

Linear Mixed Models for the Analysis of Peptide Epitope Microarray Data

Tatjana Nahtman

Institute of Mathematical Statistics, University of Tartu, Estonia

Department of Statistics, Stockholm University, Sweden

- Introduction
 - Data
 - Biological interest
 - Statistical interest

- Introduction
 - Data
 - Biological interest
 - Statistical interest
- State of the art

- Introduction
 - Data
 - Biological interest
 - Statistical interest
- State of the art
- Data Analysis: Results

- Introduction
 - Data
 - Biological interest
 - Statistical interest
- State of the art
- Data Analysis: Results
- Future Research

Introduction

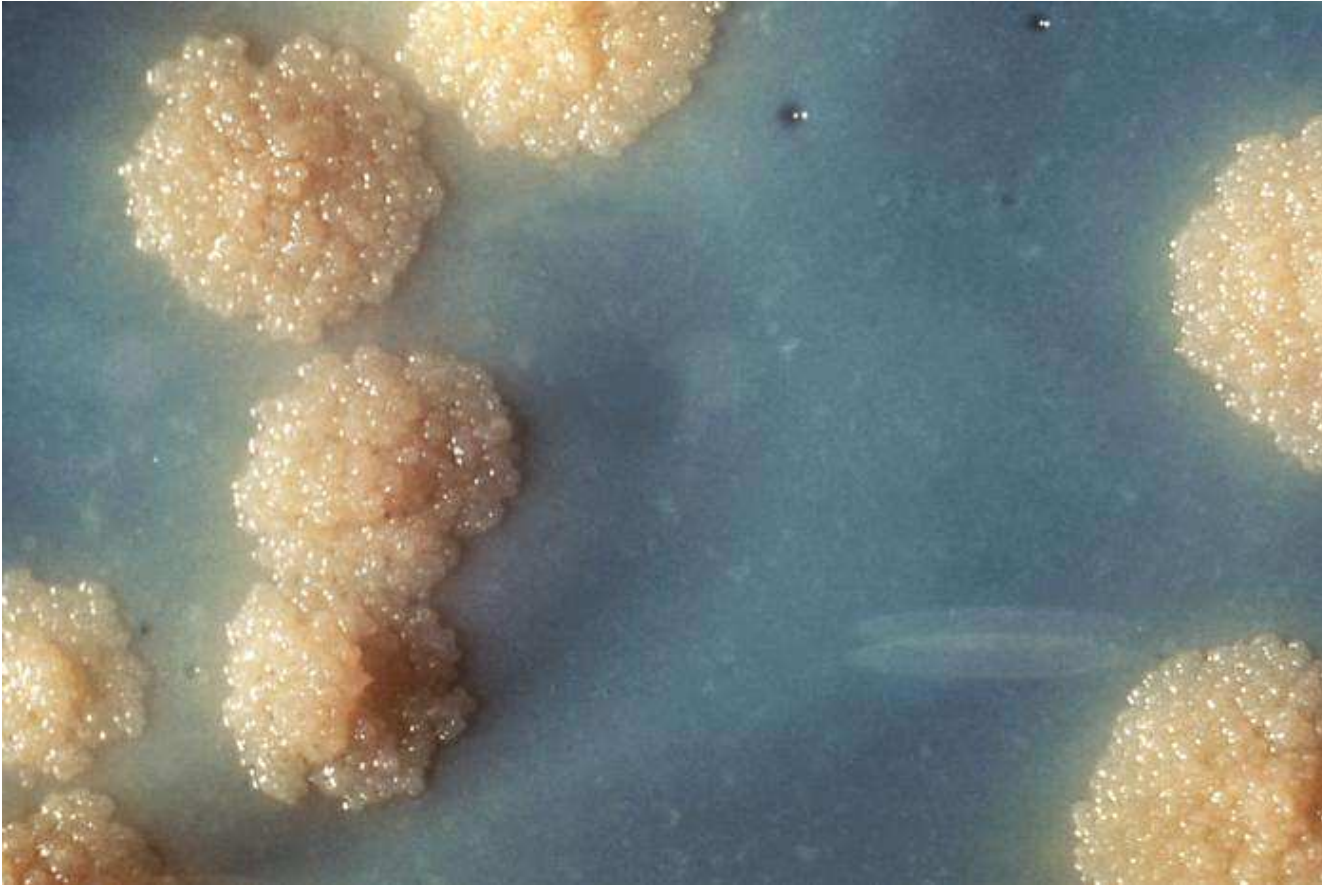
Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (MTB), remains a major threat to the human population.

The burgeoning epidemic of HIV infection in regions where tuberculosis is common has created a growing population of persons that are highly susceptible to *M. tuberculosis*. In addition, the multidrug-resistant tuberculosis continues its spread.

These unfavorable factors will cause tuberculosis to remain a major health problem in the coming decades, and increase the urgency for development of an effective vaccine.

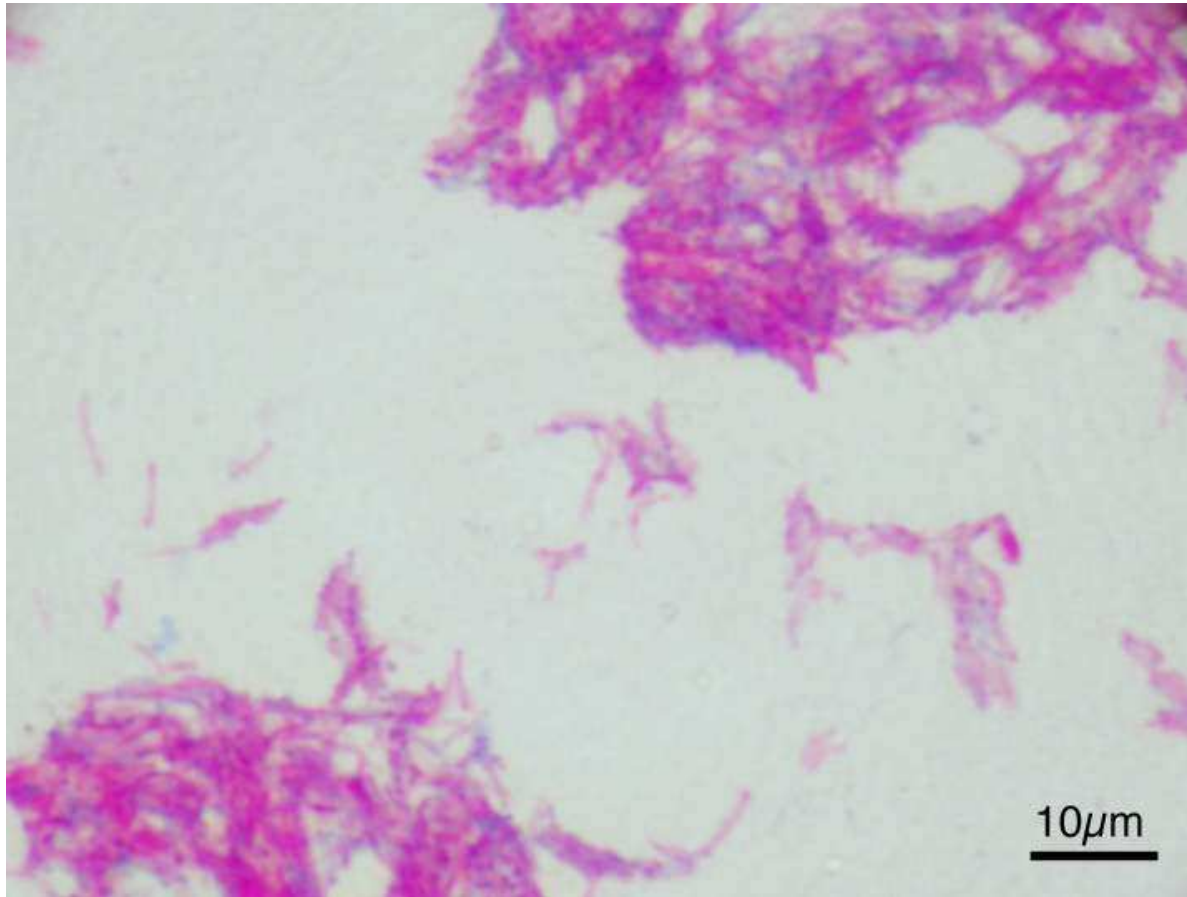
The only available antituberculosis vaccine is bacillus Calmette-Guérin (**BCG**), a live attenuated *Mycobacterium bovis* that was created in 1921.

Mycobacterium tuberculosis bacterial colonies



Mycobacterium tuberculosis is the bacterium that causes most cases of tuberculosis. It was first described on March 24, 1882 by Robert Koch, who subsequently received the Nobel Prize in physiology or medicine for this discovery in 1905; the bacterium is also known as Koch's bacillus. The *M. tuberculosis* genome was sequenced in 1998.

Mycobacterium bovis BCG



Bacille Calmette-Guérin (**BCG**) is a vaccine against tuberculosis that is prepared from a strain of the attenuated (weakened) live bovine tuberculosis bacillus, *Mycobacterium bovis*, that has lost its virulence in humans by being specially cultured in an artificial medium for years. The bacilli have retained enough strong antigenicity to become a somewhat effective vaccine for the prevention of human tuberculosis.

At best, the BCG vaccine is 80% effective in preventing tuberculosis for a duration of 15 years, however, its protective effect appears to vary according to geography.

The **antigen microarrays** consist of glass slides dotted with thousands of proteins and other molecules that are often attacked in autoimmune diseases.

Introduction

The **antigen microarrays** consist of glass slides dotted with thousands of proteins and other molecules that are often attacked in autoimmune diseases.

To use the microarray, doctors draw a blood sample from the patient and incubate it on the array.

Introduction

The **antigen microarrays** consist of glass slides dotted with thousands of proteins and other molecules that are often attacked in autoimmune diseases.

To use the microarray, doctors draw a blood sample from the patient and incubate it on the array.

Those antibodies that attack molecules on the array will locate their target and latch on. Fluorescent molecules are then added to detect the antibodies, creating colored spots on the slide. From there, it's a matter of counting the spots to see which antigens the immune system recognized.

Detection and treatment of tuberculosis infection are important measures in the fight against this epidemic.

Biological Interests

Detection and treatment of tuberculosis infection are important measures in the fight against this epidemic.

The tuberculin skin test (TST) has been the only practical means of detecting MT infection in the past century.

Unfortunately, TST has many limitations, including a high frequency of false-positive results after previous vaccination with BCG or exposure to non-tuberculous mycobacteria, and false-negative skin test results in patients with advanced TB.

Despite this considerable scientific progress, many issues remain regarding the quality, analysis and interpretation of the data the antigen microarrays produce.

Statistical Interests

Despite this considerable scientific progress, many issues remain regarding the quality, analysis and interpretation of the data the antigen microarrays produce.

There is currently no accepted approach to guide data analysis in immunology, and researchers are using a wide diversity of statistical methods as well as software tools for the analysis of immunological data.

Specific Aims

- To develop robust designs for printing peptide microarrays.

Specific Aims

- To develop robust designs for printing peptide microarrays.
- To extract quality data for analysis, by devising algorithms for screening, transforming and normalizing the raw data.

Specific Aims

- To develop robust designs for printing peptide microarrays.
- To extract quality data for analysis, by devising algorithms for screening, transforming and normalizing the raw data.
- To identify coherent groups of peptides which are differentially recognized by the antibodies of TB+ patients and healthy 'control' individuals.

Specific Aims

- To develop robust designs for printing peptide microarrays.
- To extract quality data for analysis, by devising algorithms for screening, transforming and normalizing the raw data.
- To identify coherent groups of peptides which are differentially recognized by the antibodies of TB+ patients and healthy 'control' individuals.
- To discriminate TB+ patients from vaccinated controls.

Specific Aims

- To develop robust designs for printing peptide microarrays.
- To extract quality data for analysis, by devising algorithms for screening, transforming and normalizing the raw data.
- To identify coherent groups of peptides which are differentially recognized by the antibodies of TB+ patients and healthy 'control' individuals.
- To discriminate TB+ patients from vaccinated controls.
- To characterize the immune profile in vaccinated persons over time.

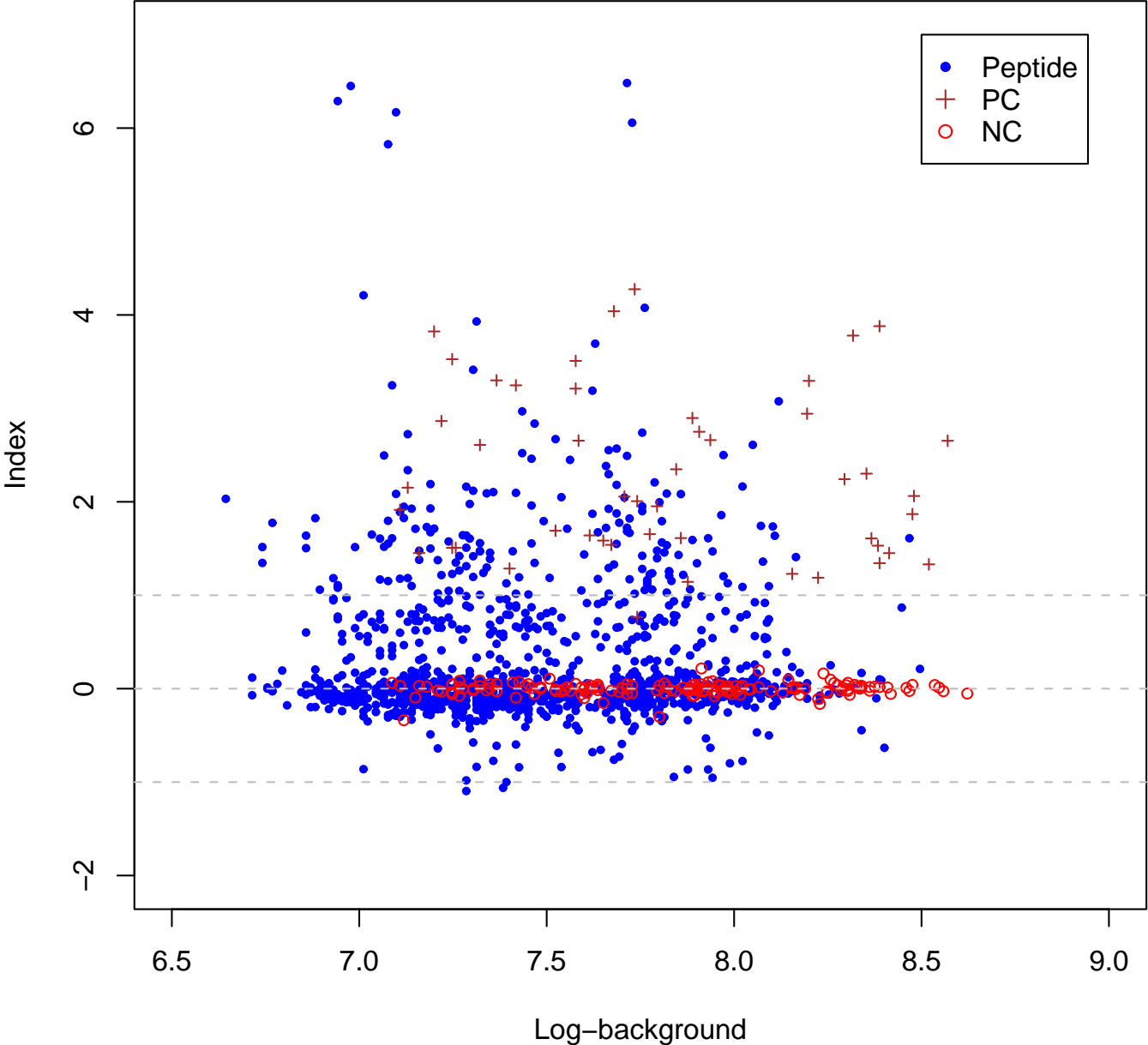
One of the major challenges of high throughput data from immunological studies (especially vaccine trials) is the repeated testing of an individual over time, resulting in repeated measures of high-dimensional data.

Data Analysis: Validation Study

In order to examine reproducibility within the same day, from day to day, and between operators (analysts), the samples were tested in duplicate, on each of two different days, by each of two different analysts, using the same batch of slides.

Thus each of the five patient specimens was tested on eight different slides, resulting in forty patient slides for analysis.

Data Analysis: Results



Results: Negative Controls

We fit the following model to the transformed responses from the negative controls:

$$Y_{ijklmn} = \mu_{NC} + I_i + A_j + D_k + R_l + B_m + \varepsilon_{ijklmn},$$

where Y_{ijklmn} is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

Results: Negative Controls

We fit the following model to the transformed responses from the negative controls:

$$Y_{ijklmn} = \mu_{NC} + I_i + A_j + D_k + R_l + B_m + \varepsilon_{ijklmn},$$

Y_{ijklmn} is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

I_i is a random effect representing the individual, $I_i \sim N(0, \sigma_I^2)$,
 $i = 1, \dots, 5$;

Results: Negative Controls

We fit the following model to the transformed responses from the negative controls:

$$Y_{ijklmn} = \mu_{NC} + I_i + A_j + D_k + R_l + B_m + \varepsilon_{ijklmn},$$

Y_{ijklmn} is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

I_i is a random effect representing the individual, $I_i \sim N(0, \sigma_I^2)$, $i = 1, \dots, 5$;

A_j is a random effect representing the effect of the j th analyst,

$A_j \sim N(0, \sigma_A^2)$, $j = 1, 2$;

Results: Negative Controls

We fit the following model to the transformed responses from the negative controls:

$$Y_{ijklmn} = \mu_{NC} + I_i + A_j + D_k + R_l + B_m + \varepsilon_{ijklmn},$$

Y_{ijklmn} is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

I_i is a random effect representing the individual, $I_i \sim N(0, \sigma_I^2)$, $i = 1, \dots, 5$;

A_j is a random effect representing the effect of the j th analyst, $A_j \sim N(0, \sigma_A^2)$, $j = 1, 2$;

D_k is a random day effect, $D_k \sim N(0, \sigma_D^2)$, $k = 1, 2$;

Results: Negative Controls

We fit the following model to the transformed responses from the negative controls:

$$Y_{ijklmn} = \mu_{NC} + I_i + A_j + D_k + R_l + B_m + \varepsilon_{ijklmn},$$

Y_{ijklmn} is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

I_i is a random effect representing the individual, $I_i \sim N(0, \sigma_I^2)$, $i = 1, \dots, 5$;

A_j is a random effect representing the effect of the j th analyst, $A_j \sim N(0, \sigma_A^2)$, $j = 1, 2$;

D_k is a random day effect, $D_k \sim N(0, \sigma_D^2)$, $k = 1, 2$;

R_l is a random effect representing the effect of the l th replicate of the experiment on the same day, $R_l \sim N(0, \sigma_R^2)$, $l = 1, 2$;

Results: Negative Controls

We fit the following model to the transformed responses from the negative controls:

$$Y_{ijklmn} = \mu_{NC} + I_i + A_j + D_k + R_l + B_m + \varepsilon_{ijklmn},$$

Y_{ijklmn} is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

I_i is a random effect representing the individual, $I_i \sim N(0, \sigma_I^2)$, $i = 1, \dots, 5$;

A_j is a random effect representing the effect of the j th analyst, $A_j \sim N(0, \sigma_A^2)$, $j = 1, 2$;

D_k is a random day effect, $D_k \sim N(0, \sigma_D^2)$, $k = 1, 2$;

R_l is a random effect representing the effect of the l th replicate of the experiment on the same day, $R_l \sim N(0, \sigma_R^2)$, $l = 1, 2$;

B_m is a fixed effect representing block, $m = 1, 2, 3$;

Results: Negative Controls

We fit the following model to the transformed responses from the negative controls:

$$Y_{ijklmn} = \mu_{NC} + I_i + A_j + D_k + R_l + B_m + \varepsilon_{ijklmn},$$

Y_{ijklmn} is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

I_i is a random effect representing the individual, $I_i \sim N(0, \sigma_I^2)$, $i = 1, \dots, 5$;

A_j is a random effect representing the effect of the j th analyst, $A_j \sim N(0, \sigma_A^2)$, $j = 1, 2$;

D_k is a random day effect, $D_k \sim N(0, \sigma_D^2)$, $k = 1, 2$;

R_l is a random effect representing the effect of the l th replicate of the experiment on the same day, $R_l \sim N(0, \sigma_R^2)$, $l = 1, 2$;

B_m is a fixed effect representing block, $m = 1, 2, 3$;

$\varepsilon_{ijklmn} \sim N(0, \sigma_e^2)$, is random error, $n = 1, \dots, n_{NC}$, where n_{NC} denotes the number of negative controls in each block of a slide.

Results: Peptides

To study variation in peptide responses we use the following model:

$$Y_{hijklmn} = \mu_P + P_h + I_i + (PI)_{hi} + A_j + D_k + R_l + B_m + (PB)_{hm} + \varepsilon_{hijklmn}$$

where $Y_{hijklmn}$ is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

Results: Peptides

To study variation in peptide responses we use the following model:

$$Y_{hijklmn} = \mu_P + P_h + I_i + (PI)_{hi} + A_j + D_k + R_l + B_m + (PB)_{hm} + \varepsilon_{hijklmn}$$

$Y_{hijklmn}$ is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

B_m is a fixed effect representing block, $m = 1, 2, 3$;

Results: Peptides

To study variation in peptide responses we use the following model:

$$Y_{hijklmn} = \mu_P + P_h + I_i + (PI)_{hi} + A_j + D_k + R_l + B_m + (PB)_{hm} + \varepsilon_{hijklmn}$$

$Y_{hijklmn}$ is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

B_m is a fixed effect representing block, $m = 1, 2, 3$;

P_h is a fixed effect for the h th peptide, $h = 1, \dots, n_P$ (the number of distinct peptides studied);

Results: Peptides

To study variation in peptide responses we use the following model:

$$Y_{hijklmn} = \mu_P + P_h + I_i + (PI)_{hi} + A_j + D_k + R_l + B_m + (PB)_{hm} + \varepsilon_{hijklmn}$$

$Y_{hijklmn}$ is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

B_m is a fixed effect representing block, $m = 1, 2, 3$;

P_h is a fixed effect for the h th peptide, $h = 1, \dots, n_P$ (the number of distinct peptides studied);

A_j is a random effect representing the effect of the j th analyst, $A_j \sim N(0, \sigma_A^2)$, $j = 1, 2$;

Results: Peptides

To study variation in peptide responses we use the following model:

$$Y_{hijklmn} = \mu_P + P_h + I_i + (PI)_{hi} + A_j + D_k + R_l + B_m + (PB)_{hm} + \varepsilon_{hijklmn}$$

$Y_{hijklmn}$ is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

B_m is a fixed effect representing block, $m = 1, 2, 3$;

P_h is a fixed effect for the h th peptide, $h = 1, \dots, n_P$ (the number of distinct peptides studied);

A_j is a random effect representing the effect of the j th analyst, $A_j \sim N(0, \sigma_A^2)$, $j = 1, 2$;

D_k is a random day effect, $D_k \sim N(0, \sigma_D^2)$, $k = 1, 2$;

Results: Peptides

To study variation in peptide responses we use the following model:

$$Y_{hijklmn} = \mu_P + P_h + I_i + (PI)_{hi} + A_j + D_k + R_l + B_m + (PB)_{hm} + \varepsilon_{hijklmn}$$

$Y_{hijklmn}$ is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

B_m is a fixed effect representing block, $m = 1, 2, 3$;

P_h is a fixed effect for the h th peptide, $h = 1, \dots, n_P$ (the number of distinct peptides studied);

A_j is a random effect representing the effect of the j th analyst, $A_j \sim N(0, \sigma_A^2)$, $j = 1, 2$;

D_k is a random day effect, $D_k \sim N(0, \sigma_D^2)$, $k = 1, 2$;

I_i is a random effect representing the individual,

$I_i \sim N(0, \sigma_I^2)$, $i = 1, \dots, 5$;

Results: Peptides

To study variation in peptide responses we use the following model:

$$Y_{hijklmn} = \mu_P + P_h + I_i + (PI)_{hi} + A_j + D_k + R_l + B_m + (PB)_{hm} + \varepsilon_{hijklmn}$$

$Y_{hijklmn}$ is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

B_m is a fixed effect representing block, $m = 1, 2, 3$;

P_h is a fixed effect for the h th peptide, $h = 1, \dots, n_P$ (the number of distinct peptides studied);

A_j is a random effect representing the effect of the j th analyst, $A_j \sim N(0, \sigma_A^2)$, $j = 1, 2$;

D_k is a random day effect, $D_k \sim N(0, \sigma_D^2)$, $k = 1, 2$;

I_i is a random effect representing the individual, $I_i \sim N(0, \sigma_I^2)$, $i = 1, \dots, 5$;

R_l is a random effect representing the effect of the l th replicate of the experiment on the same day, $R_l \sim N(0, \sigma_R^2)$, $l = 1, 2$;

Results: Peptides

To study variation in peptide responses we use the following model:

$$Y_{hijklmn} = \mu_P + P_h + I_i + (PI)_{hi} + A_j + D_k + R_l + B_m + (PB)_{hm} + \varepsilon_{hijklmn}$$

$Y_{hijklmn}$ is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

B_m is a fixed effect representing block, $m = 1, 2, 3$;

P_h is a fixed effect for the h th peptide, $h = 1, \dots, n_P$ (the number of distinct peptides studied);

A_j is a random effect representing the effect of the j th analyst, $A_j \sim N(0, \sigma_A^2)$, $j = 1, 2$;

D_k is a random day effect, $D_k \sim N(0, \sigma_D^2)$, $k = 1, 2$;

I_i is a random effect representing the individual, $I_i \sim N(0, \sigma_I^2)$, $i = 1, \dots, 5$;

R_l is a random effect representing the effect of the l th replicate of the experiment on the same day, $R_l \sim N(0, \sigma_R^2)$, $l = 1, 2$;

$\varepsilon_{hijklmn} \sim N(0, \sigma_e^2)$, is random error, $n = 1, \dots, n_P$, with n_P denoting the number of replicates of peptides within each block.

Results: Peptides

To study variation in peptide responses we use the following model:

$$Y_{hijklmn} = \mu_P + P_h + I_i + (PI)_{hi} + A_j + D_k + R_l + B_m + (PB)_{hm} + \varepsilon_{hijklmn}$$

where $Y_{hijklmn}$ is the transformed response of the i th individual recorded by the j th analyst on the k th day in the l th experiment in the m th block of the slide;

B_m is a fixed effect representing block, $m = 1, 2, 3$;

P_h is a fixed effect for the h th peptide, $h = 1, \dots, n_P$ (the number of distinct peptides studied);

A_j is a random effect representing the effect of the j th analyst, $A_j \sim N(0, \sigma_A^2)$, $j = 1, 2$;

D_k is a random day effect, $D_k \sim N(0, \sigma_D^2)$, $k = 1, 2$;

I_i is a random effect representing the individual, $I_i \sim N(0, \sigma_I^2)$, $i = 1, \dots, 5$;

R_l is a random effect representing the effect of the l th replicate of the experiment on the same day, $R_l \sim N(0, \sigma_R^2)$, $l = 1, 2$;

$\varepsilon_{hijklmn} \sim N(0, \sigma_e^2)$, is random error, $n = 1, \dots, n_P$, with n_P denoting the number of replicates of peptides within each block.

Peptides are excluded from analysis if they exhibit a high response on the slide with only buffer and secondary antibody.

Results: Peptides

To compare peptide responses from both groups we use the following model:

$$Y_{hi(j)jk} = \mu_P + P_h + I_{i(j)} + (PI)_{hi(j)} + G_j + (PG)_{hj} + \varepsilon_{hijk},$$

where $Y_{hi(j)jk}$ is the transformed response of the i th individual recorded in the j th group;

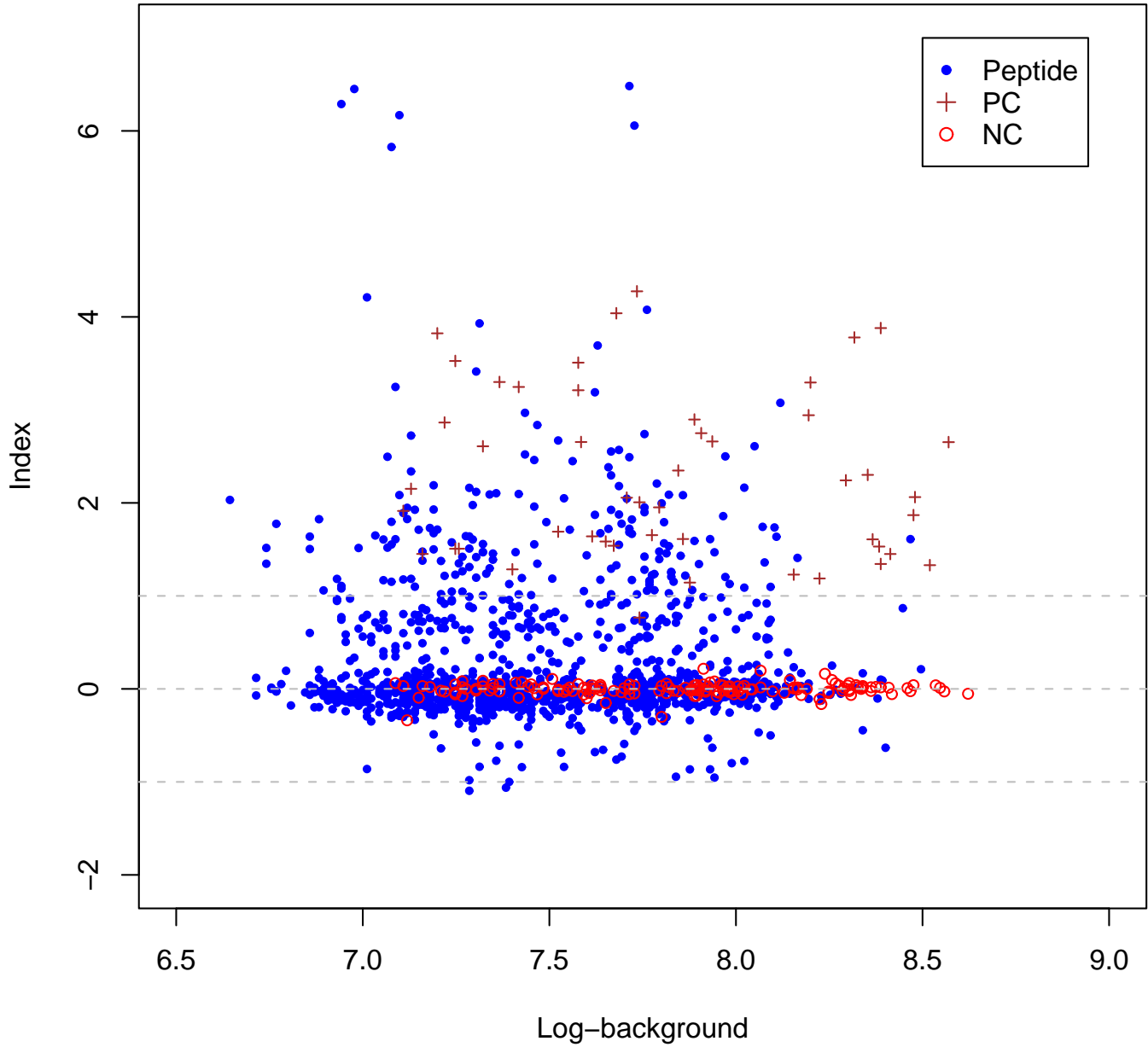
G_j is a fixed effect representing group, $j = 1, 2$;

P_h is a fixed effect for the h th peptide, $h = 1, \dots, n_P$ (the number of distinct peptides studied);

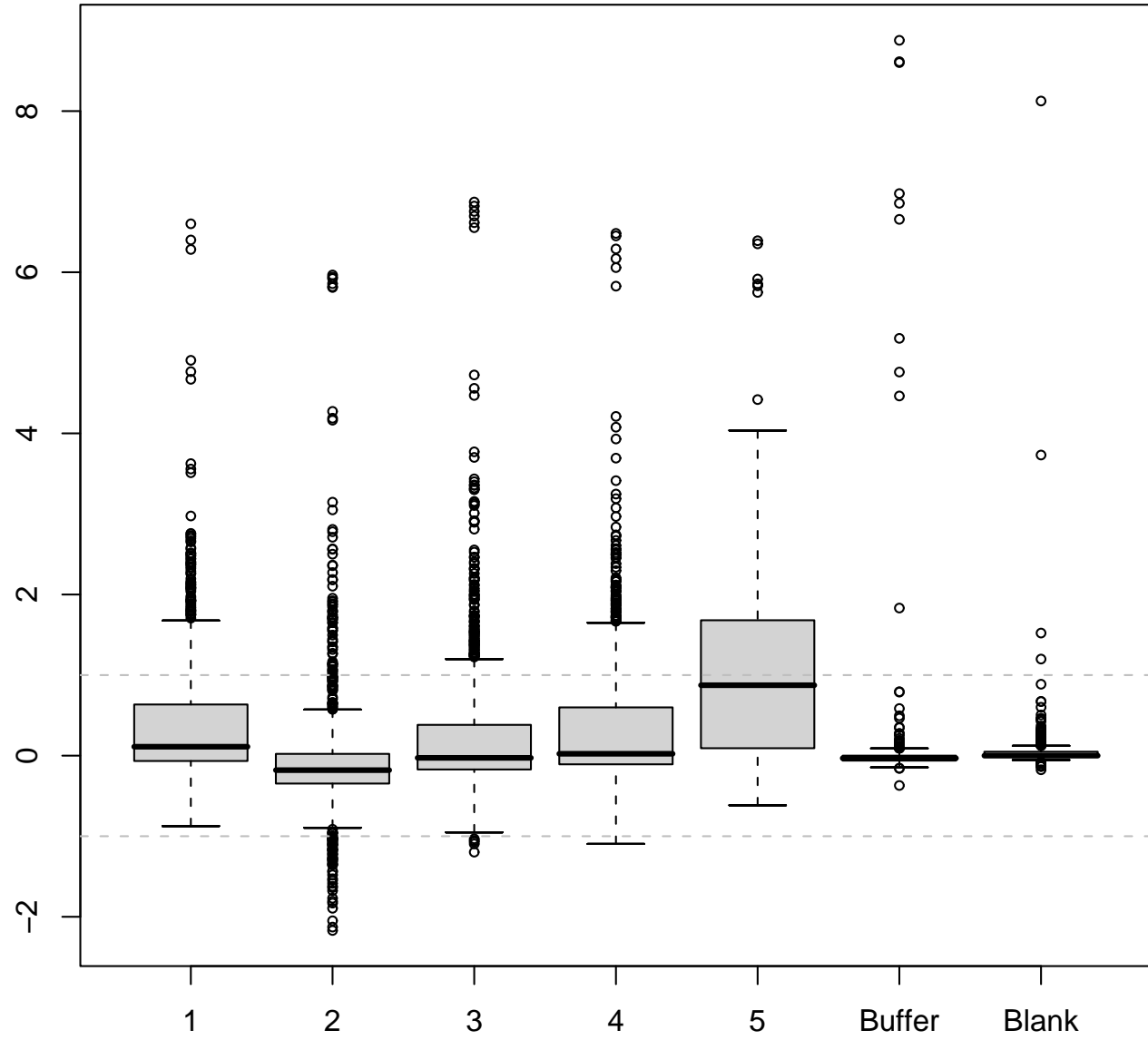
$I_{i(j)}$ is a random effect representing the individual, $I_{i(j)} \sim N(0, \sigma_I^2)$, $i = 1, \dots, 5$;

$\varepsilon_{hi(j)jk} \sim N(0, \sigma_e^2)$, is random error, $n = 1, \dots, n_P$, with n_P denoting the number of replicates of peptides within each block.

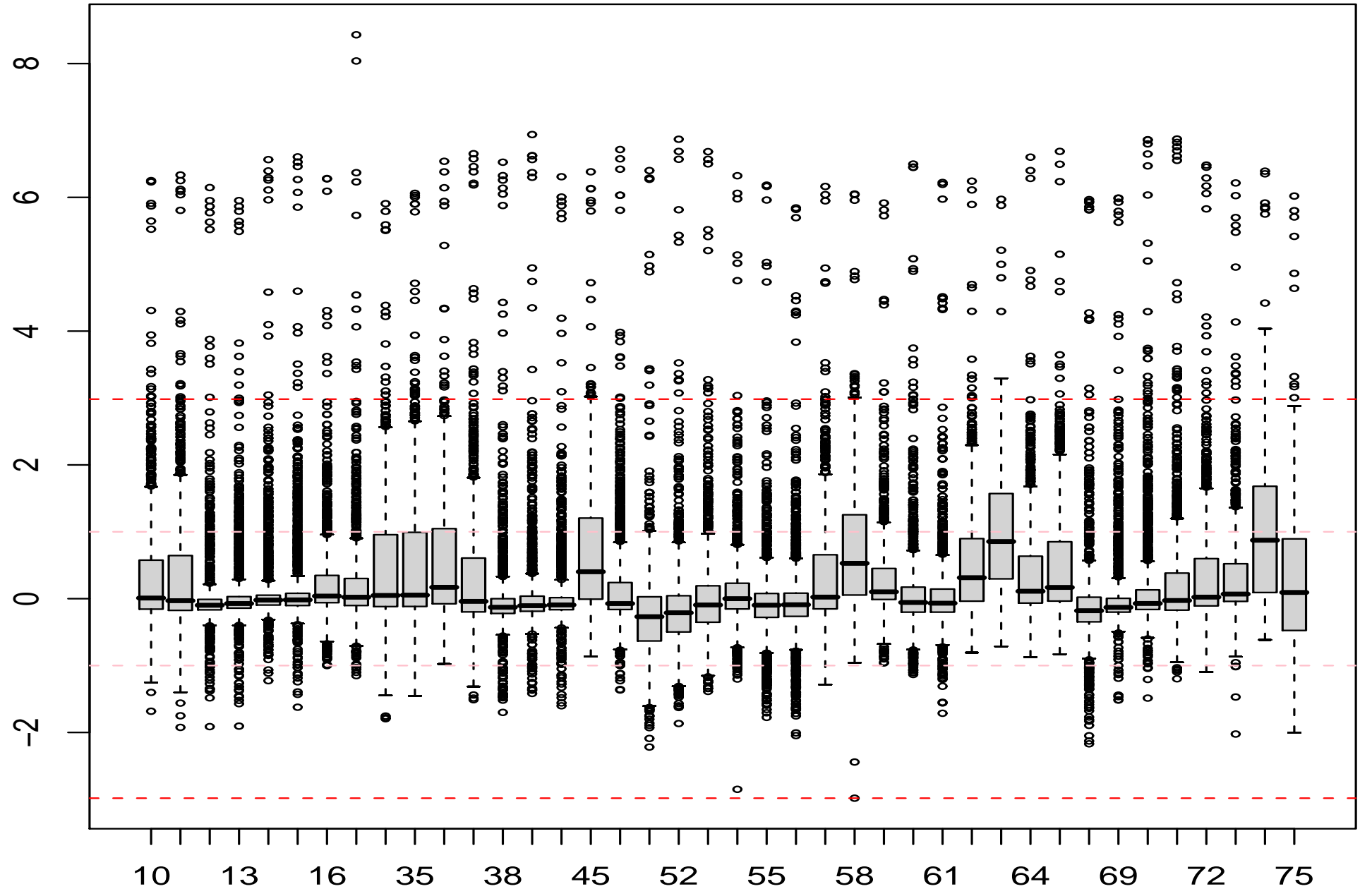
Results



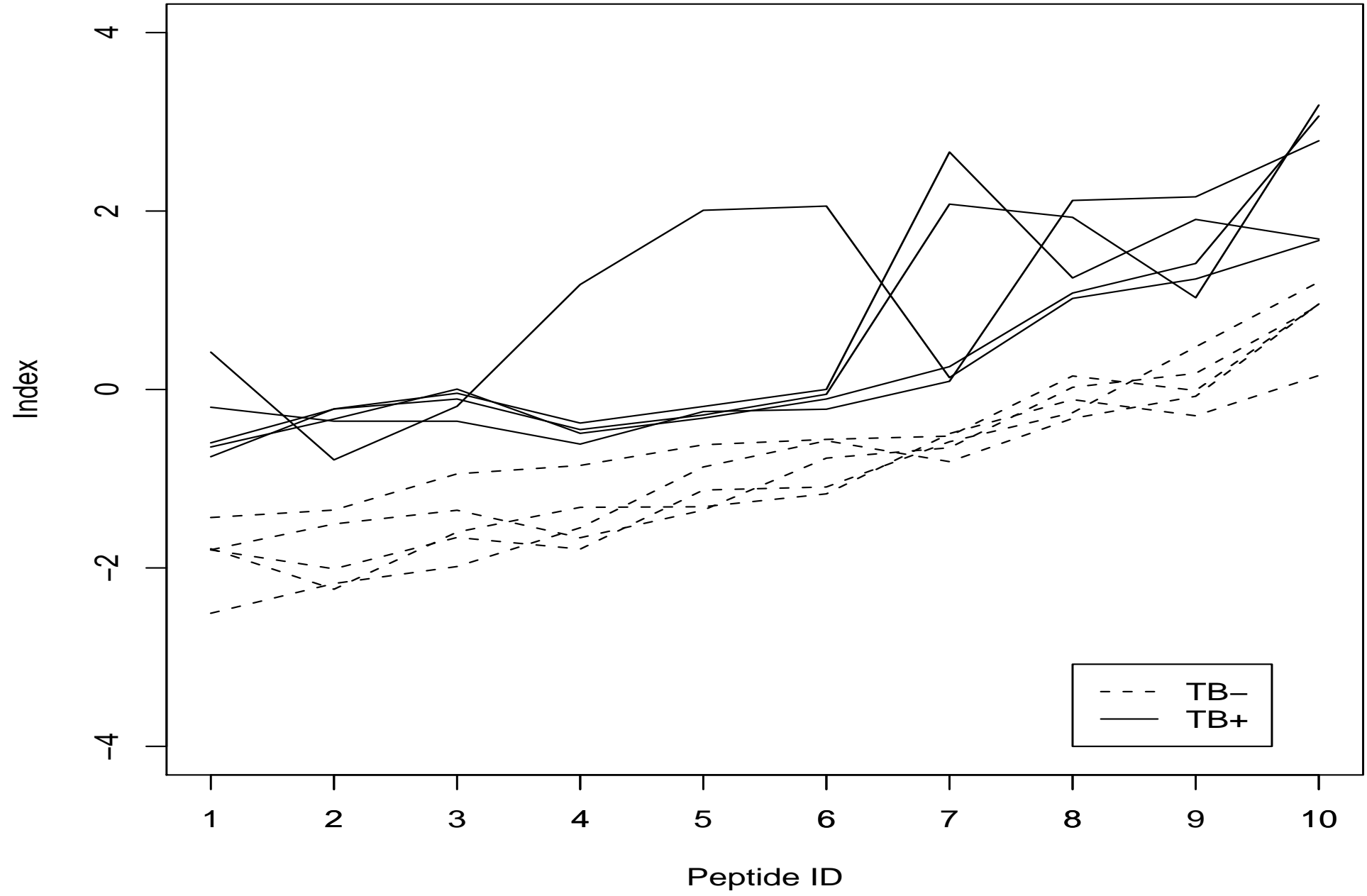
Results



Results



Results



Thank you!