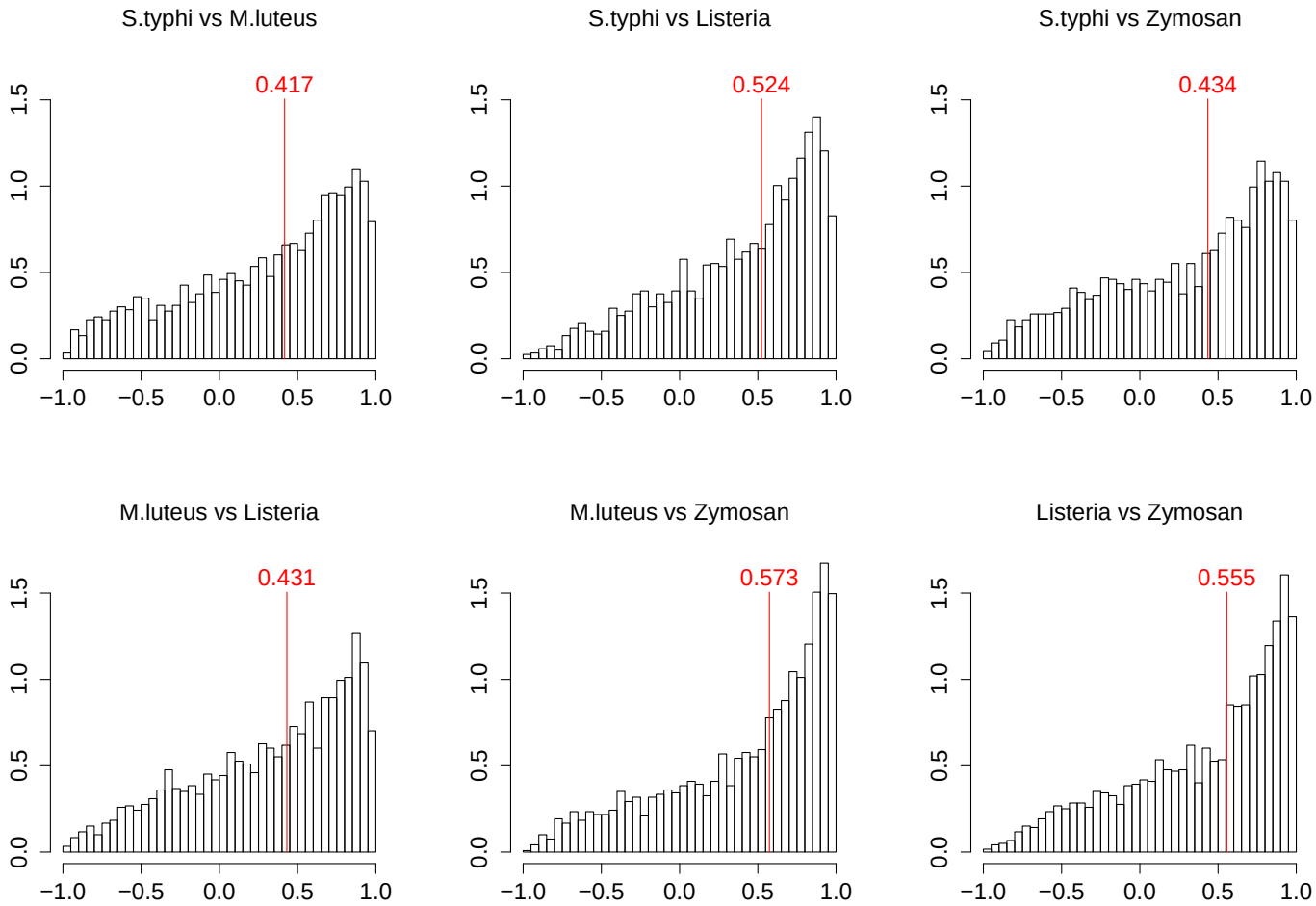
The *S. Typhi* challenge data were obtained as one of a series of experiments using (~ 15) different challenges on the same gene set. Here we look at only four:

- *S. Typhi*
- *Listeria*
- *M. Luteus*
- Zymosan (chemical)

We would expect similar patterns of regulation of immune defense genes under different challenges.

Cross-correlations

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



Extension to Clustering Model

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Extend the prior covariance structure:

$$V = \Sigma \otimes I_p$$

where Σ represents the cross-experiment dependence

- Σ is 4×4.
- $\otimes$ is the direct product.

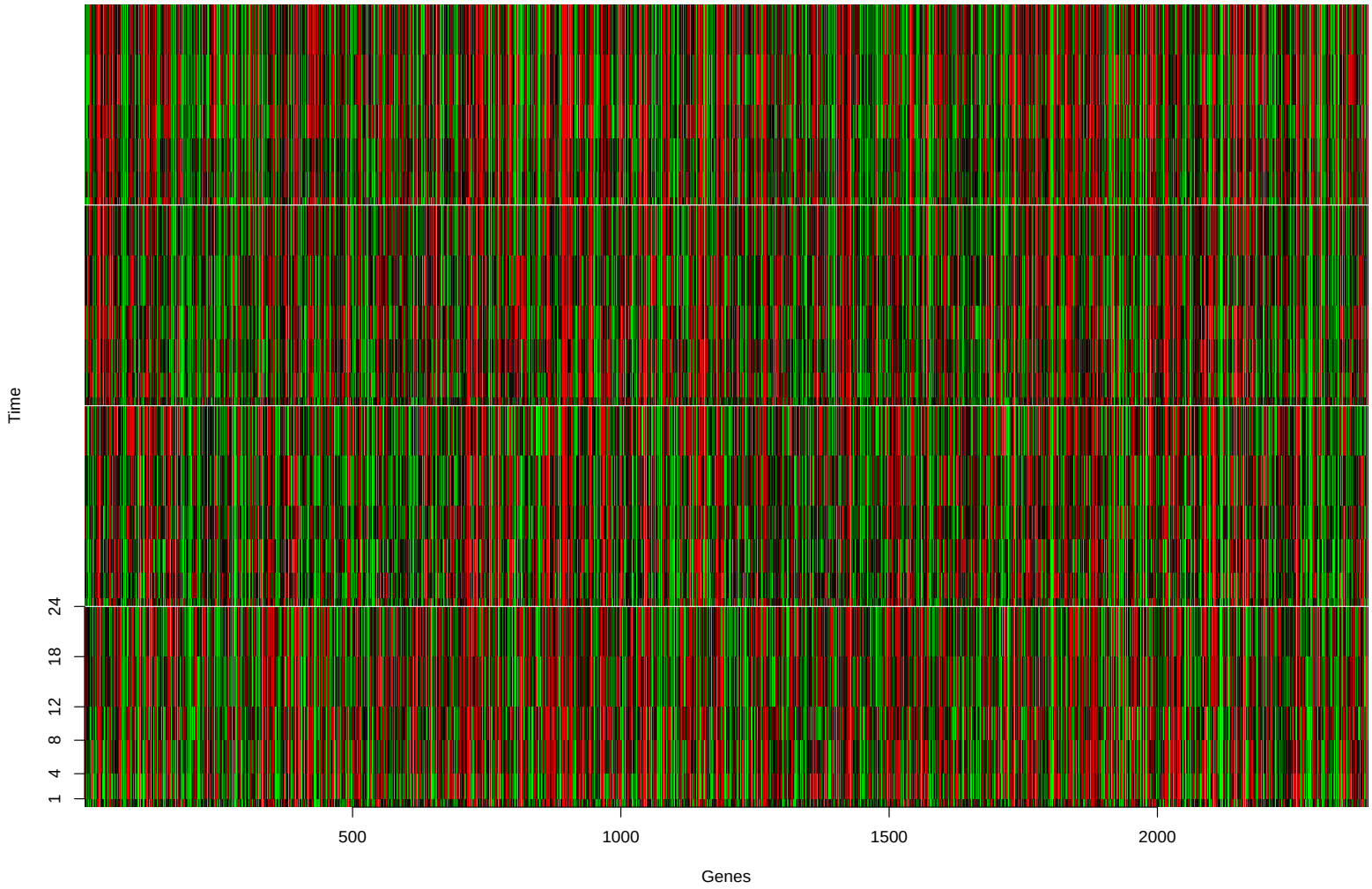
Could use a mixture prior structure (possibly no dependence)

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

where D_Σ is diagonal.

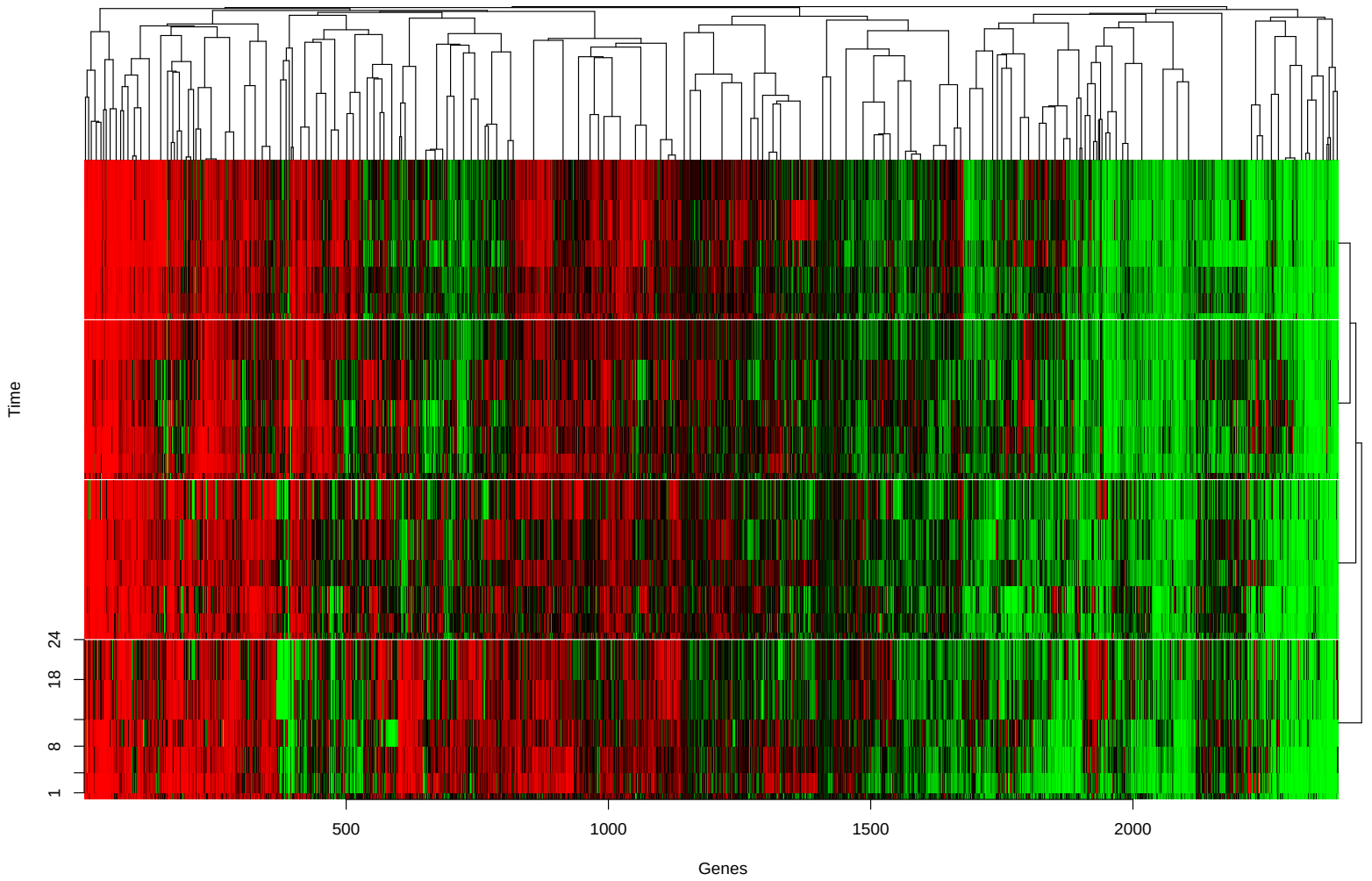
Unclustered

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



Clustered

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



Analysis Strategy: E-M

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

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

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

Step 3: Given the clustering $\mathcal{C}^*_{\Sigma,\pi}$, perform M iterations of a Metropolis Hastings random walk on W and π to obtain an approximate sample $(\Sigma,\pi)^{(1)},\ldots,(\Sigma,\pi)^{(M)}$ from the conditional posterior of $(\Sigma,\pi)|\mathcal{C}$.

Step 4: Set

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

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

Utility of co-clustering

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Measured using leave-one-out log-predictive scores.

t left out	Individual	Joint
4	-559.114	491.103
8	36.343	894.588
12	-914.612	-306.643
18	-1079.9	-977.726
Ave.	-629.321	25.331

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

Estimated cross-correlation matrix

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

	S. typhi	*Listeria*	*M. luteus*	*Zymosan*
S. typhi	1	0.782	0.624	0.614
Listeria	0.782	1	0.702	0.876
M. luteus	0.624	0.702	1	0.842
Zymosan	0.614	0.876	0.842	1

Clustering Experiments

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Using this method, we can establish a hierarchy of experiments and compare the effect of including/excluding the cross-experiment correlation:

- $\{S.typhi\}, \{Listeria\}, \{M.luteus\}, \{Zymosan\}$ (-629.321)
- $\{S.typhi\}, \{Listeria\}, \{M.luteus, Zymosan\}$ (-268.023)
- $\{S.typhi\}, \{Listeria, M.luteus, Zymosan\}$ (-41.745)
- $\{S.typhi, Listeria, M.luteus, Zymosan\}$ (25.331)

Figure in brackets is a measure of leave-one-out predictive ability; hierarchy listed in increasing order of predictive ability.

Pathway and Network Discovery

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Co-clustering can generate biologically plausible pathways.

e.g. Multiple challenge mosquito experiments

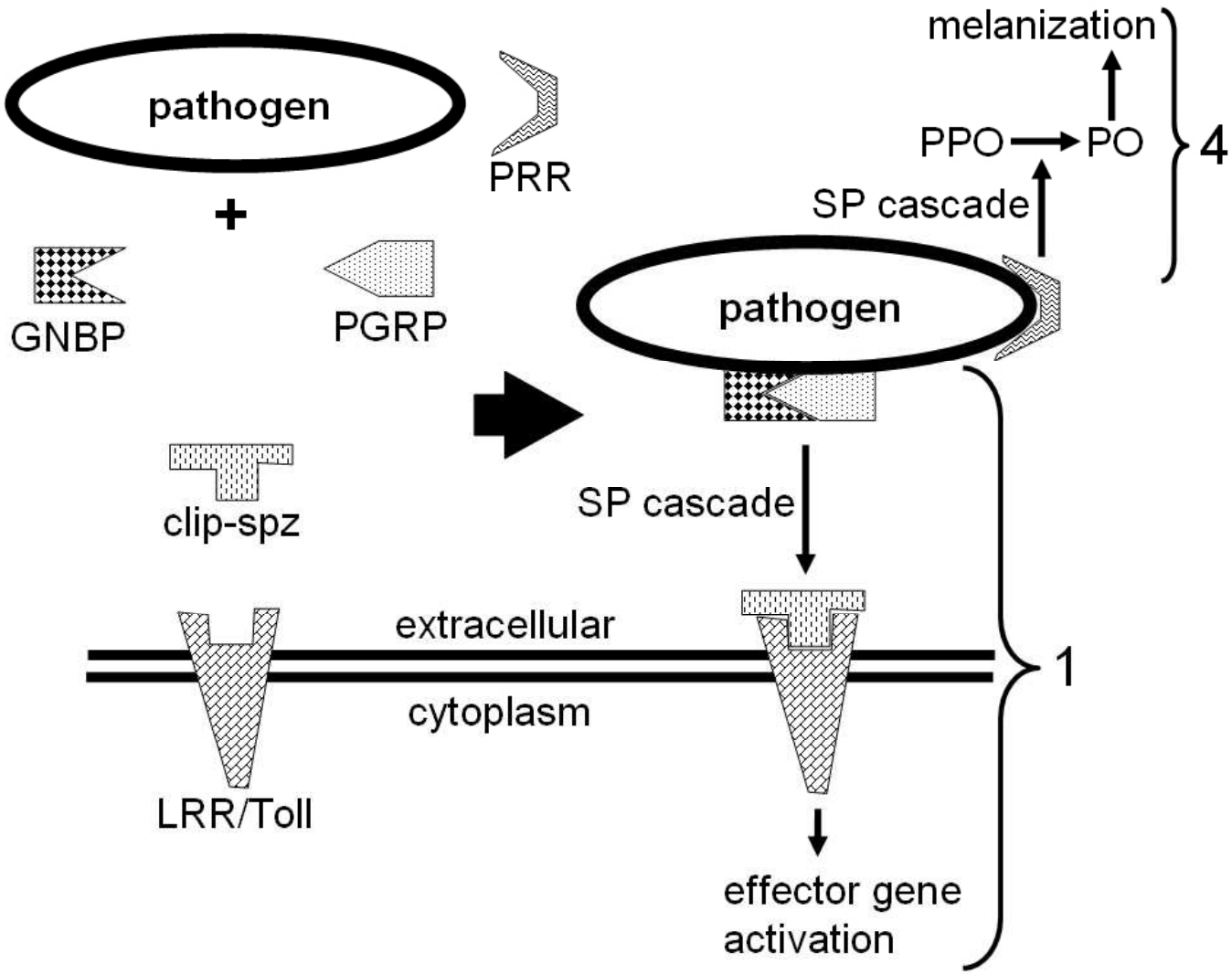
> *Cluster 1 comprises two pattern recognition receptors, GNBPB1 and PGRPLB, belonging to the gram-negative bacteria-binding gene family and the peptidoglycan recognition protein gene family, respectively. GNBPs and PGRPs have been shown to function in the same mechanism implicated in the activation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

See Heard et al. (2005).

Toll Pathway

Inferred from multiple challenge experiment:

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



Example: immune defense in humans

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

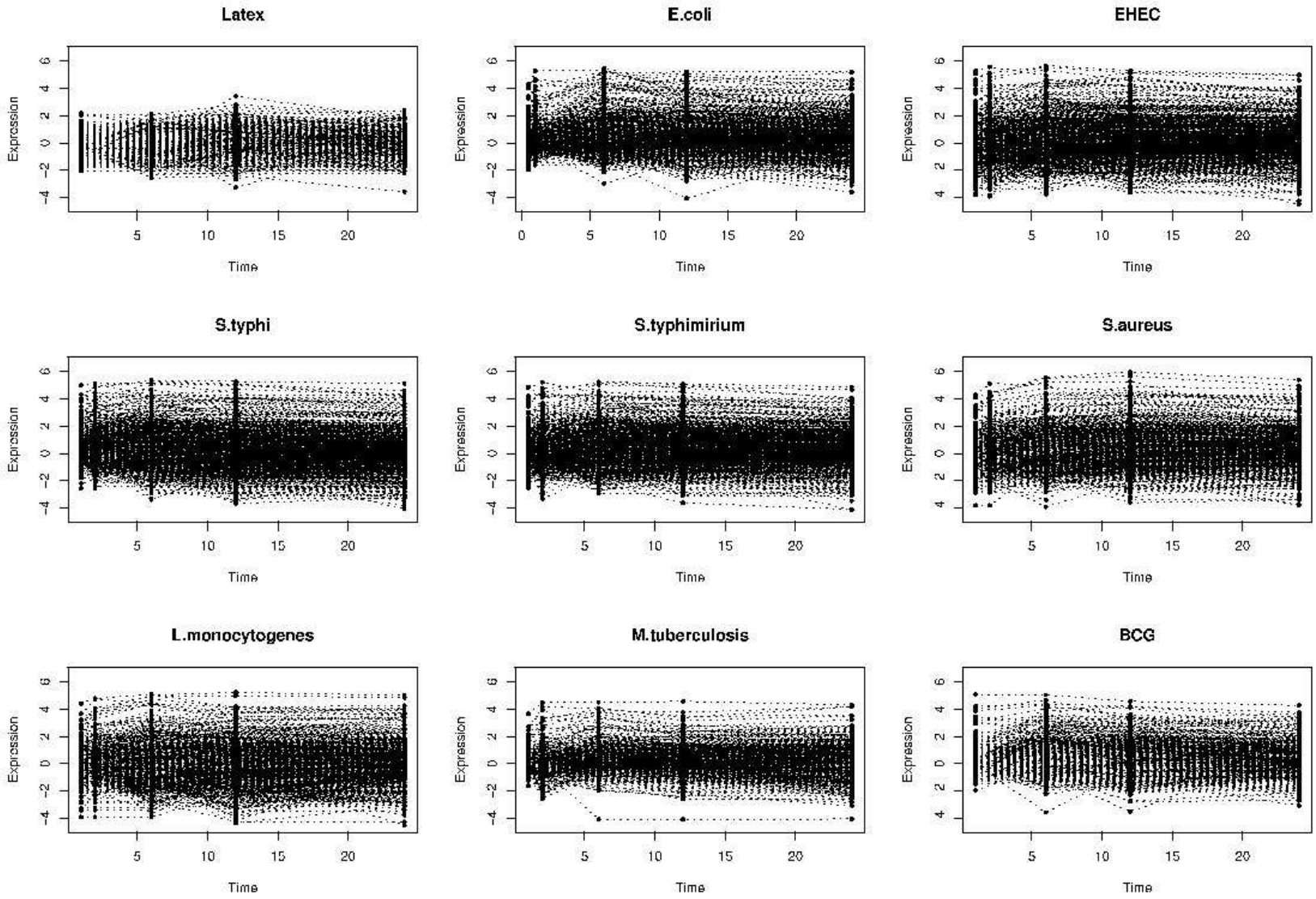
Nau *et al* (PNAS, 2002) investigated human macrophage activation induced by different bacterial pathogens.

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.

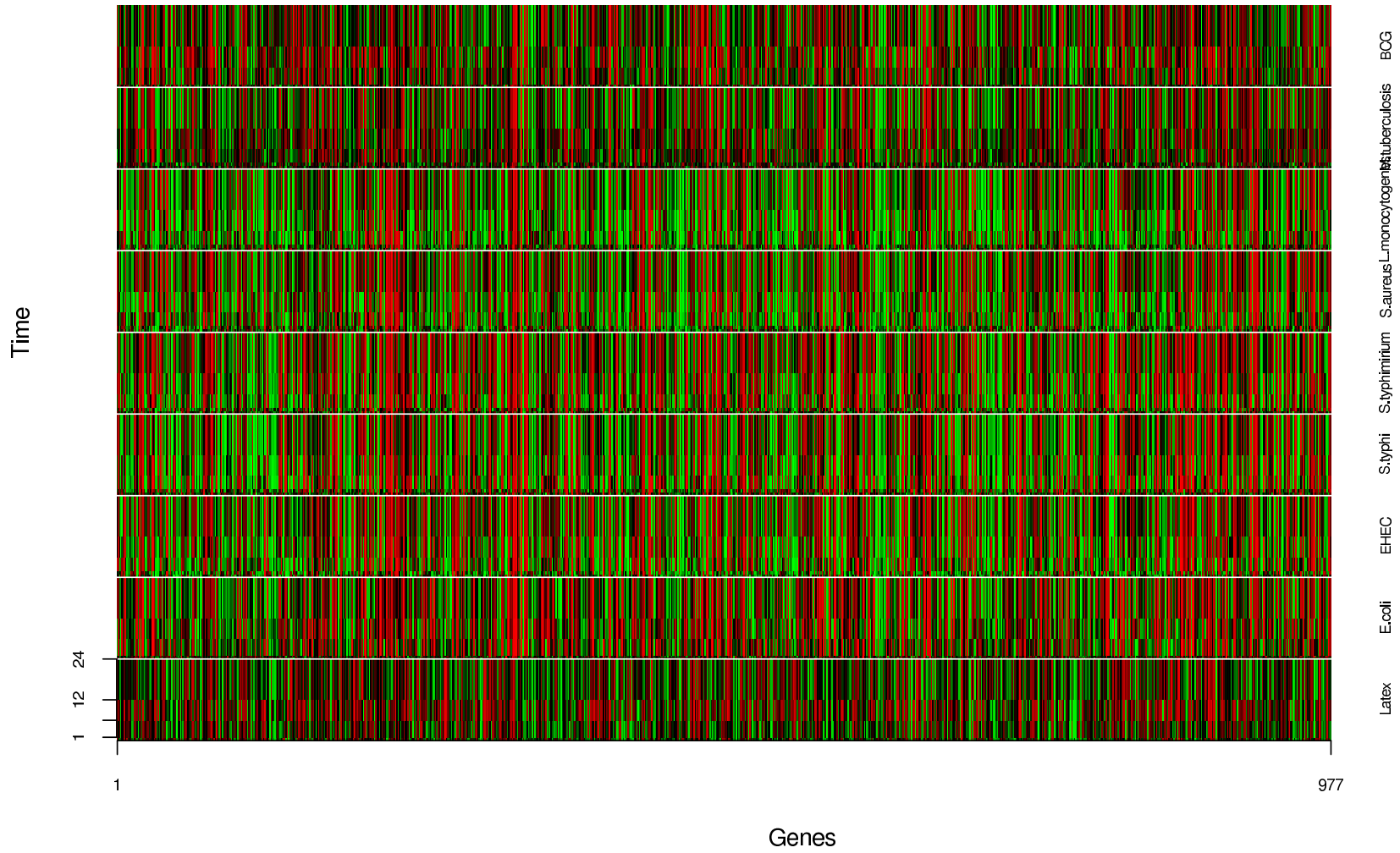
Nau *et al* data

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



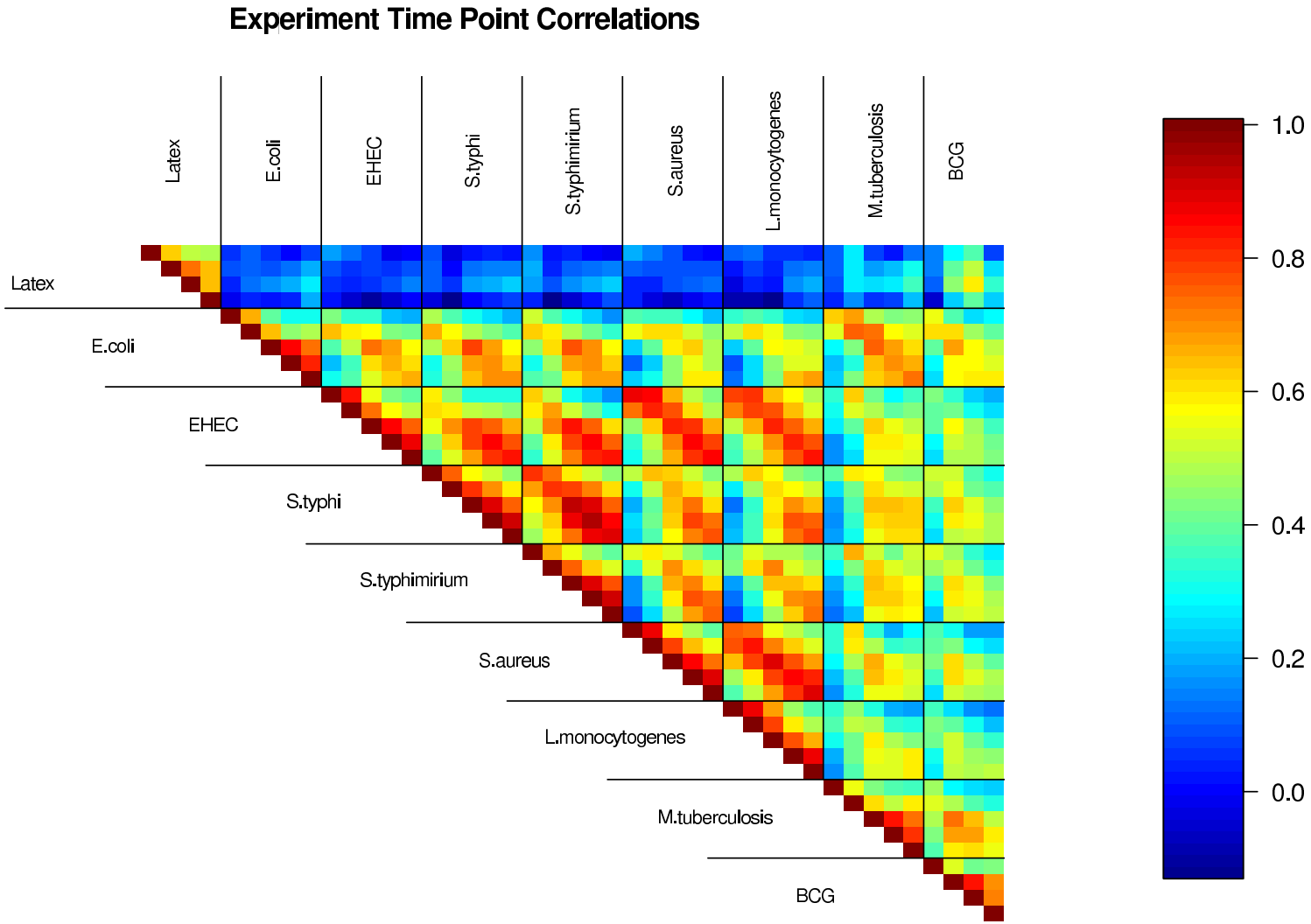
Unclustered expression data

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



Nau *et al* data, time point correlations

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



Data Fusion

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

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

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

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

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

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.

Solution: Develop Alternative model for Σ

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

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 $\{l_i\}$

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

$l_i \sim \mathrm{N}_p(0, \tau I_p) \quad \rightarrow$ 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.

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

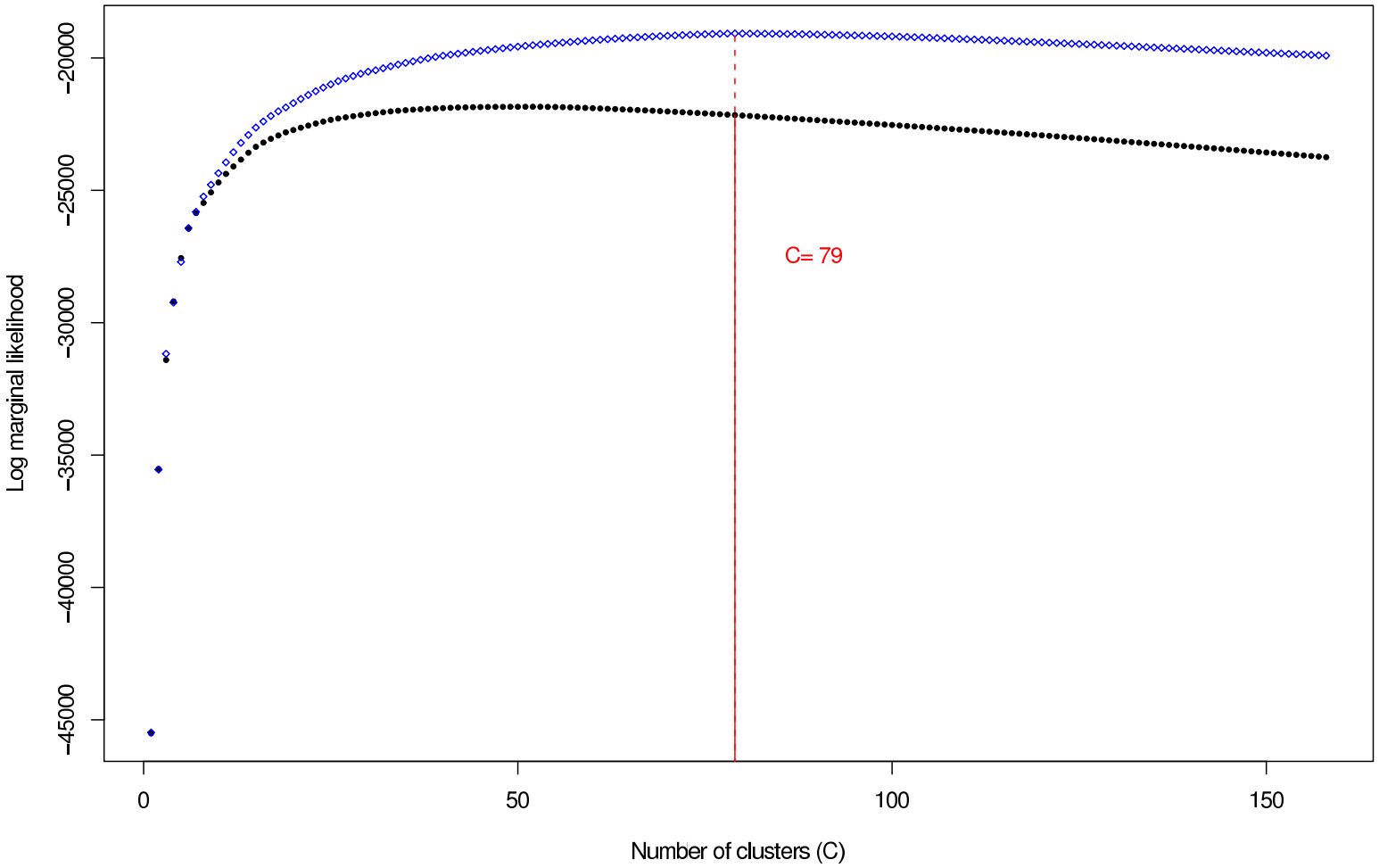
Maybe with $\lambda << \tau$.

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

Posterior score by number of Clusters

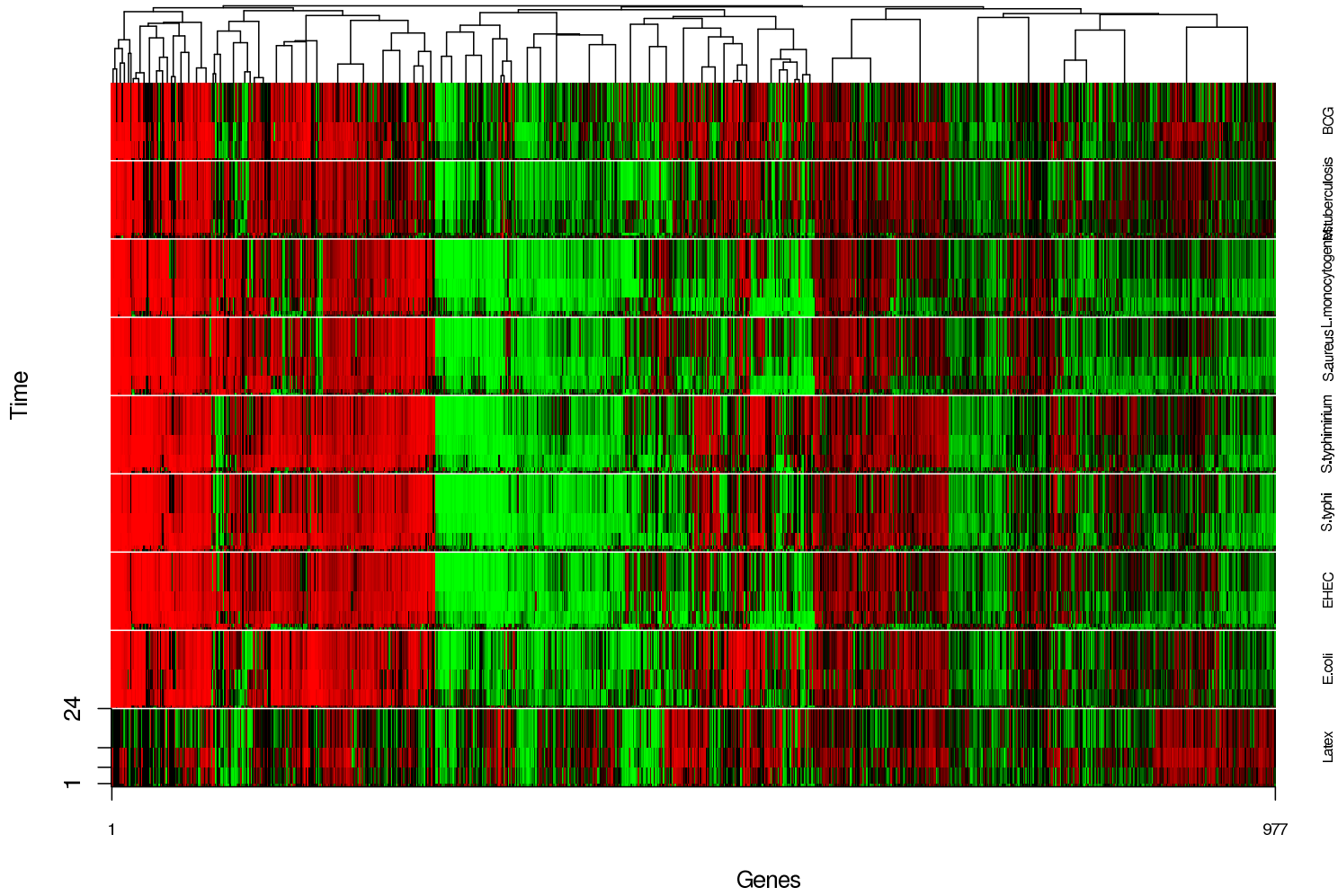
Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

With/without co-clustering.



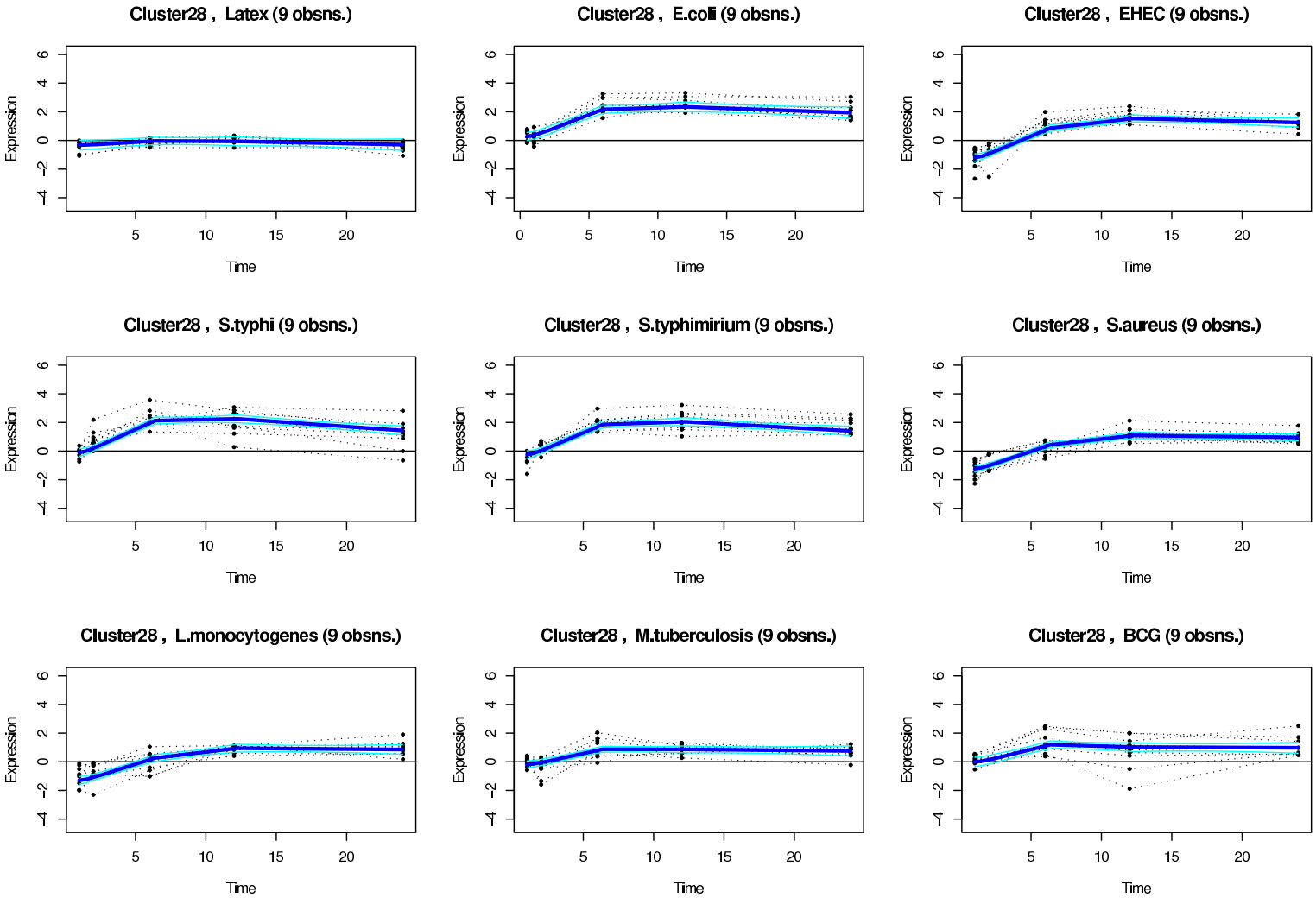
Clustered expression data

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



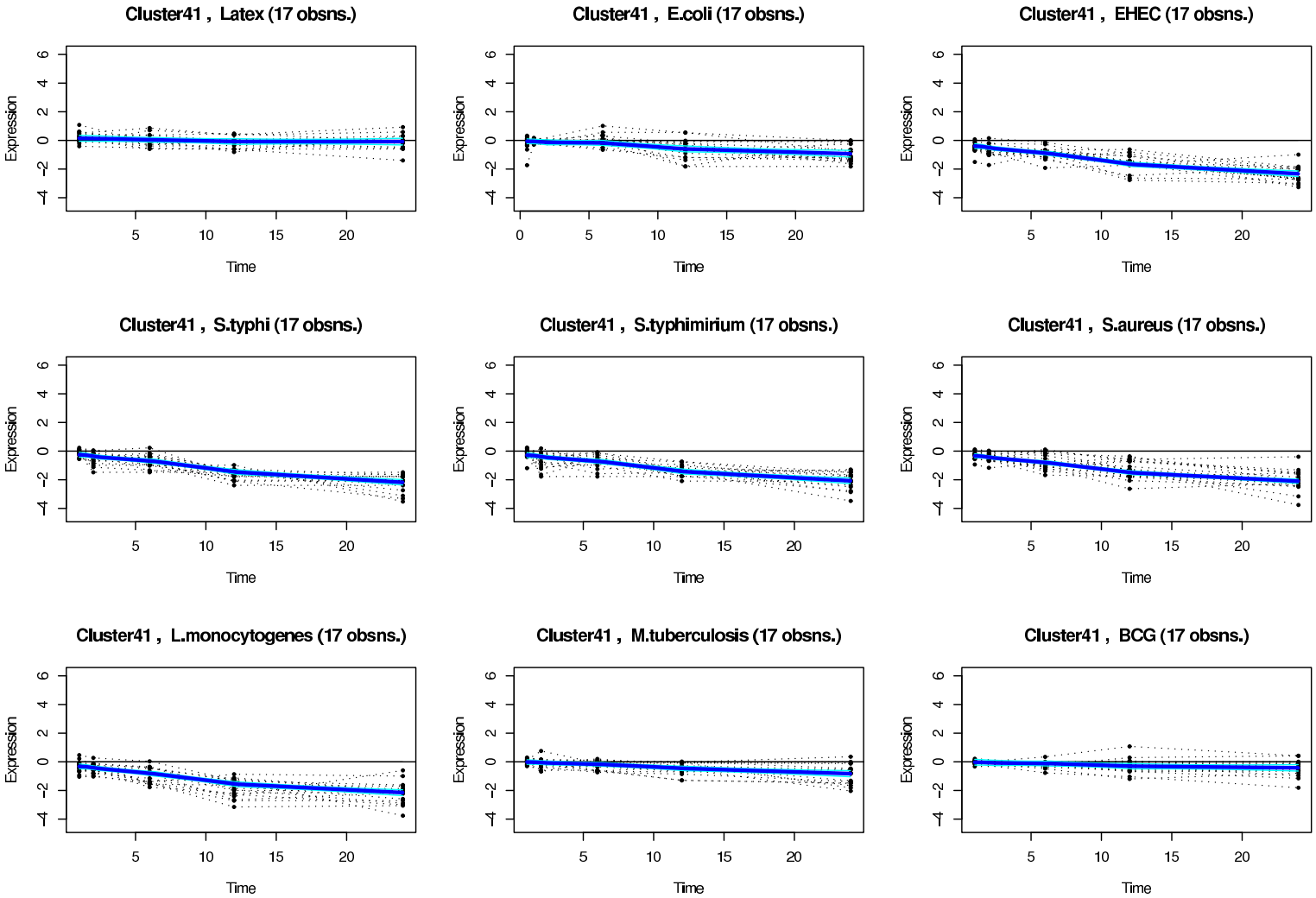
An up-regulated cluster

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



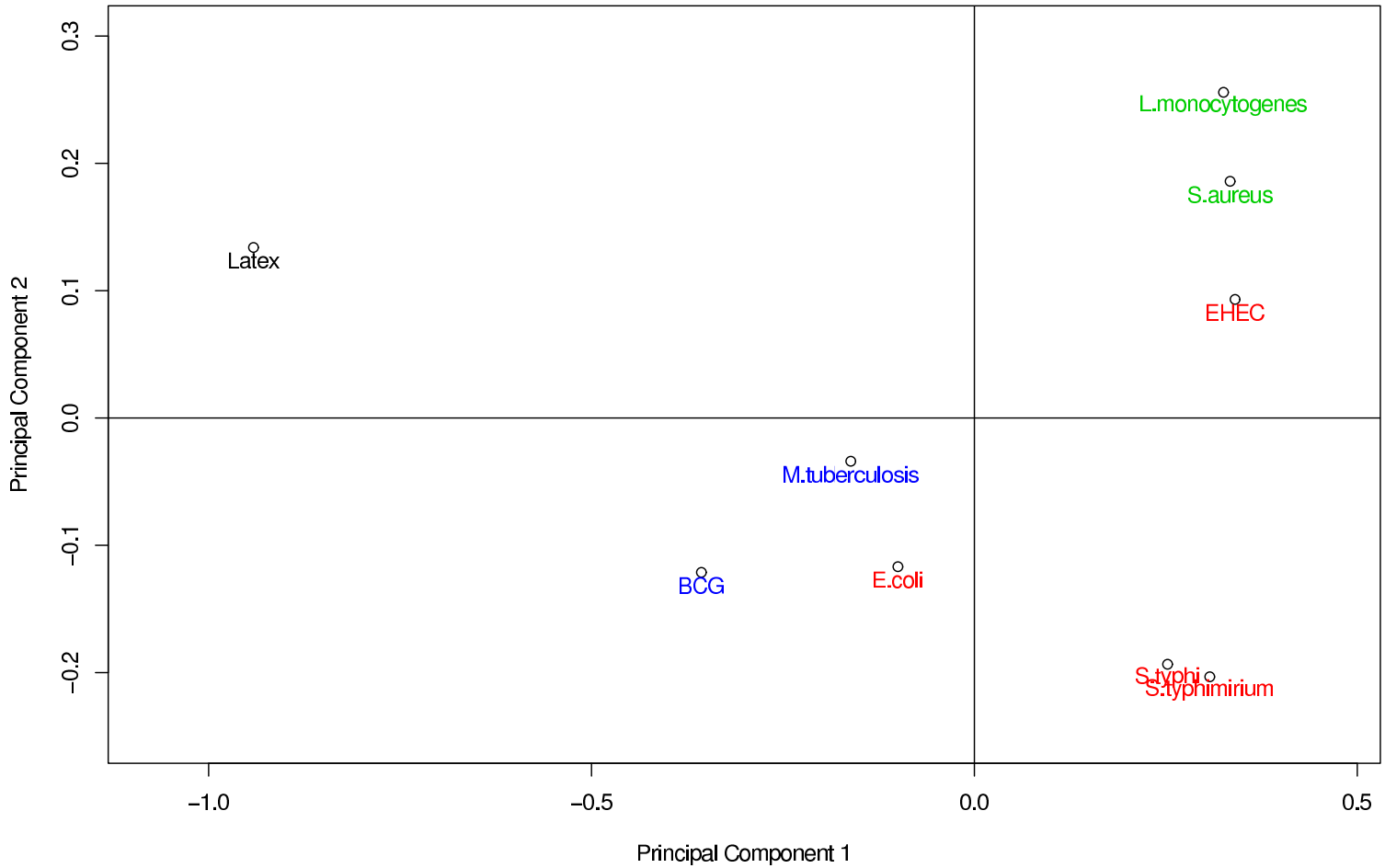
A down-regulated cluster

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



Treatment locations under hierarchical model

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal



V. Regulatory Models Derived From ODEs

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

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.

More details

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

- Related papers at
  - http://stats.ma.imperial.ac.uk/~das01/Papers/
  - http://stats.ma.imperial.ac.uk/~naheard/
- code
  http://stats.ma.imperial.ac.uk/~naheard/splinecluster

Other Bioinformatics Interests

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Include

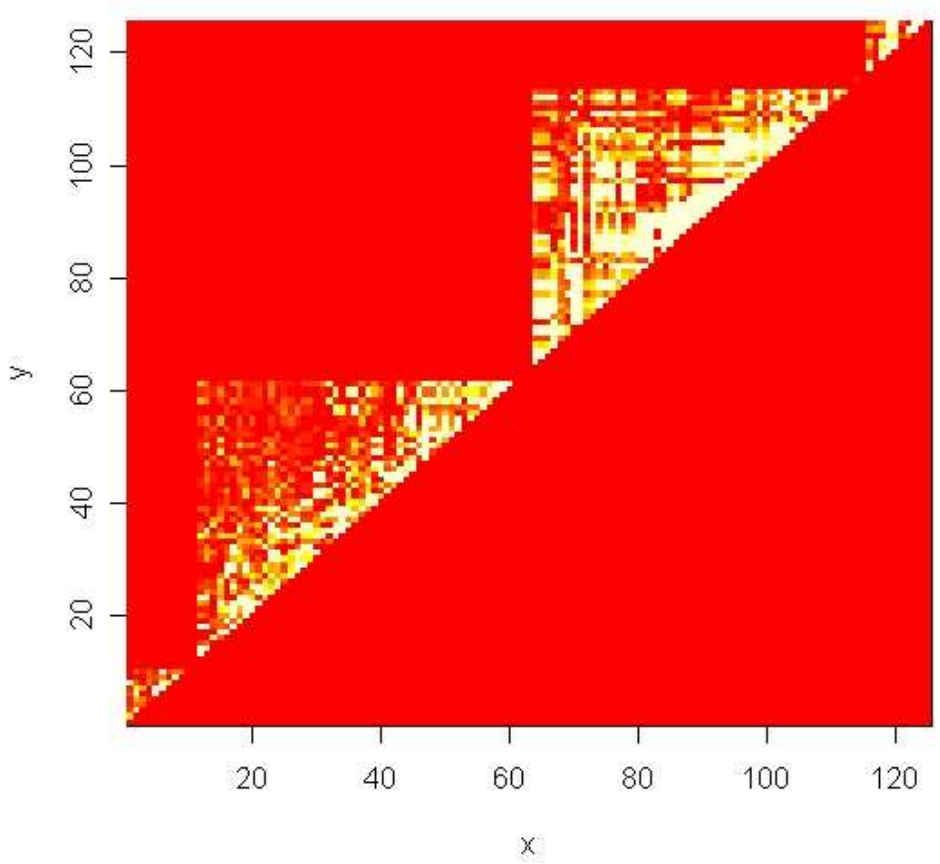
- High-throughput statistical genomics (SNPs/haplotype block structure) Pic
- Statistical analysis of biological point patterns Pic
- Metabonomics Pic
- Stochastic Trees and MCMC

End

Haplotype Blocks

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

LD Plots: Block-like structure uncovered by linkage disequilibrium map.

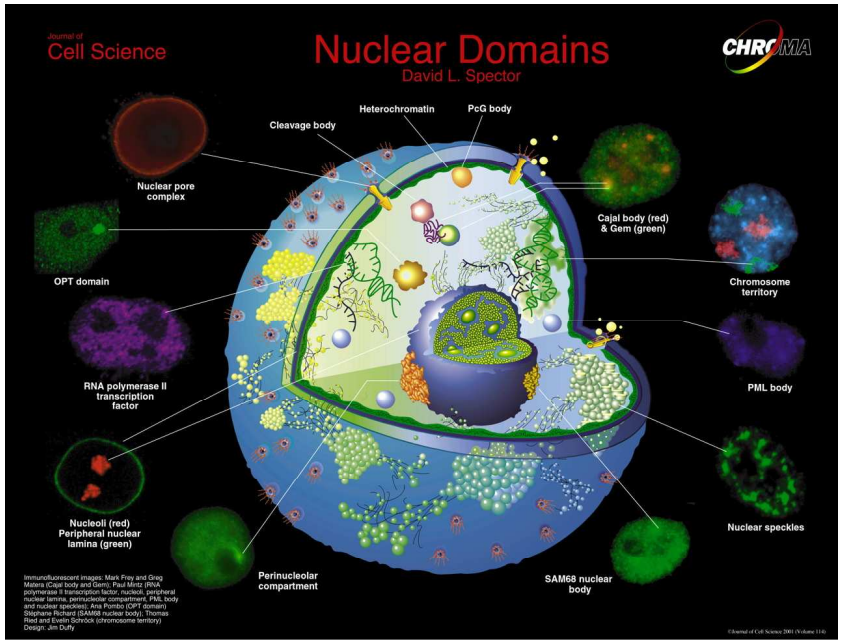


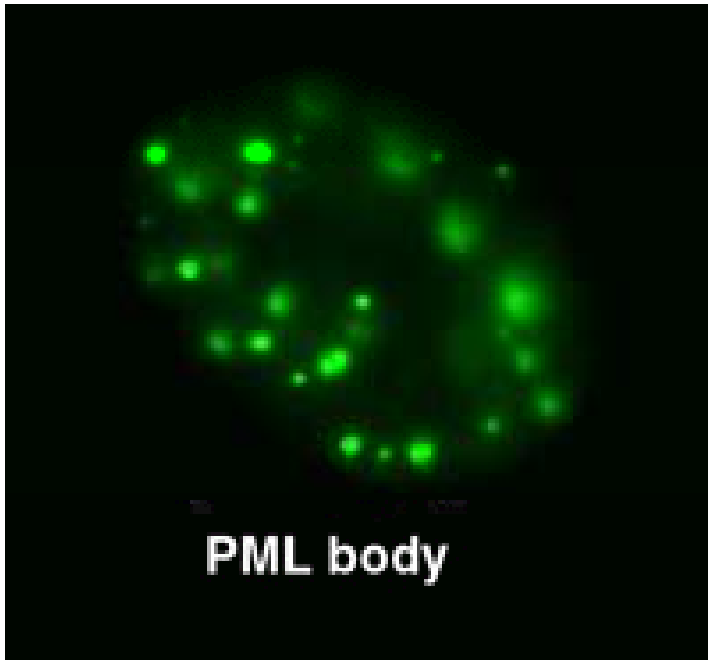
Back

SPPs

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Nuclear Domains: PMLs are one specific domain of interest.



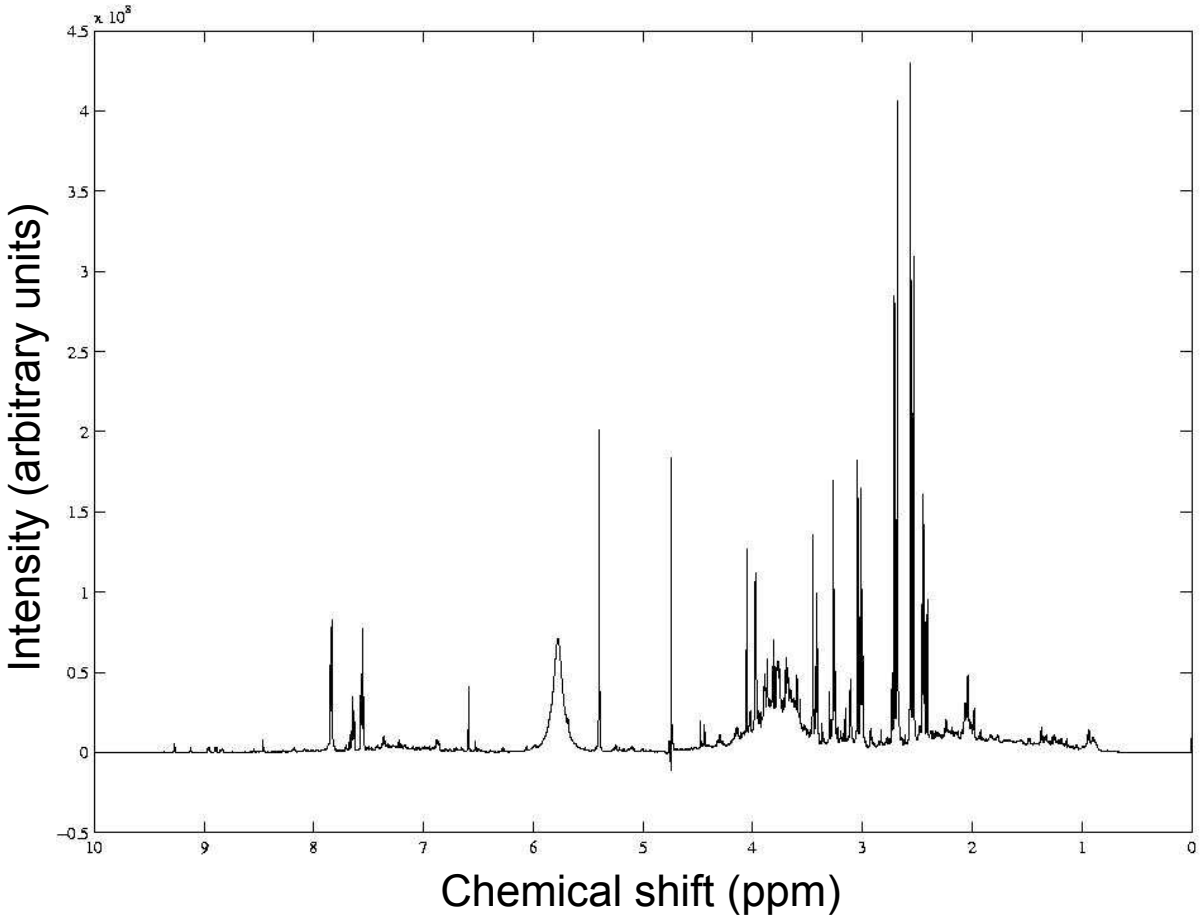


Back

Metabonomics: NMR Spectrum

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Chemical Analysis of Biofluids via NMR spectrometry



Back

Acknowledgements

Motivation
Microarrays
Statistical Analysis
Modelling
Clustering
Results
P. Falciparum
Multiple Experiments
A. Gambiae
Nau data
Future Goal

Main co-authors:

- Nick Heard, Imperial College, UK.
- Chris Holmes, University of Oxford, UK.
- George Dimopoulos, Johns Hopkins School of Medicine, US.
- Ajay Jasra, Imperial College, UK.

Funding: Wellcome Trust, BBSRC, MRC.