Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

Time Series Data in Functional Genomics

Regression, Functional Data Models and Correlated Data

David A. Stephens

[1]Department of Mathematics and Statistics, McGill University

d.stephens@math.mcgill.ca



http://www.math.mcgill.ca/~dstephens/MCB/

Outline

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

- Time course data
- Linear Regression Models
- Estimation and Inference
- Flexible Models: Splines
- Model-based Clustering
- Models for correlated data

Time Course Microarray Data

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

Microarrays are used to record the relative gene expression of many thousands of genes simultaneously.

This is useful when we wish to compare the functional activity across different genetic subgroups

- wild-type vs knockout
- heterozygote/double homozygote
- different developmental stages

Microarrays can detect differential expression across these subgroups.

Time Course Microarray Data

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

A more powerful experimental set-up is where we detect **changes in expression through time**.

A typical design would involve carrying out repeated microarray experiments on similar experimental units a different times over a number of hours or days that would allow us to measure how the differential expression changes in time.

This allows us to better understand **patterns of regulation**.

Motivation

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

We motivate the subsequent statistical models by looking at several real examples

- the principal model organism is the *Anopheles Gambiae* mosquito
- we study patterns in regulation of immune defence mounted in response to bacterial and chemical challenges
- this informs the study of malaria, and the immune defence of the mosquito to infestation by the protozoan parasite *Plasmodium Falciparum*

The Global Impact of Malaria.

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

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



Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

- Alongside HIV and tuberculosis, it represents one of the world's most damaging infectious diseases.
- Malaria affects two to three hundred million people each year, one million of whom are children living in sub-Saharan Africa.
- Globally, two thousand million people (40% of the world's population) are at risk.
- Malaria research is ongoing: a key element is to understand genetic regulation in the mosquito and in the parasite.

Mosquito and Parasite Genomics

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

- 2002: genomes of *P. Falciparum* and *A. Gambiae* mapped
- 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 systematically dissected.
- special emphasis has been placed on the study of anti-malarial responses involved in limiting the extent of infection.

Microarray studies

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares Mathematical Formulation

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

- The immune defense system of the mosquito to infestation by the parasite has come under study.
- 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.

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

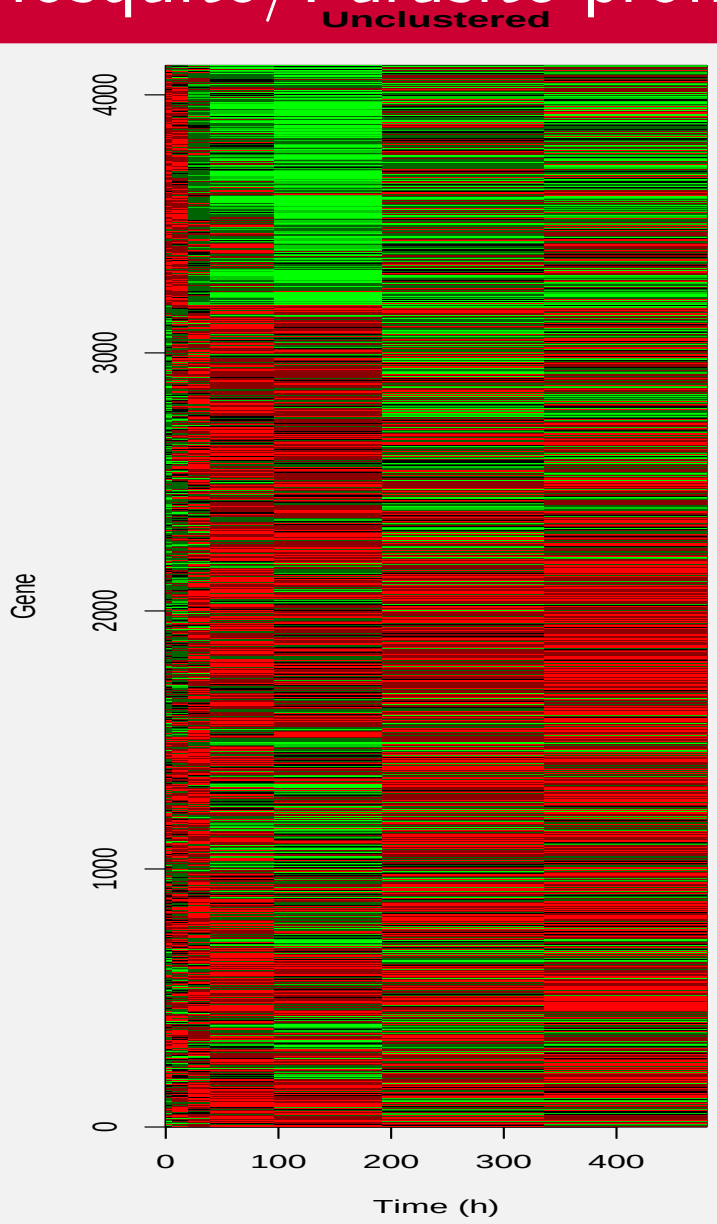
- clear correlation of immune responses with the passage of the parasite through the vector.
- the mosquito has become the organism of choice for directly studying antiparasitic innate immune responses.

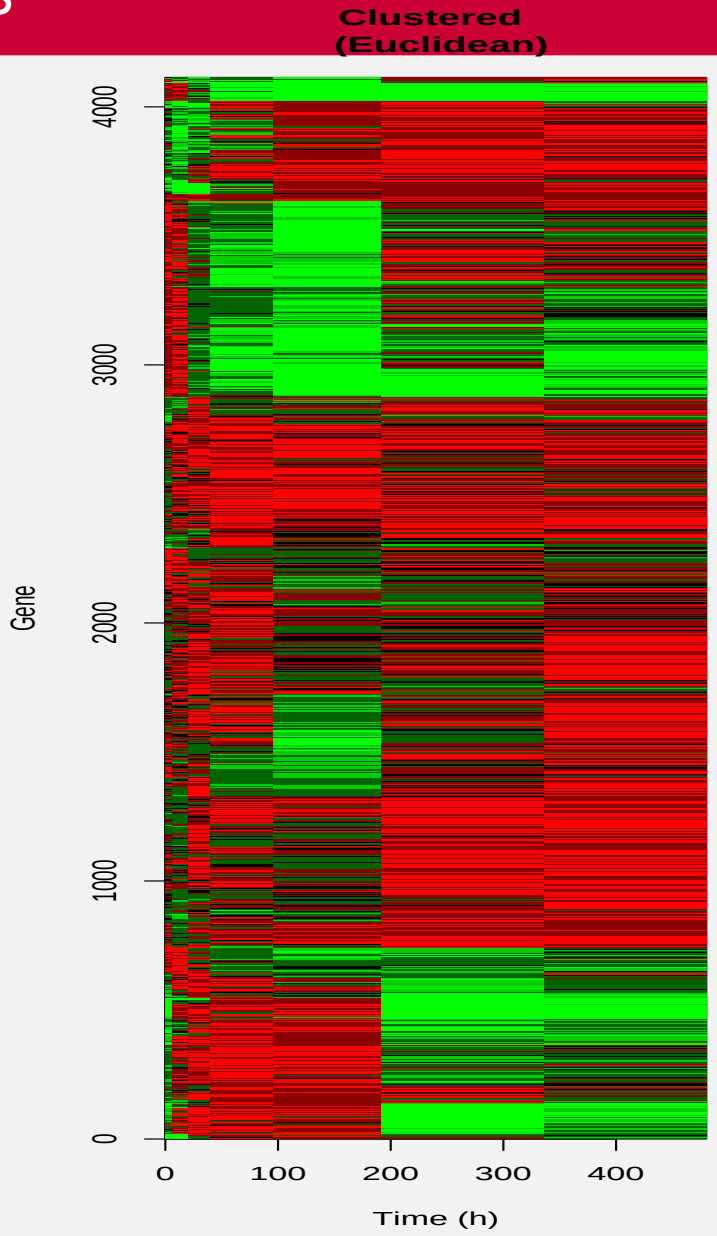
Recent genomic investigations in malaria include studies of

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

Mosquito/Parasite profiles

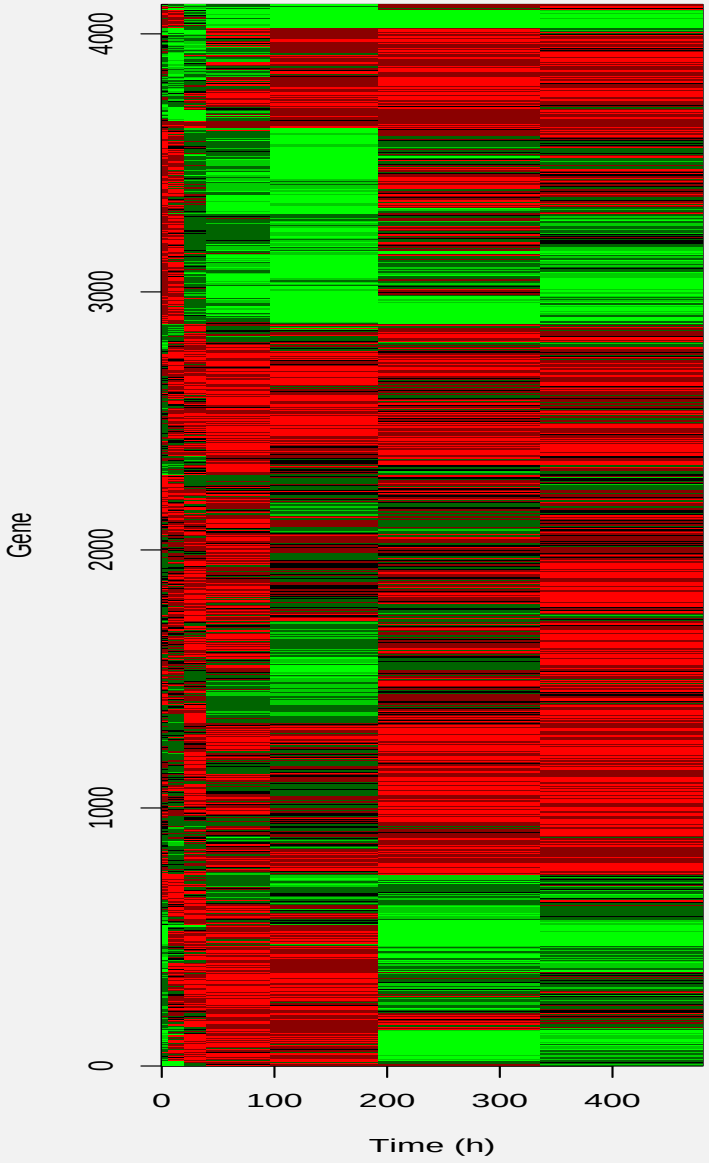
Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

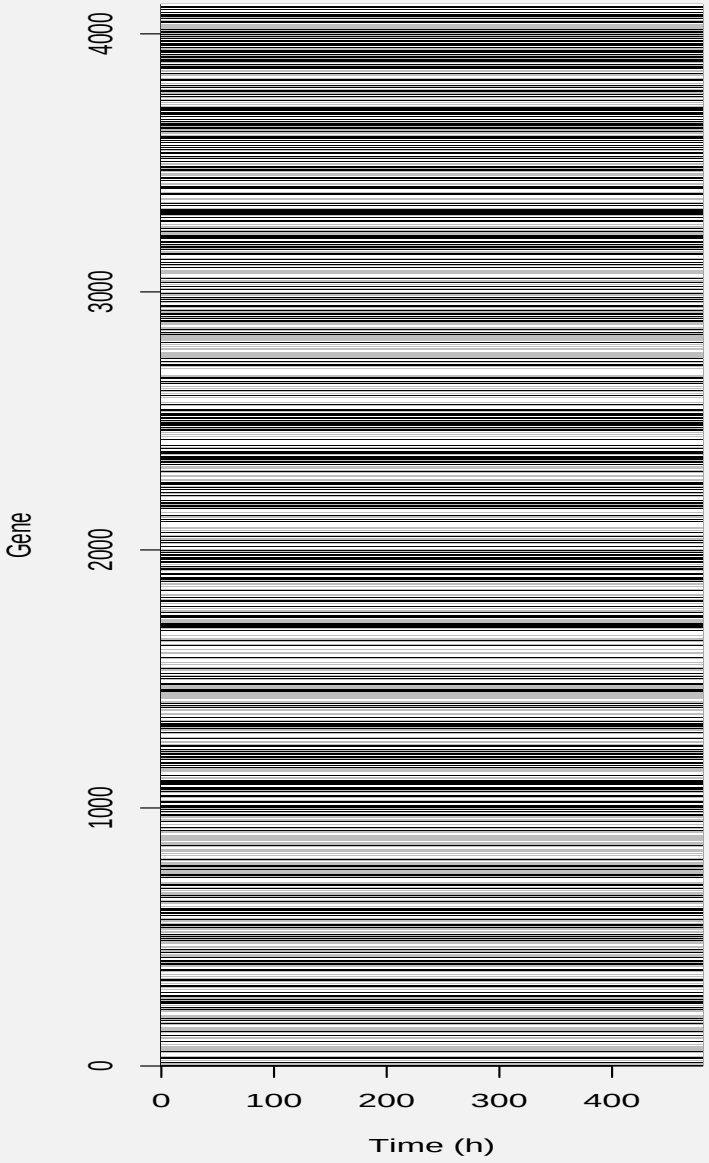




Profiles cluster within species ?

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation





Objectives

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

- 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)

Objectives

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

- to produce models and algorithms that can and will be used by researchers
- biologists often reluctant to use advanced statistical methods
- computational feasibility is an important factor

Time Course Data

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

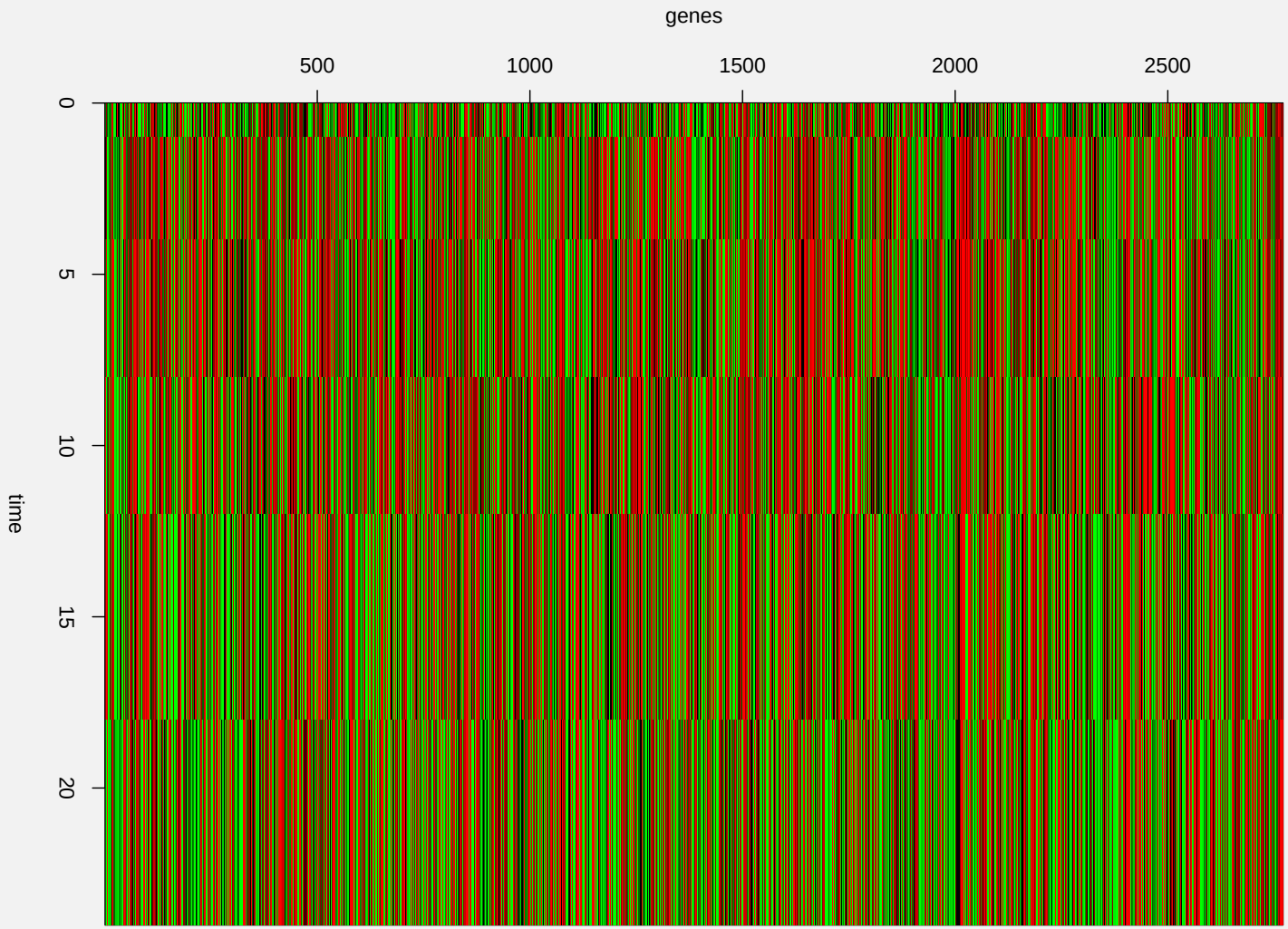
Statistical analysis of gene expression profiles obtained by microarray assays of mosquitos/cell-lines compromised by bacterial and chemical agents (challenges):

- data comprise 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.

We focus on a single bacterial challenge, *Salmonella typhi*.

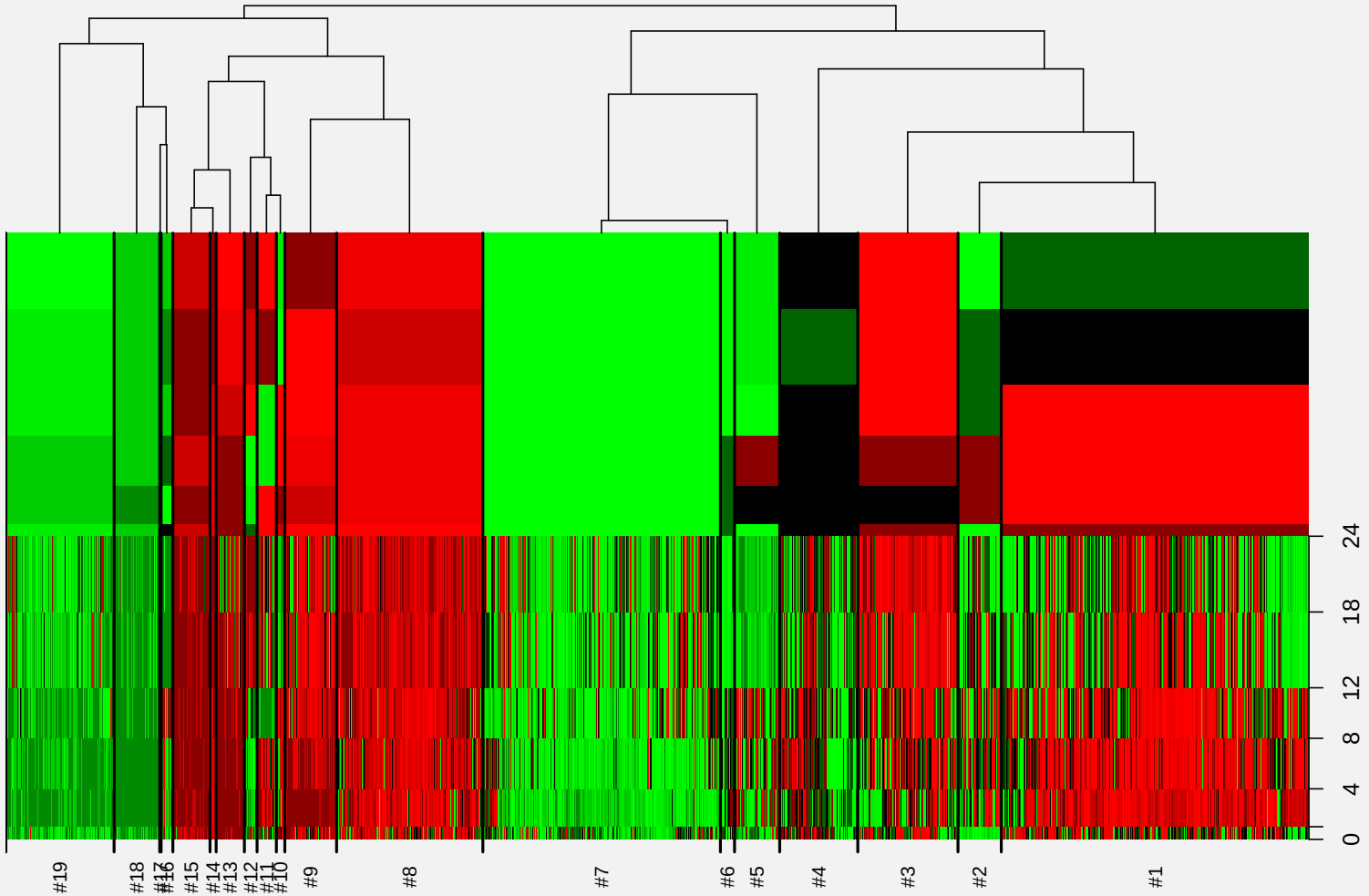
Salmonella typhi challenge data.

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



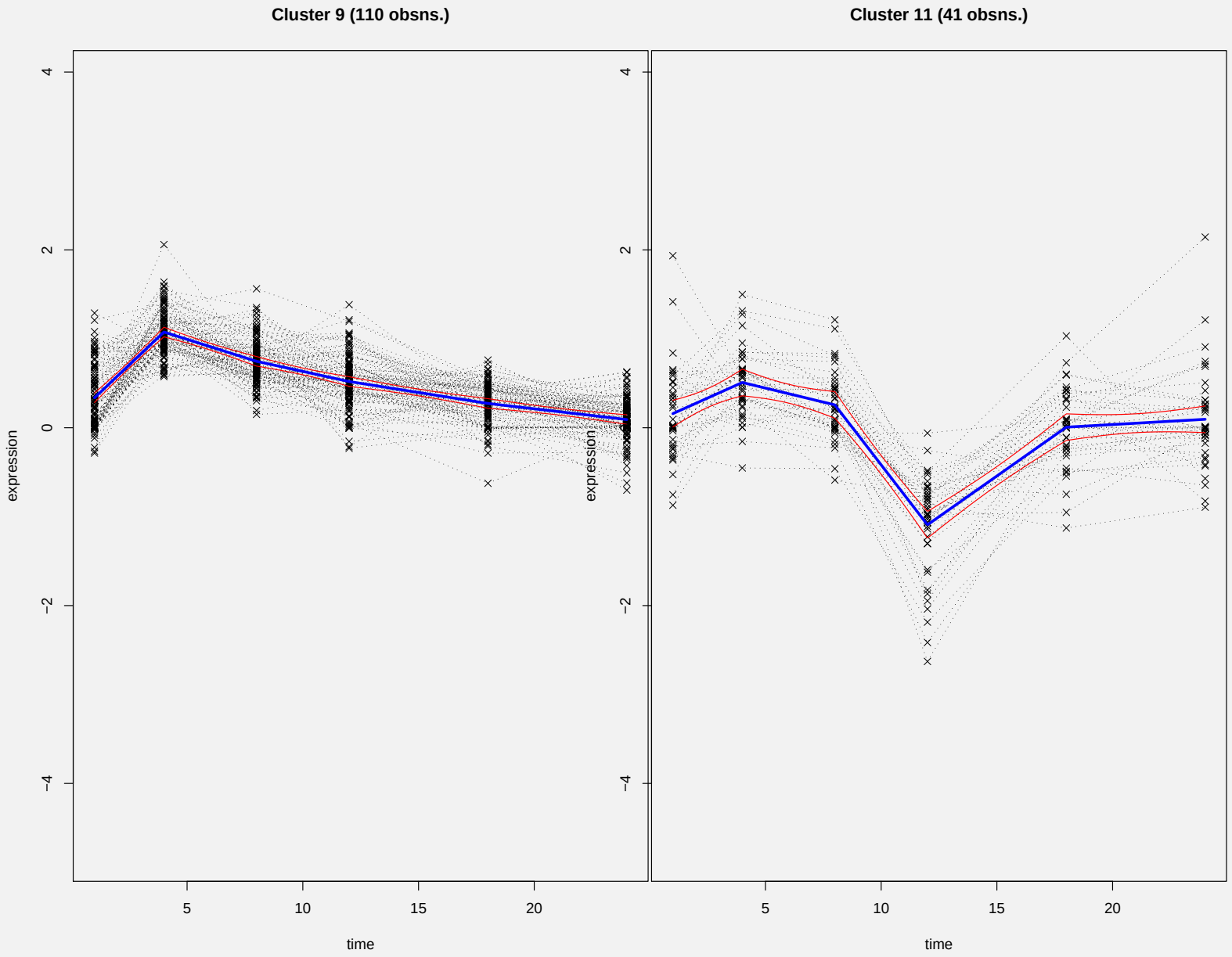
Clustering Results

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



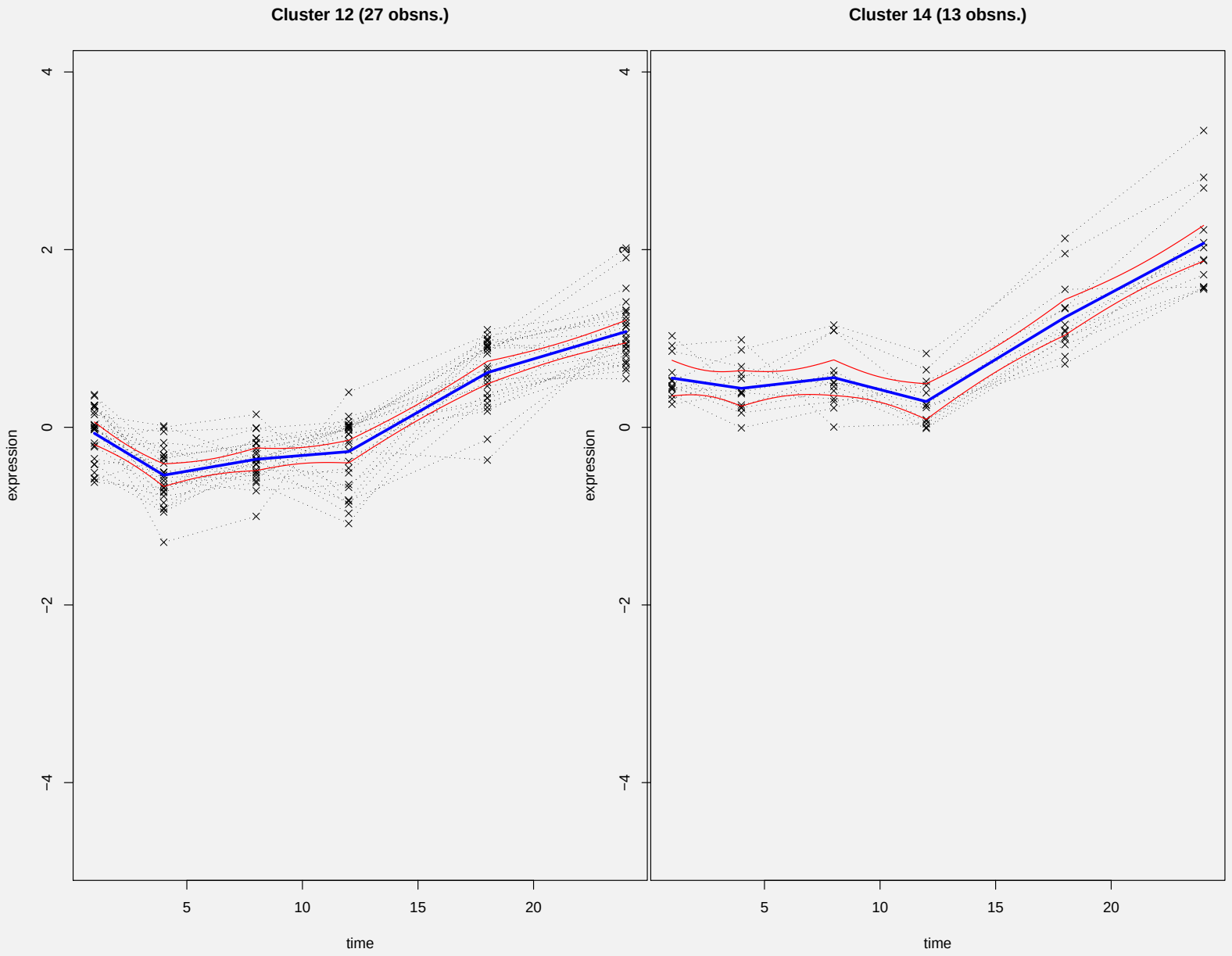
Clusters 9 and 11

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



Clusters 12 and 14

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



Co-clustering for Multiple Challenges

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

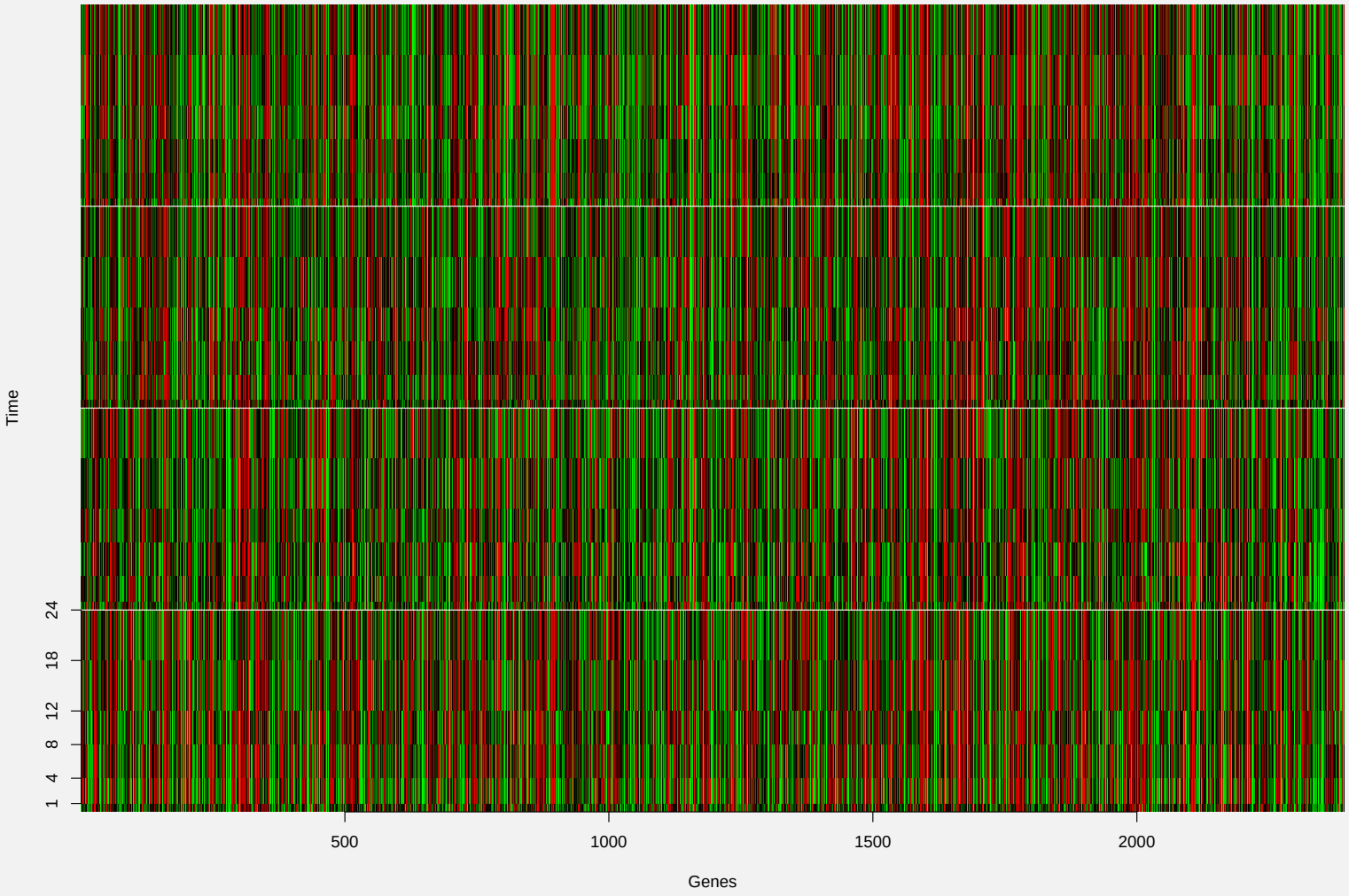
The *S. Typhi* challenge data were obtained as one of a series of experiments using (~ 15) different challenges on the same gene set. Here we look at only four:

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

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

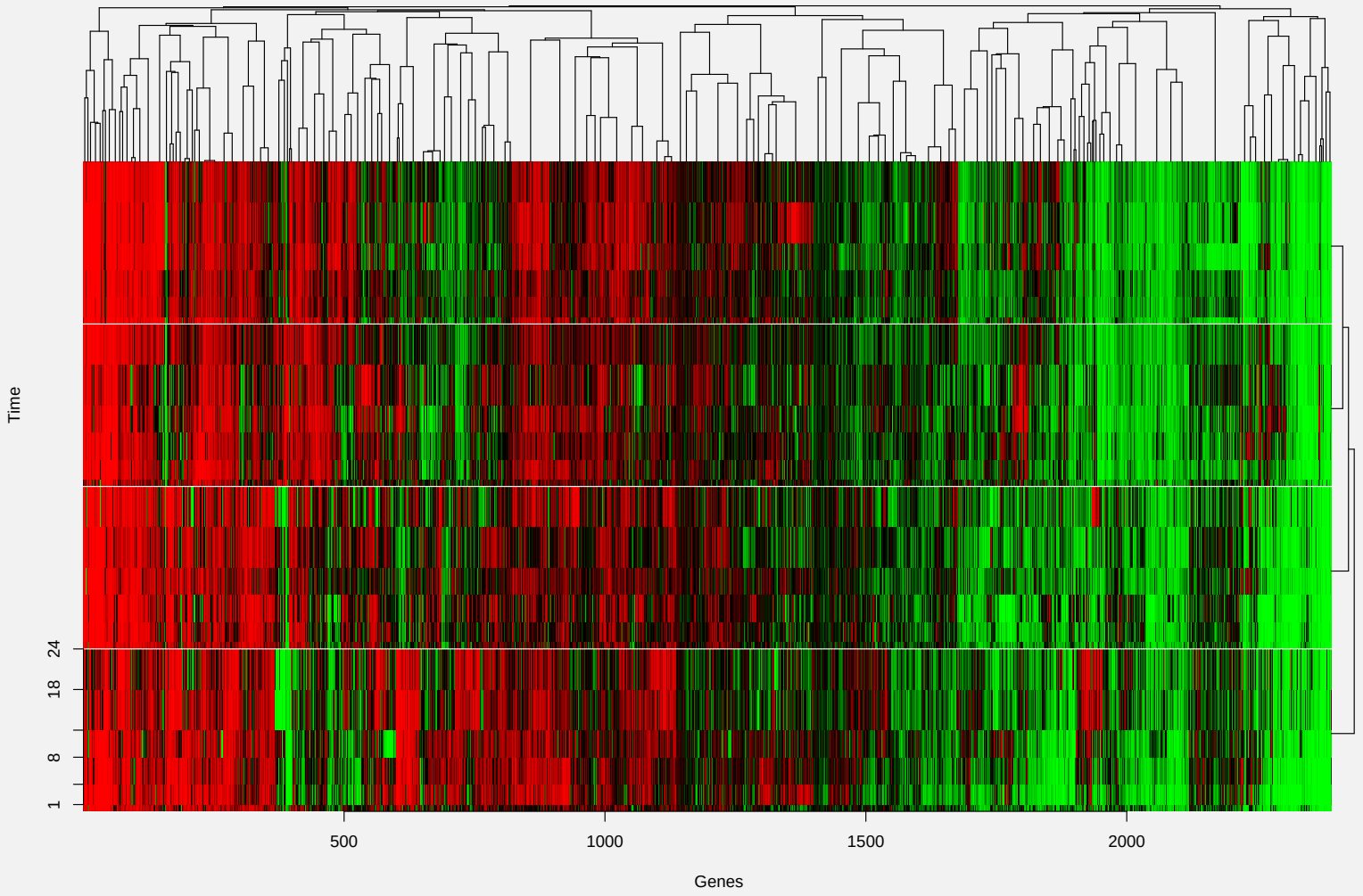
Unclustered

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum
CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



Co-clustered

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



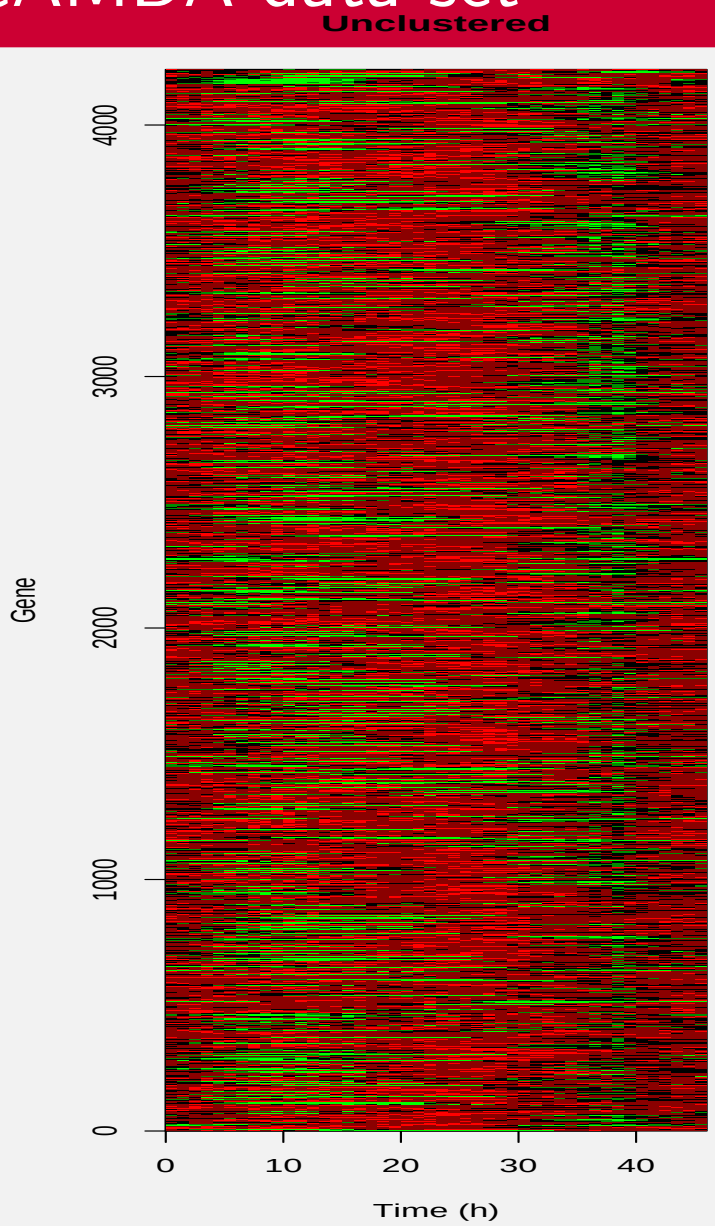
Gene Regulation in *P. Falciparum*

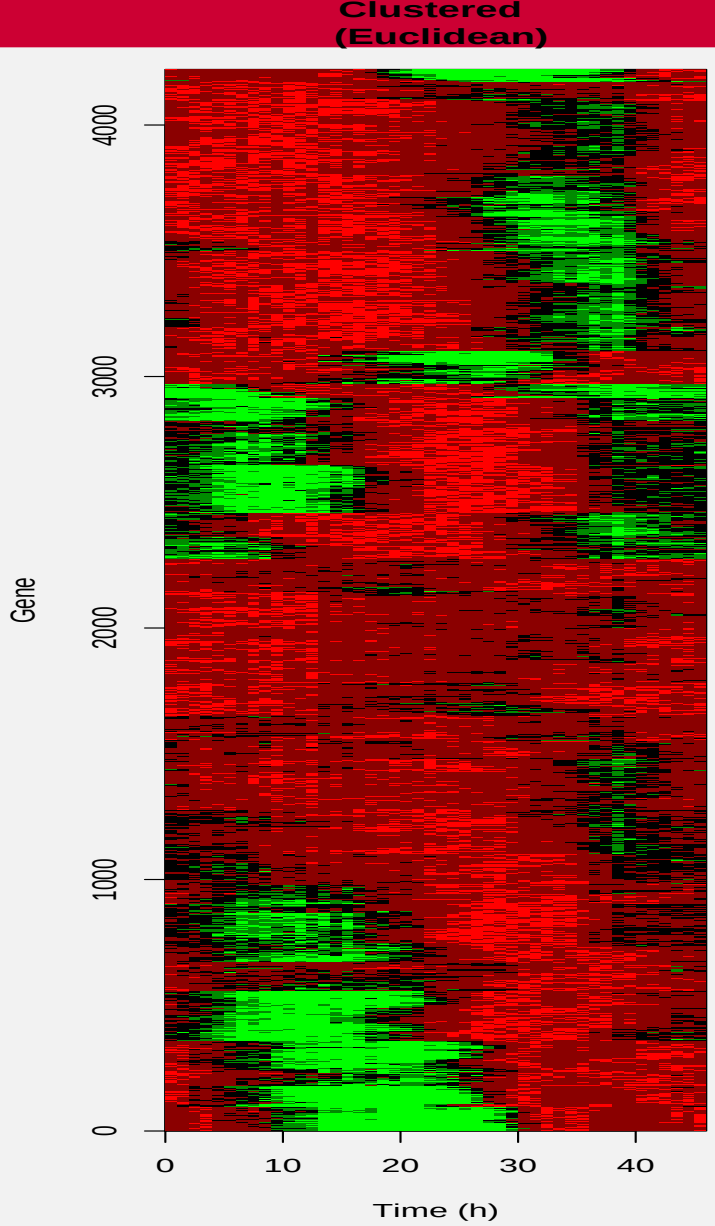
Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

- 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 data set

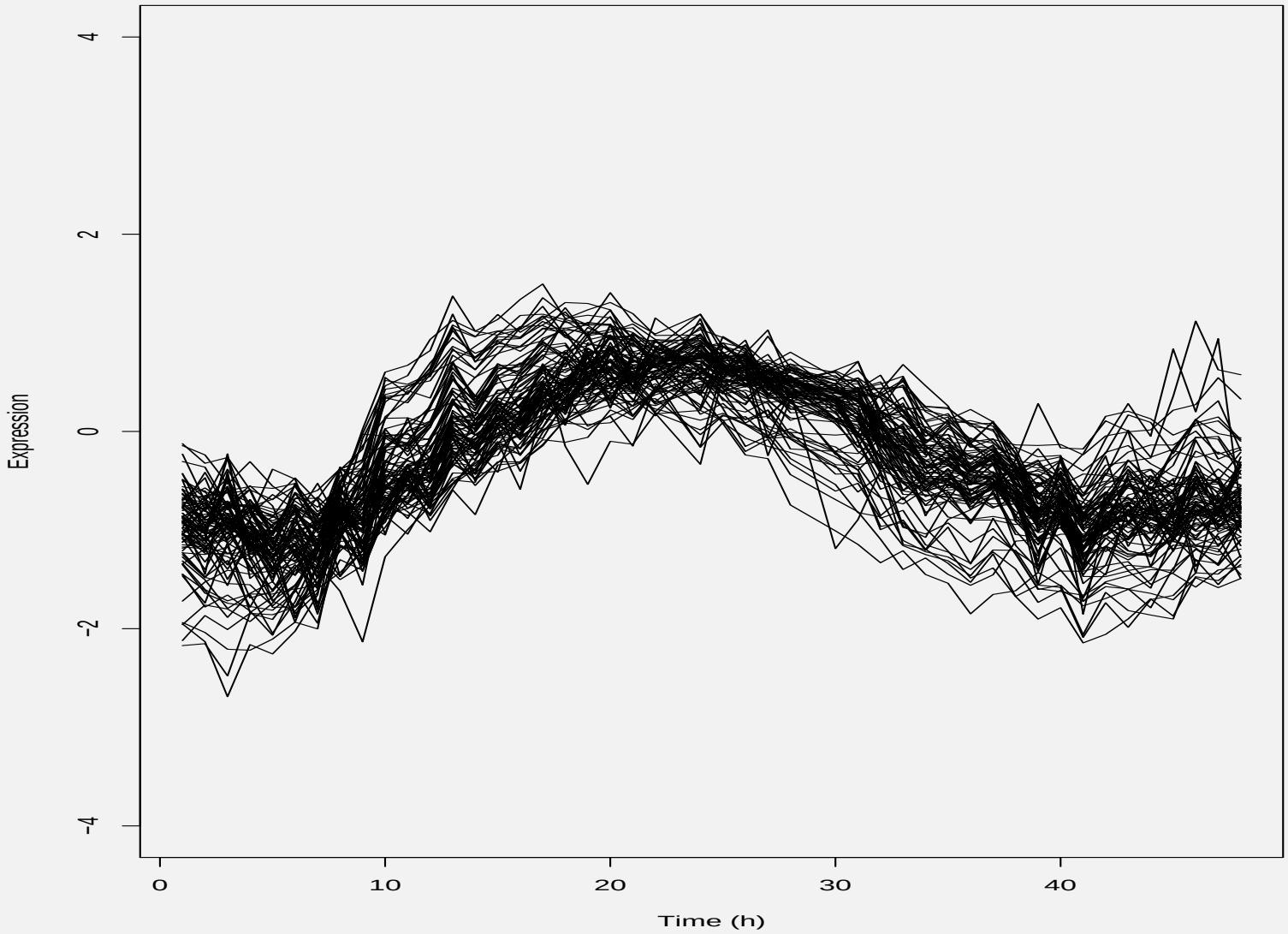
Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation





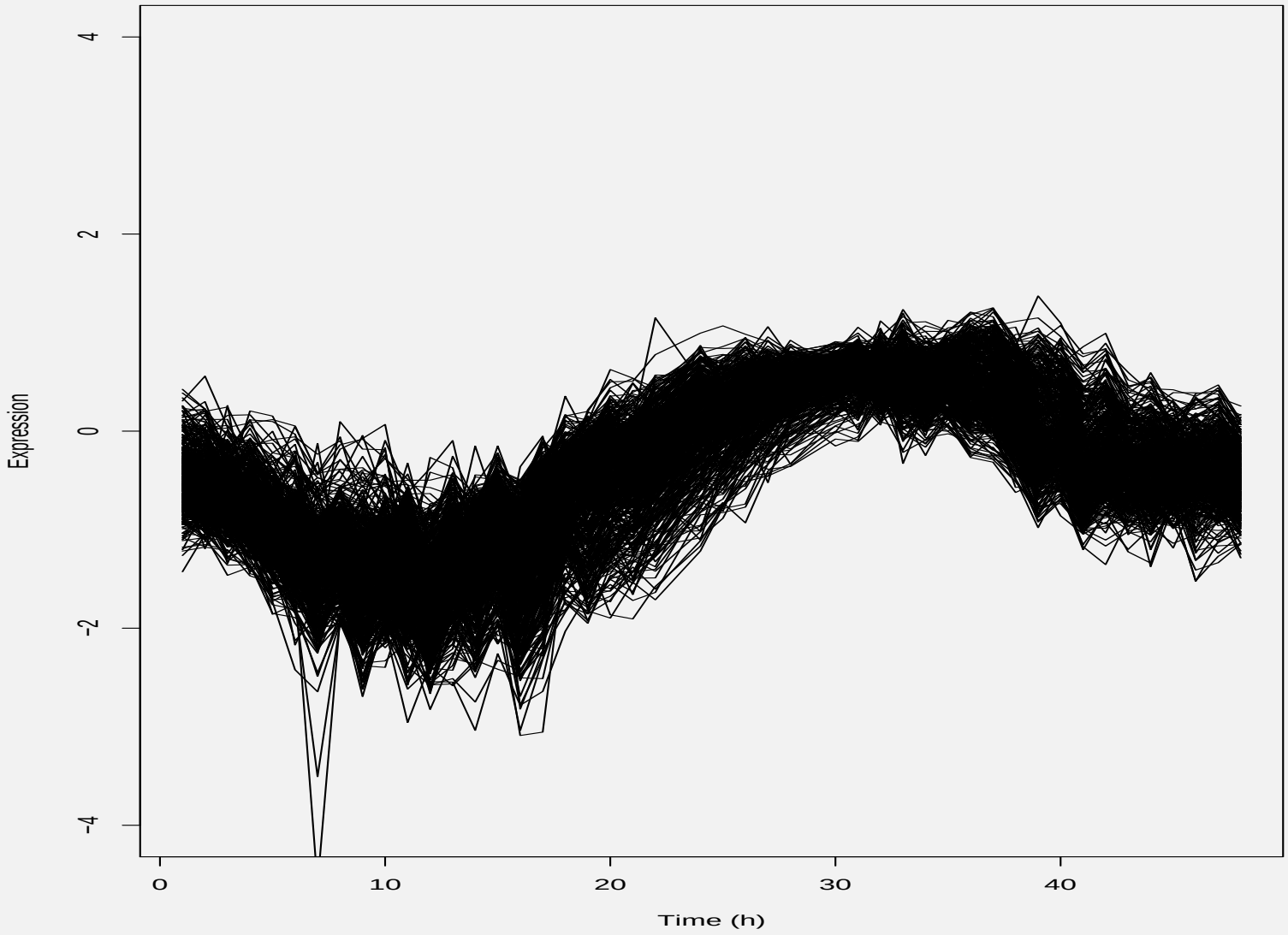
Cluster 4

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



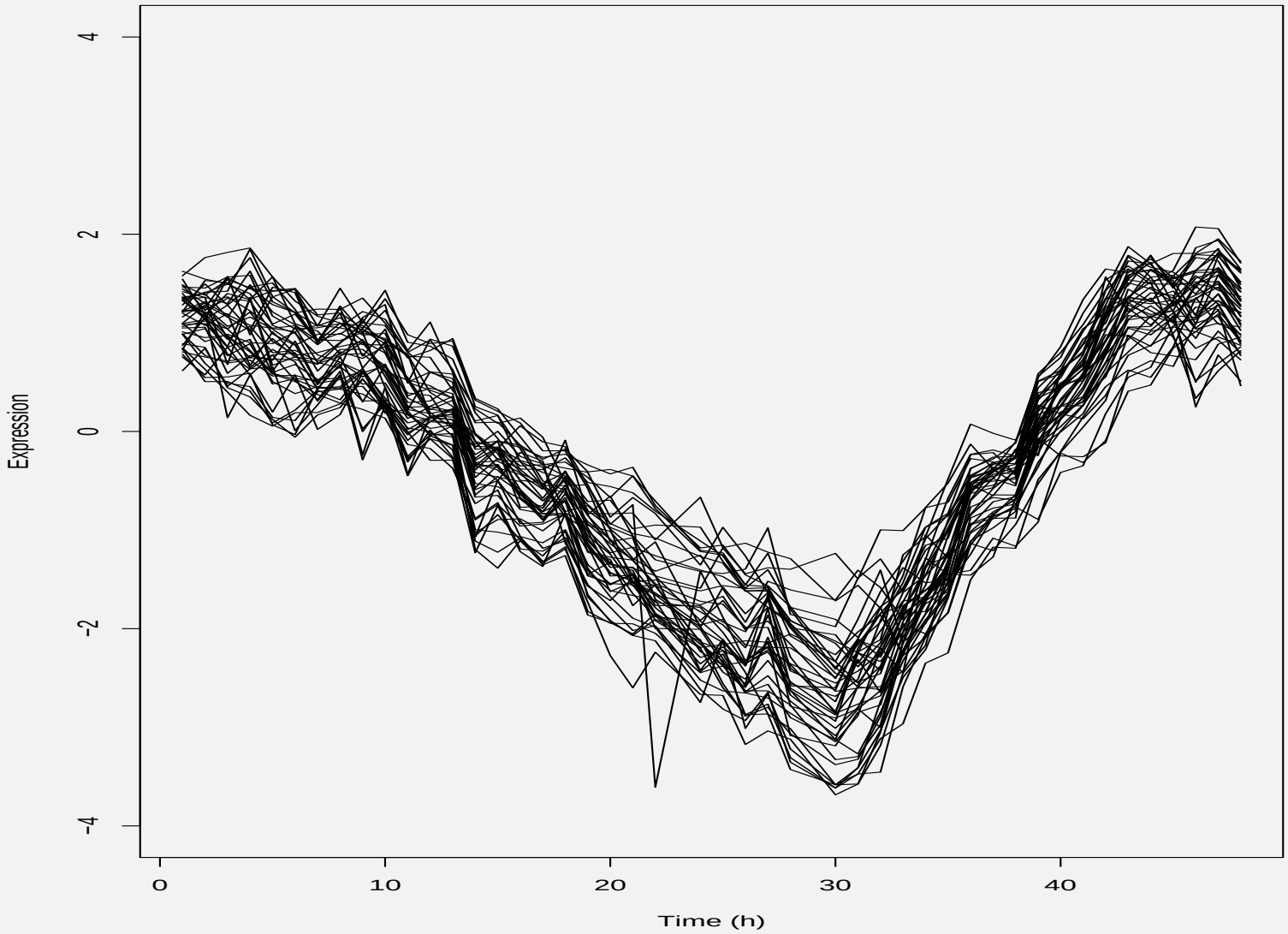
Cluster 6

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



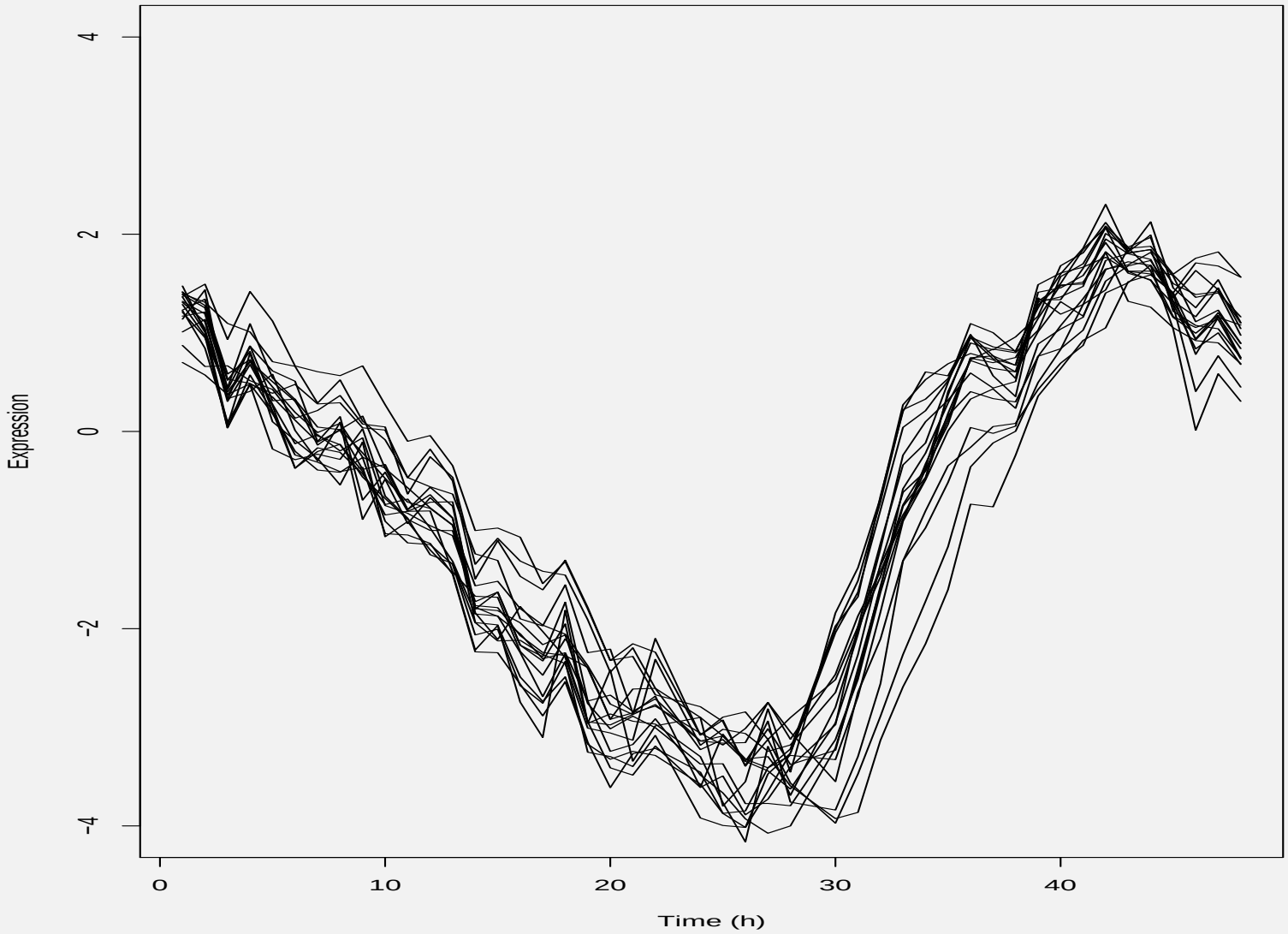
Cluster 1

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



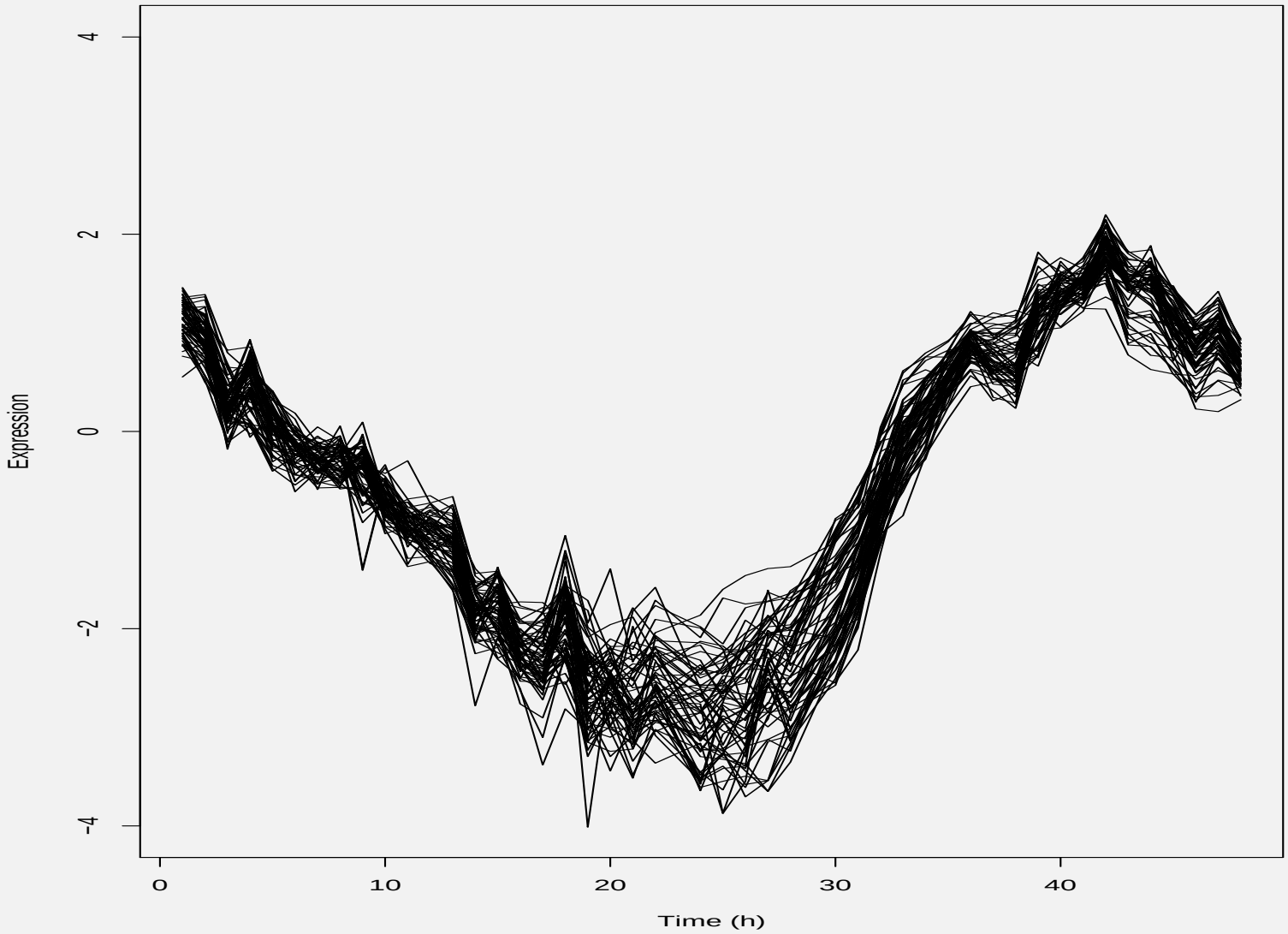
Cluster 2

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



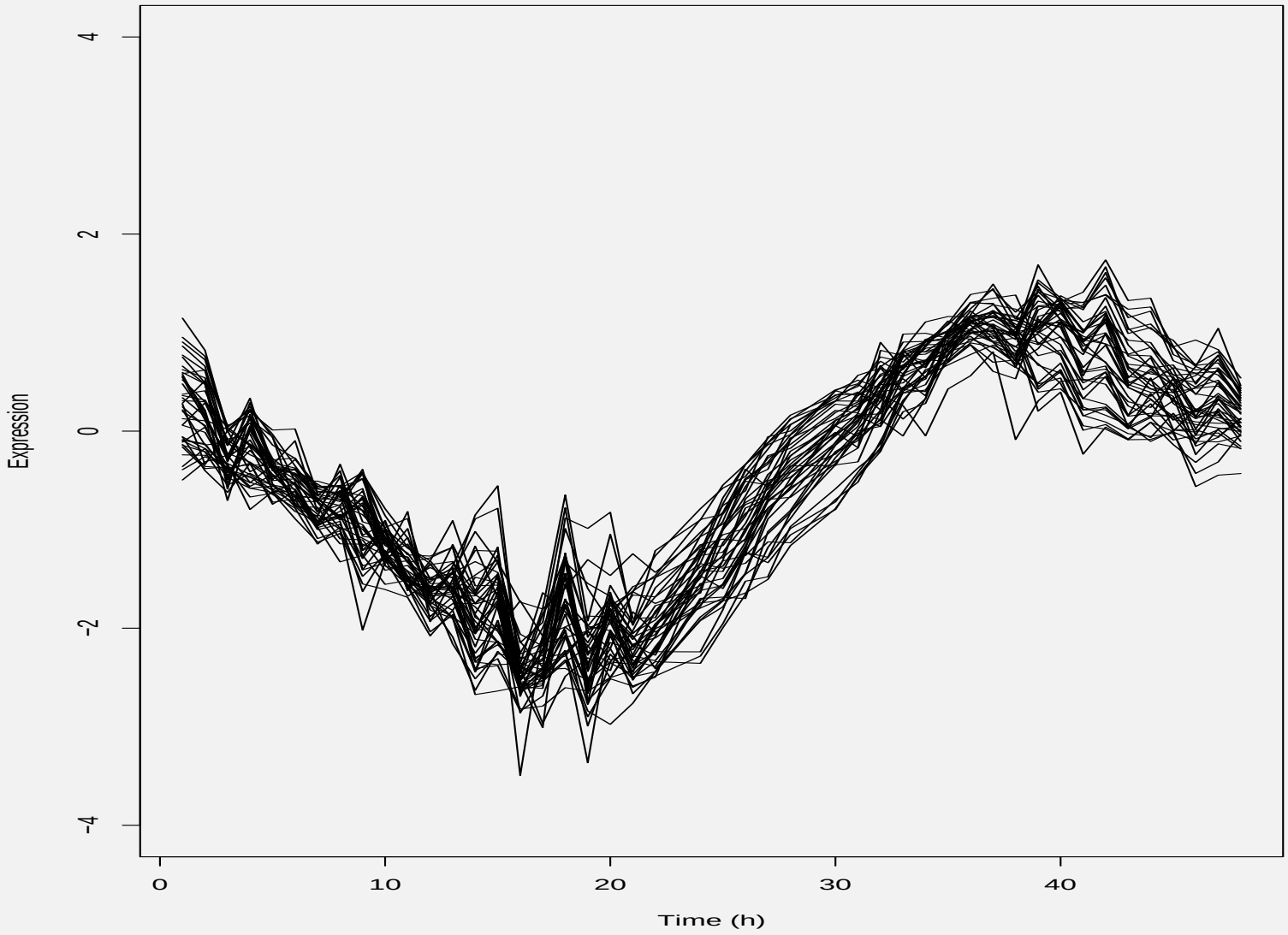
Cluster 7

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



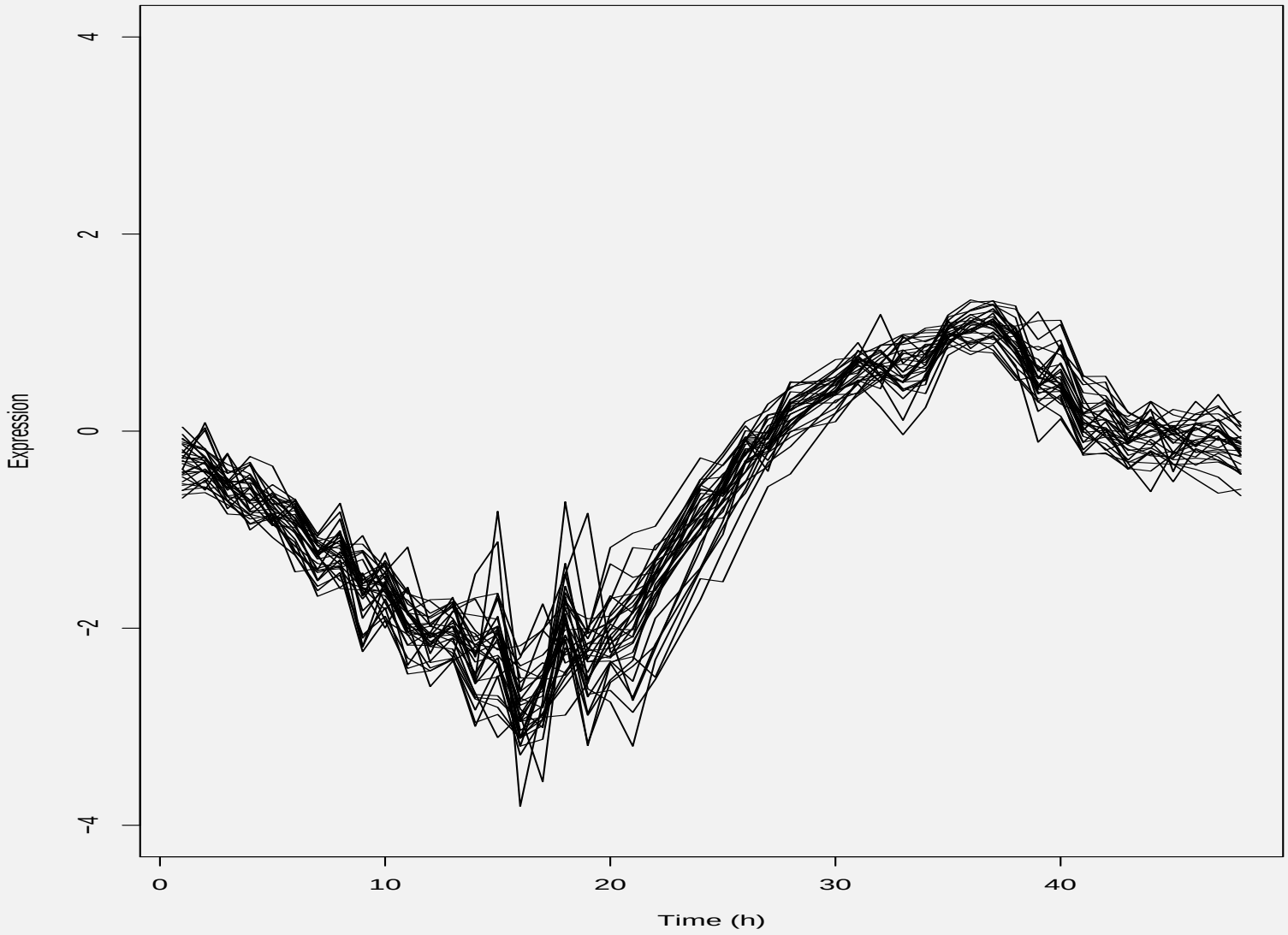
Cluster 25

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



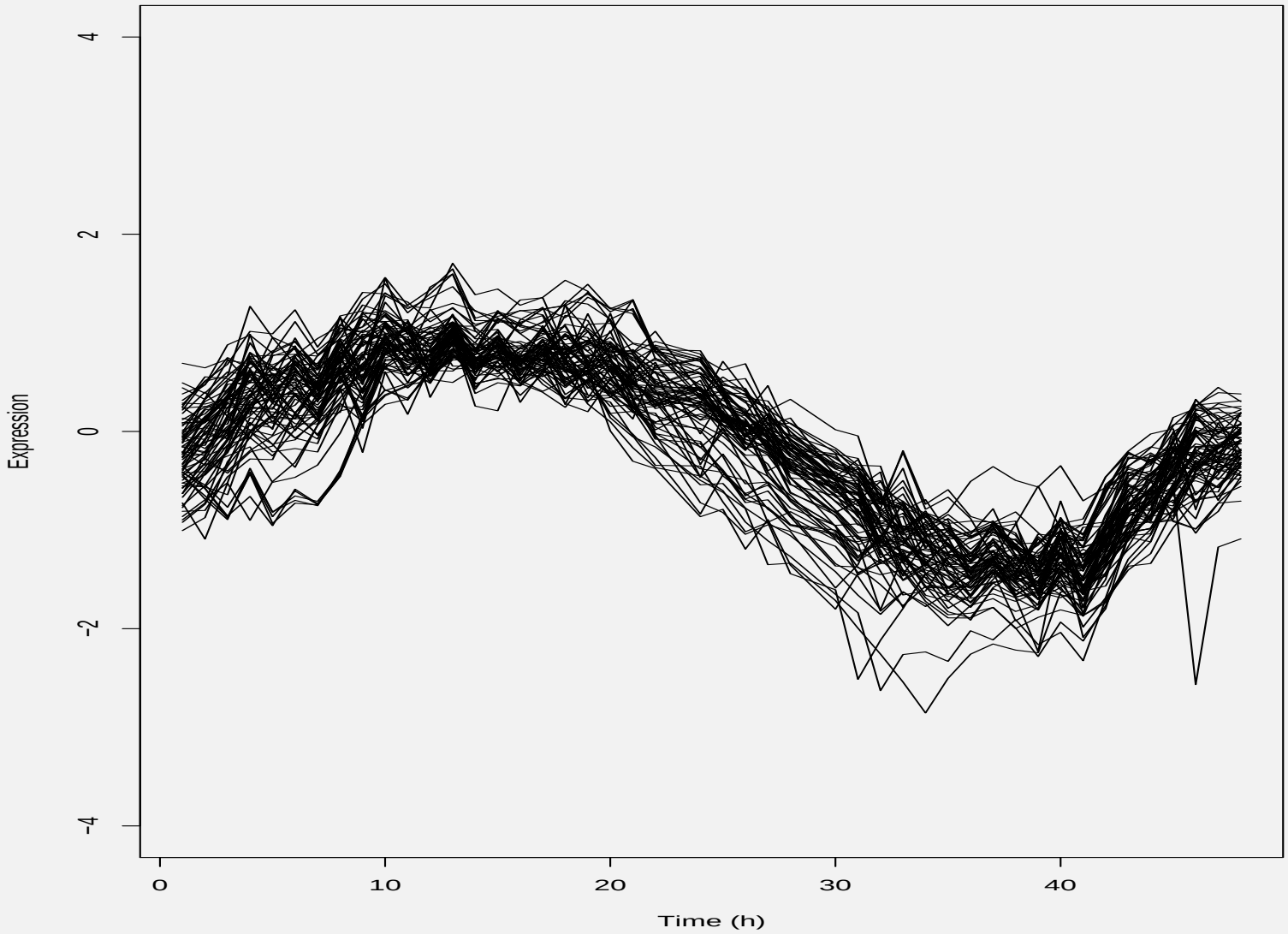
Cluster 19

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



Cluster 16

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



Mosquito and parasite interaction

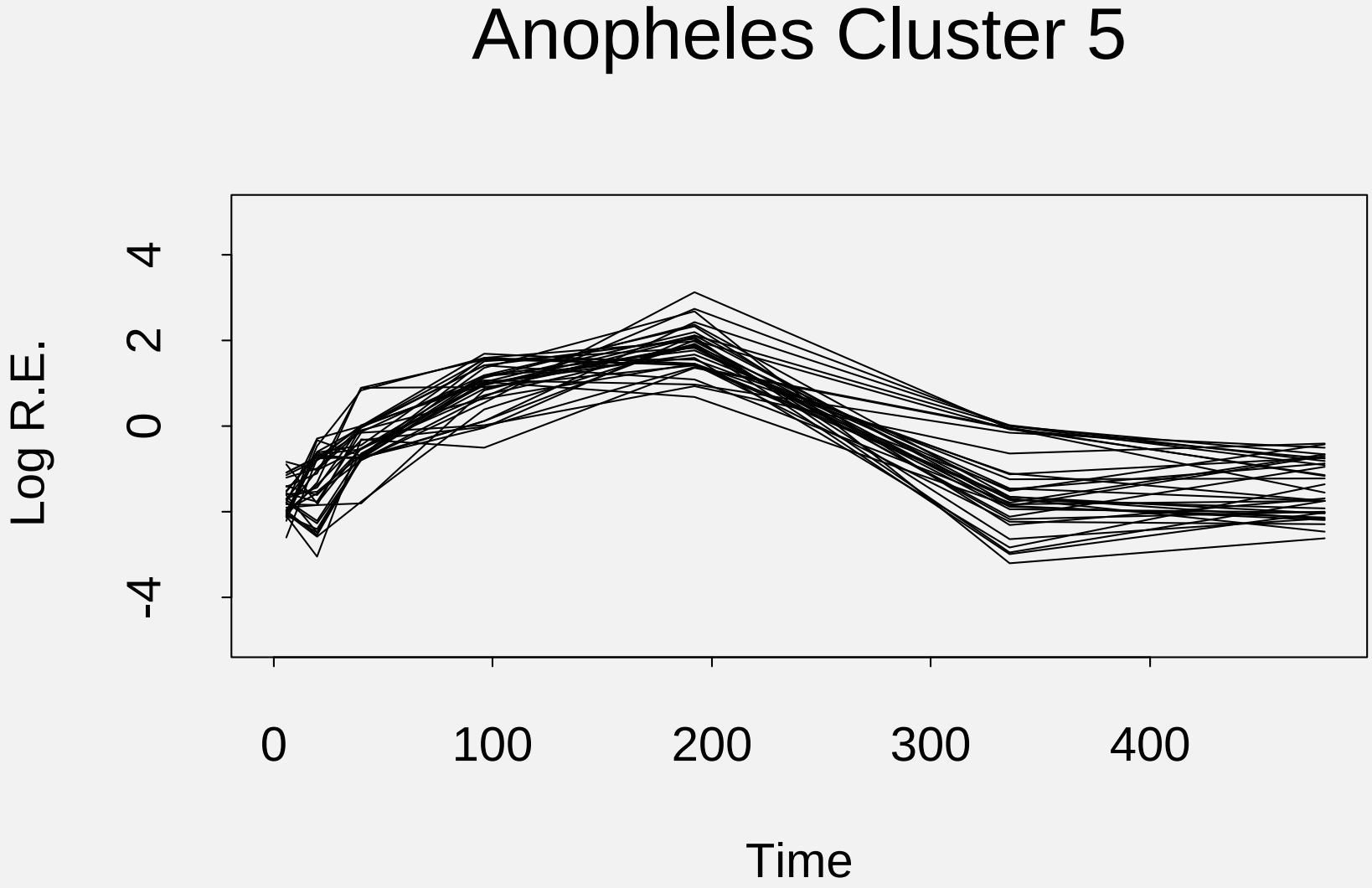
Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

Mosquito and parasite regulation studied concurrently

- mosquitos studied longitudinally over 20 days after infected blood meal
- 4200 genes/ESTs
- 1400 from each the Anopheles and Plasmodium genomes
- 1400 unidentified ESTs
- seven time points

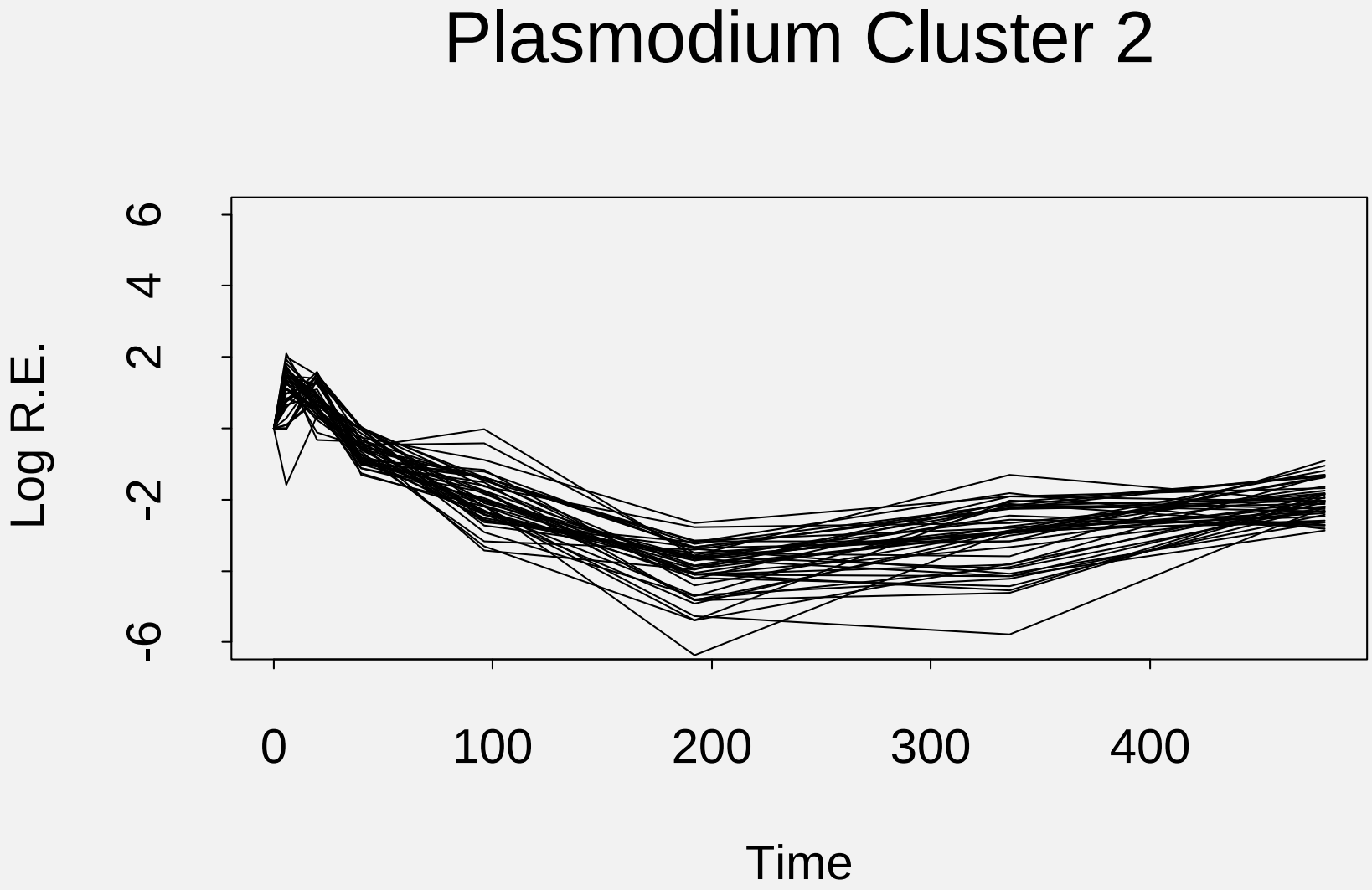
After clustering: Anopheles

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



After clustering: Plasmodium

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



Example: immune defense in humans

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

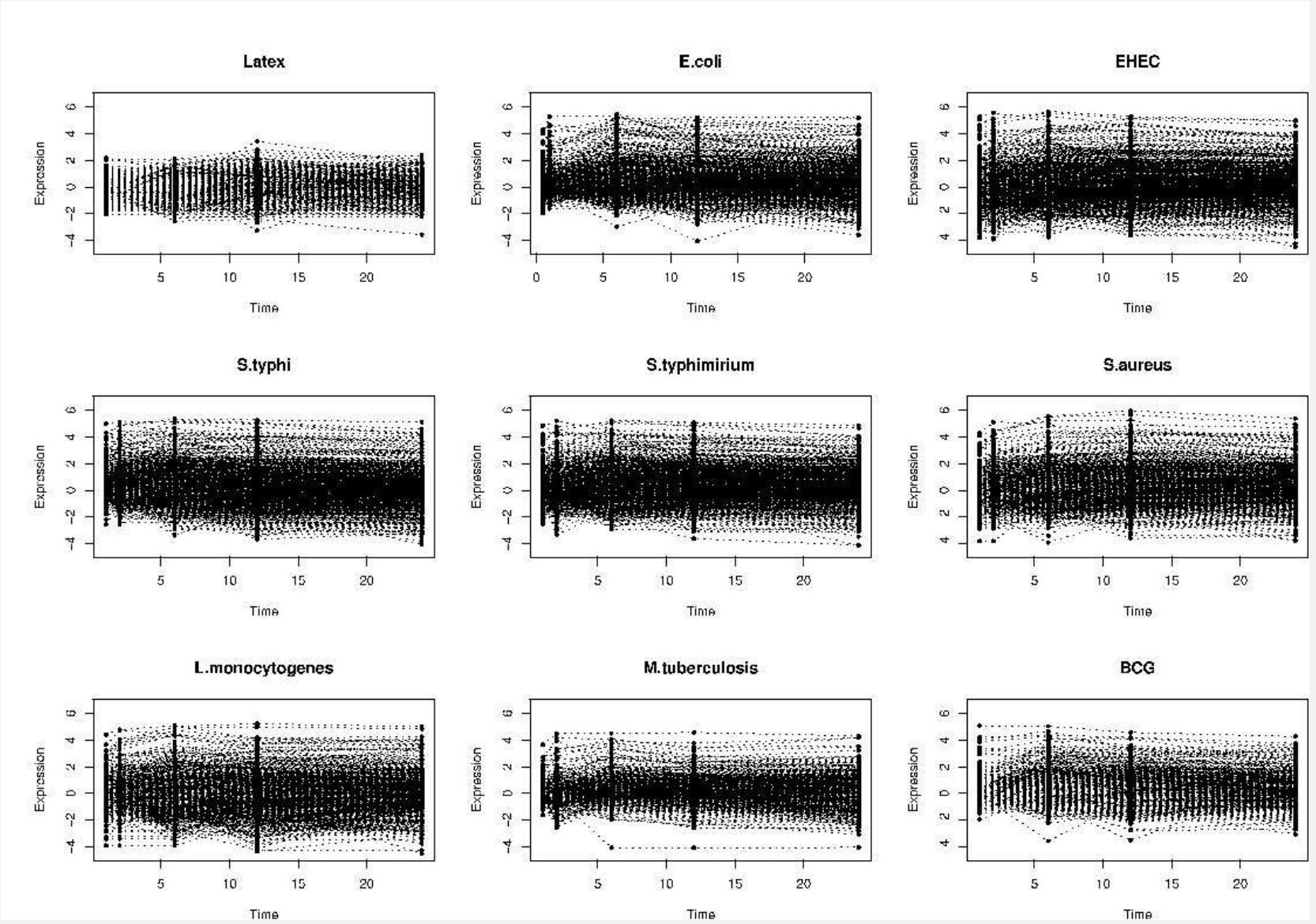
Nau *et al* (PNAS, 2002) investigated human macrophage activation induced by different bacterial pathogens.

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.

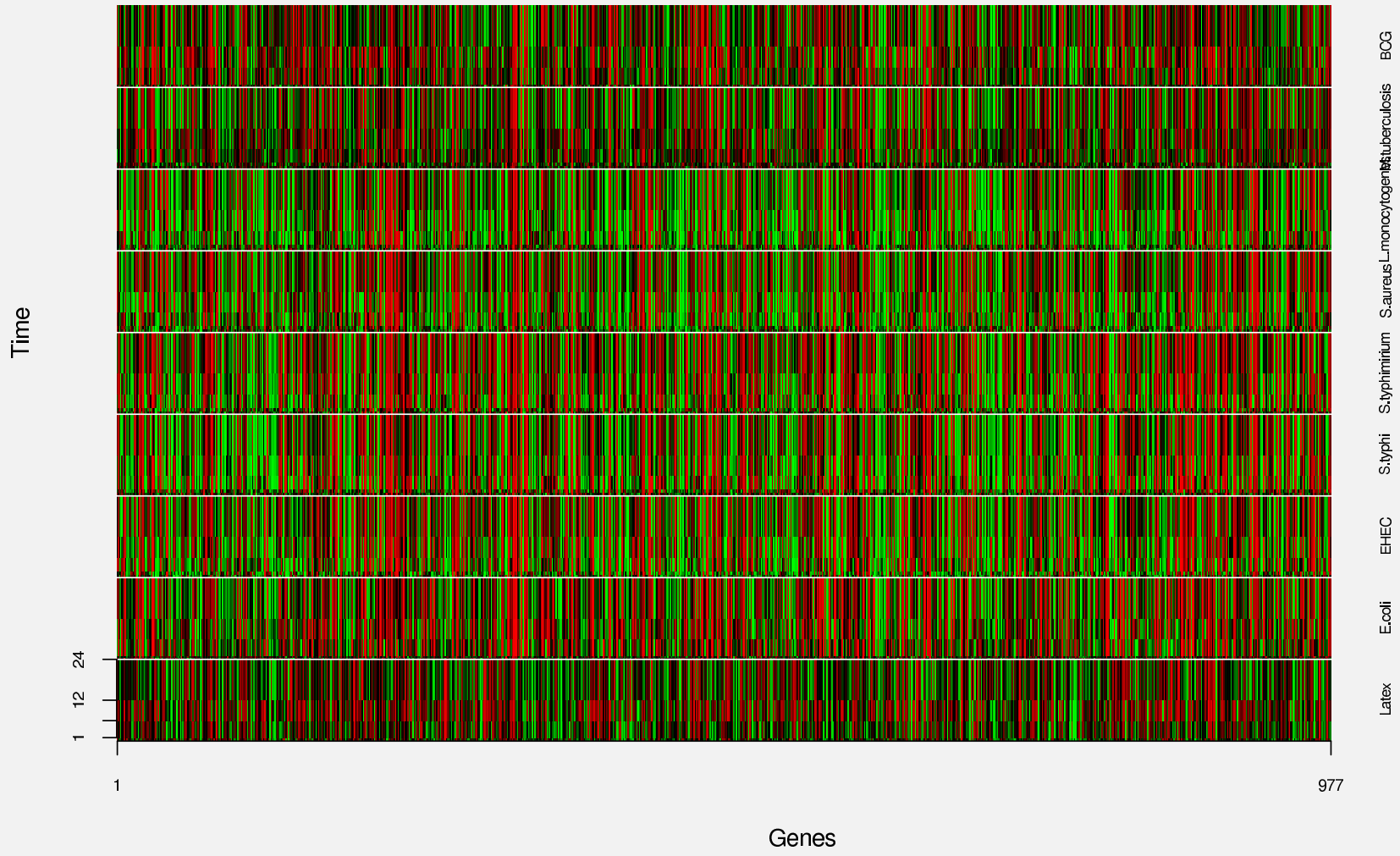
Nau *et al* data

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



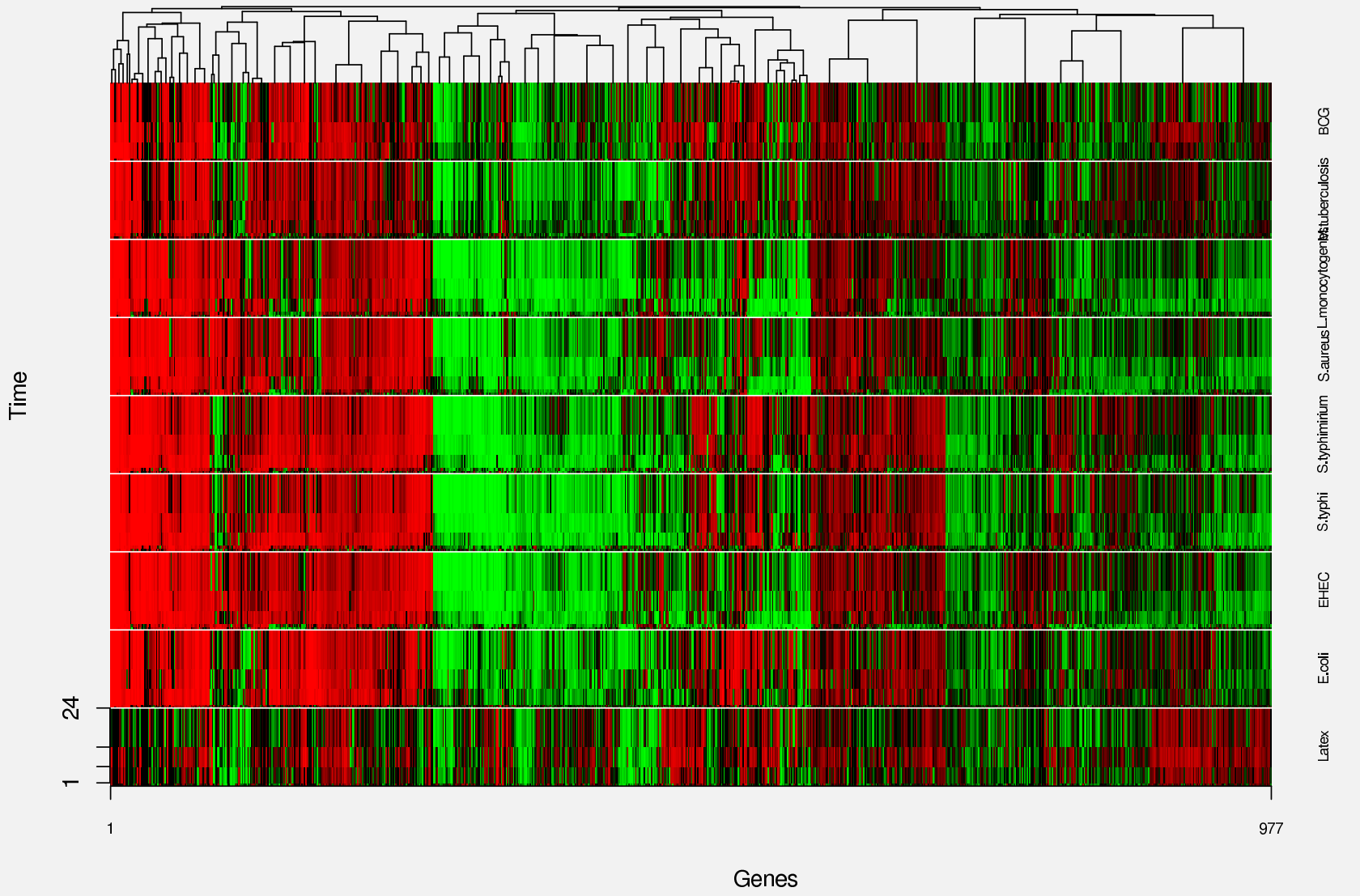
Unclustered expression data

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



Clustered expression data

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation



Regression Models

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

We wish to explain the variation of a time-varying *signal*. We will look at models of the form

$$y = f(x) + \epsilon$$

where

- x will represent time
- y will represent the measured relative expression of a gene or collection of genes
- $f(.)$ is a non-constant function
- ϵ is a random measurement error term.

Regression Models

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares Mathematical Formulation

The choice of f will crucially determine the types of patterns of regulation that can be captured.

- No differential expression: f is constant

$$f(x) = a$$

- Curvilinear Pattern: Quadratic form

$$f(x) = a + bx + cx^2$$

- Periodic Pattern: Trigonometric form

$$f(x) = a + b\cos(\lambda_1 \pi x) + c\sin(\lambda_1 \pi x)$$

We will see that each of these models can be regarded as special cases of a specific model; the

LINEAR REGRESSION MODEL

Regression Models

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

We will study the simplest form of a model in the class of Linear Regression Models where the relationship between time x and response y is a straight line.

Clearly for most real biological systems, the straight line model is an unrealistic simplification;

- the model implies that, beyond the time frame of the experiment the response increases (or decreases) over the whole range of x
- we expect periodic behaviour, or perhaps return to equilibrium after a stimulus, as time increases.

However, the properties of the model are best understood in the simplest case, and generalization to more realistic situations is then more straightforward.

Simple Linear Regression

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

We will investigate models relating two quantities x and y through equations of the form

$$y = ax + b$$

where a and b are constants (that is, a straight-line).

Note: variables x and y will not be treated exchangeably - we will regard y as being a function of x.

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

Such models are **DETERMINISTIC**, that is, if we know x (and the values of the constants), we can compute y exactly without error.

A more useful model allows for the possibility that the system is not observed perfectly, that is, we do not observe (x, y) pairs that are always consistent with a simple functional relationship.

Probabilistic Models

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

In a **probabilistic** model, we allow for the possibility that y is observed with random error, that is,

$$y = ax + b + ERROR$$

where *ERROR* is a random term that is present due to imperfect observation of the system due to (i) measurement error or (ii) missing information.

Note that we do not treat x and y exchangeably; x is a fixed observed variable that is measured *without error*, whereas y is an observed variable that is measured *with random error*.

We model the variation in y as a function of x. We observe pairs $(x_i, y_i), i = 1, \ldots, n$.

A Basic Probabilistic Model

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares Mathematical Formulation

Terminology:

- y - *Dependent variable* or *independent variable*
- x - *Independent variable*, or *predictor*, or *covariate*

The model we study takes the form

$$y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ is a random error term, a random variable with mean zero and finite variance ($E[\epsilon] = 0, Var[\epsilon] = \sigma^2$); it represents the error present in the measurement of y.

- β_0 - *Intercept* parameter
- β_1 - *Slop*e parameter

- $\beta_1 > 0$ - increasing y with increasing x
- $\beta_1 < 0$ - decreasing y with increasing x
- $\beta_1 = 0$ - no relationship between x and y

Note:

$$E[Y|x] = \beta_0 + \beta_1 x$$

where $E[Y|x]$ is the expected value of Y for fixed value of x.

Recall the notation

- Y - a random variable with a probability distribution
- y - a fixed value that the variable Y can take.

Fundamental Problem: If we believe the straight-line model with error is correct, how do we find the values of parameters β_0 and β_1. We only have the observed data $\{(x_i, y_i), i = 1, \ldots, n\}$.

Least Squares Fitting

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

We select the best values of β_0 and β_1 by minimizing the *error in fit*. For two data points (x_1, y_1) and (x_2, y_2), the errors in fit are

$$\begin{aligned} e_1 &= y_1 - (\beta_0 + \beta_1 x_1) \\ e_2 &= y_2 - (\beta_0 + \beta_1 x_2) \end{aligned}$$

respectively. But note that, potentially, $e_1 > 0$ and $e_2 < 0$ so there is a possibility that these fitting errors cancel each other out. Therefore we look at **squared** errors (as a large negative error is as bad as a large positive error)

$$\begin{aligned} e_1^2 &= (y_1 - (\beta_0 + \beta_1 x_1))^2 \\ e_2^2 &= (y_2 - (\beta_0 + \beta_1 x_2))^2 \end{aligned}$$

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

For n data, we obtain n misfit squared errors

$$e_1^2, \ldots, e_n^2$$

We select β_0 and β_1 as the values of the parameters that minimize SSE, where

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

We wish to make the total misfit squared error as small as possible.

SSE - sum of squared errors.

We could write

$$SSE = SSE(\beta_0, \beta_1)$$

to show the dependence of SSE on the parameters.

Minimization of $SSE(\beta_0, \beta_1)$ is achieved **analytically**.

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

It follows that the best parameters $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are given by

$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

where

- Sum of Squares SS_{xx}:

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

- Sum of Squares SS_{xy}:

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

$\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the **least-squares estimates**

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

is the **least-squares line of best fit**. The **fitted-values** are

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

and the **residuals** or **residual errors** are

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

Model Assumptions for Least-Squares

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

To utilize least-squares for the probabilistic model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

we make the following assumptions

1. The expected error $E[\epsilon]$ is zero so that

$$E[Y] = \beta_0 + \beta_1 x$$

2. The variance of the error, $Var[\epsilon]$, is constant and does not depend on x.
3. The probability distribution of ϵ is a **symmetric** distribution about zero; a stronger assumption is that ϵ is **Normally** distributed.
4. The errors for two different measured responses are **independent**, i.e. the error ϵ_1 in measuring y_1 at x_1 is independent of the error ϵ_2 in measuring y_2 at x_2.

Parameter Estimation: Estimating σ^2

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

Using the LS procedure, we can construct an estimate of the *error* or *residual error* variance

Recall that

$$Var[\epsilon] = \sigma^2$$

An estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2} = s^2$$

say.

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

The denominator $n-2$ is a *degrees of freedom* parameter of the form

$$\begin{array}{ccc} \text{TOTAL NUMBER} & - & \text{NUMBER OF PARAMETERS} \\ \text{OF DATA} & & \text{ESTIMATED} \end{array}$$

or $n-p$, where in the simple linear regression, $p=2$ ($\widehat{\beta}_0$ and $\widehat{\beta}_1$). Note also that

$$SSE(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_{yy} - \widehat{\beta}_1 SS_{xy}$$

where

$$SS_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

Mathematical Formulation

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

To allow for generalization of the model, we consider a matrix formulation.

For data pairs $(x_1, y_1), \ldots, (x_n, y_n)$ we form

- $\mathbf{y} = [y_1, \ldots, y_n]^T$ (an $n \times 1$ column vector)
- $n \times 2$ matrix $\mathbf{X}$, where the first column of $\mathbf{X}$ is a column of 1s, and the second column is

$$[x_1, \ldots, x_n]^T$$

  Thus the ith row is $\mathbf{x}_i = [1\ x_i]^T$.

- $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^T$ (an $n \times 1$ column vector)

The notation ^{T} means matrix transpose

Mathematical Formulation

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

We may then write the model in vector form

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where $\mathbf{X}\beta$ is the matrix multiplication product of $\mathbf{X}$ and β, which yields an $n \times 1$ column vector.

The SSE quantity is then

$$\begin{aligned} SSE(\beta) &= (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \\ &= \sum_{i=1}^{n}(y_i - \mathbf{x}_i\beta)^2 \end{aligned}$$

Least-Squares Estimates

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

It can be shown that the least-squares estimates of the model parameters are given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

where

- $\widehat{\boldsymbol{\beta}} = [\widehat{\beta}_0, \widehat{\beta}_1]^T$ are the **least-squares estimates**
- the notation $^{-1}$ means matrix inversion

Least-Squares Estimates

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

The estimate of σ^2 is given by calculation of

$$\begin{aligned} SSE(\widehat{\boldsymbol{\beta}}) &= (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y} \end{aligned}$$

and then

$$\widehat{\sigma}^2 = \frac{SSE(\widehat{\boldsymbol{\beta}})}{n-p}$$

where $p = 2$ is the dimension of $\boldsymbol{\beta}$.

Least-Squares Estimates

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

The estimated standard errors of $\widehat{\beta}$ are given by the matrix variance calculation

$$\text{Var}[\widehat{\beta}] = \widehat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$$

and the estimated standard errors are the diagonal elements of this 2×2 matrix.

These calculations look complicated, but are actually easy to compute. In the case of simple linear regression

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & S_x \\ S_x & S_{xx} \end{bmatrix}$$

where

$$S_x = \sum_{i=1}^{n} x_i \qquad S_{xx} = \sum_{i=1}^{n} x_i^2$$

Least-Squares Estimates

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

Then

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{nS_{xx} - S_x^2}\begin{bmatrix} S_{xx} & -S_x \\ -S_x & n \end{bmatrix}$$

and hence

$$\widehat{\beta} = \frac{1}{nS_{xx} - S_x^2}\begin{bmatrix} S_{xx} & -S_x \\ -S_x & n \end{bmatrix}\begin{bmatrix} S_y \\ S_{xy} \end{bmatrix}$$

where

$$S_y = \sum_{i=1}^{n} y_i \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i$$

Least-Squares Estimates

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

Hence

$$\widehat{\boldsymbol{\beta}} = \frac{1}{nS_{xx} - S_x^2} \begin{bmatrix} S_{xx}S_y - S_x S_{xy} \\ nS_{xy} - S_x S_y \end{bmatrix}$$

and

$$\widehat{\beta}_0 = \frac{S_{xx}S_y - S_x S_{xy}}{nS_{xx} - S_x^2}$$

$$\widehat{\beta}_1 = \frac{nS_{xy} - S_x S_y}{nS_{xx} - S_x^2} = \frac{nSS_{xy}}{nSS_{xx}}$$

as before.

Extending the model

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

The great advantage of writing the model in this apparently more complicated form is that the model can be **extended** from the simple straight line in a very straightforward way.

Provided the model can be written in the **Linear Model** form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\beta}$ contains the **parameters** in linear form, then the least-squares can be found easily.

Extending the model

Time Course Data
Malaria
S. Typhi
Multiple Challenges
P. Falciparum CAMDA
Both Organisms
Nau et al.
Linear Regression Models
Simple Linear Regression
Least-Squares
Mathematical Formulation

Crucially, the form of the matrix $\mathbf{X}$ does not change the basic approach to estimation. Provided we can invert $\mathbf{X}^T\mathbf{X}$, we can form the LS estimates. Thus $\mathbf{X}$ can include

- more terms in involving x ($x^2, x^3, \ldots$) (*polynomial regression*)
- non-linear functions of x ($\log x, \exp(x), \ldots$)
- other predictors if they available (*multiple regression*)