Degree project

UNIVERSITY OF SKÖVDE

Multivariate analysis and classification of pathogenic priming components in wild-type and lab mice

Master Degree Project (120 credits) in Systems Biology

Second Cycle 45 credits

Spring term 2023

Student: Armand Stoe

Supervisor: Patric Nilson, University of Skövde

Examiner: Sanja Jurcevic, University of Skövde

Abstract

Animal models have a long history of being used in research for the purpose of investigating biological processes and testing the effect of specific compounds on the functionality of biological processes. Different types of mice are used as animal models, most notably inbred and outbred strains. This study investigates the effect of certain priming conditions on the production of cytokines in wild mice and lab mice, using multivariate data analysis. This analytical study involves exploratory analysis, in the form of PCA, MANOVA and LDA, training of different classification models and their validation. Based on the conducted exploratory analysis, certain priming conditions (CD3CD28, CPG and PG) have been identified as clearly defined groups by PCA and LDA, in both wild mice and lab mice. MANOVA concluded that most of the variables tested are statistically significant in determining group association. Subsequent classification modeling determined that the Random Forest algorithm is the most accurate in predicting class, in both the wild and lab mice. The performed analysis has given insight into the major trends exhibited by the data, but further post-processing analysis could potentially extract more data. The results of this study could be used to further investigate the discovered pattern in the data or be supplemented by comparing additional mouse strains under the same experimental conditions.

Popular scientific summary

Animal models have long played a crucial role in advancing medical knowledge, aiding our understanding of diseases, vaccines, drug development, and more. Among the animal models used in research, rodents have emerged as a highly valued option due to their controlled conditions, ease of maintenance, and genetic similarity to humans. In particular, laboratory mice have been extensively employed to study genetic interactions, proteins, and other biological molecules, owing to their selectively bred strains and controlled environments.

The modern house mouse, *Mus musculus*, is a complex species comprising three lineages: *M. musculus castaneus*, *M. musculus musculus*, and *M. musculus domesticus*. These mice have coexisted with human agricultural communities, adapting to changes in diet, behavior, and pathogenic exposure. With human migration, *M. musculus* has spread across different ecosystems, leading to significant genetic variations and local adaptations. This genetic diversity, often unexplored and underutilized, holds immense potential for understanding disease associations and human diversity patterns.

The *C57BL/6J* strain is perhaps the most widely used laboratory mouse strain worldwide. Originally inbred in 1921, this strain exhibits unique characteristics such as thermic and pain sensitivity, analgesic resistance, and predisposition to certain diseases. The advantages of using genetically identical mice, such as controllability and extensive phenotypic and genomic data availability, continue to make the *C57BL/6J* strain popular in research. However, concerns arise regarding randomization and its potential impact on biological experiments. Thus, a critical examination of the *C57BL/6J* strain's advantages and limitations in comparison to other mouse derivations is warranted.

Cytokines are soluble molecules facilitating communication between immune cells that play critical roles in immune responses. Interleukins, interferons, and tumor necrosis factors are among the various cytokines involved. Chemokines, a subfamily of cytokines, are instrumental in immune cell mobilization during homeostatic and inflammatory responses.

To explore the immunological differences between laboratory mice and their wild counterparts, extensive studies have been conducted on wild mice captured in the United Kingdom. However, these studies often employ univariate approaches, overlooking potential correlations and networks between the variables. Multivariate data analysis, including techniques such as Principal Component Analysis (PCA) and Multivariate Analysis of Variance (MANOVA), allow for a more comprehensive examination of data, revealing complex trends and interactions. Unsupervised machine learning algorithms can provide classification and predictions based on collected data.

The aim of this paper is to provide an overview of multivariate data analysis on a data set and what insight these methods can provide from the data. Based on the initial analysis, certain priming conditions (CD3CD28, CPG, and PG) were more easily grouped together using statistical techniques such as PCA and LDA. This was observed in both wild and lab mice. Additionally, the MANOVA test showed that most of the variables examined were statistically significant in determining group association. To predict the group/class accurately, different classification models were used. The Random Forest algorithm was found to be the most accurate in predicting the class for both wild and lab mice. Although the analysis provided valuable insights into the main trends in the data, further analysis could potentially uncover more information. It would be beneficial to conduct additional post-processing analysis to extract more data. The results of this study can be further used to investigate the patterns discovered in the data. Likewise, comparing additional mouse strains under the same experimental conditions could supplement the findings.

Table of Contents

Abbreviations

Abbreviation	Definition
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
CD3CD28	anti-CD23/anti-CD2
CPG	Cytosine-phosphate-Guanine
DNA	Deoxyribonucleic Acid
Eotaxin	Eosinophil chemotactic protein
FLAG	Flagellin
GCSF	Granulocyte colony-stimulating factor
GMCSF	Granulocyte-macrophage colony-stimulating factor
IFN-γ	Interferon-gamma
IL-10	Interleukin-10
IL-12p40	Interleukin-12 subunit p40
IL-12p70	Interleukin-12 subunit p70
IL-13	Interleukin-13
IL-15	Interleukin-15
IL-17a	Interleukin-17a
IL-18	Interleukin-18
IL-1α	Interleukin-1 alpha
IL-1β	Interleukin-1 beta
IL-2	Interleukin-2
IL-3	Interleukin-3
IL-4	Interleukin-4
IL-5	Interleukin-5
IL-6	Interleukin-6
IL-9	Interleukin-9
KC	Keratinocyte chemoattractant
LDA	Linear Discriminant Analysis
LIF	Leukemia inhibitory factor
LOOCV	Leave-One-Out Cross-Validation
LPS	Lipopolysaccharide
MANOVA	Multiple Analysis of Variance
MCP-1	Monocyte chemoattractant protein-1
MCSF	Macrophage colony-stimulating factor
MIG	Monokine induced by gamma interferon
MIP-1a	Macrophage inflammatory protein-1 alpha
MIP-1b	Macrophage inflammatory protein-1 beta
MIP-2α	Macrophage inflammatory protein-2 alpha
Naive Bayes	Naive Bayes classifier
NLRs	NOD-like Receptors
PCA	Principal Component Analysis
PG	Peptidoglycan
PIC	Polyinosinic-polycytidylic Acid

PRRs	Pattern Recognition Receptors
RF	Random Forest
RPMI	Roswell Park Memorial Institute
SVM	Support Vector Machine
TLRs	Toll-like Receptors
TNFα	Tumor necrosis factor alpha
VEGF	Vascular Endothelial Growth Factor

Introduction

The mouse animal model

Throughout the history of medicine and research, animal models have supplied valuable information for medical knowledge, allowing a better understanding of diseases, metabolic pathways, vaccines, drug development, the origin of diseases, and more (Robinson et al., 2019). Due to factors such as controlled conditions, hygiene, ease of maintenance, and high reproduction rate, rodents have been a highly appreciated research model animal. Its small size and genetic resemblance to human DNA favors the replicability of pathogenicity (Bryda, 2013). However, a vast majority of phenotype-based diseases are strain-dependent and therefore in order to ensure reproducibility specific strains have to be used depending on the research goal (Bryda, 2013). Lab mice have a long history of being used as animal models for the study of structure, function interactions and potential dissociative conditions for genes, proteins and other biological molecules. Laboratory mice have been selectively bred and isolated from their wild type counterparts for more than eight decades (Yang et al., 2011). The selectively cultivated strains of mice used as laboratory animal models have vastly homozygous genotypes that manifest phenotypes significantly affected by recessive alleles. The genetic inheritance of laboratory mice differs greatly from their wild counterparts and manifests in numerous aspects both phenotypic and genotypic (Wade et al., 2002). Laboratory mice have a significantly bulkier build than their wild counterparts; display a rapid growth time and early maturation, whit high levels of fertility and docile behavior (Sellers et al., 2011). The environments in which the mice live are tightly controlled and are significantly different from wild environments in terms of food and water accessibility, sheltered from any atmospheric and pathogenic conditions. In stark contrast with their captive overly controlled counterpart, wild mice are constantly submitted to environmental pressure and pathogenic particles (Viney et al., 2015).

Mouse strains

The genetics of wild mice

The modern house mouse is a species complex that derives from three individual lineages: *Mus musculus castaneus* (endemic to southeastern Asia), *Mus musculus musculus* (endemic from eastern Europe to northern Asia), and *Mus musculus domesticus* (endemic from Middle East to western Europe) (Geraldes et al., 2008). The *Mus musculus* aggregate of species began assimilating into a commensalist relationship with early human agricultural communities. This commensal existence has entailed considerable adaptations in dietary and behavioural patterns, as well as pathogenic exposure (Geraldes et al., 2008). The more recent human migratory and expansionistic tendencies have dispersed *M. musculus* to all terrestrial continents and has exposed them to a wide range of ecosystems and environmental pressures (Ferris et al., 2021). There is current research into the sequencing of genomes of mice with local adaptations in order to understand the evolutionary shifts underwent in different populations of *Mus musculus* and the underlying effects (Lawal et al., 2021).

The genetic variation incorporated in the genomes of wild mice is vastly diverse. The genetic variation of current inbred mouse strains incorporates approximately ten different haplotypes at 97% of all genomic loci (Chang et al., 2017). This indicates that wild type mice contain a significantly understudied and underused genomic variation which especially disease associated potential (Bogue et al., 2019). Wild-derived inbred mice encompass a higher degree of natural

variation than typical laboratory inbred mouse strains (Chang et al., 2017). Recent development in exome analysis of wild-derived inbred strains have already identified significant genetic variation which most likely encodes for functional phenotypic effects (Phifer-Rixey et al., 2018). The effect of incorporating increased genetic variation in mouse animal models is significantly important for the translation of human adaptations and diversity patterns, specifically regarding disease risk and incidence (Saul et al., 2019).

The C57BL/6J strain

Arguably the most widely used laboratory mouse strain in the world, its history dates back to 1921. This congenic strain was originally inbred by C. C. Little from a stock of mice purchased from a reputed rodent breeder, Abbie Lathrop (Steensma et al., 2010). The *C57BL/6J* strain (figure 1.) manifests many peculiarities which have made it so popular; such as: thermic and pain sensitivity, analgesic resistance (Taniguchi et al., 1998), susceptibility to alcohol (Gentry, 1989) and morphine addiction as well as predisposition towards developing certain diseases (Crusio et al., 2013).

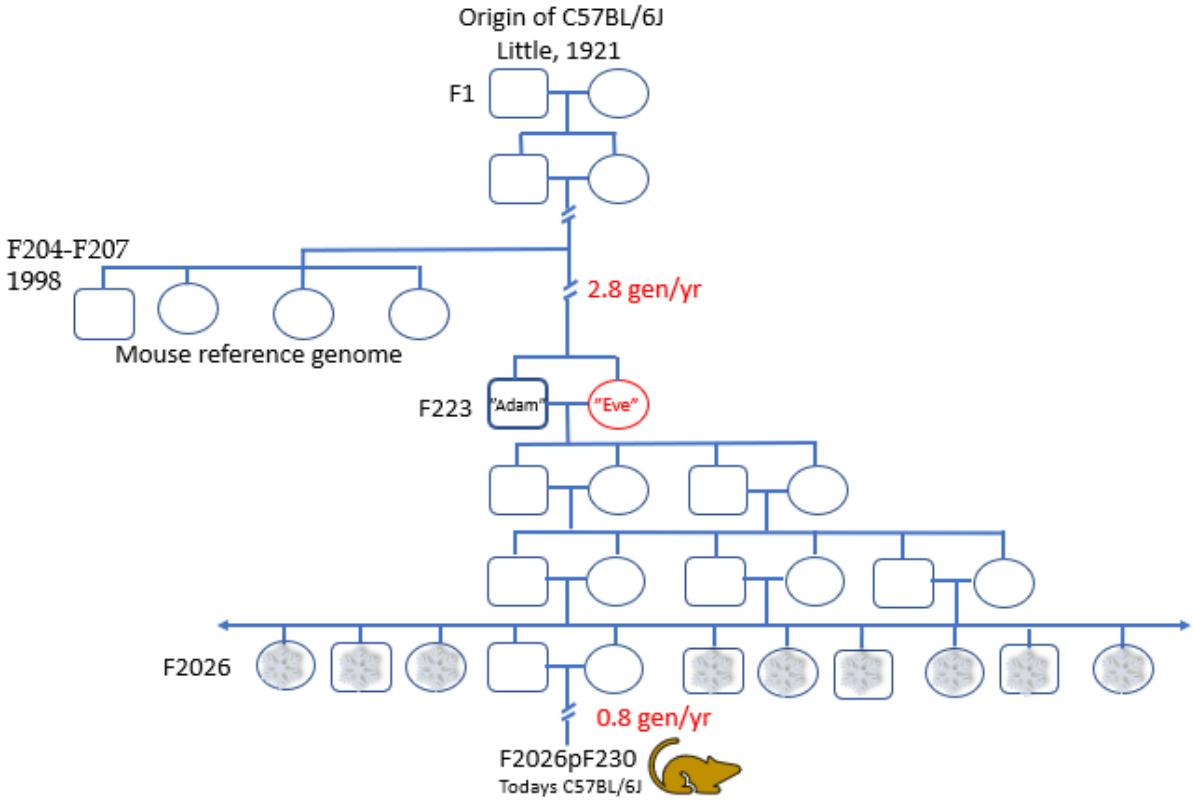

Figure 1. Diagram of the *C57BL/6J* lineage from oldest known progenitor to currently available Jackson Laboratories commercial strain (designed in PowerPoint based on JAX diagram). Each row represents a Filial generation, and snowflake symbols represent cryopreserved embryos. Notable generations are represented with an alphanumeric code starting with F. Starting from F1 representing the original stock purchased by C. C. Little up to the current commercial stock F2026pF230.

The advantages of genetically identical mice used for research endures, especially because of the controllability of their genetic background. Since commercial mouse strains have already been sequenced, including the *C57BL/6J* strain, there is a large collection of phenotypic and genomic data collected (Sarsani et al., 2019). Individuals from the isogenic laboratory *C57BL/6J* mouse strain are understood to be identical, in spite of inbreed mice being exposed to some genetic drift;

a considerable effort is undertaken to minimize false positive results (Sarsani et al., 2019). Well-developed research studies also consider the necessity of a certain degree of randomization between individuals in order to avoid accidental bias in biological experiments, which has a high significance in human trials (Suresh, 2011). The lack of randomization is a concern of the scientific community about in-vivo experiments using mice models isogenic mice (Muhlhausler et al., 2013). Addressing this concern requires a critical study of the widely used strains such as *C57BL/6J*, as well as the advantages and limitations in comparison to other mice derivations.

Cytokines and Chemokines

Cytokines are a class of soluble molecules that can communicate between cells of the immune system (Cannon, 2000). The interactions between cytokines and their affiliate receptor molecules can trigger numerous effects such as inducing the increase or decrease in the production of certain proteins inside the cell, the expression of specific surface molecules (Remick, 2002), as well as triggering certain transcriptional modifications (Leonard, 2001). The vast majority of the cytokine nomenclature is comprised out of *interleukins* (Mizel & Farrar, 1979); from the Latin *inter* - between and *leukocytes*- white blood cells. Other molecules that are also considered cytokines include interferons, haematopoietin, tumor growth factors and tumor necrosis factors (Kuby et al., 2006). Chemokines are subfamily of cytokines that have a role in the immune response, most importantly in the mobilization of immune cells (Charo & Ransohoff, 2006). Chemokines are divided into two functional groups: homeostatic – intrinsically produced by specific cells, having a role in the transmigration of basal leukocytes; and inflammatory – expressed as a response to pathological events, actively participating in recruiting immune cells to the location of the inflammatory response (Raman et al., 2011).

Priming conditions

Priming agents are substances or stimuli that are used to "prime" or activate cells, tissues, or organisms in experimental settings. Priming is a process by which a cell or organism is prepared or sensitized to respond more effectively to subsequent stimuli or challenges (Punt et al., 2019). Priming agents can include various substances, such as chemicals, drugs, molecules, or biological agents, that induce specific cellular or physiological responses (Tang et al., 2012). The choice of priming agent depends on the research objectives and the specific system being studied. The priming conditions evaluated in this study are:

Cytosine-phosphate-Guanine, or CPG, is a synthetic DNA molecule that mimics bacterial DNA and can activate the innate immune system by triggering Toll-like receptors (TLRs) on immune cells, typically TLR9 (Schnare et al., 2000; Rattanakiat et al., 2009).

Lipopolysaccharide, or LPS, is a molecule found on the outer membrane of Gram-negative bacteria that can elicit a strong immune response by activating Toll-like receptors (TLRs) on immune cells, typically TLR4. LPS is a potent endotoxin that can cause sepsis and septic shock (Rietschel et al., 1994).

CD3CD28: CD3 and CD28 are surface proteins found on T cells. CD3 is involved in signal transduction and activation of T cells, while CD28 provides a co-stimulatory signal that enhances T cell activation. The CD3CD28 complex is often used to activate T cells in vitro for research purposes (Heylmann et al., 2018).

Flagellin, or FLAG, is a protein that makes up the filamentous structure of bacterial flagella, which are the whip-like tails that some bacteria use for locomotion. Flagellin is an important antigen that

can trigger immune responses by activating Toll-like receptors (TLRs) on immune cells, typically TLR5; and play a role in bacterial pathogenesis (Hatai et al., 2016).

Peptidoglycan, or PG, is a component of bacterial cell walls that can be recognized by the innate immune system through pattern recognition receptors (PRRs) such as NOD-like receptors (NLRs) and Toll-like receptors (TLRs) on immune cells, typically TLR2 (Davis et al., 2010; Bersch et al., 2021).

Polyinosinic-polycytidylic acid, or PIC, is a synthetic RNA molecule that mimics viral double-stranded RNA and can activate the innate immune system by triggering Toll-like receptors (TLRs) on immune cells, typically TLR3. PIC is often used to study antiviral immune responses in vitro (Fortier et al., 2004).

RPMI media is a commonly used culture medium that provides a nutrient-rich environment for the growth and maintenance of cells in vitro. It contains various components such as amino acids, vitamins, glucose, and salts to support cell growth (Švajger & Jeras, 2011).

A community resource

In order to evaluate the potential difference in immunology between laboratory mice and wild mice, the research group of Abolins et al., have conducted an extensive study of morphology, immunology and genetics of a large number of wild mice captured in different parts of the United Kingdom (Abolins et al., 2017). The captured mice were screened for afflicting infections and serum concentration of inflammatory proteins. The authors of the paper wave provided their findings as a community resource. The analysis conducted did not involve multivariate data analysis which could yield significant information.

Multivariate data analysis

Intrinsic to any type of life science research is data analysis. Very often, quantitative data analysis is based on univariate approaches, which individually calculate means and standard deviations for one variable at the type and evaluate the placement of a variable's value within the threshold of a specific range. P-values are computed in order to determine the significance of difference between the values of two or more groups. However, univariate approaches to data analysis disregard potential correlations between the variables and in the context of biological measures the possibility of systems or networks between said variables (Vargason et al., 2017).

Multivariate data analysis has numerous approaches that can be used self-standing or by combining them into different frameworks. Principal component analysis (PCA) is highly useful for data visualization, but could potentially disregard incipient trends in complex data. Multiple analysis of variance (MANOVA) can enhance the statistical processing of the significance of interactions between variables (Johnson et al., 2007). Different classifiers were compared and evaluated using the microbial antigen groups as the dependent variable.

Aim

The aim of this thesis is to use a multivariate approach to data analysis on the differential expression of cytokines and chemokines after stimulation with pathogenic components, in order to determine if and what combination of immunological biomarkers can be used to identify the pathogenic stimulus. The importance of this research is generating comparative profiles of the mouse animal model in contrast with its wild counterpart. The objectives of this project are:

1. Process the data set by splitting it into a laboratory mice subset and wild type subset in order to evaluate these subsets separately.
2. Perform exploratory analysis of the data set using cluster analysis and factor analysis
3. Perform classification and predictive modelling
4. Perform model inspection and validation

The differential expression will be evaluated by splitting the data set into wild mice and lab mice and evaluating these subsets individually. The performance of different classification methods for the identification of the pathogenic stimulus will be evaluated. The response to pathogenic stimulus priming (artificial stimulation) should be directly correlated with pathogenic infection (natural stimulation). Evaluating the change in immunological biomarkers after priming with pathogenic stimuli is important for understanding how useful are laboratory mouse models in contrast with their wild type counterparts. These profiles could later on be compared with human ones therefore identifying the fitness of the laboratory mouse for the modelling of cellular responses to pathogenic stimuli.

Materials and Methods

Data set

The data set was provided as a community resource by Abolins et al. The mice were euthanized using sodium pentobarbital intraperitoneal injection. The measures of rodents were subsequently taken, followed by exsanguination via cardiac puncture and dissection. Rodent blood was stored into heparinized containers, centrifuged at 13,000 g at a temperature of 18 °C for 10 min; aliquots of plasma were created and stored at a temperature of - 20 °C. The spleens were removed in aseptic environments, measured and subsequent single cell suspensions were generated using homogenization via a 70 μm cell strainer (BD Biosciences). Using a haemocytometer, the cells were microscopically counted; living cells were determined using trypan blue exclusion. The specific data set used as the basis of this project was generated from Cytokine bead arrays and is referred to as Supplementary Data 4, by Abolins et al.

Software

The *RStudio* (version 4.2.3) integrated development environment was used for the scripting of multivariate data analysis algorithms. Individual algorithms are mentioned in the sections corresponding to each statistical procedure.

Exploratory analysis

In order to make large scale data more interpretable, as well as identifying trends and identifying potential outliers, at least one type of exploratory analysis must be performed (Filipovych et al., 2011). Once an appropriate type of exploratory analysis is performed, the compacted data is often visualized graphically in order to facilitate the interpretation (Wehrens, 2011).

Principal component analysis

Principal component analysis (PCA) is a combinatorial multivariate statistical procedure that enables an improved interpretation of large-scale data by summarizing parts of the data into indices that can be more easily manipulated and visualized (Daffertshofer et al., 2004). PCA was used in order to illustrate and identify clusters, outliers and patterns in the analyzed data set. In order to simplify the interpretation of data, PCA constructs new variables based on specific

combinations of the pre-existing variables and defines the vector of the new variable (principal component) in the direction that encompasses the highest variation (Leigh & Jackson, 1993). Each additional vector has an orthogonal orientation to the previously computed new variable with a direction encompassing the majority of the remaining variation. The main principle at the basis of PCA is that with a high degree of incidence a large percentage of variables in a data set are redundant (Hess & Hess, 2018). With the help of PCA, multi-dimensional data can be directly mapped into a reduced dimensional space without losing significant amounts of information from the original data set. The underlining understanding is that variation equates to information (Jolliffe, 1990). Although this is true in most cases, sometimes variation could also be unimportant, possibly representing noise (Jolliffe, 1993). Once PCA has computed the new "latent" variables, plotting the first two components should yield significant information. In R, the *prcomp* function was used to generate a PCA.

Multiple analysis of variance

A multiple analysis of variance statistical test or MANOVA, is a more complex version of the univariate ANOVA. MANOVA is a useful procedure in which two or more dependent variables can be modelled at the same time (Thulin, 2016). However, this is only useful in the case in which the supposition at the basis of the analysis that a linear combination of independent variables can be used to determine the dependent variables that need to be modelled (Ho-Wan Kwak, 2010). The purpose of using MANOVA in contrast to performing several univariate ANOVAs, in addition to aforementioned considerations is the mitigation of type I error rates stemming for using disparate tests (Kariya, 1978). In R, the *manova* function included in the standard suite of packages was used to perform the statistical test.

Linear discriminant analysis

Linear discriminant analysis (LDA) is a statistical procedure that is useful for dimensionality reduction. In order to enable dimensionality reduction, this supervised machine learning procedure generates a linear combination of features that discriminates between the different classes in the dataset. LDA maximizes the variance between classes, while minimizing the variance within classes (Hastie et al., 2009). The linear transformation of the aforementioned features is termed as linear discriminants. In R, the *lda* function in the *MASS* package was used to generate an LDA.

Modelling

The use of modelling algorithms or classification methods is important for the recognition of patterns. In other terms, classification allows the segregation of objects into classes and can provide a predictive framework for the incorporation of new objects. Classification algorithms generate predictive models based on training data, which is generally obtained from previous research data (Ramasubramanian & Moolayil, 2019). On par with the predictive capability, a functional classification method enables the determination of important variables (variables that can be associated with an effect), which in life sciences can be dubbed as a biomarker (Sotirov et al., 2022).

Linear discriminant analysis classifier

Linear discriminant analysis is a reductionist statistical procedure that computes linear discriminants in order to reduce the dimensionality of a data set and can enable class prediction based on the computed linear discriminants (Park & Park, 2008). The *lda* function in R was used to generate a linear discriminant analysis.

K-nearest neighbors classifier

K-nearest neighbor classifies new observations into classes that contain the highest number of previous observations that had variables with similar values (Zhang, 2016). This classifier, does not rely on assumptions about the distribution of items into classes, but rather the distances between the items, more specifically the ones that have the shortest distance between them (nearest neighbors). The k-parameter refers to the number of items taken into consideration; if k=1, the nearest neighbor will be allocated to the class of the item that is closest to it, if k>1 a majority vote of the k nearest neighbors will allocate the class to item (Cover & Hart, 1967). For the training set of data, the distance between each object is calculated using the Euclidian distance. In R, the *knn* function in the package *class was used*.

Logistic regression classifier

The applications of linear regression analysis entail establishing causal relationships between a dependent variable and a number of independent variable (regressors), generating predictions of the dependent variable based on specific values of independent variables, and screen independent variables in order to determine which variables have a higher importance for determining the dependent variable with efficiency and accuracy (Yan et al., 2009). In classification problems in contrasts with typical regression, the dependent variable is categorical instead of numeric (Stoltzfus, 2011). In order to model categorical variables, the outcome of the regression must be transformed into the log of the odds ratio (OR), representing the probability of occurrence of a specific event (Hess & Hess, 2019). In R, the *glm* (general linear model) function was used to create a logistic regression model of the data.

Random forest classifier

A random forest (RF) is formed out of a collection of decision trees that utilizes random sampling and adjustable parameters. The RF algorithm can provide accurate predictions, evaluate the importance of features and pairwise proximity between samples (Cha Zhang & Ma, 2012). The random forest algorithm was initially developed by Breiman (Breiman, 2001), and is a versatile tool that can be used for supervised classification and regression. A random forest is constructer as an ensemble of decision trees which are created by allocating training samples through a process of replacement, or so called “bagging”. The bagging approach entails the selection of certain items on multiple occasions, while certain items might not be selected even once (Biau & Scornet, 2016). Approximately two thirds of all items, considered in-bag items, are part of the sub-set used to train the decision trees. The other one third of the items, considered out-of-the bag samples are utilized as cross-validation items in evaluating the performance of the random forest model. The performance estimation method is referred to as the out-of-bag error (Belgiu & Drăguţ, 2016). The individual decision trees are generated without pruning, while nodes are generated based on a user designated number of features and are selected at random. The number of decision trees generated is based on a user designated input and exhibit a high level of variance alongside a low level of bias (Belgiu & Drăguţ, 2016). The conclusive classification of the algorithm is generated by averaging the arithmetic mean of the estimated class designation computed by each individual decision tree. In the case of new data which does not have any classification, each item is scrutinized by all the constructed decision trees and the class assignation is given based on the highest number of concordant decisions (Belgiu & Drăguţ, 2016). The *randomForest* package in R was used to generate the eponymous classifier.

Naive Bayes classifier

The Naive Bayes (NB) classifier is based on Bayes' theorem, which is a fundamental principle of probability theory. Bayes' theorem states that the probability of a hypothesis (such as a class label) given some evidence (such as a set of features) is proportional to the probability of that evidence given the hypothesis, multiplied by the prior probability of the hypothesis. In the context of classification, the Naive Bayes classifier calculates the probability of each class label given a set of features for a new instance, and then assigns the class label with the highest probability as the prediction for that instance (Stephens et al., 2017). The "naive" in the name comes from the assumption that the features are conditionally independent given the class label. In other words, the algorithm assumes that the presence or absence of one feature does not affect the probability of any other feature being present or absent. This simplifying assumption allows the algorithm to work efficiently with a large number of features.

To calculate the probabilities, the Naive Bayes classifier uses a probability distribution model for each feature, which can be either a Gaussian distribution for continuous features or a multinomial distribution for discrete features. The parameters of these probability models are estimated from the training data using maximum likelihood estimation. Once the probability models are trained, the Naive Bayes classifier can be applied to new instances by calculating the posterior probability of each class label given the observed features, using Bayes' theorem. The class label with the highest posterior probability is then assigned as the prediction (Taheri & Mammadov, 2013).

Overall, the Naive Bayes classifier is a simple and efficient algorithm for classification that can work well even with high-dimensional data. However, its performance can be limited by the "naive" assumption of feature independence, which may not hold in all cases. The *nb* function, from the *e1071* package in R was used to create a Naive Bayes model of the data.

Support Vector Machine classifier

Support vector machine (SVM) is a popular classification algorithm that can be used for both linear and non-linear classification tasks. SVM works by finding the hyperplane in a high-dimensional space that best separates the classes. In other words, it attempts to find the best possible boundary that can separate the two classes in a given dataset (Decoste & Schölkopf, 2002).

The hyperplane that SVM finds is the one that maximizes the margin between the two classes. The margin is the distance between the hyperplane and the closest data points from each class. By maximizing the margin, SVM is trying to find the hyperplane that will best separate the two classes and generalize well to unseen data. When the classes cannot be separated by a linear hyperplane in the input space, SVM uses a kernel trick to transform the input data into a higher dimensional space where the classes can be separated by a linear hyperplane. The kernel function maps the input data into a new space where it becomes easier to find a hyperplane that separates the classes. This technique is called the kernel trick, and it allows SVM to classify non-linearly separable data (Noble, 2006). The *svm* function, from the *e1071* package in R was used to create a support vector machine model of the data.

Artificial Neural Network classifier

An Artificial Neural Network (ANN) classifier is a machine learning algorithm that is inspired by the structure and function of biological neurons in the human brain. It is a type of supervised learning algorithm that can be used to classify input data into different categories or classes. An

ANN classifier consists of multiple layers of artificial neurons, each of which performs a simple computation on its inputs and passes the result to the next layer (Schmidhuber, 2015).

The first layer is called the input layer, which receives the input data, and the last layer is the output layer, which produces the classification output. Between the input and output layers, there can be one or more hidden layers that perform complex computations on the input data. Each neuron in a hidden layer takes the weighted sum of the inputs and passes it through an activation function, which produces the output of the neuron. This output is then used as input to the neurons in the next layer. During the training phase, the weights and biases of the neurons in the network are adjusted based on the error between the predicted output and the actual output. This process is known as backpropagation, and it allows the network to learn how to classify the input data correctly. Once the network is trained, it can be used to classify new input data by passing it through the network, and the output of the last layer gives the predicted class label (Krogh, 2008). The *neuralnet* package in R was used to create an artificial neural network model of the data.

Model inspection and validation

Cross-validation is a method for estimating the performance of a model by splitting a dataset into training and testing sets, and evaluating the model on each split. There are numerous cross-validation methods such as the holdout, k-fold, leave-one-out, bootstrapping and others.

The holdout method entails randomly splitting the data into a training and a testing set; this cross-validation methods drawback is that there is a possibility that the fractional data set might not properly represent a model's performance. In k-fold cross-validation, the data set is divided into k partitions, using k-1 partitions as the training and the remainder is used for testing. Repeated k-fold cross-validation extends this by repeating the process of k-fold cross-validation multiple times (Cheng et al., 2017). In each repetition, a different random partitioning of the data into k folds is performed, and the model is trained and evaluated k times on these new partitions. The final performance of the model is then calculated as the average of all the evaluations across all repetitions. Leave-one-out cross-validation (LOOCV) is essentially an extreme version of k-fold cross-validation in which k is equal with the total number of items; the training data is essentially the entirety of the data excepting one item which is used for testing, this procedure is repeated a n number of times. The major drawback of LOOCV is the high computational cost. Bootstrap resampling is a method for estimating the variability of a model by repeatedly sampling a dataset with replacement and fitting a model on each sample. Combining cross-validation with bootstrap resampling provides a more accurate estimation of the performance of a model. The basic idea is to use bootstrap resampling to create multiple training and testing sets, and then use cross-validation on each of these sets to evaluate the performance of the model (Cheng et al., 2017).

In order to establish an average accuracy of each classifier, the LOOCV, repeated k-fold and bootstrap resampling cross-validation methods were implemented, using the *caret* package.

Results

Exploratory analysis

Principal Component Analysis of the wild mice subset

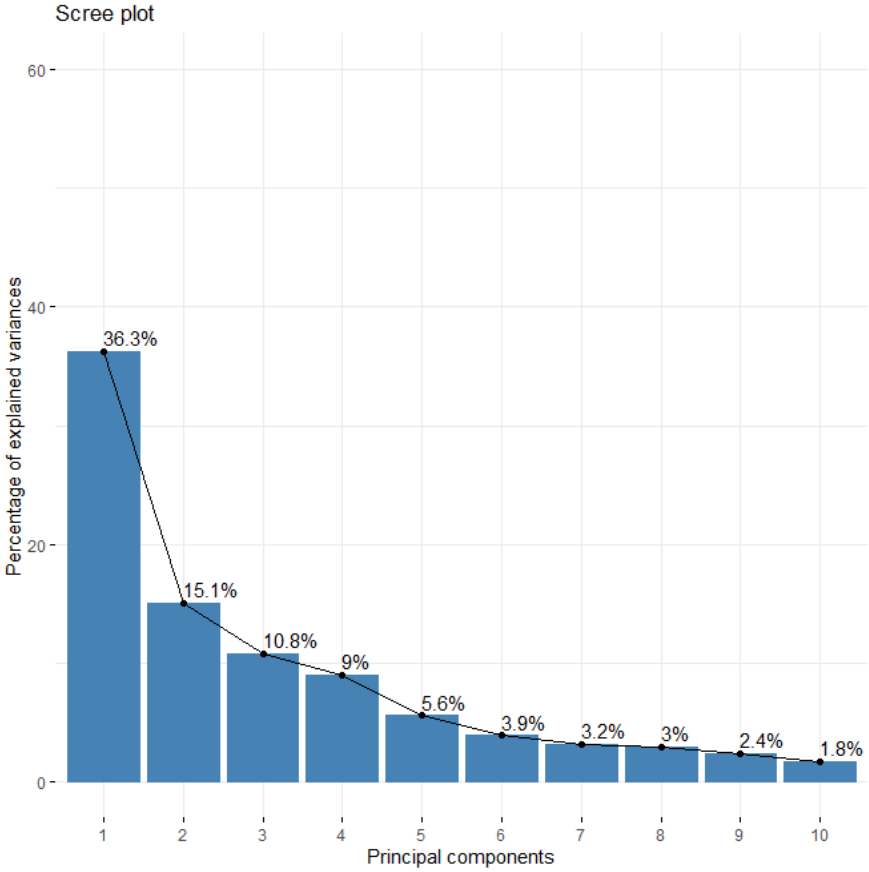

Figure 2. Scree plot of the principal components of the Wild mice subset PCA. On the x-axis the computed principal components are enumerated, while the y-axis displays the percentage of variance explained by each component.

The scree plot (figure 2.) indicates retaining the first 6 components as they are part of the steep curve of the plot (computed in R with the code block in Appendix, #Section 2).

Table 1. Importance of principal components, of the wild mice subset PCA, according to standard deviation, proportion of variance represented by each component and the cumulative proportion of variance represented by subsequent principal component.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	3.41	2.20	1.86	1.69	1.34	1.12	1.01	0.98	0.88	0.75
Proportion of Variance	0.36	0.15	0.11	0.09	0.06	0.04	0.03	0.03	0.02	0.02
Cumulative Proportion	0.36	0.51	0.62	0.71	0.77	0.81	0.84	0.87	0.89	0.91

The principal component analysis of the wild mouse strains managed to modestly reduce the complexity of the data. The total cumulative variance (table 1) accounting for 80 percent of the

data is represented by the first 6 components and 90 percent of data being encompassed by the first 10 components (computed in R with the code block Appendix, #Section 2). Taking into consideration that these two principal components (PC1 and PC2), encompass 51.4% of the variation of the data; they can provide an overview of the distribution of the data and thus enable a reduction in data complexity.

$$Variance(of\ a\ PC) = \frac{PC\ sdev^2}{\sum PC\ sdev^2}$$

Equation 1. The equation used to calculate the variance of a principal component.

$$Average\ variance = \frac{\sum Variance\ (the\ sum\ of\ each\ PCs\ variances)}{PCs(number\ of\ principal\ components)}$$

Equation 2. The equation used to calculate average variance.

Table 2. All the principal components, of the wild mice subset PCA, with a variance greater than the average variance (computed in R according to equations 1 and 2).

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**
Variance	0.363	0.151	0.108	0.090	0.056	0.039	0.032
Average	0.031						

According to the Keiser's criterion specifying the retention of the components whose variance (equation 1) is higher than the average variance (equation 2) (Jolliffe & Cadima, 2016); the first 7 components (table 2) should be retained (computed in R with the code block in Appendix, #Section 2).

Table 3. The loadings of the first eight most principal components important (according to the consensus of the scree plot, Keiser criterion and percentage of total variance represented), of the wild mice subset PCA, extracted from the total 32 generated principal components.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
Basic.FGF	0.075	0.015	-0.109	0.001	-0.207	0.66	0.135	-0.008
Eotaxin	0.212	-0.011	0.226	-0.158	-0.096	0.023	0.303	-0.148
GCSF	0.131	0.06	0.062	0.457	-0.198	-0.141	-0.066	-0.001
GMCSF	0.238	-0.071	0.14	-0.096	-0.183	-0.01	0.091	-0.207
IFN.γ	0.052	-0.285	-0.036	0.045	0.012	-0.066	-0.423	-0.417
IL.10	0.246	0.12	-0.152	-0.018	0.248	-0.056	-0.049	0.092
IL.12p40	0.196	0.094	0.112	-0.191	0.158	0.034	0.177	-0.02
IL.12p70	0.274	0.077	-0.073	-0.09	0.079	-0.053	0.086	-0.045
IL.13	0.25	-0.144	0.164	-0.012	-0.018	0.012	0.144	0.024
IL.15	0.219	0.141	-0.285	-0.027	0.142	-0.025	0.002	0.021
IL.17a	0.061	-0.324	-0.095	-0.027	-0.033	-0.142	0.076	-0.36
IL.18	0.074	0.103	-0.073	-0.289	-0.389	0.019	-0.376	0.181
IL.1α	0.241	0.07	-0.047	0.167	-0.152	0.1	-0.029	0.005
IL.1β	0.231	0.109	0.008	-0.095	-0.126	0.15	0.284	-0.08

IL.2	0.038	-0.365	-0.085	-0.029	-0.078	-0.092	0.171	0.001
IL.3	0.244	0.031	-0.224	-0.033	0.18	-0.061	-0.033	-0.02
IL.4	0.063	-0.391	-0.118	-0.001	-0.04	-0.053	0.218	0.139
IL.5	0.048	-0.343	-0.084	0.02	0.002	-0.007	0.087	0.545
IL.6	0.131	-0.001	0.366	-0.116	0.192	-0.067	-0.101	0.129
IL.9	0.072	0.089	0.009	-0.312	-0.376	-0.343	-0.215	0.1
KC	0.17	0.086	-0.011	0.426	-0.124	-0.018	0.06	0.013
LIF	0.119	-0.361	-0.185	0.005	0.032	-0.101	-0.029	0.044
MCP.1	0.215	0.136	-0.268	0.032	0.217	-0.057	0.017	0.012
MCSF	0.193	-0.196	-0.017	0.006	-0.178	0.252	-0.237	0.328
MIG	0.111	-0.21	-0.146	-0.015	0.08	0.263	-0.264	-0.278
MIP.1a	0.224	0.006	0.296	0.124	-0.012	-0.128	-0.065	0.009
MIP.1b	0.245	-0.046	0.195	0.057	0.153	-0.081	-0.163	0.019
MIP.2α	0.146	0.069	0.025	0.465	-0.174	-0.096	-0.03	0.033
PDGF.BB	0.181	0.161	-0.331	-0.073	0.145	-0.064	-0.102	0.09
RANTES	0.106	-0.026	0.243	0.012	0.201	0.39	-0.299	-0.009
TNFα	0.186	0.119	-0.096	-0.185	-0.335	-0.057	0.051	-0.165

In the case of PC1; the highest loadings (table 3) belong to IL-1a, IL-1b, IL-3, IL-10, IL-12p70, IL-13, Eotaxin, GMCSF, MCP-1, MIP-1a, MIP-1b and IL-15 are all positively correlated with the principal component (PC1). The cutoff value is arbitrary (based on observation) and in this case set at 0.2. What this means is that, essentially, by only taking into consideration the aforementioned variables, the underlying structure of PC1 can be preserved (computed in R with the code block Appendix, #Section 2). In the case of PC2; the highest loadings (table 3) belong to: IL-2, IL-4, IL-5, IL-17a, INF-γ, LIF and MIG. All of these variables have a negative correlation with the principal component (PC2). The same arbitrary cutoff value (0.2) was used (computed in R with the code block Appendix, #Section 2).

Figure 3. PCA plot of the wild mice subset illustrating the color-coded distribution of data. The coordinates of each data point are compiled from the first two principal components. The first principal component (PC1) is plotted on the x-axis, the second principal component (PC2) is plotted on the y-axis; the different pathogenic stimuli are color coded and explained in the legend. The graph was computed using the ggplot package.

The first two principal components of the wild mice subset explained 36,3% (PC1) and 15.1% (PC2) of the total variance respectively. The loadings, or coefficients of each variable in a principal component underpin a structural pattern in the data. The loadings with the highest numerical values contribute the most to understanding the meaning of each principal component. The sign of each loading (positive or negative) indicates the correlation with the principal component. Based on the magnitude of loading, variables influence on the principal component is explained, while based on the sign, the relationship of the variable with the principal component can be explained.

By plotting the first two principal components of the wild mice subset, the spread of data can be visualized allowing for a more facile interpretation, as illustrated in figure 3. In the PCA plot of the wild mice subset there is a clear separation of the CD3CD28 group, and a partial separation

between the CPG group and the PG group; the LPS, FLAG, PIC and RPMI groups are intermingled with no clear separation of the groups.

Principal Component Analysis of the lab mice subset

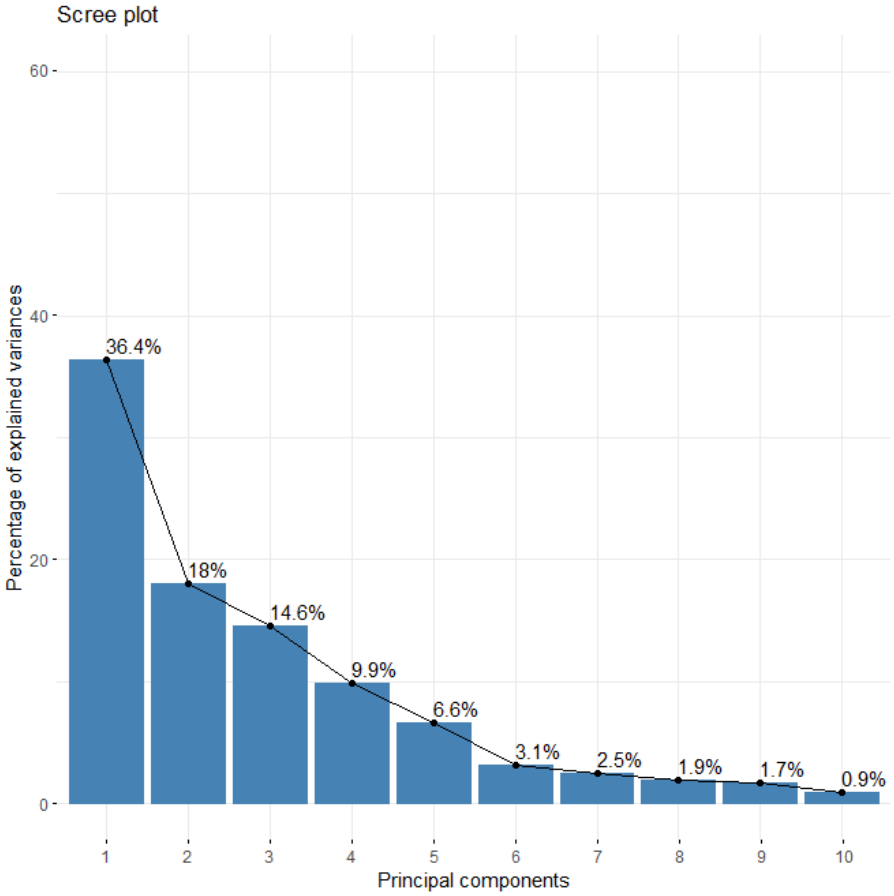

Figure 4. Scree plot of the principal components of the lab mouse strain PCA. On the x-axis the computed principal components are enumerated, while the y-axis displays the percentage of variance explained by each component.

The scree plot (figure 4.) indicates retaining the first 5 components as they are part of the steep curve of the plot (computed in R with the code block Appendix, #Section 2).

Table 4. Importance of components, of the lab mice subset, according to standard deviation, proportion of variance represented by each component and the cumulative proportion of variance represented by subsequent principal component.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	3.41	2.40	2.16	1.78	1.46	1.00	0.89
Proportion of Variance	0.36	0.18	0.15	0.10	0.07	0.03	0.02
Cumulative Proportion	0.36	0.54	0.69	0.79	0.85	0.89	0.91

The principal component analysis of the lab mouse strain, similarly, managed to modestly reduce the complexity of the data. The total cumulative variance (table 4) accounting for 80 percent of the data is represented by the first 5 components and 90 percent of data being encompassed by the first 7 components (computed in R with the code block Appendix, #Section 2). Taking into consideration that these two principal components (PC1 and PC2), encompass 54.4% of the variation of the data; they can provide an overview of the distribution of the data and thus enable a reduction in data complexity.

Table 5. All the principal components, of the lab mice subset PCA, with a variance greater than the average variance (computed in R according to equations 1 and 2).

	PC1	PC2	PC3	PC4	PC5	PC6
Variance	0.364	0.180	0.146	0.099	0.066	0.031
Average	0.031					

According to the Keiser's criterion specifying the retention of the components whose variance is higher than the average variance; the first 6 components (table 5) should be retained (computed in R with the code block Appendix, #Section 2).

Table 6. The loadings of the first eight most principal components important (according to the consensus of the scree plot, Keiser criterion and percentage of total variance represented), of the lab mice subset PCA, extracted from the total 32 generated principal components.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Basic.FGF	-0.089	-0.04	-0.128	0.17	0.03	0.782	-0.051	-0.111
Eotaxin	-0.21	0.095	0.155	-0.138	-0.186	0.072	0.227	0.336
GCSF	-0.142	-0.023	0.101	0.445	-0.119	-0.101	0.007	-0.174
GMCSF	-0.247	-0.006	0.161	-0.052	-0.098	0.205	0	0.121
IFN.γ	-0.109	-0.306	0.086	0.153	-0.075	-0.181	-0.218	-0.046
IL.10	-0.234	0.077	-0.209	-0.01	0.203	-0.074	-0.114	-0.101
IL.12p40	-0.212	0.098	0.061	-0.252	0.166	-0.063	0.151	0.073
IL.12p70	-0.266	0.067	-0.133	-0.075	0.117	-0.002	0.003	0.023
IL.13	-0.256	-0.069	0.155	-0.022	0.069	0.001	0.18	0.173
IL.15	-0.191	0.069	-0.329	-0.017	0.105	-0.038	-0.043	-0.001
IL.17a	-0.026	-0.402	-0.031	-0.065	0.02	-0.048	-0.021	0.129
IL.18	-0.055	0.07	-0.146	-0.175	-0.545	0.102	-0.169	0.163
IL.1α	-0.226	0.068	-0.011	0.178	-0.223	0.031	-0.064	0.327
IL.1β	-0.247	0.102	0.003	-0.023	-0.146	-0.02	0.309	0.2
IL.2	0.011	-0.346	-0.043	-0.104	-0.008	0.174	0.37	-0.17
IL.3	-0.222	-0.049	-0.264	-0.049	0.125	-0.085	-0.039	0.067
IL.4	-0.052	-0.367	-0.014	-0.088	0.011	0.118	0.298	-0.024
IL.5	-0.048	-0.347	0	0.032	-0.039	-0.212	-0.227	0.289
IL.6	-0.17	0.053	0.287	-0.228	0.115	-0.078	-0.008	-0.123
IL.9	-0.116	0.004	0.037	-0.259	-0.418	-0.118	-0.112	-0.494
KC	-0.186	0.036	-0.051	0.376	0.045	-0.004	0.198	-0.097
LIF	-0.07	-0.356	-0.131	-0.074	0.06	0.054	0.087	-0.109
MCP.1	-0.185	0.068	-0.319	0.007	0.199	-0.05	-0.009	-0.014
MCSF	-0.204	-0.179	0.116	0.07	-0.198	0.01	-0.181	-0.036
MIG	-0.076	-0.359	-0.068	-0.063	0.048	-0.044	-0.173	0.063

MIP.1a	-0.233	0.036	0.25	0.113	0.017	-0.063	0.027	-0.083
MIP.1b	-0.252	0.018	0.186	-0.011	0.134	-0.075	-0.078	-0.187
MIP.2α	-0.167	0.001	0.037	0.445	-0.065	-0.08	0.072	-0.129
PDGF.BB	-0.155	0.077	-0.362	-0.052	0.035	-0.098	-0.14	0.045
RANTES	-0.091	0.006	0.268	-0.028	0.249	0.334	-0.511	0.116
TNFα	-0.197	0.039	-0.196	-0.152	-0.24	0.109	-0.069	-0.286

In the case of PC1; the highest loadings (table 6) belong to IL-1a, IL-1b, IL-3, IL-10, IL-12p40, IL-12p70, IL-13, Eotaxin, GMCSF, MIP-1a and MIP-1b are all negatively correlated with the principal component (PC1), using the arbitrary cutoff value of 0.2. These variables underpin the structure of PC1 (computed in R with the code block Appendix, #Section 2).

In the case of PC2; the highest loadings (table 6) belong to: IL-2, IL-4, IL-5, IL-17a, INF-γ, LIF and MIG. All of these variables have a negative correlation with the principal component (PC2). The same arbitrary cutoff value (0.2) was used (computed in R with the code block Appendix, #Section 2).

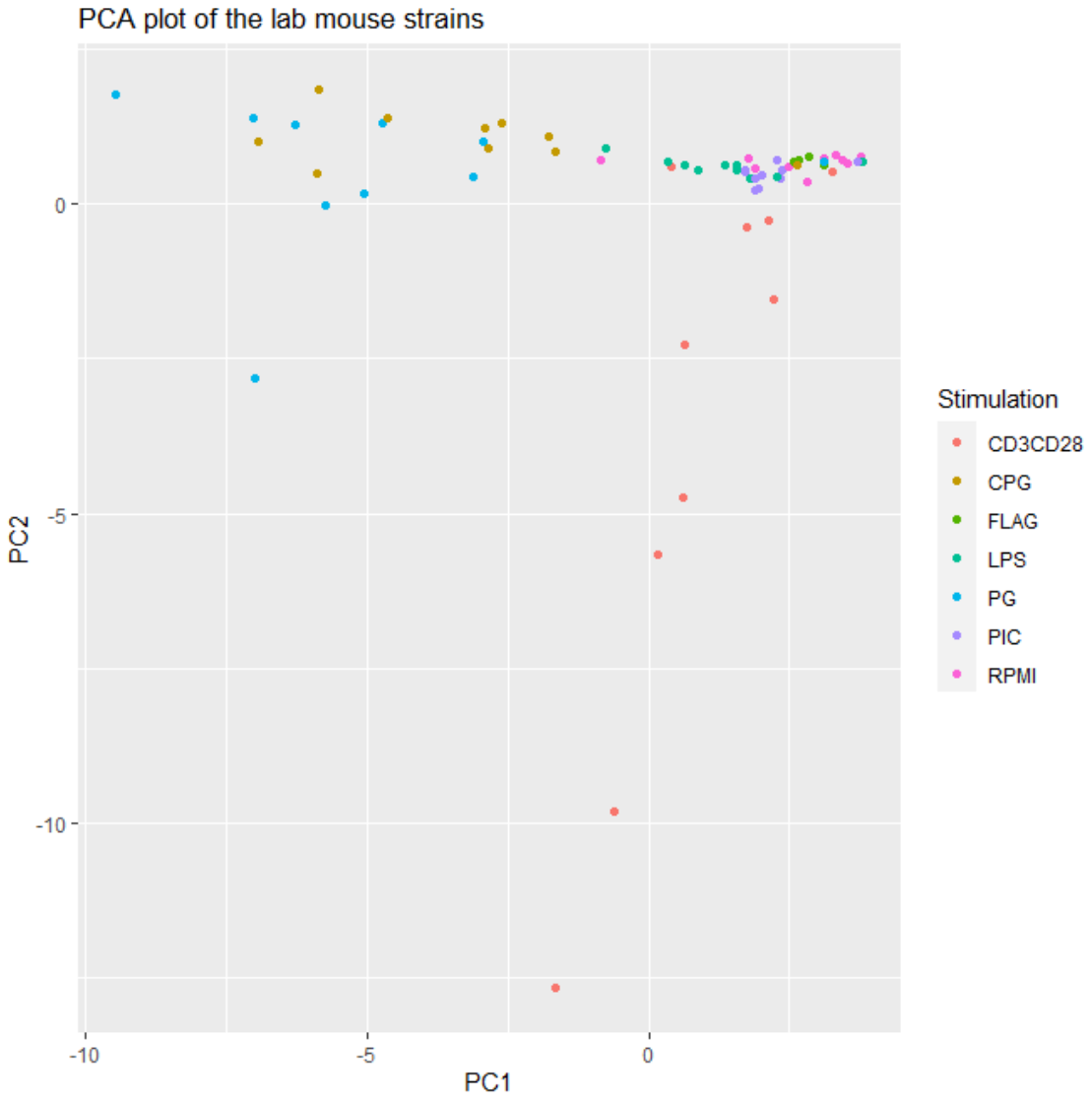

Figure 5. PCA plot of the lab mouse strain illustrating the color-coded distribution of data. The coordinates of each data point are compiled from the first two principal components. The first principal component (PC1) is plotted on the x-axis, the second principal component (PC2) is plotted on the y-axis; the different pathogenic stimuli are color coded and explained in the legend. The graph was computed using the ggplot package.

The first two principal components of the lab mice subset explained 36,4% (PC1) and 18% (PC2) of the total variance respectively. By plotting the first two principal components of the lab mice subset, the spread of data can be visualized allowing for a more facile interpretation, as illustrated in figure 5. In the PCA plot of the lab mice subset there is also a clear separation of the CD3CD28 group, a partial separation between the CPG group and the PG groups, well as some clustering of LPS group; the FLAG, PIC and RPMI groups are intermingled with no clear separation of the groups.

Multiple Analysis of Variance

After performing a MANOVA statistical test of the subset of wild mouse strains, using the Stimulation variable as a dependent variable and the Cytokines as independent variables; a statistically significant difference ($p < 0.05$) of p= 2.2e-16 (table 7) was identified between the different types of Stimulation.

Table 7. Results of the MANOVA function on the wild mice subset.

Term	Df.	Pillai	approx F	num Df	den Df	p-value
Stimulation	6	2.94	4.95	192	990	2.2e-16
Residuals	191					

The full output of the MANOVA analysis of the wild mice subset is collated in table 8.

Table 8. MANOVA input variables and residuals from the wild mice subset.

Terms	Stimulation	Residuals
Basic-FGF	12.19	184.81
Eotaxin	79.13	117.87
GCSF	70.25	126.75
GMCSP	82.56	114.44
IFN-γ	26.69	170.31
IL-1	56.32	140.68
IL-10	100.02	96.98
IL-12p40	79.27	117.73
IL-12p70	99.7	97.3
IL-13	97.44	99.56

IL-15	84.77	112.23
IL-17a	43.03	153.97
IL-18	17.53	179.47
IL-1a	88.36	108.64
IL-2	114.85	82.15
IL-3	80.39	116.61
IL-4	91.63	105.37
IL-5	34.95	162.05
IL-6	79.12	117.88
IL-9	25.23	171.77
KC	93.66	103.34
LIF	58.37	138.63
MCP-1	110.49	86.51
MCSF	39.85	157.15
MIG	48.67	148.33
MIP- 2a	87.72	109.28
MIP-1	100.28	96.72
MIP-1a	108.62	88.38
PDGF-BB	86.73	110.27
RANTES	51.57	145.43
TNFa	34.09	162.91
VEGF	77.25	119.75
Deg. of Freedom	6	191

In order to identify where there is a statistically significant difference between the different Stimulations, a series of univariate ANOVAs were performed as a post-hoc analysis. The statistically significant differences have been collated in the second column of table 9. The majority of the tested variables were determined as statistically significant in the wild mice subset with Basic-FGF being the only variable not determined as statistically significant.

Table 9. The collated p-values of the multiple ANOVA post-hocs; in which the first column contains the individual explanatory variables, the second column contains the p-value of the ANOVA post-hoc of the Wild mice subset, the third column contains the p-value of the ANOVA post-hoc of the lab mice subset (Significance level codes: >0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1).

ANOVA post-hoc explanatory variable	Wild mice subset (p-values)	Lab mice subset (p-values)
Basic-FGF	5.51e-02 .	1.51E-01
Eotaxin	2.20e-16 ***	2.09e-08 ***
GCSF	3.18e-16 ***	6.13e-06 ***
GMCSF	2.20e-16 ***	5.63e-09 ***
IFN-γ	9.02e-05 ***	6.87e-02 .
IL-10	2.20e-16 ***	1.76e-08 ***
IL-12p40	2.20e-16 ***	4.67e-08 ***
IL-12p70	2.20e-16 ***	1.14e-09 ***
IL-13	2.20e-16 ***	2.13e-12 ***
IL-15	2.20e-16 ***	1.01e-05 ***
IL-17a	1.45e-08 ***	4.85e-04 ***
IL-18	6.27e-03 **	8.51E-01
IL-1α	2.20e-16 ***	3.37e-10 ***
IL-1β	4.40e-12 ***	1.36e-09 ***
IL-2	2.20e-16 ***	1.79e-15 ***
IL-3	2.20e-16 ***	1.61e-06 ***
IL-4	2.20e-16 ***	1.17e-08 ***
IL-5	1.29e-06 ***	1.51E-01
IL-6	2.20e-16 ***	1.96e-09 ***
IL-9	1.84e-04 ***	2.55e-03 **
KC	2.20e-16 ***	7.49e-14 ***
LIF	1.15e-12 ***	2.93e-05 ***
MCP-1	2.20e-16 ***	2.04e-06 ***
MCSF	8.83e-08 ***	2.04e-03 ***
MIG	5.21e-10 ***	1.56e-02 *
MIP-1a	2.20e-16 ***	4.86e-10 ***
MIP-1b	2.20e-16 ***	6.82e-10 ***
MIP-2α	2.20e-16 ***	7.02e-11 ***
PDGF-BB	2.20e-16 ***	6.35e-04 ***
RANTES	8.85e-11 ***	3.45e-03 **
TNFα	2.05e-06 ***	5.89e-03 **
VEGF	2.20e-16 ***	1.43e-09 ***

In the MANOVA output of the laboratory mouse strains, a statistically significant difference of p= 1.708e-11 (table 10) was identified between the different types of stimulation. The identified statistically the significant differences are illustrated in column three of table 9. Similar to the wild mice subset, in the lab mice subset, many of the variables were deemed statistically significant, excepting IL-5, IFN-γ, IL-18 and Basic-FGF. The variables that have p-values lower that 0.05 could be excluded from further analysis, in order to reduce the dimensionality of the data.

Table 10. Results of the MANOVA function on the lab mice subset.

Term	Df.	Pillai	approx F	num Df	den Df	p-value
Stimulation	6	4.41	2.68	192	186	1.708e-11
Residuals	57					

The full output of the MANOVA analysis of the lab mice subset is collated in table 11.

Table 11. MANOVA input variables and residuals from the lab mice subset.

Terms	Stimulation	Residuals
Basic-FGF	9.31	53.69
Eotaxin	34.54	28.46
GCSF	27.76	35.24
GMCSP	35.89	27.11
IFN-γ	11.35	51.65
IL-1	37.27	25.73
IL-10	34.72	28.28
IL-12p40	33.68	29.32
IL-12p70	37.44	25.56
IL-13	42.67	20.33
IL-15	27.08	35.92
IL-17a	21.21	41.79
IL-18	2.77	60.23
IL-1a	38.56	24.44
IL-2	47.25	15.75
IL-3	29.51	33.49
IL-4	35.15	27.85
IL-5	9.32	53.68
IL-6	36.92	26.08
IL-9	18.28	44.72
KC	44.99	18.01
LIF	25.58	37.42
MCP-1	29.21	33.79
MCSF	18.69	44.31
MIG	14.71	48.29
MIP- 2a	39.92	23.08
MIP-1	37.92	25.08
MIP-1a	38.22	24.78
PDGF-BB	20.75	42.25
RANTES	17.72	45.28
TNFa	16.69	46.31
VEGF	37.23	25.77
Deg. of Freedom	6	57

Linear Discriminant Analysis of the wild mice subset

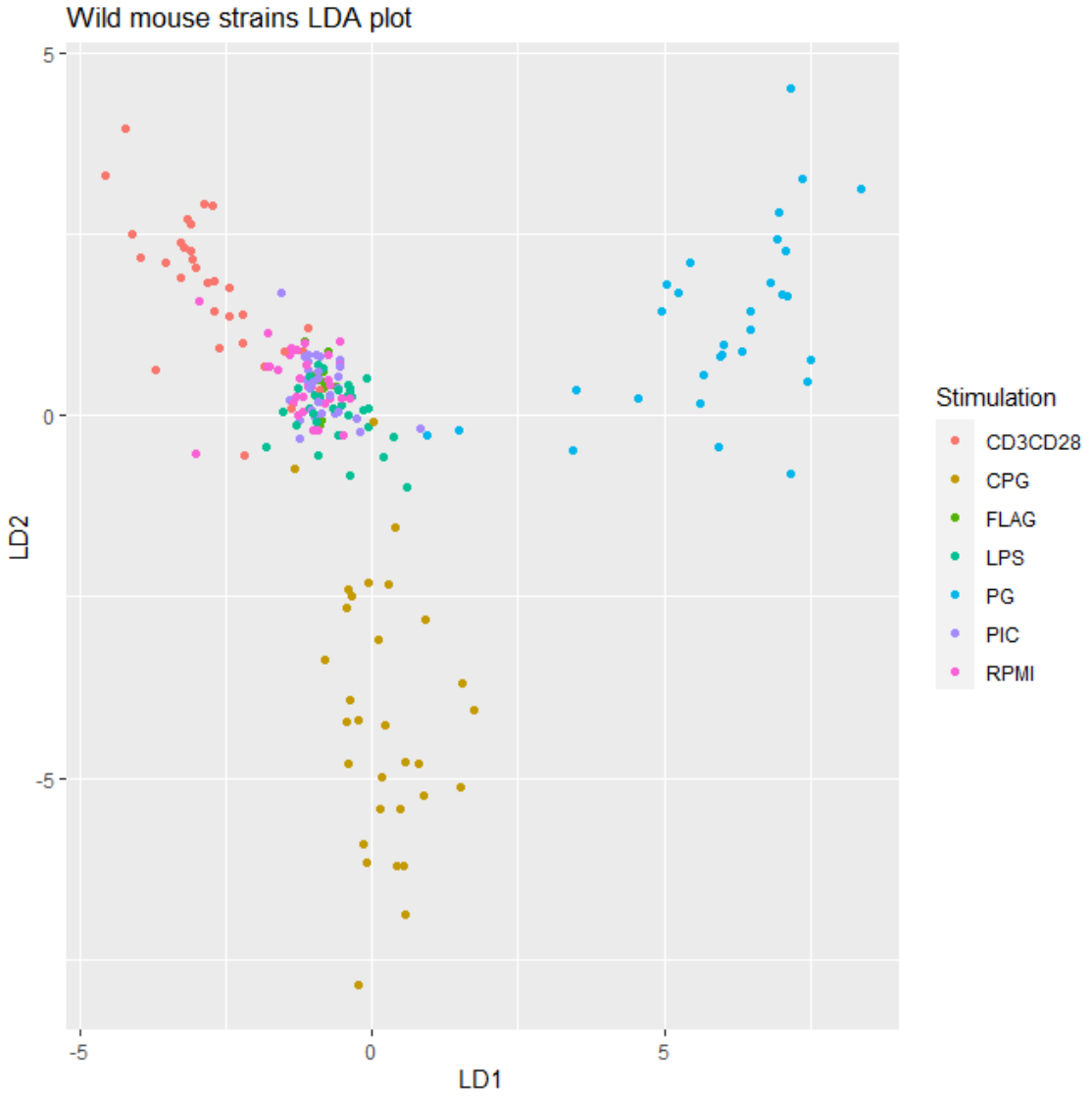

Figure 6. LDA plot of the wild mouse strains illustrating the color-coded distribution of data. The coordinates of each data point are compiled from the first two linear discriminants. The first linear discriminant (LD1) is plotted on the x-axis, the second linear discriminant (LD2) is plotted on the y-axis; the different pathogenic stimuli are color coded and explained in the legend. The graph was computed using the ggplot package.

Linear discriminant analysis attempts to maximize the between-class variance while minimizing the within-class variance, or in other words enable separation of classes while maintaining the highest level of variation possible inside these classes (Martinez & Kak, 2001). The linear discriminant analysis of the wild mice subset computed six linear discriminants that reduce the dimensionality of the data. The coefficients assigned to each independent variable are indicative of their importance and direction in the separation of classes. The first three linear discriminants account for 93.8% of the total explained between-group variance; with LD1 accounting for 49.2%, LD2 for 25.5% and LD3 for 19.1% of explained between-group variance (computed in R with the code block Appendix, #Section 4).

Similar to PCA, the coefficients of the linear discriminants, the numeric value gives information about the importance of a component, while the sign indicates the relationship with the linear

discriminant. A positive coefficient for a variable indicates that an increase in the value of that variable contributes to moving the data point towards one class, while a negative coefficient suggests that an increase in the value of the variable moves the data point towards the opposite class.

In this case, using an arbitrary cutoff value of 1.0, for LD1, the following variables have the highest importance in the linear discriminant: IL-13, MCP-1, MIP-1a, IL-15, PDGF-BB with positive coefficients, while IL-3 and IL-4 have negative coefficients. In LD2, the highest importance variables are: IL-10 with a positive coefficient, IL-3, IL-12p70 and MIP-1a with a negative coefficient.

The plot generated using the first two linear discriminants of the wild mice subset provide an overview of the data distribution, as illustrated in figure 6. The plot displays quite clear clustering for the CD3CD28, CPG and PG groups, with diffuse overlapping for the LPS, PIC, FLAG and RPMI groups.

Linear Discriminant Analysis of the lab mice subset

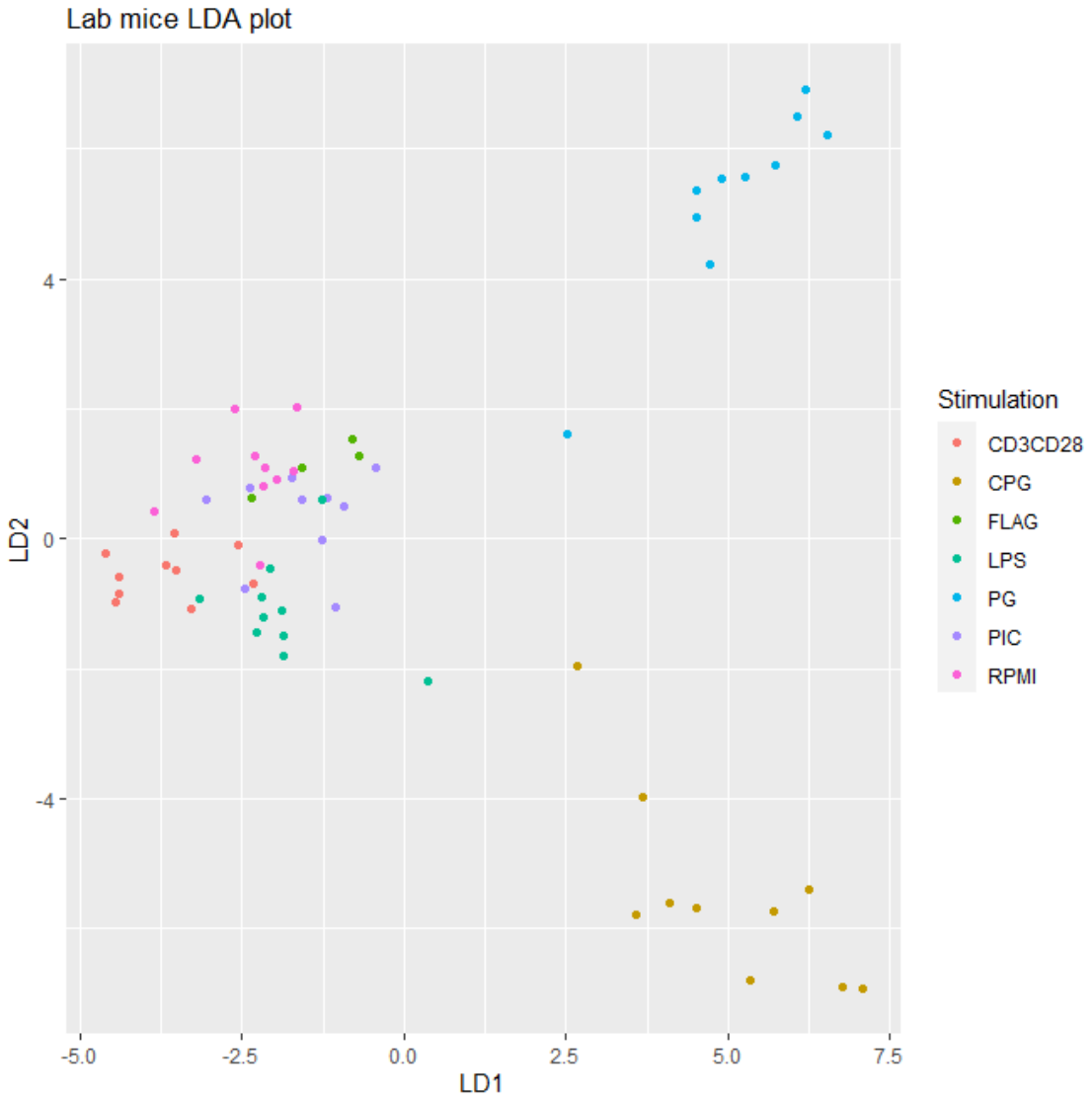

Figure 7. LDA plot of the lab mouse strain, illustrating the color-coded distribution of data. The coordinates of each data point are compiled from the first two linear discriminants. The first linear discriminant (LD1)

is plotted on the x-axis, the second linear discriminant (LD2) is plotted on the y-axis; the different pathogenic stimuli are color coded and explained in the legend. The graph was computed using the ggplot package.

In the case of the lab mice subset, six linear discriminants were generated. The first three linear discriminants account for 95.2% of the total explained between-group variance; with LD1 accounting for 36.2%, LD2 for 28.8%, LD3 for 17.3% and LD4 for 12.9% of explained between-group variance (computed in R with the code block Appendix, #Section 4).

With the same arbitrary coefficient of 1.0, LD1 of the lab mice subset is highly influenced by: IL-3, GMCSF, MCP-1, MIP-1a, IL-15 and MIP-2a with positive coefficients; IL-2, IL-10, IL12p40, IL-13, GCSF, KC, MIP-1b, RANTES, MCSF and PDGF-BB with negative coefficients. In the case of LD2, the most influencing variables are:IL-6, IL-13, IL-17a, MCP-1, MIP-1b, TNFa, IL-15 and MIP-2a with positive coefficients; IL-1, IL-3, IL-5, IL-12p40, IL-12p70, MIG and VEGF with a negative coefficient (computed in R with the code block Appendix, #Section 4).

The plot generated using the first two linear discriminants of the lab mice subset provide an overview of the data distribution, as illustrated in figure 7. In contrast with the LDA plot for the wild mice subset, the LDA plot of the lab mice subset display clearer delimitation of groups for all of the groups. The most separated grouping of data points is for the CPG and PG stimulations.

Model construction and cross-validation

Thanks to the caret package, multiple iterations of the model with different tweaks to their relevant parameters are generated and compared to each other.

The accuracy of the constructed models, using the wild mice subset, was evaluated accounting for optimal model setting using the caret package. For the purpose of conciseness, the accuracy metrics from each model were compiled and displayed in a table in order to enable easy visualization. Each of the models with the different cross-validation methods involved have been computed in R, using the code blocks from Appendix, I, #Sections 5 – 11.

Table 12. Accuracy ratings of the evaluated classifiers for the wild mice subset. In which the values are represented as decimals (1.0 = 100%).

Classifier	LOOCV	Repeated k-fold	Bootstrap	Mean Accuracy
LDA	0.67	0.66	0.62	0.65
KNN	0.58	0.59	0.55	0.57
NB	0.60	0.63	0.57	0.60
RF	0.75	0.75	0.71	0.74
SVM	0.71	0.71	0.69	0.70
GLM	0.72	0.72	0.69	0.71
ANN	0.74	0.75	0.71	0.73

The highest accuracy rating of each cross-validation method of the wild mice subset models is displayed in table 12. The most accurate classification algorithm for the wild mice subset is Random Forest with an accuracy of 74%; followed by the Artificial Neural Network with 73% accuracy and General Linearized Model with 71% accuracy.

Table 13. Accuracy ratings of the evaluated classifiers for the lab mice subset. In which the values are represented as decimals (1.0 = 100%).

Classifier	LOOCV	Repeated k-fold	Bootstrap	Mean Accuracy
LDA	0.67	0.63	0.44	0.58
KNN	0.58	0.62	0.48	0.56
NB	0.59	0.63	0.50	0.57
RF	0.83	0.84	0.76	0.81
SVM	0.72	0.73	0.61	0.69
GLM	0.81	0.80	0.69	0.77
ANN	0.75	0.75	0.71	0.74

The accuracy of the constructed models, using the lab mice subset, was evaluated accounting for optimal model setting using the caret package. The highest accuracy rating of each cross-validation method is displayed in table 13. The most accurate classification algorithm for the lab mice subset is Random Forest with an accuracy of 81%; followed by General Linearized Model with 77% accuracy and the Artificial Neural Network with 74% accuracy.

Discussion

Exploratory analysis

Principal component analysis

When investigating the loadings of the first two principal components (PC1 and PC2), the same variables were identified as significant in both the wild and the lab mice. For PC1: IL-1a, IL-1b, IL-3, IL-10, IL-12p70, IL-13, Eotaxin, GMCSF, MIP-1a and MIP-1b; for PC2: IL-2, IL-4, IL-5, IL-17a, INF-γ, LIF and MIG (table 3; table 6). The loading of IL-12p40 in wild mice (table 3) was excluded due to the set cutoff, similarly to the MCP-1 loading from the lab mice (table 6). This is why PCA is considered an exploratory analysis because it allows scientists to make decisions based on the distribution of data. Cutoff parameters can be used to reduce complexity of data and/or eliminate potential outliers (Jolliffe & Cadima, 2016). What this means experimentally is that the particular variables that have been identified as important can be used to a similar effect as using all the variables. Depending on the experimental set-up this can reduce overhead costs or computational requirements (Ringnér, 2008). The PCA plot of the wild mice subset illustrated clear clustering of datapoints for the CD3CD28, and a partial clustering between the PG and CPG groups (figure 3). Similarly, the PCA plot of the lab mice subset illustrated clear clustering of datapoints for the CD3CD28, a partial clustering between the PG and CPG groups, as well as some clustering of the LPS group (figure 5).

Multiple analysis of variance

The MANOVA of the wild mice set determined that all of the variables, excepting Basic-FGF are statistically significant (table 9.) in determining group differences. In the case of the lab mice, a vast majority of the variables were deemed statistically significant (table 9.), excepting IL-5, IFN-γ, IL-18 and Basic-FGF. Variables that are not statistically significant could be discarded and maintain the same integrity of the original data. The expanded output of MANOVA (table 8; table 11) indicates the coefficient of each independent variable (cytokine) representing the average shift in the dependent variable (stimulation) based on a one-unit change in a specific independent

variable whilst all other independent variables are maintained constant. The residuals quantify the difference between observed and expected values for individual observations in the analysis. They represent unexplained variation in the dependent variables that is not accounted by the effects of the independent variable (Grice et al., 2009).

That is why MANOVA is used as a dimensionality reduction method in multivariate analysis. Some researchers, claim that using MANOVA for dimensionality reduction can potentially have negative impacts on the quality of the data (Everitt & Torsten, 2011; Tabachnick & Fidell, 2019).

Linear discriminant analysis

In the LDA plot of the wild mice subset (figure 6), the clear clustering of datapoints for the CPG, PG and CD3CD28 groups is illustrated. In contrast with the LDA plot for the wild mice subset, the LDA plot of the lab mice subset (figure 7) displays better delimitation of groups for all of the groups. The most clearly separated grouping of data points is for the CPG and PG stimulations.

When investigating the coefficients of the linear discriminants, and comparing individual linear discriminants of the wild mice subset with their counterparts in the lab mice subset, numerous differences were observed, both in the magnitude of the coefficients and their sign.

In the case of both PCA and LDA, clear grouping of data points for the CPG, PG and CD3CD28 stimulations were observed. This computational outcome could potentially imply significant ramifications about the biology of immune response. Although both PCA and LDA indicate a similar pattern in the data, LDA visualization provides a clearer separation of stimulation groups.

Why is it that many researchers use both PCA and LDA? Both procedures have their pros and cons (Martinez & Kak, 2001; Choubey et al., 2020); PCA is a type of unsupervised machine learning, but it does not take into consideration class labels. Although PCA reduces data dimensionality, the resulting principal components can be difficult to interpret as they are linear combinations of the original features (variables). Because PCA focuses on containing the maximum variation of the data, the between groups variation may not align with the total variation in the data (Zamora Saiz et al., 2020).

LDA is a type of supervised machine learning, that takes into consideration class labels and thus maximizing class separation, but it is sensitive to class imbalance, assuming a multivariate normal class distribution and classes with similar covariance matrices (Park & Park, 2008; Zhao et al., 2020). The dimensionality reduction in LDA maximizes variation between groups while minimizing variation within groups, in case in which the number of samples is low in comparison with the number of variables, LDA can be prone to overfitting (over-representing certain data). LDA can allow the identification of important variables for class separation, but is limited to linear separation (Hastie et al., 2009).

Taking in to consideration the aforementioned pros and cons of PCA and LDA it is important to investigate the results of both procedures. The graphical visualization of both LDA and PCA could be potentially improved by adding the third linear discriminant (LD3)/ principal component (PC3) on a z-axis in order to create a three-dimensional model.

In a similar study conducted by Fonseca dos Reis et al. (2021), using the same public data set provided by Abolins et. al, PCA did not indicate any significant patterns. In their study, Fonseca dos Reis et al., use 120 immune measurements including: antibodies, serum proteins, spleen measurements, mean fluorescence index measurements of immune cells and fluorescence

activated cell sorted (FACS)-measurements of immune cells. Their research suggests that network analysis is significantly better than PCA for observing patterns in the data, but suggest that network analysis should be used in conjunction with other statistical procedures. In their research, Fonseca dos Reis et al., have excluded numerous data points from the analysis due to missing values. When confronted with the same challenges, the research presented in this paper followed the same procedure as Abolins et al., which replaced out of range values with a nominal value of 0.001, allowing for the retention of a significant number of data points which have been discarded by Fonseca dos Reis et al. As limitations of their study, Fonseca dos Reis et al., also quote the sample size of the data set and comment on the impact of additional data on the potential increase in the accuracy of the findings. Indeed, further data could substantially increase the quality of findings; either by corroborating previous analysis or determining a different outcome than what was previously expected.

Biological context

But why is it that these priming stimuli (CPG, PG and CD3CD28) elicited such distinct responses in contrast with the others (LPS, PIC, FLAG and RPMI)?

In experiments involving priming with substances like LPS, PIC, PG, FLAG, CD3CD28, or CPG, a control group is typically included to compare the effects of the priming agent. The control group is typically treated with a vehicle control or a mock treatment that mimics the experimental conditions but lacks the active component being tested. This control group allows researchers to assess the specific effects of the priming agent by comparing it to the baseline response observed in the control group.

Different methods of exploratory analysis have identified clustering of date points especially for the CPG, CD3CD28 and PG groups.

In the case of the data set provided by Abolins et al., both the wild mice and the lab mice responded to CPG, PG and CD3CD28 priming more consistently than any of the other stimulations. Because the wild mice have a burden of exposure, vast genetic diversity and varying environmental exposure; in contrast with the lab mice with very low genetic diversity, no burden of exposure and stable environmental conditions and the aforementioned stimulations have had such a profound effect on cytokine production, the observation can be made that it is likely that such stimuli are less likely to naturally occur or that rodents in mice in general have a high susceptibility to these particular stimuli. The diffuse pattern of the other priming conditions (LPS, PIC, FLAG and RPMI) could potentially mean that these stimuli do not elicit a particularly strong response due to the intensity of the stimuli or a high degree of similarity between the stimuli.

Previously conducted research, has determined that CPG is detected by TLR9 and triggers a signaling cascade that leads to a release of IL-6 and IL-10 (Latz et al., 2004), IL-12, TNFa and type I interferons such as: INF-α and INF-β (Zhu et al., 2009). In the case CD3CD28, specific components of T cell receptors are activated by CD3 molecules and co-stimulated by CD28 which after triggering a signaling cascade leads to the release of IL-2, IFN-γ, TNFa, IL-4 and IL-10 (Wang et al., 2004; Weiss & Littman, 1994). PG affects the TLR2 receptor triggers the release of IL-1β, TNFa, IL-6, IL-10 and IFN-γ (Irazoki et al., 2019; Keestra-Gounder & Tsolis, 2017).

LPS priming affects TLR4 and can influence the production of TNF-α, IL-1β, IL-6, IL-10, and IFN-γ (Page et al., 2022). In the case of flagellin, the TLR5 receptor is affected and priming triggers the release of IL-1β, TNF-α, IL-6, IL-10, and IFN-γ (Hajam et al., 2017). PIC is detected by TLR3 and

priming leads to the release of IL-1β, TNF-α, IL-6, IL-10, and IFN-α (He et al., 2021). RPMI can potentially influence cytokine production in certain cells but studies are not conclusive. Some researchers (Daigneault et al., 2010; Kaur & Dufour, 2012; Klein & Flanagan, 2016) have determined that using RPMI as a culture medium has an impact on the production of certain cytokines such as: IL-1β, IL-6, IL-10, and TNF-α.

It is very interesting that previous research concluded that stimulation with RPMI, PIC, LPS and FLAG has a potent influence on the production of IL-1β, IL-6, IL-10, and TNF-α. This seems to validate the results of the current exploratory analysis on the dataset provided by Abolins et al.

Model construction and cross-validation

In the case of both the wild mice subset and the lab mice subset, the most accurate model (table 12; table 13) was Random Forest (74% and 81% respectively). In the case of the wild mice subset (table 12), the second most accurate model was Artificial Neural Network (73%), with Generalized Linear Model as the third (71%). The second most accurate model of the lab mice subset (table 13) was the Generalized Linear Model (77%), followed by the Artificial Neural Network as third (74%).

There are numerous papers claiming that a certain classification algorithm is better than another, but in the vast majority of cases, it depends on the data used to train the model and the tunning parameters of the model. That is why generally multiple models are trained and evaluated of individual data sets. Generally, SVM, ANN, Random Forest, Naïve Bayes and more recently Gradient Boosting Methods are considered the most robust and accurate machine learning classifiers available (Wehrens, 2020).

Using the *caret* package to build and test models for classification and regression problems allows for a streamlined and time-saving procedure. The default settings when employing a model are based on previously confirmed optimal parameters with are programed into the package, any of the setting available for the individual algorithm can also be tunned in the caret version. Because the package contains many built in cross-validation methods it is easy to determine the accuracy of the model by selecting and optimizing cross-validation techniques (Kuhn, 2008). The difference in accuracy between the different cross-validation methods is due to the mechanism they employ to split the data in to training and test data. Using an average of different cross-validation methods should provide a robust overall accuracy. It is important to test classification models in order to prepare for incoming data. In the case in which the class of a sample is not known, using such a model should provide the real class of the sample (LeCun et al., 2015).

Scientific context

Wild mice and lab mice differ significantly when used as disease models; here are some key differences between the two genetic diversity wild mice exhibit much higher genetic diversity compared to lab mice (Latham & Mason, 2004). This diversity can influence disease susceptibility, response to treatments and the progression of diseases. Lab mice on the other hand are typically bred from a limited number of inbred strains which reduces genetic variability and environmental exposure (Svenson et al., 2012). Wild mice are exposed to a wide range of environmental factors such as pathogens parasites and varying diets which can impact disease development and progression. Lab mice are reared in controlled environments with standardized diets and minimal exposure to pathogens ensuring consistent and controlled experimental conditions (Speakman et al., 2007). In the context of disease prevalence; wild mice may encounter a different set of diseases

compared to their lab counterparts. The exposure to naturally occurring infections and diseases that are not typically encountered in lab mice can be both an advantage and a limitation depending on the specific research goals (Latham & Mason, 2004). In terms of reproducibility, lab mice offer a higher level of reproducibility in experiments due to controlled breeding, standardized housing conditions and reduced genetic variability that enables an easier comparison of results across different studies. Wild mice can exhibit greater variability in disease outcomes making it more challenging to achieve consistent results. Considering practicality, lab mice are easier to handle, manipulate and breed in controlled settings. Lab mice have well characterized genetic backgrounds, known health histories and are readily available from established breeding colonies (Casellas, 2011). Summarizing, wild mice offer a more diverse genetic background and exposure to natural pathogens which may be advantageous for studying certain diseases or environmental interactions; however, lab mice provide greater control, reproducibility and practicality for disease modeling experiments, making them the preferred choice for most research studies (Svenson et al., 2012).

Conclusion

It is necessary to use multiple types of exploratory analyses as all of them have their strengths and shortcomings. In this thesis, both PCA and LDA managed to identify the same clearly defined groups (CPG, CD3CD28 and PG), while the other stimulations (LPS, FLAG, PIC) and the control (RPMI) exhibited diffuse clustering. These findings seem to be validated by previously conducted research but might require further analysis to confirm. Looking at the coefficients from both PCA and LDA seems to indicate that the vast majority of the variables have an impact on the reduced components (PCs and LDs), but certain components have a higher influence than others and further post-processing might be required. The MANOVA finding also indicate that the majority of the variables are statistically significant in determining group separation, with the notable exception of Basic-FGF. Out of the seven classification models tested, the Random Forest model exhibited the highest level of accuracy.

Ethical aspects

In terms of ethical implications, this project is exempt of scrutiny due to the purely analytical nature of the research and the lack of contact with samples. In terms of the dataset used, provided as a community resource by Abolins et al., the study was conducted with the approval of the University of Bristol's Animal Welfare and Ethical Review Board.

Using wild mice for research purposes raises ethical concerns due to potential harm inflicted on wild populations. Lab mice, on the other hand, are bred specifically for research purposes and their use is subject to ethical guidelines and regulations.

The *C57BL*/6 strain of pathogen-free laboratory mice were purchased from Charles River Laboratories (Margate, UK), Harlan (UK) or B&K Universal (UK) and were accommodated into groups of five same-sex individuals and were subjected to a 12h light/12 dark diurnal cycle within an enriched environment with a typical rodent diet (Keenan, 1999). The wild mice were obtained by trapping in various sites in the southern United Kingdom. The mice were trapped using Longworth traps using a bait of mixed oats and carrots, overnight followed by daily diurnal inspection of the traps. The captures rodents were shuttled to a housing facility and reared under identical conditions to their laboratory counterparts, the only exception being individual containment. The wild rodents were captured in the period March 2012 to April 2014.

In terms of sex and gender perspectives, this study was gender blind. A mixed pool of male (144 observations) and female (118 observations) mice cytokine profiles were studied together. Further analysis could be performed on the dataset in order to investigate gender differences by applying a filter ($Sex) to separate the data set into male ("M") and female ("F") individuals.

The providers of the dataset (Abolins et al.), declare no competing financial interest and were comprised of a team of both male and female researchers with no declared gender identities.

The future implications of this research are confirming and contributing to designing better research frameworks for immunological studies using mouse animal models. The importance of realistic modelling of human immunological processes in animal models is paramount for improving scientific acumen of diseases as well as the basis for therapeutic procedures.

Future perspectives

In further analysis, the aforementioned stimulations (CPG, CD3CD28 and PG) could be hidden from the data set in order to observe how the other groups are represented in their absence. Also taking into consideration that many of the different stimulations affect IL-1β, IL-6, IL-10, and TNF-α production, either removing these variables or focusing solely on them should elucidate if they are important for separating the groups. Another direction altogether, could be evaluating different stimulations in the same experimental conditions.

From an experimental stand point, it would be interesting to determine how humanized mice would fare under the same priming conditions. Understanding how the same stimuli can affect different animal models, such as wild mice, lab mice and humanized mice would be interesting to observe from the different stand points such as: efficiency of representation, ease of maintenance and of course cost.

The use of humanized mice under the same experimental conditions should provide a clearer comparison between the usefulness of either wild mice or lab mice as disease models. The cost of humanized mice is significantly higher and time consuming than that any other available options (Theocharides et al., 2015). Humanized mice are a type of laboratory mice that has been genetically modified or implanted with human cells, tissues, or genes in order to mimic certain aspects of the human immune system or physiology. Creating humanized mice enables more accurate models for the study of human diseases and testing potential treatments or therapies (Brehm et al., 2014). The process of creating such humanized mice involves one or a combination of procedures, such as: injecting human immune cells into the mouse, implanting human tissues or organs, or introducing specific human genes into the mouse genome. Inserting these modifications in mice models allow the mice to exhibit characteristics or responses that are more similar to those observed in humans (Fujiwara, 2017).

From a purely analytical point of view, it would be interesting to know which animal model is the most representative of human immunology and cellular biology, but from a practical point of view, such a comprehensive analysis could be considered overtaxing and thus frivolous. Nevertheless, is cost reduction could be achieved, the increase in accuracy of representation could be priceless in understanding, treating and preventing diseases in humans.

Acknowledgments

I take this moment to express my deepest gratitude and appreciation to each and every one of you for your significant roles in the completion of my master's thesis. Your unwavering support, encouragement, and invaluable contributions have been instrumental in making this journey a resounding success.

To my beloved parents, Walid and Carmen, I am forever indebted to you for your unconditional love, unwavering belief in me, and constant encouragement throughout my academic pursuit. Your unwavering support and sacrifices have been the driving force behind my achievements. I am immensely grateful for your guidance, wisdom, and unwavering belief in my abilities. Without you, I would not have reached this milestone.

To my partner Alicia, I extend my heartfelt thanks for your unwavering support and understanding throughout the entire duration of my research. Your patience, insightful suggestions, and continuous encouragement have been invaluable. Your unwavering belief in my abilities and your constant motivation have played a vital role in keeping me focused and determined to overcome any obstacles I encountered. I am truly fortunate to have you by my side.

To my dear son, Sebastian, your presence in my life has been a constant source of inspiration and motivation. Your innocent smile and unwavering faith in me have driven me to push my limits and strive for excellence. I deeply appreciate your understanding and patience during those long hours of studying. You are a remarkable source of joy, and I am grateful for your unwavering support.

I would also like to express my gratitude to my professors and mentors for their guidance, expertise, and invaluable insights throughout this research journey. Your knowledge, constructive feedback, and willingness to share your expertise have been indispensable in shaping the outcome of this thesis. Your dedication to education and commitment to my growth have made a profound impact on my academic and personal development.

In conclusion, this thesis would not have been possible without the unwavering belief, support, and inspiration provided by each and every one of you. Your contributions and encouragement have been the cornerstone of my success. I am forever grateful for your presence in my life and for the role you have played in this significant achievement.

References

Abolins, S., King, E. C., Lazarou, L., Weldon, L., Hughes, L., Drescher, P., Raynes, J. G., Hafalla, J. C. R., Viney, M. E., & Riley, E. M. (2017). The comparative immunology of wild and laboratory mice, Mus musculus domesticus. *Nature Communications*, *8*(1). https://doi.org/10.1038/ncomms14811

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bersch, K. L., DeMeester, K. E., Zagani, R., Chen, S., Wodzanowski, K. A., Liu, S., Mashayekh, S., Reinecker, H. C., & Grimes, C. L. (2021). Bacterial Peptidoglycan Fragments Differentially Regulate Innate Immune Signaling. *ACS Central Science*, *7*(4), 688–696. https://doi.org/10.1021/acscentsci.1c00200

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Bogue, M. A., Philip, V. M., Walton, D. O., Grubb, S. C., Dunn, M. H., Kolishovski, G., Emerson, J., Mukherjee, G., Stearns, T., He, H., Sinha, V., Kadakkuzha, B., Kunde-Ramamoorthy, G., & Chesler, E. J. (2019). Mouse Phenome Database: a data repository and analysis suite for curated primary mouse phenotype data. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz1032

Brehm, M. A., Wiles, M. V., Greiner, D. L., & Shultz, L. D. (2014). Generation of improved humanized mouse models for human infectious diseases. *Journal of Immunological Methods*, *410*, 3–17. https://doi.org/10.1016/j.jim.2014.02.011

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Bryda, E. C. (2013). The Mighty Mouse: the impact of rodents on advances in biomedical research. *Missouri Medicine*, *110*(3), 207–211. ncbi. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3987984/

Cannon, J. G. (2000). Inflammatory Cytokines in Nonpathological States. *Physiology*, *15*(6), 298–303. https://doi.org/10.1152/physiologyonline.2000.15.6.298

Casellas, J. (2011). Inbred mouse strains and genetic stability: a review. *Animal*, *5*(1), 1–7. https://doi.org/10.1017/s1751731110001667

Cha Zhang, & Ma, Y. (2012). *Ensemble Machine Learning*. Springer.

Chang, P. L., Kopania, E., Keeble, S., Sarver, B. A. J., Larson, E., Orth, A., Belkhir, K., Boursot, P., Bonhomme, F., Good, J. M., & Dean, M. D. (2017). Whole exome sequencing of wild-derived inbred strains of mice improves power to link phenotype and genotype. *Mammalian Genome*, *28*(9-10), 416–425. https://doi.org/10.1007/s00335-017-9704-9

Charo, I. F., & Ransohoff, R. M. (2006). The Many Roles of Chemokines and Chemokine Receptors in Inflammation. *New England Journal of Medicine*, *354*(6), 610–621. https://doi.org/10.1056/nejmra052723

Cheng, H., Garrick, D. J., & Fernando, R. L. (2017). Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. *Journal of Animal Science and Biotechnology*, *8*(1). https://doi.org/10.1186/s40104-017-0164-6

Choubey, D. K., Kumar, M., Shukla, V., Tripathi, S., & Dhandhania, V. K. (2020). Comparative Analysis of Classification Methods with PCA and LDA for Diabetes. *Current Diabetes Reviews*, *16*(8), 833–850. https://doi.org/10.2174/1573399816666200123124008

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/tit.1967.1053964

Crusio, W. E., Sluyter, F., Gerlai, R. T., & Pietropaolo, S. (2013). *Behavioral genetics of the mouse*. Cambridge University Press.

Daffertshofer, A., Lamoth, C. J. C., Meijer, O. G., & Beek, P. J. (2004). PCA in studying coordination and variability: a tutorial. *Clinical Biomechanics*, *19*(4), 415–428. https://doi.org/10.1016/j.clinbiomech.2004.01.005

Daigneault, M., Preston, J. A., Marriott, H. M., Whyte, M. K. B., & Dockrell, D. H. (2010). The Identification of Markers of Macrophage Differentiation in PMA-Stimulated THP-1 Cells and Monocyte-Derived Macrophages. *PLoS ONE*, *5*(1), e8668. https://doi.org/10.1371/journal.pone.0008668

Davis, K. M., & Weiser, J. N. (2010). Modifications to the Peptidoglycan Backbone Help Bacteria To Establish Infection. *Infection and Immunity*, *79*(2), 562–570. https://doi.org/10.1128/iai.00651-10

Decoste, D., & Schölkopf, B. (2002). Training Invariant Support Vector Machines. *Machine Learning*, *46*(1/3), 161–190. https://doi.org/10.1023/a:1012454411458

Everitt, B., & Torsten , T. (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer Science & Business Media.

Ferris, K. G., Chavez, A. S., Suzuki, T. A., Beckman, E. J., Phifer-Rixey, M., Bi, K., & Nachman, M. W. (2021). The genomics of rapid climatic adaptation and parallel evolution in North American house mice. *PLOS Genetics*, *17*(4), e1009495. https://doi.org/10.1371/journal.pgen.1009495

Filipovych, R., Resnick, S. M., & Davatzikos, C. (2011). Semi-supervised cluster analysis of imaging data. *NeuroImage*, *54*(3), 2185–2197. https://doi.org/10.1016/j.neuroimage.2010.09.074

Fonseca dos Reis, E., Viney, M., & Masuda, N. (2021). Network analysis of the immune state of mice. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-83139-7

Fortier, M.-E., Kent, S., Ashdown, H., Poole, S., Boksa, P., & Luheshi, G. N. (2004). The viral mimic, polyinosinic:polycytidylic acid, induces fever in rats via an interleukin-1-dependent mechanism. *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology*, *287*(4), R759-66. https://doi.org/10.1152/ajpregu.00293.2004

Fujiwara, S. (2017). Humanized mice: A brief overview on their diverse applications in biomedical research. *Journal of Cellular Physiology*, *233*(4), 2889–2901. https://doi.org/10.1002/jcp.26022

Gentry, R. Thomas. (1989). Nutritional determinants of alcohol intake in c57bl/6j mice. *Appetite*, *12*(1), 71. https://doi.org/10.1016/0195-6663(89)90073-1

Geraldes, A., Basset, P., Gibson, B., Smith, K. L., Harr, B., Yu, H.-T., Bulatova, N., Ziv, Y., & Nachman, M. W. (2008). Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Molecular Ecology*, *17*(24), 5349–5363. https://doi.org/10.1111/j.1365-294x.2008.04005.x

Grice, J. W., & Iwasaki, M. (2009). A Truly Multivariate Approach to Manova. *Applied Multivariate Research*, *12*(3), 199. https://doi.org/10.22329/amr.v12i3.660

Hajam, I. A., Dar, P. A., Shahnawaz, I., Jaume, J. C., & Lee, J. H. (2017). Bacterial flagellin—a potent immunomodulatory agent. *Experimental & Molecular Medicine*, *49*(9), e373–e373. https://doi.org/10.1038/emm.2017.172

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning*. Springer.

Hatai, H., Lepelley, A., Zeng, W., Hayden, M. S., & Ghosh, S. (2016). Toll-Like Receptor 11 (TLR11) Interacts with Flagellin and Profilin through Disparate Mechanisms. *PLOS ONE*, *11*(2), e0148987. https://doi.org/10.1371/journal.pone.0148987

He, Y., Taylor, N., Yao, X., & Bhattacharya, A. (2021). Mouse primary microglia respond differently to LPS and poly(I:C) in vitro. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-89777-1

Hess, A. S., & Hess, J. R. (2018). Principal component analysis. *Transfusion*, *58*(7), 1580–1582. https://doi.org/10.1111/trf.14639

Hess, A. S., & Hess, J. R. (2019). Logistic regression. *Transfusion*. https://doi.org/10.1111/trf.15406

Heylmann, D., Badura, J., Becker, H., Fahrer, J., & Kaina, B. (2018). Sensitivity of CD3/CD28-stimulated versus non-stimulated lymphocytes to ionizing radiation and genotoxic anticancer drugs: key role of ATM in the differential radiation response. *Cell Death & Disease*, *9*(11), 1–17. https://doi.org/10.1038/s41419-018-1095-7

Ho-Wan Kwak. (2010). Precautions in Applying Multivariate Statistics: Analysis of Variance and Multiple Regression Analysis. *Korean Journal of Cognitive and Biological Psychology*, *22*(2), 247–259. https://doi.org/10.22172/cogbio.2010.22.2.008

Irazoki, O., Hernandez, S. B., & Cava, F. (2019). Peptidoglycan Muropeptides: Release, Perception, and Functions as Signaling Molecules. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.00500

Johnson, H. E., Lloyd, A. J., Mur, L. A. J., Smith, A. R., & Causton, D. R. (2007). The application of MANOVA to analyse Arabidopsis thaliana metabolomic data from factorially designed experiments. *Metabolomics*, *3*(4), 517–530. https://doi.org/10.1007/s11306-007-0065-3

Jolliffe, I. T. (1990). PRINCIPAL COMPONENT ANALYSIS: A BEGINNER'S GUIDE - I. Introduction and application. *Weather*, *45*(10), 375–382. https://doi.org/10.1002/j.1477-8696.1990.tb05558.x

Jolliffe, I. T. (1993). Principal component analysis: A beginner's guide - II. Pitfalls, myths and extensions. *Weather*, *48*(8), 246–253. https://doi.org/10.1002/j.1477-8696.1993.tb05899.x

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kariya, T. (1978). The General MANOVA Problem. *The Annals of Statistics*, *6*(1), 200–214. https://doi.org/10.1214/aos/1176344079

Kaur, G., & Dufour, J. M. (2012). Cell Lines. *Spermatogenesis*, *2*(1), 1–5. https://doi.org/10.4161/spmg.19885

Keenan, K. (1999). Diet, caloric restriction, and the rodent bioassay. *Toxicological Sciences*, *52*(2), 24–34. https://doi.org/10.1093/toxsci/52.2.24

Keestra-Gounder, A. M., & Tsolis, R. M. (2017). NOD1 and NOD2: Beyond Peptidoglycan Sensing. *Trends in Immunology*, *38*(10), 758–767. https://doi.org/10.1016/j.it.2017.07.004

Klein, S. L., & Flanagan, K. L. (2016). Sex differences in immune responses. *Nature Reviews Immunology*, *16*(10), 626–638. https://doi.org/10.1038/nri.2016.90

Krogh, A. (2008). What are artificial neural networks? *Nature Biotechnology*, *26*(2), 195–197. https://doi.org/10.1038/nbt1386

Kuby, J., Goldsby, R. A., Kindt, T. J., & Osborne, B. A. (2006). *Kuby Immunology.* W H Freeman.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5). https://doi.org/10.18637/jss.v028.i05

Latham, N., & Mason, G. (2004). From house mouse to mouse house: the behavioural biology of free-living Mus musculus and its implications in the laboratory. *Applied Animal Behaviour Science*, *86*(3), 261–289. https://doi.org/10.1016/j.applanim.2004.02.006

Latz, E., Schoenemeyer, A., Visintin, A., Fitzgerald, K. A., Monks, B. G., Knetter, C. F., Lien, E., Nilsen, N. J., Espevik, T., & Golenbock, D. T. (2004). TLR9 signals after translocating from the ER to CpG DNA in the lysosome. *Nature Immunology*, *5*(2), 190–198. https://doi.org/10.1038/ni1028

Lawal, R. A., Arora, U. P., & Dumont, B. L. (2021). Selection shapes the landscape of functional variation in wild house mice. *BMC Biology*, *19*(1). https://doi.org/10.1186/s12915-021-01165-3

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Leigh, S., & Jackson, J. E. (1993). A User's Guide to Principal Components. *Technometrics*, *35*(1), 84. https://doi.org/10.2307/1269292

Leonard, W. J. (2001). Cytokines and immunodeficiency diseases. *Nature Reviews Immunology*, *1*(3), 200–208. https://doi.org/10.1038/35105066

Martinez, A. M., & Kak, A. C. (2001). PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(2), 228–233. https://doi.org/10.1109/34.908974

Mizel, S. B., & Farrar, J. J. (1979). Revised nomenclature for antigen-nonspecific T-cell proliferation and helper factors. *Cellular Immunology*, *48*(2), 433–436. https://doi.org/10.1016/0008-8749(79)90139-4

Muhlhausler, B. S., Bloomfield, F. H., & Gillman, M. W. (2013). Whole Animal Experiments Should Be More Like Human Randomized Controlled Trials. *PLoS Biology*, *11*(2), e1001481. https://doi.org/10.1371/journal.pbio.1001481

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Page, M. J., Kell, D. B., & Pretorius, E. (2022). The Role of Lipopolysaccharide-Induced Cell Signalling in Chronic Inflammation. *Chronic Stress*, *6*, 24705470221076390. https://doi.org/10.1177/24705470221076390

Park, C. H., & Park, H. (2008). A comparison of generalized linear discriminant analysis algorithms. *Pattern Recognition*, *41*(3), 1083–1097. https://doi.org/10.1016/j.patcog.2007.07.022

Phifer-Rixey, M., Bi, K., Ferris, K. G., Sheehan, M. J., Lin, D., Mack, K. L., Keeble, S. M., Suzuki, T. A., Good, J. M., & Nachman, M. W. (2018). The genomic basis of environmental adaptation in house mice. *PLOS Genetics*, *14*(9), e1007672. https://doi.org/10.1371/journal.pgen.1007672

Punt, J., Stranford, S., Jones, P. P., & Owen, J. A. (2019). *Kuby Immunology* (8th ed.). Macmillan Education.

Raman, D., Sobolik-Delmaire, T., & Richmond, A. (2011). Chemokines in health and disease. *Experimental Cell Research*, *317*(5), 575–589. https://doi.org/10.1016/j.yexcr.2011.01.005

Ramasubramanian, K., & Moolayil, J. (2019). *Applied Supervised Learning with R*. Packt Publishing Ltd.

Rattanakiat, S., Nishikawa, M., Funabashi, H., Luo, D., & Yoshinobu Takakura, Y. (2009). The assembly of a short linear natural cytosine-phosphate-guanine DNA into dendritic structures and its effect on immunostimulatory activity. *Biomaterials*, *30*(29), 5701–5706. https://doi.org/10.1016/j.biomaterials.2009.06.053

Remick, D. (2002). Pharmacology of Cytokines. *Cytokines, Cellular & Molecular Therapy*, *7*(1), 38–39. https://doi.org/10.1080/13684730216403

Rietschel, E. T., Kirikae, T., Schade, F. U., Mamat, U., Schmidt, G., Loppnow, H., Ulmer, A. J., Zähringer, U., Seydel, U., & Di Padova, F. (1994). Bacterial endotoxin: molecular relationships of structure to activity and function. *The FASEB Journal*, *8*(2), 217–225. https://doi.org/10.1096/fasebj.8.2.8119492

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*(3), 303–304. https://doi.org/10.1038/nbt0308-303

Robinson, N. B., Krieger, K., Khan, F. M., Huffman, W., Chang, M., Naik, A., Yongle, R., Hameed, I., Krieger, K., Girardi, L. N., & Gaudino, M. (2019). The current state of animal models in research: A review. *International Journal of Surgery*, *72*, 9–13. https://doi.org/10.1016/j.ijsu.2019.10.015

Sarsani, V. K., Raghupathy, N., Fiddes, I. T., Armstrong, J., Thibaud-Nissen, F., Zinder, O., Bolisetty, M., Howe, K., Hinerfeld, D., Ruan, X., Rowe, L., Barter, M., Ananda, G., Paten, B., Weinstock, G. M., Churchill, G. A., Wiles, M. V., Schneider, V. A., Srivastava, A., & Reinholdt, L. G. (2019). The Genome of C57BL/6J "Eve", the Mother of the Laboratory Mouse Genome Reference Strain. *G3 Genes|Genomes|Genetics*, *9*(6), 1795–1805. https://doi.org/10.1534/g3.119.400071

Saul, M. C., Philip, V. M., Reinholdt, L. G., & Chesler, E. J. (2019). High-Diversity Mouse Populations for Complex Traits. *Trends in Genetics*, *35*(7), 501–514. https://doi.org/10.1016/j.tig.2019.04.003

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schnare, M., Holt†, A. C., Takeda, K., Akira, S., & Medzhitov, R. (2000). Recognition of CpG DNA is mediated by signaling pathways dependent on the adaptor protein MyD88. *Current Biology*, *10*(18), 1139–1142. https://doi.org/10.1016/s0960-9822(00)00700-4

Sellers, R. S., Clifford, C. B., Treuting, P. M., & Brayton, C. (2011). Immunological Variation Between Inbred Laboratory Mouse Strains. *Veterinary Pathology*, *49*(1), 32–43. https://doi.org/10.1177/0300985811429314

Sotirov, S. S., Pencheva, T., Janusz Kacprzyk, Krasimir Atanasov, Evdokia Sotirova, & Galya Staneva. (2022). *Contemporary Methods in Bioinformatics and Biomedicine and Their Applications*. Springer Nature.

Speakman, J., Hambly, C., Mitchell, S., & Król, E. (2007). Animal models of obesity. *Obesity Reviews*, *8*(s1), 55–61. https://doi.org/10.1111/j.1467-789x.2007.00319.x

Steensma, D. P., Kyle, R. A., & Shampo, M. A. (2010). Abbie Lathrop, the "Mouse Woman of Granby": Rodent Fancier and Accidental Genetics Pioneer. *Mayo Clinic Proceedings*, *85*(11), e83. https://doi.org/10.4065/mcp.2010.0647

Stephens, C. R., Huerta, H. F., & Linares, A. R. (2017). When is the Naive Bayes approximation not so naive? *Machine Learning*, *107*(2), 397–441. https://doi.org/10.1007/s10994-017-5658-0

Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Suresh, K. (2011). An Overview of Randomization techniques: an Unbiased Assessment of Outcome in Clinical Research. *Journal of Human Reproductive Sciences*, *4*(1), 8–11. https://doi.org/10.4103/0974-1208.82352

Švajger, U., & Jeras, M. (2011). Optimal Dendritic Cell Differentiation in RPMI Media Requires the Absence of HEPES Buffer. *Immunological Investigations*, *40*(4), 413–426. https://doi.org/10.3109/08820139.2011.556172

Svenson, K. L., Gatti, D. M., Valdar, W., Welsh, C. F., Cheng, R., Chesler, E. J., Palmer, A. A., McMillan, L., & Churchill, G. A. (2012). High-Resolution Genetic Mapping Using the Mouse Diversity Outbred Population. *Genetics*, *190*(2), 437–447. https://doi.org/10.1534/genetics.111.132597

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Boston Pearson.

Taheri, S., & Mammadov, M. (2013). Learning the naive Bayes classifier with optimization models. *International Journal of Applied Mathematics and Computer Science*, *23*(4), 787–795. https://doi.org/10.2478/amcs-2013-0059

Tang, D., Kang, R., Coyne, C. B., Zeh, H. J., & Lotze, M. T. (2012). PAMPs and DAMPs: signal 0s that spur autophagy and immunity. *Immunological Reviews*, *249*(1), 158–175. https://doi.org/10.1111/j.1600-065x.2012.01146.x

Taniguchi, K., Sugiyama, F., Kakinuma, Y., Uehara, S., Nishijho, N., Tanimoto, K., Murakami, K., Fukamizu, A., & Yagami, K. I. (1998). Pathologic characterization of hypotensive C57BL/6J-agt: angiotensinogen-deficient C57BL/6J mice. *International Journal of Molecular Medicine*. https://doi.org/10.3892/ijmm.1.3.583

Theocharides, A. P. A., Rongvaux, A., Fritsch, K., Flavell, R. A., & Manz, M. G. (2015). Humanized hemato-lymphoid system mice. *Haematologica*, *101*(1), 5–19. https://doi.org/10.3324/haematol.2014.115212

Thulin, M. (2016). Two-sample tests and one-way MANOVA for multivariate biomarker data with nondetects. *Statistics in Medicine*, *35*(20), 3623–3644. https://doi.org/10.1002/sim.6945

Viney, M., Lazarou, L., & Abolins, S. (2015). The laboratory mouse and wild immunology. *Parasite Immunology*, *37*(5), 267–273. https://doi.org/10.1111/pim.12150

Wade, C. M., Kulbokas, E. J., Kirby, A. W., Zody, M. C., Mullikin, J. C., Lander, E. S., Lindblad-Toh, K., & Daly, M. J. (2002). The mosaic structure of variation in the laboratory mouse genome. *Nature*, *420*(6915), 574–578. https://doi.org/10.1038/nature01252

Wang, D., Matsumoto, R., You, Y., Che, T., Lin, X.-Y., Gaffen, S. L., & Lin, X. (2004). CD3/CD28 Costimulation-Induced NF-κB Activation Is Mediated by Recruitment of Protein Kinase C-θ, Bcl10, and IκB Kinase β to the Immunological Synapse through CARMA1. *Molecular and Cellular Biology*, *24*(1), 164–171. https://doi.org/10.1128/mcb.24.1.164-171.2003

Wehrens, R. (2011). *Chemometrics with R*. Springer Science & Business Media.

Wehrens, R. (2020). *Chemometrics with R : Multivariate Data Analysis in the Natural and Life Sciences*. Springer Berlin Heidelberg, Imprint Springer.

Weiss, A., & Littman, D. R. (1994). Signal transduction by lymphocyte antigen receptors. *Cell*, *76*(2), 263–274. https://doi.org/10.1016/0092-8674(94)90334-4

Yan, X., Su, X., & Ebrary, I. (2009). *Linear regression analysis : theory and computing*. World Scientific.

Yang, H., Wang, J. R., Didion, J. P., Buus, R. J., Bell, T. A., Welsh, C. E., Bonhomme, F., Yu, A. H.-T., Nachman, M. W., Pialek, J., Tucker, P., Boursot, P., McMillan, L., Churchill, G. A., & de Villena, F. P.-M. (2011). Subspecific origin and haplotype diversity in the laboratory mouse. *Nature Genetics*, *43*(7), 648–655. https://doi.org/10.1038/ng.847

Zamora Saiz, A., Quesada González, C., Hurtado GilL., & Mondéjar RuizD. (2020). *An introduction to data analysis in R : hands-on coding, data mining, visualization and statistics from scratch*. Springer.

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*, *4*(11), 218–218. https://doi.org/10.21037/atm.2016.03.37

Zhao, H., Wai Keung Wong, Leung, H., & Zhang, X. (2020). Linear Discriminant Analysis. *Feature Learning and Understanding*, 71–85. https://doi.org/10.1007/978-3-030-40794-0_5

Zhu, P., Liu, X., Treml, L. S., Cancro, M. P., & Freedman, B. D. (2009). Mechanism and Regulatory Function of CpG Signaling via Scavenger Receptor B1 in Primary B Cells. *J Biol Chem.*, *284*(34), 22878–22887. https://doi.org/10.1074/jbc.m109.018580

Appendix

```
# R script

# Section 1

library(readxl)

library(caret)

library(MASS)

library(ggplot2)

library(caTools)


data<- read_excel("D:/Opera/Supplementary_Data_3.xlsx")

df <- data.frame(rbind(data))

df[] <- lapply(df, function(i) if(is.numeric(i)) ifelse(is.infinite(i), 0,
i) else i)

df[] <- lapply(df, function(i) if(is.numeric(i)) ifelse(is.na(i), 0, i) else
i)


#Data frames

df1=df[1:198,]

df1[,5:36]<-scale(df1[,5:36])

x1 <- subset(df1[, 5:36])

y1<-df1$Stimulation

y1<-factor(y1)

dfc1<-data.frame(cbind(y1,x1))


df2=df[199:262,]

df2[,5:36]<-scale(df2[,5:36])

x2 <- subset(df2[, 5:36])

y2<-df2$Stimulation

y2<-factor(y2)

dfc2<-data.frame(cbind(y2,x2))


#LOOCV

LOOCVcontrol <- trainControl(method="LOOCV")

#Repeated k-fold Cross-validation
```

```
RKFcontrol <- trainControl(method="repeatedcv", number=10, repeats=3)
#Bootstrap re-sampling cross-validation
bootcontrol <- trainControl(method="boot", number=100)

#Section 2
# PCA of the wild mouse strains
pca=prcomp(df1[,5:36],scale. = TRUE)
pca
summary(pca)
#PCA plot of the first PCs (wild mouse strains)
pca_plot1 <- cbind(df1, pca$x)
ggplot(pca_plot1, aes(PC1, PC2)) +
  geom_point(aes(color = Stimulation)) +
  ggtitle("PCA plot of the wild mouse strains")
# compute total variance
variance = pca$sdev^2 / sum(pca$sdev^2)
variance
mean(variance)
library("factoextra")
fviz_eig(pca, addlabels = TRUE, xlab = "Principal components",  ylim = c(0,
60))

#PCA of the lab strain
pca2=prcomp(df2[,5:36],scale. = TRUE)
pca2
summary(pca2)
#PCA plot of the first PCs (lab mouse strain)
pca_plot2 <- cbind(df2, pca2$x)
ggplot(pca_plot2, aes(PC1, PC2)) +
  geom_point(aes(color = Stimulation)) +
  ggtitle("PCA plot of the lab mouse strains")

# compute total variance
variance = pca2$sdev^2 / sum(pca2$sdev^2)
```

```
variance

mean(variance)

library("factoextra")

fviz_eig(pca2, addlabels = TRUE, xlab = "Principal components",  ylim = c(0,
60))


#Section 3

#MANOVA of the wild mouse strains

MANOVA1                <-                manova(cbind(IL.1α,IL.1β,IL.2,IL.3,
IL.4,IL.5,IL.6,IL.9,IL.10,IL.12p40,IL.12p70,IL.13,IL.17a,Eotaxin,GCSF,GMCSF
,IFN..U.03B3.,KC,MCP.1,MIP.1a,MIP.1b,RANTES,TNFα,IL.15,IL.18,Basic.FGF,LIF,
MCSF,MIG,MIP.2α,PDGF.BB,VEGF)~Stimulation,data=df1)

summary(MANOVA1)

MANOVA1

summary.aov(MANOVA1)


#MANOVA of the lab strain

MANOVA2                                <-manova(cbind(IL.1α,IL.1β,IL.2,IL.3,
IL.4,IL.5,IL.6,IL.9,IL.10,IL.12p40,IL.12p70,IL.13,IL.17a,Eotaxin,GCSF,GMCSF
,IFN..U.03B3.,KC,MCP.1,MIP.1a,MIP.1b,RANTES,TNFα,IL.15,IL.18,Basic.FGF,LIF,
MCSF,MIG,MIP.2α,PDGF.BB,VEGF)~Stimulation,data=df2)

summary(MANOVA2)

MANOVA2

summary.aov(MANOVA2)


#Section 4

#Linear Discriminant Analysis of the wild mouse strains

LDA1 <- lda(y1~., data=dfc1)

LDA1

#Linear Discriminant Analysis of the lab strain

LDA2 <- lda(y2~., data=dfc2)

LDA2


#LDA plot of the wild mouse strains

lda1_plot <- cbind(df1, predict(LDA1)$x)

ggplot(lda1_plot, aes(LD1, LD2)) +
```

```
  geom_point(aes(color = Stimulation)) + ggtitle ("Wild mouse strains LDA
plot")


#LDA plot of the lab mice

lda2_plot <- cbind(df2, predict(LDA2)$x)

ggplot(lda2_plot, aes(LD1, LD2)) +

  geom_point(aes(color = Stimulation))+ ggtitle ("Lab mice LDA plot")


#Section 5


# LOOCV

LDA.model1 <- train(y1~., data=dfc1,trControl=LOOCVcontrol, method="lda")

print(LDA.model1)

# Repeated K-fold

LDA.model2<- train(y1~., data=dfc1,trControl=RKFcontrol, method="lda")

print(LDA.model2)

# Bootstrap re-sampling

LDA.model3<- train(y1~., data=dfc1,trControl=bootcontrol, method="lda")

print(LDA.model3)


# LOOCV

LDA.model4 <- train(y2~., data=dfc2,trControl=LOOCVcontrol, method="lda")

print(LDA.model4)

# Repeated K-fold

LDA.model5<- train(y2~., data=dfc2,trControl=RKFcontrol, method="lda")

print(LDA.model5)

# Bootstrap re-sampling

LDA.model6<- train(y2~., data=dfc2,trControl=bootcontrol, method="lda")

print(LDA.model6)


#Section 6

#KNN classifier(s) for the wild mice strains

library(e1071)

library(caTools)
```

```
library(class)

#LOOCV
KNN.model1<-train(y1 ~ ., data=dfc1 ,method='knn',trControl=LOOCVcontrol)
print(KNN.model1)
# Repeated K-fold
KNN.model2<- train(y1~., data=dfc1,trControl=RKFcontrol, method="knn")
print(KNN.model2)
# Bootstrap re-sampling
KNN.model3<- train(y1~., data=dfc1,trControl=bootcontrol, method="knn")
print(KNN.model3)

#KNN classifier(s) for the lab mouse strain
#LOOCV
KNN.model4<-train(y2 ~ ., data=dfc2 ,method='knn',trControl=LOOCVcontrol)
print(KNN.model4)
# Repeated K-fold
KNN.model5<- train(y2~., data=dfc2,trControl=RKFcontrol, method="knn")
print(KNN.model5)
# Bootstrap re-sampling
KNN.model6<- train(y2~., data=dfc2,trControl=bootcontrol, method="knn")
print(KNN.model6)

#Section 7

# Naive Bayes classifier for the wild mice strains
library(e1071)
# LOOCV
NB.model1 <- train(y1~., data=dfc1,trControl=LOOCVcontrol, method="nb")
print(NB.model1)
# Repeated K-fold
NB.model2<- train(y1~., data=dfc1,trControl=RKFcontrol, method="nb")
print(NB.model2)
```

```
# Bootstrap re-sampling
NB.model3<- train(y1~., data=dfc1,trControl=bootcontrol, method="nb")
print(NB.model3)

# Naive Bayes Models for the lab mouse strain
# LOOCV
NB.model4 <- train(y2~., data=dfc2,trControl=LOOCVcontrol, method="nb")
print(NB.model4)
# Repeated K-fold
NB.model5<- train(y2~., data=dfc2,trControl=RKFcontrol, method="nb")
print(NB.model5)
# Bootstrap re-sampling
NB.model6<- train(y2~., data=dfc2,trControl=bootcontrol, method="nb")
print(NB.model6)

#Secttion 8
#  Random Forest classifier for the wild mice strains

library(randomForest)

# LOOCV
RandomForest.model1<-train(y1 ~ ., data=dfc1
,method='rf',trControl=LOOCVcontrol)
print(RandomForest.model1)

# Repeated K-fold
RandomForest.model2<-train(y1 ~ ., data=dfc1
,method='rf',trControl=RKFcontrol)
print(RandomForest.model2)

# Bootstrap re-sampling
RandomForest.model3<-train(y1 ~ ., data=dfc1
,method='rf',trControl=bootcontrol)
print(RandomForest.model3)
```

```
# Random Forest classifier for the lab mouse strain


# LOOCV

RandomForest.model4<-train(y2 ~ ., data=dfc2
,method='rf',trControl=LOOCVcontrol)

print(RandomForest.model4)


# Repeated K-fold

RandomForest.model5<-train(y2 ~ ., data=dfc2
,method='rf',trControl=RKFcontrol)

print(RandomForest.model5)


# Bootstrap re-sampling

RandomForest.model6 <-train(y2 ~ ., data=dfc2
,method='rf',trControl=bootcontrol)

print(RandomForest.model6)


#Section 9

#Support Vector Machine classifier for the wild mice strains

library(e1071)


# LOOCV

SVM.model1 <- train(y1~., data=dfc1,trControl=LOOCVcontrol,
method="svmLinear")

print(SVM.model1)


# Repeated K-fold

SVM.model2<- train(y1~., data=dfc1,trControl=RKFcontrol, method="svmLinear")

print(SVM.model2)


# Bootstrap re-sampling

SVM.model3<- train(y1~., data=dfc1,trControl=bootcontrol,
method="svmLinear")

print(SVM.model3)
```

```
#Support Vector Machine classifier for the lab mouse strains


# LOOCV

SVM.model4      <-      train(y2~.,      data=dfc2,trControl=LOOCVcontrol,
method="svmLinear")

print(SVM.model4)


# Repeated K-fold

SVM.model5<- train(y2~., data=dfc2,trControl=RKFcontrol, method="svmLinear")

print(SVM.model5)


# Bootstrap re-sampling

SVM.model6<-          train(y2~.,          data=dfc2,trControl=bootcontrol,
method="svmLinear")

print(SVM.model6)


#Section 10


#Generalized linear model classifier for the wild mice strains


# LOOCV

GLM.model1<- train(y1~., data=dfc1,trControl=LOOCVcontrol, method="glmnet")

print(GLM.model1)

# Repeated K-fold

GLM.model2<- train(y1~., data=dfc1,trControl=RKFcontrol, method="glmnet")

print(GLM.model2)

# Bootstrap re-sampling

GLM.model3<- train(y1~., data=dfc1,trControl=bootcontrol, method="glmnet")

print(GLM.model3)


#Generalized linear model classifier for the lab mouse strain


# LOOCV
```

```
GLM.model4<- train(y2~., data=dfc2,trControl=LOOCVcontrol, method="glmnet")
print(GLM.model4)
# Repeated K-fold
GLM.model5<- train(y2~., data=dfc2,trControl=RKFcontrol, method="glmnet")
print(GLM.model5)
# Bootstrap re-sampling
GLM.model6<- train(y2~., data=dfc2,trControl=bootcontrol, method="glmnet")
print(GLM.model6)

#Section 11

#Artificial neural network classifier of the wild mice strains
library("neuralnet")

# LOOCV
ANN.model1<- train(y1~., data=dfc1,trControl=LOOCVcontrol, method="nnet")
print(ANN.model1)
# Repeated K-fold
ANN.model2<- train(y1~., data=dfc1,trControl=RKFcontrol, method="nnet")
print(ANN.model2)
# Bootstrap re-sampling
ANN.model3<- train(y1~., data=dfc1,trControl=bootcontrol, method="nnet")
print(ANN.model3)

#Artificial neural network classifier of the lab mouse strain

# LOOCV
ANN.model4<- train(y2~., data=dfc2,trControl=LOOCVcontrol, method="nnet")
print(ANN.model4)
# Repeated K-fold
ANN.model5<- train(y1~., data=dfc1,trControl=RKFcontrol, method="nnet")
print(ANN.model5)
# Bootstrap re-sampling
```

```
ANN.model6<- train(y1~., data=dfc1,trControl=bootcontrol, method="nnet")
print(ANN.model6)
```