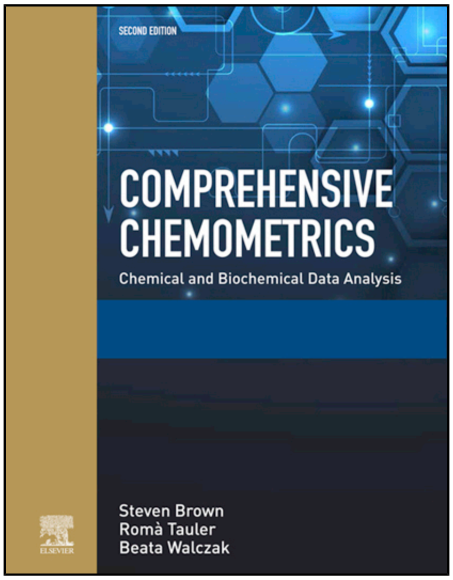

SECOND EDITION
COMPREHENSIVE CHEMOMETRICS
Chemical and Biochemical Data Analysis
Steven Brown
Romà Tauler
Beata Walczak
ELSEVIER

From Mosleth, E. F.; McLeod, A.; Rud, I.; Axelsson, L.; Solberg, L.; Moen, B.; Gilman, K.; Færgestad, E. M.; Lysenko, A.; Rawlings, C.; Dankel, S. N.; Mellgren, G.; Barajas-Olmos, F.; Orozco, L. S.; Sæbø, S.; Gidskehaug, L.; Oust, A.; Kohler, A.; Martens, H.; Liland, K. H. Analysis of Megavariate Data in Functional Omics. In Comprehensive Chemometrics: Chemical and Biochemical Data Analysis; Brown, S., Tauler, R., Walczak, B., Eds., Elsevier, 2020; pp 515–567.
ISBN: 9780444641656

4.22 Analysis of Megavariate Data in Functional Omics☆

EF Mosleth, A McLeod, I Rud, L Axelsson, LE Solberg, and B Moen, Nofima, Norwegian Institute for Food, Fisheries and Aquaculture Research, Ås, Norway
KME Gilman, Faculty of Science and Technology, Norwegian University of Life Sciences (NMBU), Ås, Norway; and Nofima, Norwegian Institute for Food, Fisheries and Aquaculture Research, Ås, Norway
EM Færgestad, Department of Chemistry, University of Oslo, Norway
A Lysenko, Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
C Rawlings, Department of Computational and Analytical Sciences, Rothamsted Research, Harpenden, United Kingdom
SN Dankel, Center for Diabetes Research, Department of Clinical Science, University of Bergen, Bergen, Norway; MOHN Nutrition Research Laboratory, Department of Clinical Science, University of Bergen, Bergen, Norway; and Hormone Laboratory, Haukeland University Hospital, Bergen, Norway
G Mellgren, Center for Diabetes Research, Department of Clinical Science, University of Bergen, Bergen, Norway; MOHN Nutrition Research Laboratory, Department of Clinical Science, University of Bergen, Bergen, Norway; and Hormone Laboratory, Haukeland University Hospital, Bergen, Norway
F Barajas-Olmos and LS Orozco, Immunogenomics and Metabolic Diseases Laboratory, National Institute of Genomic Medicine. Mexico City, Mexico
S Sæbø, Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway
L Gidskehaug, Camo Analytics, Oslo Science Park, Oslo, Norway
A Oust, Nofima, Norwegian Institute for Food, Fisheries and Aquaculture Research, Ås and Biotechnology and Chemistry, Oslo and Akershus University, College of Applied Sciences, Oslo, Norway
A Kohler, Nofima, Norwegian Institute for Food, Fisheries and Aquaculture Research, Ås, Norway; and Faculty of Science and Technology, Norwegian University of Life Sciences (NMBU), Ås, Norway
H Martens, Nofima, Norwegian Institute for Food, Fisheries and Aquaculture Research, Ås, Norway; and Department of Engineering Cybernetics, NTNU, Trondheim, Norway
KH Liland, Nofima, Norwegian Institute for Food, Fisheries and Aquaculture Research, Ås, Norway; and Faculty of Science and Technology, Norwegian University of Life Sciences (NMBU), Ås, Norway

This is an update of E.M. Færgestad, Ø. Langsrud, M. Høy, K. Hollung, S. Sæbø, K.H. Liland, A. Kohler, L. Gidskehaug, J. Almergren, E. Anderssen, H. Martens, 4.08 - Analysis of Megavariate Data in Functional Genomics, in Comprehensive Chemometrics, edited by Steven D. Brown, Romá Tauler, Beata Walczak, Elsevier, 2009, https://doi.org/10.1016/B978-044452701-1.00011-9.

☆*Change History*: January 2020. EM Færgestad, EF Mosleth, KM Erikson Gilman, A Lysenko, C Rawlings, S Dankel, G Mellgren, LE Solberg, F Barajas-Olmos, LS Orozco, S Sæbø, L Gidskehaug, B Moen, A Oust, A Kohler, H Martens and K H Liland updated the article and figures.

Comprehensive Chemometrics, Second Edition, 2020, 515–567

4.22.1 Introduction

The science of functional omics aims at investigating pools of biological molecules that translate into the structure, function, and dynamics of one or more organisms. For all living organisms the genes carry the information necessary for reproduction and regulation. Each individual has its own genetic makeup, giving rise to differences between individuals. Whereas the sequences of the genetic code are mostly constant throughout the whole life period of an organism, the genes are turned on and off in a dynamic process in response to external/internal environmental factors and the developmental stage of the organism. When a gene is turned on, a copy of the gene is produced containing the code for a specific protein.

The genetic makeup, which is usually inherited from the parents, defines which proteins that can be synthesized, and a number of different regulating mechanisms control the expression of the genes.

The proteins have different functions in the cells, such as being enzymatic proteins executing chemical reactions, hormones that give signals to other cells, or structural proteins used as building blocks for cells or organs. The activation of genes followed by protein synthesis will thereby in turn determine the metabolic activity in the cells, and the resulting physiology of the organism.

From the genetic code to the final phenotype of the organism there is a chain of causalities, and at the same time there are feedback regulation mechanisms at all stages. External and internal environment factors as well as the developmental stages of the organisms affect all levels of the process from the activation of genes to the final functionality of the organism. The metabolic complexity is further increased by the fact that each gene may give rise to several proteins, and one protein may act on different processes in the cells. Thus, a very large number of genes and their products along with the environmental conditions function in a complicated and orchestrated way to regulate the metabolic processes. Hence, the scientific task of obtaining insight into the regulatory mechanisms from the gene activation to the final functionality of the organism is a complex and challenging exercise. The fundamental understanding of biological systems, comes from understanding not only each piece of the puzzle, but to understand how all the elements together describe the functionality of the organisms. This calls the need for a multivariate approach.

Modern science has easy availability of instruments and measurement techniques. The amount of data generated in functional omics studies can be enormous. Examples are metagenomics and metatranscriptomics analyses, which focus on studying the genomic content and the gene expression, respectively, of microbial communities. Even more challenging is the fact that the sources of variability are numerous. Traditional chemometric tools, which have given major success in a number of different fields, were originally developed for simpler situations for example the use of spectroscopy to quantify the main constituent in the samples, which is fat, protein and carbohydrates. To measure for example fat content of salmon, a number of wavelengths are used by chemometric tools, but the underlying variation, which is the variation in fat content, is simple. Omics data is by definition observation of all molecules in the cell, and the aim is to unravel all important variation related to the purpose of the study, for example all changes in the cells related to a particular disease, which by itself may be heterogenous. In functional omics we face a situation where a very large number of genes are constantly activated and repressed in a dynamic process giving rise to a large number of underlying sources of variability. Functional genomics can therefore be characterized as being of megavariate size with a multidimensional latent structure.

Due to the complexity of functional omics, the link between the biologist and the data analyst is even more crucial than for most other areas. This chapter is written with the intention of contributing to building a bridge across the gap between the two scientific communities. The readers we have in mind are therefore both experts from the areas of molecular biology and biochemistry (etc.) who are not trained in data analysis, and experts in and data analysts who lack biological training.

The choice of strategy for analyzing data from functional omics must be based on understanding of the system under study. We therefore start with a description of crucial aspects of functional omics and the data typically generated as the empirical basis of functional omics studies. Thereafter, attention is paid to various aspects that must be taken into consideration before choosing the strategy for analyzing the data. Finally, we go through some practical approaches for data analysis where we use a set of experimental omics data for method demonstration and discussion. In the appendix are more detailed mathematical descriptions and program codes available for interested readers, but not needed for the readers of the main text.

4.22.2 Molecular Basis of Functional Omics

An overview of the flow of information from the genetic makeup to the final phenotypic expression are displayed in Fig. 1. Below we go through each of these sources of information. More comprehensive information on these topics are given elsewhere.[1–5]

4.22.2.1 Genome

The genome is usually composed of deoxyribonucleic acids (DNAs) (Fig. 2), which are long polymeric molecules of nucleotides. Each nucleotide consists of a phosphate group, a sugar molecule, and a cyclic nitrogen-containing base. DNA consists of two polymeric strands of nucleotides helically wound around each other to form a DNA double helix. In DNA there are four different bases; thymine (T), adenine (A), cytosine (C), and guanine (G), and the two DNA strands are connected by hydrogen bonds between base pairs. The chemical structure of the bases allows A and T to be connected, and likewise C and G. That is, there is a one-to-one relationship between the two DNA strands.

A gene is a defined base pair sequence along the DNA ranging from a hundred to several thousand base pairs, which is used as a template when activated to give a copy (a transcript) of that particular gene. The position on the DNA housing a gene is called a locus. At a given locus there are different variants of the gene (alleles) with slightly differing base pair combinations, which may encode proteins with different properties. Allelic variation is the genetic basis for phenotypic variation between individuals. The genome is the complete collection of genetic information of an organism, and usually the genome contains several DNA molecules organized into structures called chromosomes. In addition to protein coding gene sequences, DNA contains regulatory elements and other intervening nucleotide sequences. Although diploid organisms with DNA-based genomes are most common, there are exceptions such as haploids (one gene copy) or polyploids (several copies), and organisms with ribonucleic acid (RNA) as carrier of the genetic code rather than DNA.

The unique and essential property of DNA is its ability to reproduce itself, which is the essential fundament of life. During normal cell division (mitosis) the two DNA strands are split, and each strand serves as a template for the synthesis of a new strand (see Fig. 3). The base C will then find a free nucleotide with base G to link up with, the base A will find a new T, etc. Thereby the two daughter DNA strands will be identical to the mother DNA strand. Prokaryotes (bacteria and archaea) have a single circular DNA, whereas eukaryotic organisms have long linear DNA macromolecules. In eukaryotes, the DNA is wrapped around proteins called histones located in distinct nucleus. The complex of DNA plus histones and other structural proteins is called chromatin. Under cell division, the chromatin condens and break up into separate, linear pieces called chromosomes. Prokaryotes do not have nucleus. The genetic materials are in prokaryotes free-floating in the cells. Each species has its own characteristic number of chromosomes. Humans, for instance, have 46 chromosomes in a typical body cell. Humans are diploid, which means that chromosomes come in matched sets known as homologous pairs. The 46 chromosomes of a human cell are organized into 23 pairs, and the two members of each pair are said to be homologues of one another (with the slight exception of the chromosomes called X and Y which determines the sex). Homologues chromosomes carry the same type of genetic information, and they have the same genes in the same locations. One of these homologous gene pair is inherited from the mother, the other from the father.

The process of DNA copying is, however, not perfect; occasionally, errors do occur, giving rise to mutations. By mutation, one base may be shifted over to another. Such changes occur more frequently at particular sites. Sites where mutations occur frequently are often called hot spots. Although mutations in general are rare, they constitute the basis for biological diversity.

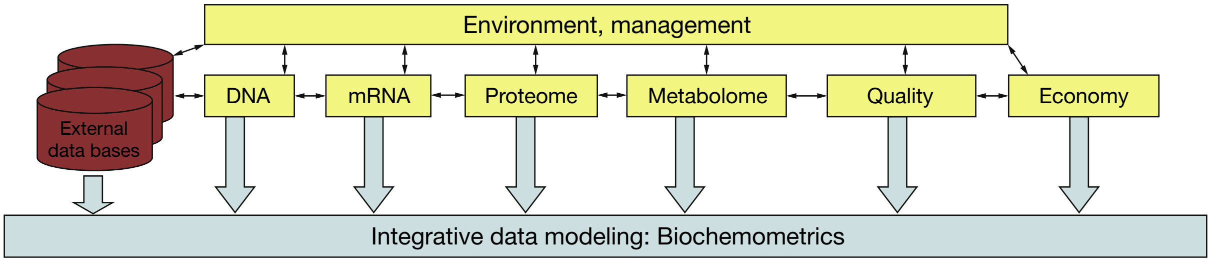


Fig. 1 Overview of the flow of data in functional omics from DNAs (deoxyribonucleic acids, the Genome), the transcriptome (mRNA messenger single-stranded ribonucleic acid), proteome, metabolome, to the final phenotypic expression. The final phenotypic expression may be called quality "phenome" or "end-product." The final phenotypic expression may for example be utilized to produce end-products with higher quality and thereby enhance its market value and this might be organized as data tables as displayed in this figure. Information from external databases may also shed light on all these complex data blocks.

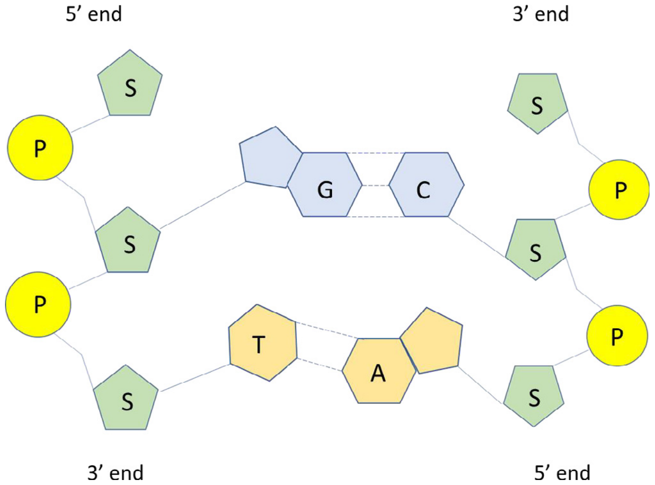


Fig. 2 Base pairing in DNA. The DNA bases are thymine (T), adenine (A), cytosine (C), and guanine (G). Two hydrogen bonds connect T to A; three hydrogen bonds connect G to C. A sugar (S)-phosphate (P) backbone (gray) run anti-parallel aligned to each other, as seen by the opposite directions of the 3′ and 5′ indications in the figure. The 3′ and 5′ refer to the carbon number in the sugar molecule in the backbone.

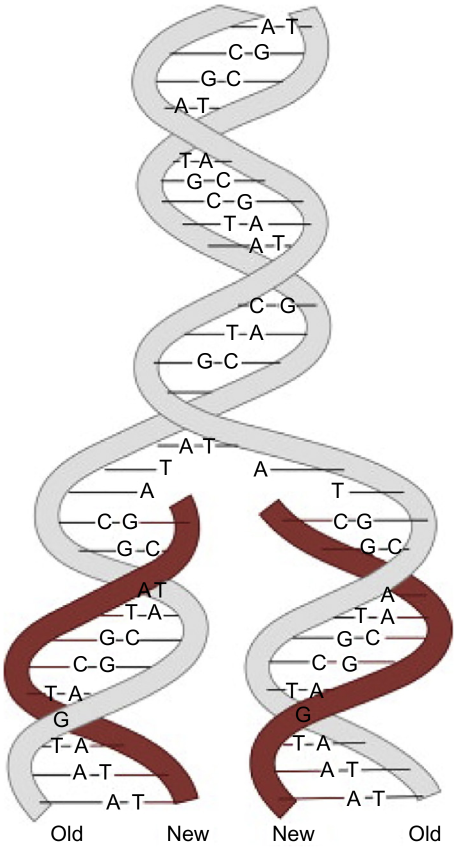


Fig. 3 DNA duplication during cell division (mitosis). Reproduced with permission from Singer, M; Berg, P. *Genes and Genomes: A Changing Perspective*. Mill Valley, CA: University Science Books, 1991.

Prokaryotes mostly reproduce by cell division, whereas eukaryotic organisms usually have sexual reproduction. The production of germ cells (meiosis) consists of two steps. In the first step, homologue pairs separate, and under the second step, sister chromatids separate yielding haploid germ cells. When two germ cells meet, one from a female and one from a male, two chromatids are again paired to constitute the diploid chromosome set of a normal cell. Before the division of the two chromatids in meiosis, a very important process of chromatid sequence exchange occurs. The germ cell will thereby not consist of a DNA chromatid identical to that in the parental cell, but a mix of the two with several crossings over from one DNA chromatid to the other. This may sometimes lead to duplication or deletion of DNA sequences. The crossing-over does not occur randomly along the DNA chromatids, but rather at specific sites. Some genes will therefore be tightly linked together and inherited as such, from one generation to the next. The particular combination of nucleotides within one such region is called haplotype, and can be regarded as a genotype in miniature.

For some genes where two different alleles are present in a diploid species, one allele may dominate over the other with respect to the resulting phenotype; that is, one allele is dominant and the other is recessive. A classic example is eye color where a person carries one allele for brown eyes and one for blue eyes. The allele for brown eyes will normally dominate over the blue. In situations with total dominance, only the dominant phenotype will be expressed. In other situations, both genes may contribute giving an average phenotype of the two, or a phenotypic expression deviating from the mean.

Another important characteristic of the action of genes is that, one gene may affect several phenotypic characteristics. This is called pleiotropy. This implies that in genetics it is not a simple one-to-one relationship between the genes and the functionality of the cell.

Even more challenging is the important phenomenon that the action of a gene suppressing the action of allelic variation of other genes.[6–8] Such interacting effects among genes, which is called epistasis,[7,8] is a major challenge to address. This can be illustrated by considering a situation comprised of two genes (**Table 1**), each with two alleles where one gene is dominant over the other, and the dominant allele of one of the genes is epistatic over the allelic variation at the other gene. The dominant gene is often assigned with large letter A and B, and their resistive counterpart is here assigned with small letter a and b.

Whenever the gene A is present, A will determine the phenotypic expression. Thus, a genotype that carries the combination AA and Aa will express the same phenotype since A will dominate over a. Likewise, B is dominant over b and both combinations BB and Bb express similar phenotype. However, here comes the complex challenge to address; if A is epistatic over B, then the *B* phenotype will only be expressed in the absence of the gene A, and the phenotype *b* is only expressed in the absence of both genes A and B. Thus, the 16 alternative allele combinations will express phenotypic variation in the ratio 12:3:1. With this situation, from a data analytical point of view, it may be easy to detect the dominant effect of A over a, as many genotypes will reflect this effect. The dominant effect of B over b is also reflected in several genotypes so it can be detected. The epistatic effect of A over allelic variation at the other gene, is, however, obscure and difficult to detect as this gene interaction relies on only 1 out of the 16 alternative genetic makeups. With more genes involved and more alternative alleles, the complexity increases. A crucial feature with such a situation is that the main information is governed in the highest order of interaction and only a few genotypes reflect the information needed to reveal the genetic interactions. As a consequence, although interactions between genes have long been recognized in the literature, addressing such interactions is indeed challenging. This brings in a very important point, which is the need for large data sets to unravel complex genetic interacting effects. In cell biology, interacting effects are likely to be important. As example, some persons smoke heavily throughout their life and yet reach a high age, although it is well recognized that smoking is one of the main risks of early death. It is the combination of the genetic makeup along with the environmental impacts that determine the final phenotypic characteristics. And that is what we aim to uncover in the data analysis of functional omics.

Genomic analysis is the identification, measurement or comparison of genomic features at a genomic scale. Next generation sequencing (NGS) technologies revolutionized genomic research by enabling sequencing of millions of small fragments of DNA in parallel. Bioinformatics analyses are used to piece together these fragments by mapping the individual sequence reads.

4.22.2.2 Transcriptome

When a gene is activated, a transcript of that particular gene is made as a single-stranded ribonucleic acid (RNA) molecule. The nitrogenous bases in RNA are adenine (A), guanine (G), cytosine (C), and uracil (U), which replaces thymine (T) in DNA. The transcriptome is the collection of transcripts from all genes that have been turned on at a given time in the cell or tissue under study. The transcriptome is thus a global way of looking at gene expression patterns. There are various types of RNA. The major type, messenger RNA (mRNA), carries codes of proteins. The composition of the transcriptome can be analyzed by microarray techniques and by RNA sequencing (RNA-Seq) technologies. RNA-Seq presents several advantages over microarrays, including higher specificity and sensitivity. Information on e.g. small noncoding RNA molecules (an RNA molecule that is not translated into a protein), which may target mRNA for destruction and regulation of gene expression may also be obtained.

The transcriptome gives dynamic snapshots into a short period of time for the organism, tissue, or cell from which the transcripts are collected. The snapshot reflects the genes that are turned on and that have resulted in transcripts at the particular time investigated. This provides valuable information on the regulation of activated genes and the proteins that may be expressed. When

Table 1 Allelic combinations for two genes, each with two alternative alleles, A vs. a and B vs. b, and the resulting phenotype. The gene A is dominant over a, B is dominant over b, and A is epistatic over the allelic variation of the other gene. The phenotypes are expressed as cursive as resulting from the gene *expression*.

	BB	Bb	bB	bb
AA	*A*	*A*	*A*	*A*
Aa	*A*	*A*	*A*	*A*
aA	*A*	*A*	*A*	*A*
aa	*B*	*B*	*B*	*b*

transcriptome experiments are conducted according to an experimental design, the aim is usually to identify RNAs (most often primarily the mRNAs) that are differently expressed as a result of the experimental conditions. As the activation of genes and the actions of mRNA are dynamic events changing over time, it is highly relevant to include a time-course study in such experiments searching for changes occurring over a period of time. It is then crucial to take the circadian rhythm into account.[9,10]

4.22.2.3 Proteome

Proteins are chains of amino acids connected by peptide bonds. The order of the base pairs on mRNA comprises a code for the synthesis of proteins as a sequence of three. Three base pairs on the mRNA constitute the code for one given amino acid. The order of amino acids is, thus, directly connected to the order of the base pairs of the gene coding for the given protein (Fig. 4). In total there are 20 different amino acids, and the structure and functionality of the protein is determined by the order and the properties of the amino acids.

The proteome is the collection of all proteins in a cell at a given time. Some proteins are metabolic enzymes involved in energy metabolism, others may act as hormones or signaling molecules; some may act as structural building blocks keeping the cell or tissue structures together and others may act as protecting molecules or defense molecules. The protein composition in a cell will therefore reflect the whole metabolic activity of the cells, and thereby the final physiology and phenotype of the organism.

Proteomics involves the applications of technologies for the identification and quantification of overall proteins present in the cell, tissue or an organism. Such mapping is an ambitious goal that is seldom achieved in complex biological samples, but a large number of proteins can be separated by various approaches, e.g. by mass spectrometry with LC-MS-MS, gel electrophoresis and MALDI-TOF/TOF. By some techniques, such as LC-MS, the proteins are often cleaved by specific enzymes to shorter amino acid chains called peptides. Whereas other techniques, such as gel electrophoresis, analyze the intact proteins, where also protein modifications, such as changes by addition of molecules to the proteins, can be observed.

Although the proteome is made from translation of mRNA, studying the link between the transcriptome and the proteome is not straightforward, for instance due to posttranslational modifications, i.e. changes that are regulated on the protein level rather than on the gene expression level. Furthermore, the proteins have a longer turnover rate than the mRNA.

4.22.2.4 Metabolome

The metabolome has been defined as the qualitative and the quantitative collection of all low-molecular-weight molecules (metabolites) present in the cell that are participants in general metabolic reactions and that are required for the maintenance, growth, and normal function of a cell.[11] The metabolome is a result of the biochemical reactions being catalyzed by enzymes which are proteins. This in turn determines the biological structure and function of the final phenotype of the organism.

The metabolome includes, among other compounds, amino acids, fatty acids, carbohydrates, vitamins, and lipids. The number of the different molecules in the metabolome varies depending on the organism being studied, and it is constantly changing due to all the chemical reactions occurring in the cell. Thus, metabolite profiling aims to identify and quantify metabolites, for example by using sensitive chromatographic methods like GC-MS and LC-MS,[12] to give a snapshot of the physiology of a cell.

Analyzing metabolite networks may be more challenging than analyzing the transcriptome and the proteome. It is only possible to extract and analyze a smaller fraction of all the metabolites, and there are often more missing or unreliable data in metabolome and in transcriptome and proteome data. Furthermore, there are more steps in harvest procedures, sample preparation, and analysis where artifacts (e.g. loss or transformation of metabolites; errors in identification or quantification) can occur.

4.22.2.5 Redundancy

Biological redundancy is a frequent and important mechanism, which means that two or more genes/proteins/metabolites are performing the same function and that inactivation of one of these features has little or no effect on the biological phenotype. Redundancy is widespread, and even more so at the metabolome level than on the genetic level. This is the fundament for the survival mechanisms of any living organisms, as changes in one feature may be compensated by an alternative molecular route. However, for the scientists this gives major challenges in identifying the features involved in any biological process under study.

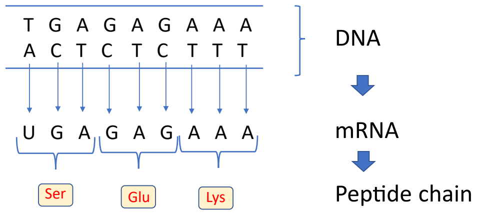


Fig. 4 Synthesis of polypeptide chain. The information in DNA is transferred to a messenger RNA (mRNA) (transcription), and the mRNA sequence is used as a template to assemble a chain of amino acids that form a protein (translation). The transcription from DNA to mRNA is indicated by arrows for each base, note that U is replaced by T in mRNA.

4.22.2.6 Metagenome and Metatranscriptome

Metagenomics is the study of metagenomes, genetic material recovered directly from environmental samples. Metatranscriptomics is the study of the function and activity of the complete set of transcripts from environmental samples. Recent studies use modern NGS technologies to get largely unbiased samples of all genes from all the members of the sampled communities. Metagenomics and metatranscriptomics offer a powerful lens for obtaining knowledge of the microbial world.

4.22.2.7 Environmental Impact and Genotype-Environmental Interaction

In addition to genetic effects, environmental effects also play a crucial role for the development and control of organisms. The environmental effects constitute both the external environment, such as temperature, nutrition and drugs, and also the internal environment in the organism. The environment can have a direct effect on the phenotype, but it can also influence how a genotype responds. Genotype–environment interaction refers to situations in which different genotypic groups respond differently to the same array of environments. The observed variation of the phenotype (V_P) can be divided into three parts: variation related to the genotype (V_G), variation related to the environment V_E, and variation related to different responses upon the environment for the different genotypic groups V_{G*E}.

$$V_P = V_G + V_E + V_{G*E}$$

The concept of heritability (h^2) is the measure of the proportion of the observed variation that is related to the genetic effects.

$$h^2 = V_G/V_P$$

However, if genotype-environment effects (V_{G*E}) are present, the relationship between the genetic variation and the phenotypic variation is not linear, and it is not obvious how to calculate and understand the heritability. This has been a debated issue for over a century. The debate started between the statistician and geneticist R.A. Fisher, one of the founders of population genetics and the creator of the statistical ANalysis Of VAriance (ANOVA), and L. Hogben, an experimental embryologist and statistician.[13]

R.A. Fisher first considered genotype-environment interactions to be of "potential, but unproved, importance",[13] whereas Hogben claimed they were "standard and fundamentally important for understanding variability". Fisher later recognized the complications raised by the "non-linear interaction of environment and heredity", and he developed an approach to handle this for the summing of variances, by the biometric concept of genotype–environment interaction, or $G^* E_B$. Hogben considered different sources of variability in a population where he recognized a genotype-environment as a result of development, and he introduced the concept development genotype–environment interaction, or $G * E_D$.

Phenotypic plasticity, which is essential for survival of an organism, is another important characteristic of many quantitative traits. It reflects the ability to adapt to changes in the environment.[14]

Taken together, the environment and the genotype-environment interactions are known to have complex effects in cell biology and must be carefully considered in the data analysis of functional omics, and it affects all levels along the chain from the gene regulation to the final phenotype.

Interaction between genetic makeup and environment, brings us over to the topic of epigenetics which imply that the impact of the environment can also be inherited from one generation to the next.

4.22.2.8 Epigenome

Epigenetics is a word starting with the Greek prefix epi- (πι-"over, outside of, around"), which implies features that are "on top of" or "in addition to" the traditional genetic basis for inheritance. It is heritable changes in gene expression (active versus inactive genes) that do not involve changes to the underlying DNA sequence. This includes modification of DNA, modification of histones (proteins the DNA is wound around) by molecules such as methyl groups and acetyl groups, and it includes non-coding RNA (ncRNA) (**Fig. 5**).

Epigenomics is the study of the epigenome, the complete set of epigenetic modifications on the genetic material of a cell. Epigenetic modifications are reversible modifications on a cell's DNA, or the proteins at the DNA entity (histones), that affect gene expression without altering the DNA sequence.[16] Unlike the underlying genome, which is largely static within an individual, the epigenome can be dynamically altered by environmental conditions. Epigenetic modifications regulate gene expression, and play significant roles in growth, development, and disease progression. The study of epigenetics on a global level has been made possible through genomic high-throughput sequencing and assays.

4.22.2.9 Phenotypic Consequences

The metabolic activity occurring in the cell affects the total chemical composition, the physical structures, the appearance of diseases, the quality, etc. We can call these phenotypic consequences.

How these different sources of information are best organized is a question of concern for each individual study. In some situations, the phenotypic consequences in mind are complex traits characterized by a combination of several types of measurements, whereas in other situations one single response parameter has the primary focus as, for example, the presence or absence of one

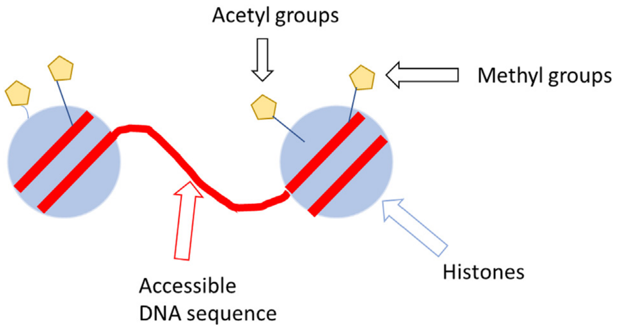


Fig. 5 A schematic diagram of DNA (in *red*) wrapped around proteins called histones (*blue*). The methyl groups bonded on the DNA strand and acetyl groups bounded on the histone are epigenetic modifications to the DNA and the histones, respectively.[15]

particular disease. In many situations, it is important to extend the view of the final phenotypic consequences as being more than one single variable. Based on the pleiotropic actions of genes, a broader picture of the final phenotypic consequences may shed light on the total action of the genes under study.

4.22.2.10 Interactome

Comprehensive cellular processes take part in maintaining biological systems through molecular interaction networks. One way to express the interactions is by the interactome defined as a network consisting of nodes representing individual molecules and connections between nodes (edges) which reflect physical (direct) or functional (indirect) interactions between molecules.[17]

The interactions may represent direct interactions between molecules that can be observed in experiments such as protein-protein interactions. Interactions may also be established indirectly, by computational approaches. For example, if two genes both are always active in one set of samples and inactive in another set, one might conclude that the two genes may be functionally related.

Numerous methods have been developed to establish interactomes to pinpoint interactions that can through light on omics data, and several bioinformatic tools are available. Typically, omics features selected from a study under considerations can be plugged into relevant bioinformatic tools to visualize interactions that may be relevant.

4.22.2.11 Use of Background Information in Data Analysis

Scientists in all research fields often use background information (accumulated knowledge and experience) both in the planning phase of new experiments and when interpreting the results of a finalized study. New and flexible multivariate data analytical tools now enable the incorporation of background information directly during the analytical phase of the study. Connecting background information with observed data may lead to improved insight into the biological system under study as well as improved predictions. As data generation proceeds at a high speed, a large amount of information on different organism and different molecules is available (genes, proteins, metabolites, etc.), which constitutes an important source of information that can be utilized. Such information is often collected and available in public libraries (e.g. databases of biochemical pathways, gene ontologies, and gene regulation sequences). There is a growing awareness of the necessity of linking experimental data with independent background information on the features to deal with some of the data redundancy in functional omics.

4.22.3 Important Considerations in Functional Omics

4.22.3.1 Scientific Strategy

The traditional approach for statistical research has been to investigate one-to-one relationships between input features and some response features. Such an approach would be useful if the onset of one gene resulted in one phenotypic trait with no interaction between genes, no feedback-regulating mechanisms, and no pleiotropic effects. However, if this was the case, the survival of the organisms would be very poor. Instead, there is a complex network of feedback-regulating mechanisms at all stages from genome to final phenome guiding development and ensuring a certain level of robustness to changes in external and internal environmental factors. The causality in the functional omics chain can therefore better be characterized by many-to-many relationships.

One consequence of this is that it is highly relevant to bring in several sources of data at the same time to shed light on one property of interest. How to attack the problem from a data analytical point of view is not straightforward. Even basic questions like defining the input and the output of a data analysis model are not obvious beforehand.

Alternative and seemingly contradictory approaches may be chosen to achieve scientific progress. One strategy is to use a confirmative approach to investigate a restricted and concrete question. This contrasts with a complementary explorative approach where the starting point is to observe the system of interest with a "wide angle" without putting restrictions or limitations to the observations based on previous assumptions. Having in mind the complexity of functional omics, a fruitful combination of these two complementary approaches would be highly useful. Scientific research can be viewed as a cycle of processes, where one answer will generate a new question, etc. The more complex the research field, the more important is this realization. Alternation between an explorative phase and confirmative phase is a useful strategy to gradually unwrap the secrets of nature (Fig. 6).

The strength of classical chemometric techniques lies in the explorative phase, where observations from a high number of sources can be viewed simultaneously using pragmatic approaches. However, the distinction between the explorative phase and the confirmative phase is not primarily related to the choice of data analysis.

Confirmative studies come to their full strength after sufficient explorative research has been conducted to ensure that "one is digging where gold is to be found." In functional omics, we often face the situation of "searching for a needle in a hay stack." A humble attitude to this complexity, with a stepwise process alternating between an explorative phase and a confirmative phase, will gradually increase knowledge about the underlying physical mechanisms. To obtain this insight, a pragmatic, patient attitude is useful.

4.22.3.2 Prediction Versus Insight

Research can be conducted with different aims. In some situations, we are interested in generating a model that can be used to predict the outcome of future samples. An example is to develop a data model to predict the probability of cancer from transcriptome records. In other situations, we want to investigate a system to gain insight into the factors and the mechanisms involved in creating variability.

In prediction settings the typical aim is to minimize the prediction error for future samples. The immense task facing the data analyst is to identify the predictor and the set of features with best prediction performance. Too complex models with many input features may give overfitting. Therefore, it is desirable to keep the predictor as simple as possible. Hence, the balance to be considered is between model complexity and prediction power. The analyst should thus seek the smallest possible subset of features yielding maximum prediction performance.

In megavariate data sets where there may be multiple, equal sized, disjoint subsets of features obtaining similar predictive power. The choice of subset may apparently not be crucial for the performance of the predictor as evaluated within one data set. The critical question of concern is how this prediction will behave for future samples. Having in mind the large number of independent sources of variability that may arise in functional genomics, there is a risk of building a model based on features randomly correlated to the response parameters within the data set investigated, whereas the correlation might not hold for future samples. If this is the case, the prediction may fail when used on future samples. Therefore, if possible, a reliable prediction model should be based on the following:

- features that are directly linked to the response by causality;
- features that are genetically linked to the causality factors and can also be expected to be so for future samples (i.e., genes inherited together as a haplotype);
- features biologically linked to some causal factors as they appear as part of the same metabolic chain and can therefore be expected to be linked for future samples.

However, in many situations, the knowledge of the topic under study has not reached the level needed to build a prediction model based on these criteria. In these cases it would be wise to find a more pragmatic predictor. The most important aspect to keep in mind is that it is necessary to perform a long period of testing of the predictor on new samples, as well as to continuously monitor the model. Although this is a general rule for any predictor, the importance of this aspect increases with the complexity of the case under study.

Data modeling with the aim of understanding a biological system is driven by a very different goal than the creation of a prediction model, as the focus is on obtaining insight into all features that can through light on the underlying biology that have given rise to the observed variation in the data. Features that are excluded from a prediction model, as they are not necessary to give a stable prediction model, might still constitute a crucial part of the causality chain leading to the response investigated. When building a model for understanding the system under study, we therefore want to capture all relevant features.

It is also important to consider the phase of the study when choosing validation criteria and significance boundaries for testing the effect of the individual features. In an early phase, the primary focus might be on ensuring that all features of possible relevance are identified, whereas the risk of selecting false-positive features would be less in focus. It may initially be beneficial to allow a large number of false positives to minimize the loss of true positives. Such a prefiltering may help to rule out obviously irrelevant features and reduce the dimensions of the data set before more elaborate methods are applied. This contrasts to a later, more concluding phase of the study, where one might be most concerned with avoiding false-positive results. Furthermore, it may be better to be quite liberal towards false positive if the features are to be interpreted in the light of background knowledge later.

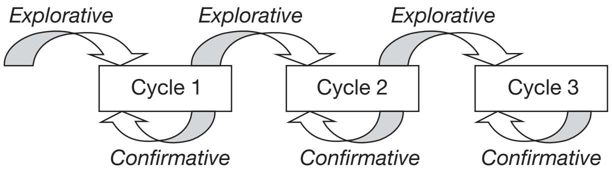


Fig. 6 Design strategy.

4.22.3.3 Considerations on the Experimental Design

Setting up the design of an experiment is critical in all scientific fields. This needs careful consideration[18,19] in view of the complexity of functional omics.[20]

The simplest form of an experimental design would be to vary only one factor and keep everything else as constant as possible. Although such experiments can be very useful, there are important limitations to this approach that need to be considered. The conclusions drawn from such experiments are, in principle, only valid for the particular setting in which the experiment is conducted. As an example, if the aim is to investigate the effects of oxygen availability on gene expression in a bacterial species, and the experiment is conducted under constant conditions on all other factors, for example temperature, the conclusion would then be valid only for the particular environment and the particular temperature used. If there are interactions between the oxygen availability and the temperature, which means that the effect of oxygen availability differs depending on the temperature, the conclusion about the effect of oxygen availability obtained at one temperature cannot be transferred to its effect at another temperature. It would then be more valuable to conduct experiments where both factors were varied according to an experimental plan. One alternative is a full factorial design where all levels of both factors are combined.[18] Then main effects of both factors as well as interaction effects between the two factors can be revealed. Yet, there will be limitations to the interpretation of the results as the conclusion is valid for. The particular environmental setting of other factors used, for example humidity may influence the results, and other genotypes than those investigated may have other genetic make ups that is not covered in the experiment. It is important to keep in mind that a very large number of factors and interactions among factors might be influential on the data generated; hence, there will always be limitations to how general the conclusions drawn from one experiment can be.

A useful strategy for capturing variation in a number of factors while still aiming at keeping the number of samples at a low level is screening designs. A systematic reduction in the experimental plan according to a fractionated factorial plan[18] can be conducted. The cost is that the effect of some factors cannot be separated (confounded effects); the factors that are confounded are defined by the experimental plan. Often only higher order interactions are chosen to be confounded, whereas main effects and lower order interactions are estimated without confounding.

In many situations, we might not be able to directly control the factors we wish to investigate. As an example, we might be interested in testing the effect of one particular gene, potentially involved in the onset of a type of cancer. Obviously, we cannot generate a number of humans constant for all genes except for that particular gene. In some situations, we might not even know whether a phenotypic characteristic of interest has a genetic origin or not, and answering that question might be the first goal of the study.

In situations where the factor of interest is not under direct control, samples are collected to represent the variation of interest, whereas the samples as such might not have any interest. Collecting samples randomly is one strategy, but there are some important pitfalls to that strategy. Consider, for example, research conducted to investigate factors affecting baking quality of wheat. When samples are collected randomly one will typically end up with a correlation between the amount of protein and protein quality given by the compositions of the proteins, as this correlation reflects the market demand for wheat. However, from a scientific point of view we want to break such correlations to investigate the impact of each underlying factors individually. Attention should be paid to avoid selecting samples that create correlations, which are not biologically linked.

Due to the large cost of omics data, there is a need for efficient strategies to capture samples representing variation in the factors of interest. One strategy is to perform high-throughput analysis on a large number of samples and apply some multivariate analysis on this data to select features that span the most relevant information for further in-depth analysis.

An important remark with regard to experimental design, which unfortunately is too often overlooked, is the need for setting up a randomized plan for the experimental execution. In addition to the factors that are systematically varied, there are always a number of uncontrollable factors that can influence the results. These can be temperature fluctuations during the day, day-to-day variability, operator-related factors, technical-related factors, etc. The run order of the samples should be randomized in such a way that the potential influence of systematic, yet uncontrollable factors is minimized. In some cases, there are practical or technical restrictions on the level of randomization, whereas in other cases it may be useful to put deliberate restrictions on randomization for other reasons. Such restrictions on randomization must be taken into consideration when formulating statistical tests for the effects of the experimental factors. Examples of technical restrictions can be illustrated for electrophoresis where 12 gels can be run simultaneously as one batch. The experiment is then performed in blocks of 12, and the experiment is a so-called split-plot experiment.[19] The uncontrolled variation is typically smaller within batches than the variation from one batch to another. If achievable, one might want to put together the samples that one primarily wishes to compare within batches since this will increase the power of the test.

Another important consideration is biological vs. technical replicates. Technical replicates are replicates where the same samples are analyzed multiple times. This is useful for validation of the reproducibility of the analyses techniques that are used, but useless for validation of the experimental factors. Biological replicates, on the other hand, which are the collection of multiple samples of the same material (i.e. the same experimental factor levels) gives a foundation for validation of the experimental factors. More attention should therefore be paid to biological replicates rather than technical replicates.

4.22.3.4 Challenges Related to the Size of Data

The high dimension of the data tables typically generated in functional omics studies can be viewed at two different levels. One is the number of features created by the instruments, which causes technical challenges. The other aspect is the dimension of the subspace reflecting independent sources variability. Functional omics data may be multidimensional at both levels.

Large data sizes create difficulties, which can be classified into three categories related to (1) hardware, (2) operating systems, and (3) software. Hardware-based problems are usually dependent on internal central processing unit and memory management architecture. The second limitation is the operating system itself. Even though the processor may be able to handle large memory spaces, the operating system may not allow it. In addition, some operating systems can also take up a significant part of the memory by themselves, but this can usually be set and tuned manually. Finally, the software being used may simply not be able to handle large arrays. The challenge with a high number of features can be overcome by compressing the information. A far more serious challenge is the high dimension of the sources of variability spanning independent variation. With this in mind the search for the relevant information is not trivial as it would require a large number of samples, which is most often restricted for economic reasons.

4.22.3.5 Multicollinearity

Multicollinearity describes a situation where different features reflect related variation. Multicollinearity is an important aspect of all multivariate analysis. In omics data, many features are highly correlated, regulated by common mechanisms. Any approaches for multivariate analysis have to deal with this multicollinearity. This is often described as the "multicollinearity problem." Typical chemometric approaches are, on the other hand, developed to see this as an advantage as it has strategies to utilize the multicollinearity, as here briefly described.

In regression analysis a response parameter (e.g. the production of ethanol in a fermentation process) is described as a function of some input variables. For simplicity, we consider the expression level of two proteins, x_1 and x_2, where a linear regression analysis of ethanol production is made as a function of the expression levels of the two proteins can be described as:

$$y = \beta_0 + \beta_0 x_1 + \beta_2 x_2 + \text{Res}$$

We consider two situations: one is if the two proteins vary independently of each other, which will give the situation illustrated in Fig. 7A, and the other situation is if the two proteins are strongly positively correlated Fig. 7B, which means they are coordinately regulated. If we imagine sitting in a room, in the former case we will have points spread around in the room whereas in the latter case we will have all points along a line. Performing the regression analysis of a response parameter from these data points, may be viewed graphically as putting a two-dimensional plane on the data points. In Fig. 7A and B, where the samples are seen as dots, they represent nails where the axis of the response parameters is pointed upward from the plot, and therefore not visible as we here look

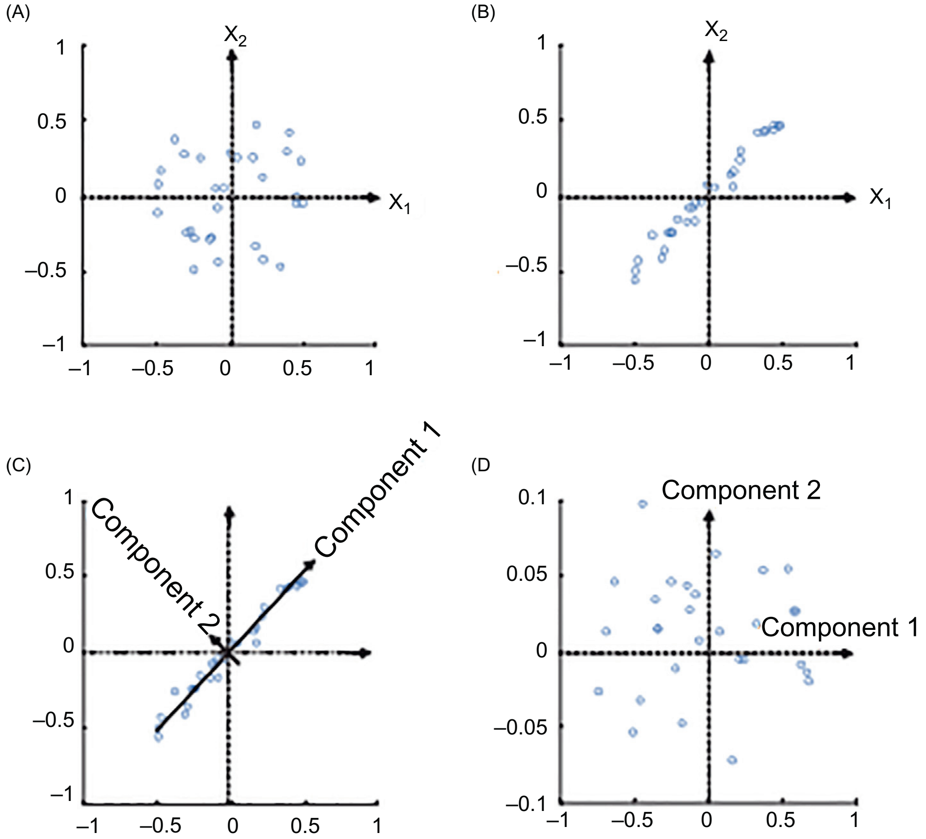


Fig. 7 Illustration of collinearity between two variables, x_1 and x_2, and the changes to the new axes where the new axes spans, in decreasing order, the variability in the data. Here the new axes are orthogonal and the illustrated changes of axis correspond to principal component analysis (see also Fig. 11).

straight down on the x_1 and x_2 axes. If the heights of the nails increase towards the upper right corner, the plane will have a slope reflecting positive relation to both x_1 and x_2. When x_1 and x_2 are correlated, as illustrated in Fig. 7B, the regression will be like balancing a plane on a fence, which will be stable along the fence, but very unstable outside the fence.

A solution to this is to transform the original variables into new variables (Fig. 7C), called component 1 and component 2, where the first typically contains most of the variation. These new variables can be defined as being orthogonal. A regression can subsequently be based on the new orthogonal variables (see Fig. 7D). Thus, we avoid using two highly correlated features in a regression analysis. The transformation of the original variables onto the new variables (components) is mathematically called projection. The components can be considered as reflecting underlying mechanisms, or latent variations, which for example might be a common transcriptional factor that determines the expression of these two proteins. An important realization that lay the fundament for such projection methods is that what we observe is often indirect observations of some underlying phenomenon that gives rise to the observed variation in the data.

Multicollinearity will be the situation whenever the number of features (e.g. the proteome or the gene transcripts) is larger than the number of samples (e.g. patients). The classical statistical methods for data modeling usually require more samples than features, whereas this does not apply to modern functional omics.

As will be described below in this book chapter, the chemometric solution to this is to transform the features into new variables which describe the majority of the variation in the data. The multicollinear problem is by this approach turned over to be an advantage as several features describing the same phenomenon will stabilize the regression. Furthermore, several features together may further unravel the fundamental understanding of the system under study, compared to what is obtained when selecting only a subset of the features prior to data analysis. This focuses on the interpretation and statistical treatment of the underlying phenomenon that gives rise to the variability rather than on the observed features.

In functional omics, we typically have multicollinearity arising both when, for example, instruments give several features on one property, and multicollinearity exists among the different properties analyzed. Regions on the chromosomes are inherited as linked groups of genes (haplotypes) giving correlated responses of different mRNAs, proteins, metabolites, and phenotypic characteristics, and the activities of genes downstream from the genes towards the phenotypic consequences are linked together in complex regulatory or metabolic networks.

By projecting the data to new latent features, we can describe, validate, and interpret some of the underlying common factors giving rise to the variation in the observed features. With a very large number of features, it is nevertheless also important to eliminate irrelevant information and to validate the significance of the observed features. In this chapter, we therefore go through both projecting based approaches to view the underlying common sources of variation, as well as various approaches for judging the relevance of each of the observed features.

All processes in the cells are extremely well controlled. This is a crucial character of omics data compared with multivariate data of non-biological origin, which has major implications on how the data are considered. Often for omics data, only a few features, e.g. a few transcriptional factors or other regulating mechanisms may control the paths from genes to the final phenotype.

Chemometrics is a scientific field that has its origin as a multivariate approach to analyze multivariate data in chemistry, which started back in the 1970s.[21,22] The main aspect of chemometrics is to take a pragmatic approach to model multivariate data, where visualization and interpretation play a key role.[23] Chemometrics is data-driven modeling, which intends to identify, quantify and display the essential relationships - expected or unexpected - within and between data tables.[23] The main modeling tools have been dimension reduction of the data by projection of the original observed features onto a smaller set of features, which are linear combinations of the original features, such as Principal Component Analysis (PCA). These methods treat intercorrelation between measured features as a stabilizing advantage, not as a "collinearity problem," to search for underlying latent features that can be interpreted to unravel underlying mechanisms that drive the observed variation in the data at hand.[23] The strength of the chemometric approach is for such complex omics data more relevant than ever before, which brings the mindset of chemometrics into a new dimension.

In modern technologies, a very large number of features can be measured. The resulting behavior of the system under consideration is often a combination of all the observed parameter that influence the system, and data analytical scientists search for the underlying drivers. The chemometric field has since its origin gone through expansion and development into different scientific fields, including the functional omics field.[24]

In the present book chapter we consider different multivariate data analytical methods that can be valuable for omics data. We put emphasis on maintaining the main definition of chemometrics as a mind-set of pragmatically unraveling underlying structures in the data where visualization and interpretation are key elements. Focus is on methodology that transforms the observed data to underlying pattern of variation that can be interpreted to search for mechanisms that have given rise to the observed data.

4.22.3.6 Causality

An objective for functional omics research is to obtain an understanding of the causal dependencies regulating a biological system. In some situations, such fundamental insight is obtained, however, in most situations, there may be limitations to the extent to which we can obtain understanding of the causal relationships. This problem was discussed as early as 1921 by Wright[25]:

> "The ideal of science is the study of indirect influence of one condition on another in experiments in which all other possible causes of variation are eliminated. Unfortunately, causes of variation often seem to be beyond control. In the biological sciences, especially, one often has to deal with a group of characteristics or conditions which are correlated because of a complex of interacting, uncontrollable, and often obscure causes. The degree of correlation between two features can be calculated by well-known methods, but when it is found it gives merely the resultant of all connecting paths of influence."

The multivariate approach to functional omics is to move gradually towards insight and interpretation of the mechanistic causality with the aim of capturing the relevant factors and latent structures.

Although being complex, an advantage in functional omics data analysis is that the features (genes, proteins, etc.) are extremely well regulated, and the technical noise in the modern omics data may be limited for some of the omics data. The pragmatic attitude with alternation between explorative and hypothesis driven confirmative research can be useful when aiming to obtain the fundamental knowledge from omics experiments.

4.22.4 Data Analysis

All omics data are multivariate by nature. Univariate analyses of omics data are, however, still dominating in the literature, but multivariate approaches increase steadily.[24,26–48]

The measurements from each of the different omics approaches can usually be organized as a data table. The transcriptome of *n* samples may, for instance, be studied using a platform comprised of *p* different spotted transcripts. The data can thus be put into a data table of *n* rows and *p* columns, and we call this a data block. The same samples may further be studied with regard to the proteome, metabolome etc., yielding new blocks of data with *n* rows representing the *n* samples and a varying number of columns, which may be called "variables" or "features." Moreover, the experiments may be conducted according to some prespecified design (a design block) and finally, a set of categorical or continuous phenotypic measures (a phenotype block) may also be available; see Fig. 8.

Occasionally, we want to connect two blocks of data, whereas in other situations we want to connect a whole chain of blocks in an integrated data analysis, as illustrated in Fig. 8. The latter is a multiblock situation where several blocks of data are to be connected. In this example, the blocks $\mathbf{X}_1$–$\mathbf{X}_6$ are aligned along a line where all blocks have the same number of rows representing the different samples. The arrows in Fig. 8 reflect a natural ordering of the blocks along the chain from gene information to phenotypic output.

It is also highly relevant to consider situations of having several blocks of data spanning in different directions. Such situations arise, for example, when we have additional information on the individual columns of one or several blocks of data. This could be background information of the individual genes or transcripts, for example interactomes between genes or proteins, as indicated by the additional blocks $\mathbf{X}_7$ and $\mathbf{X}_8$.

A typical objective for data analysis is to explain observed variability in some response features ($\mathbf{Y}$) by a functional relationship with a set explanatory features ($\mathbf{X}$). For example, in a bacteria culture the explanatory variable may be the expression of different proteins and the response may be the quantity of ethanol that it produces, and we want to see how the expression of different proteins affects the ethanol production. The model may then be expressed as:

$$\mathrm{Y} = \mathrm{f}(\mathrm{X}) + \mathrm{Res}$$

The variation that cannot be explained by a function of X, is the residual term (Res). Some refer to the residual as the "error," but the name may be somewhat misleading since these are not necessarily "errors" in a biological or technical sense. In our model, we might not succeed in explaining all observed variability in the response features, and this excess variability is captured as residuals. By including more samples or more relevant features, the model may be improved and the proportion of the variation in the residual may be reduced. It is important to keep in mind that our data models will always be heavy simplifications. Important factors are likely to be overseen, and non-influential factors are likely to influence our observed data.

4.22.4.1 Preprocessing of the Data

Prior to analyzing the data, there are several preprocessing steps that should be considered. In data analysis, the phrase "garbage in, garbage out" is often used to highlight the importance of proper preprocessing. Meaning you cannot get good, meaningful results

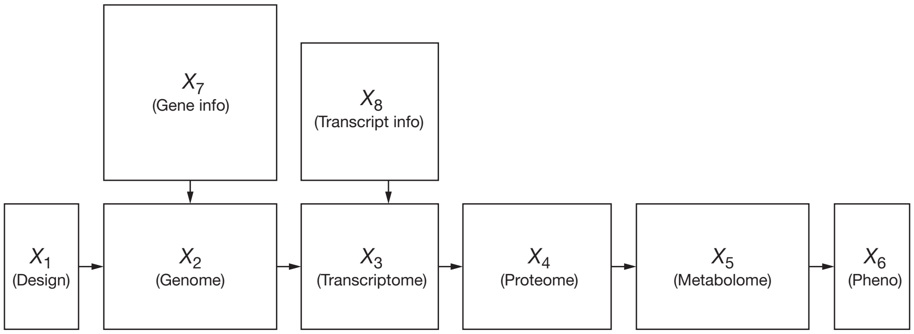


Fig. 8 A flowchart of different blocks of data in functional omics, from the experimental design, through the genome, transcriptome, proteome and metabolome to the final phenome (Pheno).

without quality data. Preprocessing typically includes handling of missing values, scaling the data, encoding categorical features, etc. There are many ways to implement the preprocessing methods described in this section, examples of how to implement them in Python is included in Additional file A3. Datasets in omics data tend to have far more variables than samples. For example, genomic data will contain tens of thousands of genes, but often only tens or hundreds of patients. When analyzing datasets of this kind, the model will often overfit to the input data. This means that the model will be very good at predicting the input data but may not generalize to new data. In other words, the results found may only apply to the patients in the analysis, not the population as a whole. To ensure that the results are generalizable, it is important to have a separate dataset that is not used in the analysis but is used afterward to evaluate the performance of the model. Collecting new data to evaluate model performance is always ideal; however, this is not always a practical possibility. The data can instead be divided into a training set and a test set. The test set should be kept completely separate in order to avoid information leakage that could cause the model to fit somewhat to the test data. In other words, the evaluation of the model may appear to generalize better than it actually does. An example of how to divide the data into sets is shown in the appendix. Once the two sets are created, preprocessing and analysis can be performed on the training set.

Most data analytical methods assume that the residual term is the same at all levels of the input features. Within one feature, this means that the higher quantities within the feature do not have more error than the error obtained at small quantities of the same feature. It also implies that the random error is the same for highly abundant features that have large variation in the expression pattern as for features that display little variation. That is most often not the case as highly abundant proteins most often also have larger random variation. This is an important issue that needs to be handled. We therefore pay attention to this here. If the random variation increases with the magnitude of the protein expression, there is a multiplicative effect between the random error and the expression of proteins that might be expressed as X*Res. One approach to solve this is to use a log function. By applying logarithmic tools, a multiplicative expression turns into sums:

$$\log(\mathrm{X} * \mathrm{Res}) = \log(\mathrm{X}) + \log(\mathrm{Res})$$

By log transformation we may achieved a model where the residual term is independent on the expression level of the proteins. Such models, called homoscedastic models (same variance), are much simpler from a data analytical point of view than models where the residual term, for instance, interacts in a multiplicative manner with the deterministic part of the model. The latter is called heteroscedastic models.

Log transformation is very often used in omics data. One important issue to be aware of is that the log function cannot be applied on zeros, and small values between 0 and 1 can be blown up in magnitude. One solution that may be used when the majority of the data have large values where only a few are below 1, is to add a small value to all numbers, for example 1 or 1.5.

Zeros in the data needs special attention. Many omics instruments leave the value 0 when it fails to detect if a protein or gene transcripts, etc., is present or not. Thus, the value 0 might not mean that the protein or gene transcript truly is observed to be absent in that particular sample. It is important to be aware that 0 is a number that might have major impact. In fact, it is the most extreme value on the scale if all other values are positive. The first thing to do on any type of data is to identify all zeros and reflect whether these zeros are true absence of a feature, or if it is failure in observing the true value that might be there. If the latter is the case, then the zeros should be replaced by missing values, or one may impute a relevant value. One alternative is the mean over other replicates of the same sample. Imputing a value may be useful as many data analytical tools to not allow missing values.

Another common transformation in omics data is the square root transformation. The square root transformation is able to help correct data that is skewed right, bringing it closer to a normal distribution. Unlike log transformations, square roots are able to handle zeros and small numbers (between 0 and 1) without any trouble. A square root transformation will be a less extreme than a log transformation. When the residual increases by the magnitude of the expression level of a feature, a power transform is a family of functions that are applied to create a monotonic transformation of data. This can stabilize the variance, make the data more normal distribution-like, and improve the validity of measures of association between different features. One type of power transformation is called Box-Cox transformation which have a tuning parameter lambda to switch between square root, log and other transformations. When lambda is zero, the result will be a log transformation; and when lambda is 0.5 it will be a square root transformation. Using an optimization function, lambda can be optimized to make the data as close to a normal distribution as possible. Log transformation, square root and optimization of these by Box-Cox transformation is illustrated in Fig. 9 on data that have distinct different distributions.

The data are often mean centered by subtracting the means of each columns for all elements in the data table. Mean centering prior to multivariate analysis implies that we consider changes relative to the mean. Graphically this can be seen as the data points then circles around zero, as was applied in the illustration of collinearity in Fig. 7. If there are missing values in the data, it is important to be aware that centering the data may introduce errors, and not just remove common offset.[49] To what extent this will affect the results depends on how severely the data are embedded with missing values.

Two main types of scaling include normalization and standardization. Normalization refers to rescaling each feature to a range from 0 to 1, but features will not have unit variance. Standardization uses centering of each feature with a mean of zero and unit variance which results in z-scores. For example, in proteomics, one may consider changes in proteins that are small in abundance to be equally important as changes in abundant proteins. Enzymatic proteins are small in abundance but may have critical controlling function on the biological processes. If proteomes are analyzed without scaling to unit variance, the more abundant proteins, such as structural proteins, muscle proteins etc., may dominate the resulting pattern. If the features are on different scales, e.g. some are measured on the level of gram and others on the levels of kilogram, one has to perform scaling prior to multivariate analysis. In other situations, one may consider the abundance of the features to be more important than features with low expression. The decision of scaling or not must be considered for each data at hand and according to the purpose of the study. Scaling can be applied after a log or square root transformation, but not before.

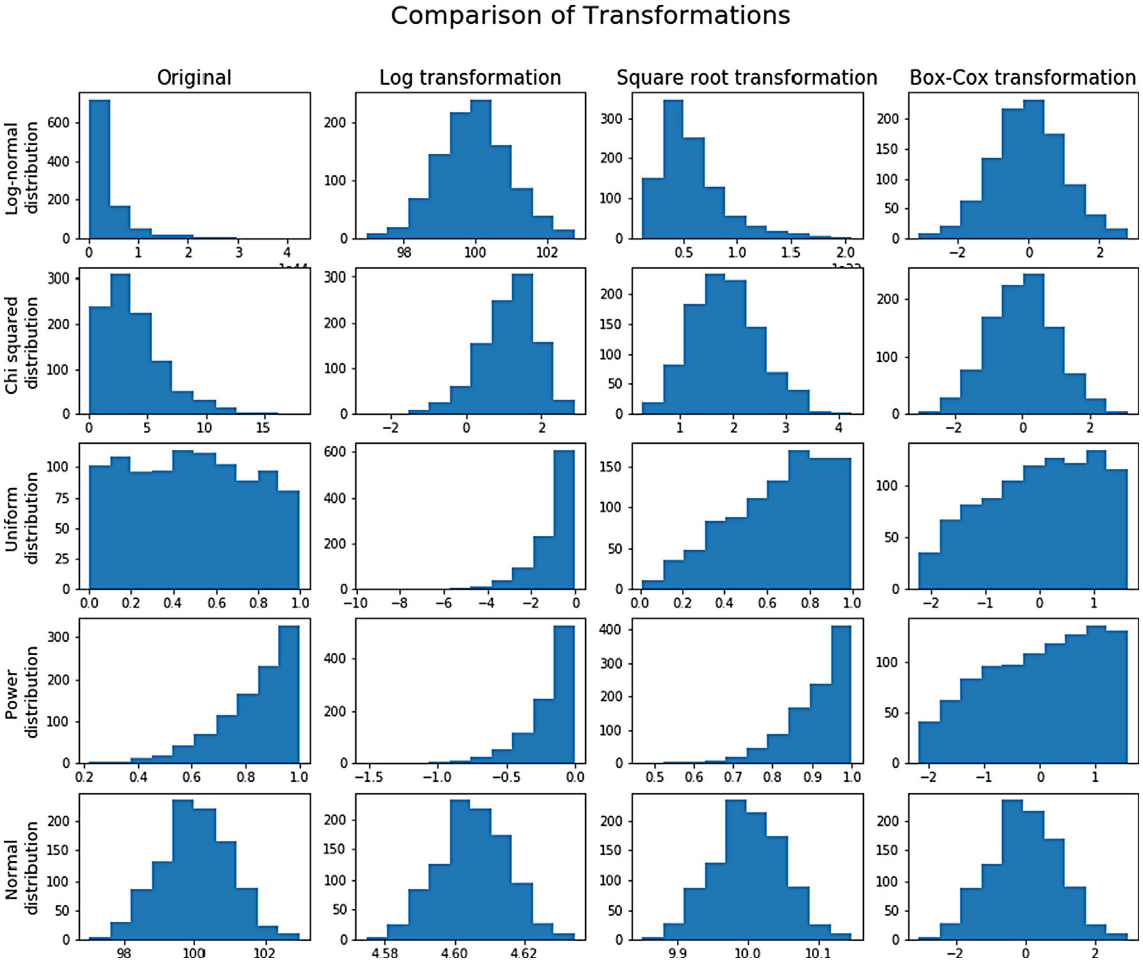


Fig. 9 Illustration of the effects of transformations on various distributions. Column one shows the original distributions in ascending order: log-normal, chi-squared, uniform, power, and normal distribution. Columns two through four show the resulting distribution after a log, square root, or Box-Cox transformation has been performed on the original data.

4.22.4.2 Data Used for Illustration of Data Analysis

In the following, we will use a set of experimental data (Box 1) to illustrate some of the methodological approaches. For practical guidance, a script is provided in r that can be copied into R and run through (Additional file A2).

4.22.4.3 Exploring the Variation Patterns Within a Block of Data

Before relating different blocks of data to each other it is advisable to investigate the structure within each data block. In the *Campylobacter jejuni* data (Data set 1, Box 1) a data table of the transcriptome display in each element the abundance of a particular gene transcript for a particular sample. As the microarray data are ratios of transcript, the value of 1, which is equal to zero for log transformed data, would reflect no changes in gene activity. Larger values implies genes being upregulated compared to the starting day, day zero, as response to the change in the environmental factors. Likewise, smaller values imply downregulated genes.

On the *Camphylobacter Jejuni* data we first made explorative analysis, which implies that the data program is not provided with any information in advance about the experiments or any hypothesis, and the program itself finds the connections in the data. Explorative analysis was first performed as cluster analysis. Cluster analysis is a useful approach for studying similarities and differences among samples,[42] where the samples are clustered into different groups according to some measure of similarity across the observed features. In a similar way, features are clustered according to their similarity across the samples, which results in two-way clustering, often visualized as a heat map. Expression heat map[56] is frequently applied clustering approach and visualization tool for omics data from transcriptomic, proteomic or metabolomic experiments. For the *Campylobacter jejuni* data (Data set 1, Box 1), a two-way cluster analysis is performed and the results displayed as a heat map in Fig. 10. The order of the columns reflects the degree of similarity between the samples, and likewise, the order of the rows reflects by the similarity between genes.

Box 1 Data sets used for illustration of the methods

Data set no. 1. A study of the bacterium *Campylobacter jejuni*—a bacterium of major concern for food safety. The experiment was conducted to unravel molecular mechanisms of the survival as that is of fundamental importance in preventing transmission of this bacterium through the food chain.

- Transcriptome and FTIR measurements of the food-borne pathogen *Campylobacter jejuni*.[50,51]
- *C. jejuni* is a microaerophilic and thermotolerant pathogen.
- Optimal growth conditions are low oxygen tension and relatively high temperature (microaerophilic conditions at 42 °C), adapted to its niche, which is the birds' intestinal tract.
- During transmission to the human host, the bacteria is exposed to atmospheric oxygen (in the environment, after poultry slaughter, during transformation, etc.). It is also exposed to reactive oxygen inside the host because macrophages use reactive oxygen as a defense against pathogenic microbes.
- The study was conducted to study survival mechanisms under stress conditions by observing changes in transcriptome and spectroscopic data in a time record over 1 week under different stressed conditions.
- The experimental design factors were: 2*3*3 full factorial design: 2 temperature levels (5 and 25 °C), 3 levels of oxygen availability (anaerobic, microaerobic, and aerobic), 3 time points of data collection (2, 4, and 7 days), 3 biological replicates, and 2 technical replicates. A full data set where means are taken over the technical replicates would be of size 3*3*2*3 = 54 samples. Some data points are, however, missing due to survival challenges, leaving 43 samples in the experiment.
- DNA microarray data are given as a ratio where the level of expression is divided by the expression at day 0. If the data are log transformed, the ratio is a subtraction. Ratios cannot be taken when the value is 0 at day 0.

Data set no. 2. Two human studies on obesity, diabetes and bariatric surgery. The aim of the study was to identify novel candidate genes that may regulate adipose tissue function, and the implications of diabetes.

- Transcriptome and clinical observation of obese patients (BMI, kg/m2) with and without diabetes from a Norwegian cohort[52] and a Mexican cohort[53] subjected to bariatric surgery.
- In the Norwegian cohort there were 8 patients with and 8 patients without diabetes and samples of the fat layer beneath the skin (the subcutaneous fat tissue) were taken before and 1 year after the bariatric surgery. Clinical data were observed on Body Mass Index (BMI), High-density lipoprotein cholesterol (HDL) and triglyceride (TG).
- In the Mexican cohort there were 14 patients with and 13 patients without diabetes, and the sampling of the subcutaneous fat tissue for transcriptome analysis was taken under the bariatric surgery.

Data set no. 3. A study of *Lactobacillus sakei*, a bacterium used in food fermentation. The aim of the study was to achieve molecular understanding of differential behavior of two different strains of *Lactobacillus sakei* upon reduced glucose availability.

- Transcriptome, proteome and end-products observed from an experiment with two strains of *Lactobacillus sakei*, LS25 and 23K, cultivated in a continuous fermentation set-up (chemostat) at two different flow rates, where glucose was set as the limiting nutrient.
- The aim of the study was to identify molecular mechanisms of the different behaviors of the two strains[54] in relation to glucose availability.
- The experimental design factors were a 2*2 full factorial design with the design parameters: 2 strains (23K and LS25) and 2 growth conditions (high and low flow rate in a glucose limited chemostat), with 3 biological replicates without any missing values. This gave a total of 12 samples for the data analysis.

Data set no. 4. A study of Salmon pancreas disease caused by salmonid alphaviruses. The aim of the study was to identify the biological basis for genetic resistance to the disease.

- RNA-seq was performed to detect differential gene expression to test differential responses to injection (IP) by the virus and co-habitation with the virus (CH) in animals with high and low genomic breeding values for disease resistance.[55]
- Here we present results of genes that gave more than 1.75-fold change in high vs. low genomic breeding value fish both in the naive (uninfected) state, and 4 weeks after injection and co-habitation with the salmonid alpha virus causing pancreas disease.

As the optimal condition for this bacteria is microaerobic condition at 42 °C, all conditions in this experiment imply stress for the bacteria to study molecular survival mechanisms under stressed conditions. The data are analyzed as a change compared with the initial non-stressed condition by considering fold change of the observed value vs the initial value. The majority of genes were downregulated as result of the stress conditions applied in this experiment, whereas a minority of genes involved in energy metabolism, cell envelope, transport, binding, and genes encoding chaperones were elevated under most conditions (Fig. 10).[50] The interpretation is that a possible survival mechanism for this bacterium is an active process where as many genes as possible are downregulated to save energy and to prioritize genes involved in energy metabolism and modification of the cell wall components. There were a strong interacting pattern between temperature and atmosphere condition. Aerobe condition at 25 °C, had the highest expression of a number of genes. Under this temperature, the bacteria survived only for the first harvest day (to the left in the heat map). This contrast to anaerobe condition at 25 °C, which lead to the strongest downregulation of most genes as seen by the green color for most genes at this condition. The differences related to the atmosphere condition was smaller for the bacteria grown under the lowest temperature of 5 °C.

To further visualize the main variation in the transcriptome data, the multivariate explorative data analysis, Principal Component Analysis (PCA)[42,47,57,58] was conducted. PCA is graphically illustrated in Fig. 11. PCA can be estimated by the NIPALS

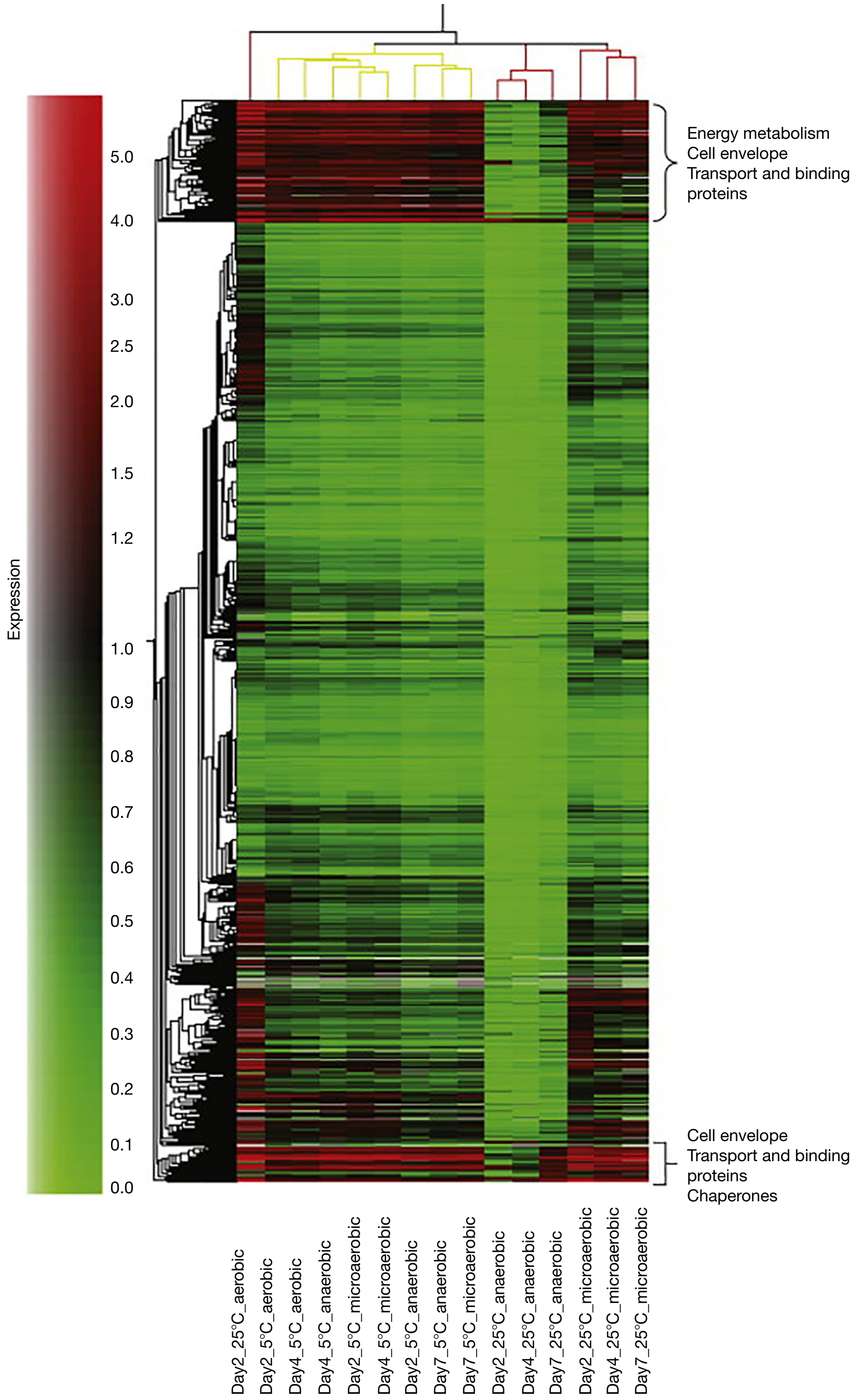


Fig. 10 *Campylobacter jejuni* (Data set 1, **Box 1**). 2D clusters of transcriptome and growth condition. Along the vertical lines are the samples and their respective experimental conditions marked along the axis below the figure. Vertically are all the expression levels of each gene. Horizontal are the experimental conditions. Each element thereby displays the expression of a given gene which is clustered according to their similarity as indicated on the left-hand side of the figure, and for each sample which are grouped according to similarity as indicated above the figure. The yellow lines on top of the diagram represent samples from 5 °C, and the red lines to the right represent samples from 25 °C. The diagram is a heat map where upregulated genes are shown in red color, repressed genes in green, and unchanged genes in black. Correlations were used as the measure of similarity. Reproduced with permission Moen, B. et al. Explorative Multifactor Approach for Investigating Global Survival Mechanisms of *Campylobacter jejuni* under Environmental Conditions. *Appl. Environ. Microbiol.* 71, 2086 (2005).

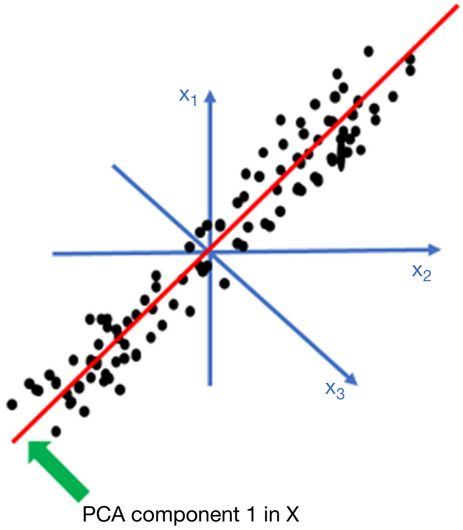


Fig. 11 Graphical illustration of the explorative multivariate analysis, Principal Component Analysis (PCA) performed within the data block X, which for example could be the expression of three proteins (x_1, x_2 and x_3). The red solid line within each data block shows a latent variable (component) estimated as a linear function that describe maximum of the variation within the data block. The latent variables are in PCA often called Principal Component (PC) or just component. After subtracting the first component, a new component can be estimated.

algorithm as explained in Box 2, or it can be mathematically calculated using Singula Value Decompositin (SVD), as described in Box 3. The estimation of PCA by NIPALS considers variances based on a statistical framework, whereas calculation of PCA by SVD is based in mathematical science. The solution is the same when there are no missing values. PCA is an explorative method where linear combinations of the original variables, called Principal Components (PC) or just components, are estimated to describe in decreasing order the main variation in the data. After estimating the PC that describes the maximum of the variation in the data, this component is subtracted, and a new component is estimated in the same way to describe the maximum of the variation that is left etc. In a given data table the number of PCs that can be estimated is one less than the minimum the number of columns or rows in the data table; the smallest number sets the limitation. In a data table of 12 samples of 700 proteins, one may estimate $12-1=11$ PCs. This reflects a very important point. Although 700 features are observed, the true dimension of the data is limited by the number of rows. This means, most of the observed features are closely connected, reflecting some underlying phenomenon that have given rise to the observed data. The aim of investing such highly correlated data should therefore search for obtaining understanding of this underlying phenomenon. As the PCs are estimated to describe in decreasing order the variation in the data, the first PCs will cover most variation and may be the most relevant, although that might not always be the case. Sometimes rare phenomena in the data may account for only a small proportion of the total variation in the data table. The different PCs can be considered as latent features that describe different phenomenon in the data. The PCs are linear both with respect to the features (the columns in the data table) and with respect to the samples (the rows in the data table), hence the name of this family of methods is bilinear modeling. The bilinear methods move the attention from each observed feature to the pattern of variation the observed features together describe.

A mathematically description of a linear model and a bilinear model is given in Box 2. A linear model describes the relation between the input variable (**x**) and response variable (**y**) as linear, which means that a change in the input gives a change in the response which is linearly related to the input in a magnitude given by the regression coefficient *b*. It is here assumed that the input and output variables are mean centred. If there are multiple input variables, for example different proteins, and the input is a data table with samples as rows and proteins as columns there will be one regression coefficient for each input variable i.e. one *b* for each column in the data table. The linear model describes a linear relation between the columns in the input data table and the response parameter (Box 2).

A bilinear model of one block of data X estimates components that are linearly related both to the columns and to the rows (Box 2) with two parameters to be estimated. One approach is to first select a random column, and consider this as a first estimate of the parameter **t**. With the parameter **t** considered as fixed, the other parameter **p** is estimated. And then set the estimate of the parameter **p** as fixed to estimate the first parameter **t**. This can iterate until the algorithm finds a stable solution for both parameters. **t** is a parameter of the rows, often called scores, and **p** is a parameter of the columns often called loadings. When the data table has samples as rows and features as columns, the scores are parameter for the samples and loadings are corresponding parameter for the observed features. This approach to estimate the coefficients is called NIPALS algorithm.[21]

The NIPALS algorithm has a useful property where there are missing values, as the missing values are estimated under iteration of the estimation of **t** and **p**. PCA performed by NIPALS can therefore tolerate missing values. Nevertheless, missing values might affect the results if they are abundant and important, so care must be taken in the interpretation when the data contains many missing values. Another method for estimation of PCA is singular value decomposition (SVD). This method will stop if there are missing

Box 2

A linear model

$$\mathbf{y} = \mathbf{x} * b + \text{residuals}$$

x and **y** are two columns of data, input (**x**) and output (**y**), respectively.
y is linearly related to **x**
b is a regression coefficient estimated from the data, which describes how **y** are related to **x**
b is the increase/decrease in **y** when **x** increases by one unit
Given a column of **x** and a column of **y** => find *b* as: *b* = **y**/**x**

A bilinear model – Principal Component Analysis (PCA) by NIPALS

$$\mathbf{X} = \mathbf{t} * \mathbf{p} + \text{residuals}$$

X is a data table of observed values
Both **t** and **p** are parameters estimated from the observed data
t is a parameter called "scores" of the rows in X
p is a parameter called "loadings" of the columns in X
The parameters can be calculated by different methods, one method is NIPALS.
PCA explained by the principles of the algorithm NIPALS.
NIPALS is iterative, switching between estimation of loadings of the features, and scores of the samples

$$\text{Start with a random column of } \mathbf{t} \text{ estimate } \mathbf{p} \text{ as : } \mathbf{p} = \mathbf{X}/\mathbf{t} \text{ in matrix notation : } \left(\mathbf{t}'^{*}\mathbf{t}\right)^{-1} \mathbf{t}'/\mathbf{X}$$

$$\text{Use the estimated } \mathbf{p} \quad \text{estimate } \mathbf{t} \text{ as : } \mathbf{t} = \mathbf{X}/\mathbf{p} \quad \text{in matrix notation : } \mathbf{X}^{*}\mathbf{p}\left(\mathbf{p}'^{*}\mathbf{p}\right)^{-1}$$

Use the estimated **t** to improve the estimate of **p** and
use the new estimate of **p** to improve the estimate of **t**
Repeat this until convergence.

The solution is the first Principal Component (PC): $\mathbf{t_1}^*\mathbf{p_1}$
which graphically describe the largest variation in the data.

The variation along the direction described by the first PC is subtracted from the data,
and the same procedure is repeated to find the second PC.

The number of PC that can be estimated is limited to the smallest number of rows/columns in the data.

The PCs can be interpreted as describing underlying phenomenon in the data based on the consideration that features and samples that are closely correlated eflect some common underlying causes of the variation

values. The conversion from SVD to PCA is given in Box 3. By SVD the parameters **t** and **p** are calculated as eigenvectors. Althoug bilinear methods are linear methodologies, they can solve non-linear challenges by including more components to account for the non-linearity.

PCA performed on the gene expression of the *Campylobacter jejuni* data (Data set 1, Box 1) is displayed in Fig. 12A and B. The data were mean centred to eliminate offset of the features, which puts the focus on relative changes among the samples. As some of the combinations of factors were missing because the bacteria did not survive, one needs to be aware of the missing values under the interpretation.

The plot of scores of the samples for the two first PCs will give a view into the main structure of variability of the data. Likewise, the corresponding loadings of each feature are visualized. The position of the samples in the score plot and the position of the features in the loading plots should be interpreted together.

PCA of the transcriptome data display negative loadings for the first and most important PC. The loadings of nearly all gene transcripts are located towards the left. In the corresponding score plot most samples are located towards the same side as the majority of the gene transcripts. As the data were centered prior to PCA, this must be interpreted in relative terms. As revealed by the histogram and the cluster diagram, most genes were downregulated. What we see in Fig. 12 is a variation in the general expression of the genes compared with the means of all samples. In the score plot of this PC, one of the growth conditions deviated

Box 3 Singular-value decomposition (SVD) and Principal Component Analysis (PCA)

SVD of a matrix **X** of size (*n* x *p*) is described as $\mathbf{X} = \mathbf{USV'}$, where
U is the eigenvectors of **XX'**, where **XX'** is of size *n* x *n*
V is the eigenvectors of **X'X**, where **X'X** is of size *p* x *p*
S is the common singular values sorted in decreasing order, where **S** is diagonal(**s**) of length the minimum of *n* and *p*
U captures information on the rows of **X**
V captures information on the columns of **X**
S describes the eigenvalue of each eigenvector, i.e., the relative size of the different components
$\mathbf{X} = u_1s_1v_1 + u_2s_2v_2 + u_2s_2v_2 + \ldots$
Transforming SVD to PCA:
the scores **T**, describe information of the rows
the loadings **P**, describe information of the columns:
$\mathbf{T} = \mathbf{US}$
$\mathbf{P} = \mathbf{V}$
Each score vector by convention is scaled to also capture information on the relative proportion of the variation accounted for by the respective PCs, whereas the loading vectors are all scaled to the length of 1.

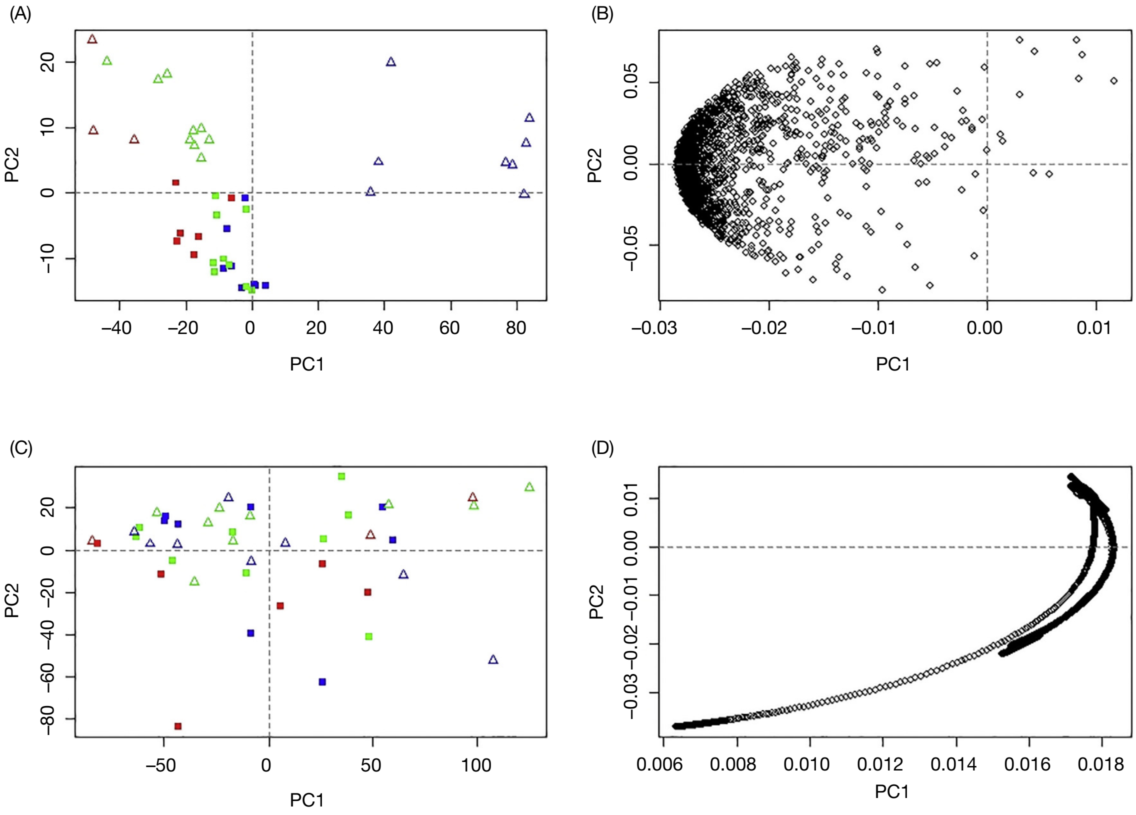


Fig. 12 *Campylobacter jejuni* (Data set 1, **Box 1**). Principal Components (PCs) PCA of (A, B) transcriptome records and (C, D) FITR data. (A, C) score plots of the samples for the first two PCs where the atmosphere is color-coded by anaerobic (*blue*), micro aerobic (*green*), and aerobic (*red*) at temperature 5 °C (filled squares) and 25 °C (open triangles), at 2, 4, and 7 days after the onset of the experiments with three biological replicates. (B, D) loading plots of the transcriptome data (C) and of the FTIR spectra (D), which is to be interpreted along with their corresponding scores (A and C), respectively.

from the other conditions, which was the combination of anaerobic condition at 25 °C. This gives the interpretation that bacteria grown under anaerobe conditions at 25 °C gave stronger reduction of the expression of most genes compared with other conditions, and this corresponds to the pattern of variation seen in the heat map for this growth condition (Fig. 10). The second PC reflects effects of temperature on the gene expression that is independent of the pattern of variation described by the first component where samples grown at 25 °C are located upwards and samples grown at 5 °C are located downwards (Fig. 12). The second PC, which accounts for only a smal part of the variation in the transcriptome, cannot easily be seen in the heat map (Fig. 10) as it was not a domination pattern in the expression data. Thus, rare patterns were better visualized by PCA than by heat map, as the first dominating pattern, described by the first PC, is subtracted, which gives a view into the remaining pattern of variation of the data.

PCA plots of the Norwegian cohort of the human study (Data set 2, Box 1) are displayed in Fig. 13 after $\log_2$ transformation, mean centering and scaling to unit variance. In the score plot of this study most samples prior to bariatric surgery are located at one side of the plot (to the *left*) and all samples taken post bariatric surgery are located to the opposite side of the score plot (to the *right*). Thus, changes in the gene expression as a consequence of the bariatric surgery is the most dominating pattern of variation in these data. Among the genes spanning variation in this direction was the expression of IL6, which encodes the gene interleukin-6, located towards the left along with the samples pre-surgery (located among the cloud of genes to the left loadings, this particular gene is not visualized in the plot). The expression of this gene was strongly reduced upon the bariatric surgery. Interleukin-6 is known to contribute to the immune reactions. It is also known that dysregulated continual synthesis of interleukin-6 has a pathological effect on chronic inflammation and autoimmunity. The expression of this gene is described to be linked to obesity.[59] Obesity is known to promote low-grade inflammation by activation of immune cell subsets in the white adipose tissue, which includes elevated expression of the gene IL-6. Thus, the bariatric surgery improved the situation with respect to the expression of this gene. Although we here only mention one gene, in any multivariate data, the combined action of genes should be interpreted to unravel the underlying molecular pattern of the observed changes. The changes that are here observed from the gene expression studies resulting from the bariatric surgery, is reduction in inflammatory condition that was high in the obese condition.

Surprisingly, PCA did not display any pattern of variation related to diabetes (red vs. blue marks in the score plot). As will be presented at the end of this book chapter, there were strong relevant effects of diabetes on specific genes, but supervised multivariate modeling was needed to detect these effects. Interestingly, univariate analysis revealed the strong effects of bariatric surgery on the gene expressions,[52] whereas the univariate method failed to detect any differences in the gene expression related to diabetes vs. non-diabetes.

In the *Lactobacillus sakei* case (Data set 3, Box 1), where bacteria are grown at two growth conditions with different glucose availability, the fermentation shifted from homolactic fermentation with mostly production of lactate, to heterolactic fermentation with a mixture of lactate, formate, acetate and ethanol. For the mean centred data, lactate is reduced when glucose availability is reduced, whereas the formate, acetate and ethanol increases. This change was stronger for the strain LS25 than the strain 23K, thus there is an interaction pattern between strain and growth condition. PCA of the end-product mirrored this pattern of interaction by the first PC, which accounted for as much as 98.9% of the total variation in the end-products. The score plots of the end-product (Fig. 14A and B) revealed that the strain LS25 changed more than the strain 23K along the first PC, which is seen by the larger distance between low and high growth rate for the strain LS25 (in red) than for the strain 23K (in blue). In the score plot, samples grown under the initial glucose availability (displayed as filled squares), are all located towards the left, whereas the samples grown under reduced glucose availability (displayed as open triangles) are located towards the right, and in the corresponding loading plot lactate is located towards the left and the others towards the right. This visualizes that at the initial growth conditions lactate was high, whereas at reduced glucose availability lactate decreased and the other end-products increased. This change was more pronounces for the strain LS25 than for the strain 23K.

Transcriptome and proteome were performed to investigate the molecular mechanism for the differential behavior of the responses of the two bacteria upon reduced glucose availability. As presented in the original publication of this data,[54] PCA, which

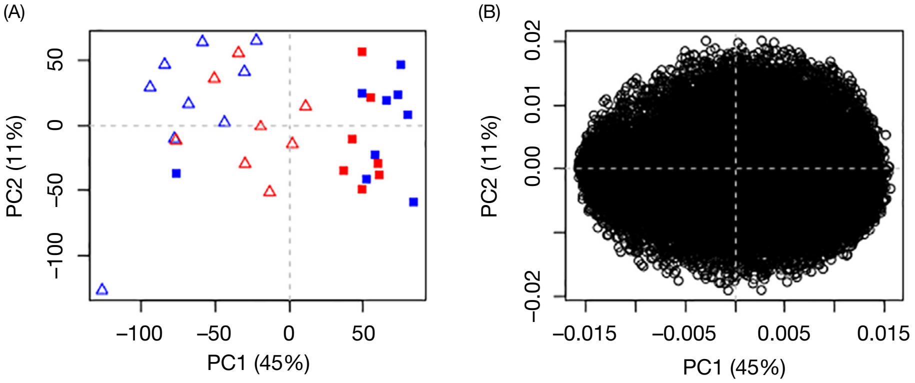


Fig. 13 Human data on obesity, diabetes and bariatric surgery (Data set 2, Box 1). PCA of transcriptome records for the Human data of obesity, diabetes and bariatric surgery. (A) 2D scatter plot of the score plots of the samples for the first two PCs, where the patients are color-coded by diabetes (in *red*) vs. non-diabetes (in *blue*) of samples taken before bariatric surgery (open triangles) and 1 year after the bariatric surgery (filled squares). (B) the corresponding loading plot of the transcriptome data for the same PC. For interpretation, the genes located towards the right are higher expressed in samples located towards the right in the score plot (i.e. elevated after the bariatric surgery), whereas genes located towards the left are highest expressed in the samples located towards the left in the score plot (i.e. downregulated by the bariatric surgery). The data were mean centered and scaled to unit variance prior PCA to let all transcripts have the same impact on the analysis, and all the transcripts are circled around zero in the loading plot.

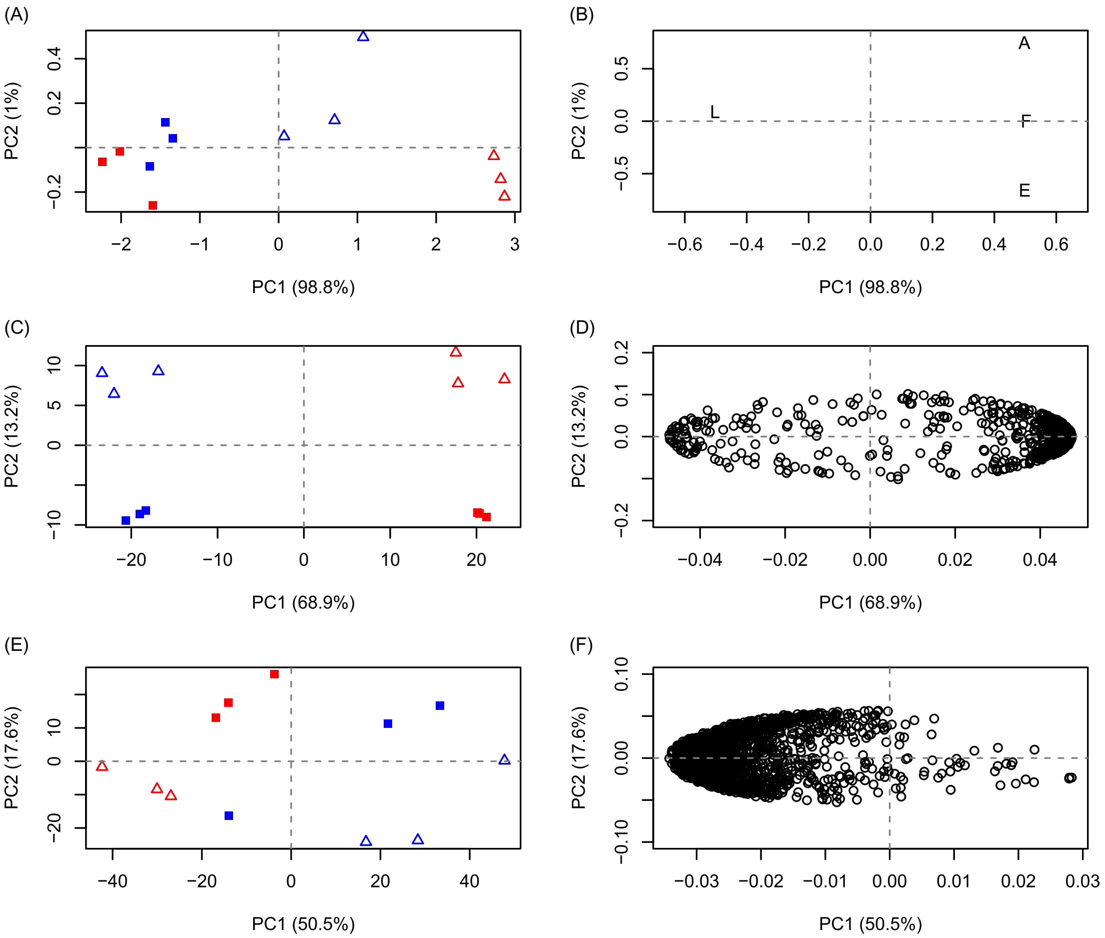


Fig. 14 *Lactobacillus sakei* (Data set 3, **Box 1**). (A and B) PCA of the end-products, (C and D) PCA of the proteome, (E and F) PCA of the transcriptome. (A, C and E) Score plots of the samples, (B, D and F) loading plots of the variables. In (B): Lactate (L), formate (F), acetate (A), and ethanol (E). The score plots of the samples display strain 23K (*blue*) and the strain LS25 (red) grown at high (filled squares) and low (open triangles) glucose availability. The percentages on the axis describe the proportion of the variability described by the different PCs.

describes the main variation within each data block did, however, not display the same interacting pattern of variation as the end-product by any of the PCs. Fig. 14C and D display the two first PCs of PCA of the proteome, and Fig. 14E and F display the two first PCs of PCA of the transcriptome. This suggests that the differential behavior of the two strains, which is so strong in the end-product, is scantly represented in the transcriptome and in the proteome. Only a few features responded to changed growth conditions as shown in the original publication of this data.[54] This shows an important aspect of omics data, which is the phenomenon that a few key features may dominate the downstream processes.

PCA may be conducted on any data table with meaningful rows and columns. The data set on Atlantic salmon pancreas disease was for an experiment using RNA-Seq to test differential gene responses in the heart to injection by the virus (IP) and co-habitation with the virus (CH) in individuals with high and low genomic breeding values for pancreas disease resistance (Data set 4, **Box 1**). The data table we created included all genes identified in the original publication of this data[55] to have a fold change above 1.75 in the comparisons of the breeding values, i.e. the ratio of the observation 4 weeks after injection and cohabitation with the virus versus the initial starting values. There were 42 genes that fulfilled this criteria. In the present analysis we let the fold change values under the three experimental conditions; naïve, IP and CH be the column and the genes be the rows in the data table. The first component, which accounted for the majority of the variation in the data, shows that all the fold change values for breeding values were higher for the genes located towards the right in the PCA plot than those located towards the left (**Fig. 15**). Furthermore, a difference between the experimental condition was detected along the second PCA. Towards the top of the loading plot is the injection approach and among the genes in the corresponding score plot is mmp13, which is an extracellular matrix degrading enzyme, that is consistently activated in pathogen infected Atlantic salmon.[55] The PCA plot suggest that the elevation of this gene (and other genes located in the same direction on the PCA plot) is more prominent after injection (located upward along PC2) than in the naive state (located downward along PC2), and also when compared with infection by co-habitation (located slightly downward near the centre along PC2).

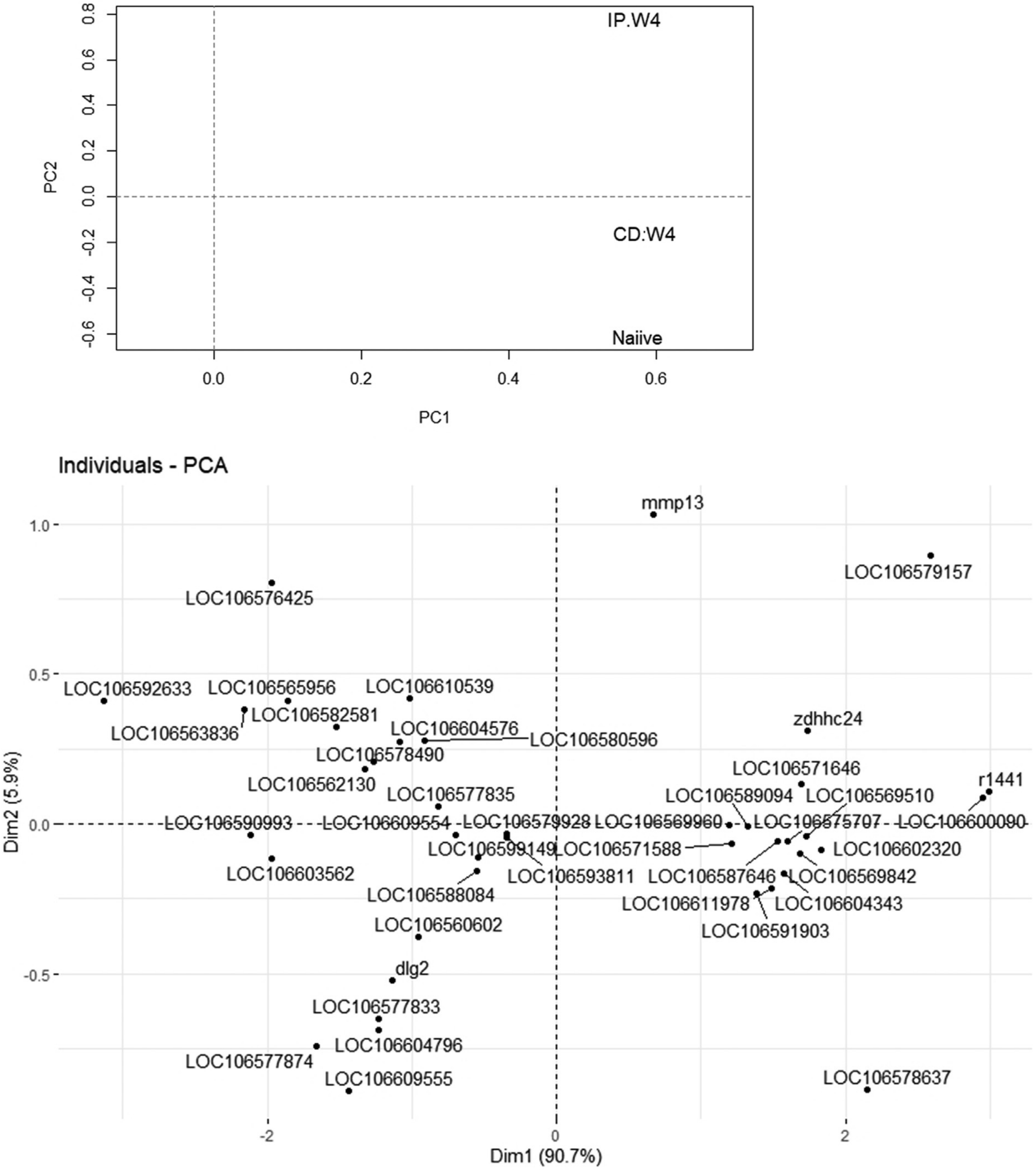


Fig. 15 PCA genes differentially expressed, as analyzed by RNA-Seq, in animals with high and lowbreeding values for pancreas disease resistance in response to salmonid alphavirus infection causing the disease in Atlantic salmon. Salmon were tested while in the naive state, 4 weeks after injection by the virus (IP) and 4 weeks after co-habitation with the virus (CH). The upper plot shows loadings of the columns in the data table (fold change values under the different experimental conditions) and the lower plot shows the scores of the specific genes.

4.22.4.4 Exploring the Variation Patterns between Different Blocks of Data

Exploring variation between different blocks of data is highly relevant in functional omics. A typical situation could be to analyze the relation between variation in transcriptome, proteome or metabolome data and some response parameter(s) of interest.

In studies of two-block relations, one approach is to study the relation between the explanatory features and the response. The input data block may be called "independent data block," "explanatory variable," "regressor features" or "predictor features" and the output data block may be called "response data block" or "regressand data block." The input and the output may reflect causal relationships, but that does not have to be the case. Even when causality between data blocks is reasonable to assume, one might be interested in reversing input and output as that may reveal information that is otherwise lost. The input data are often assigned as **X** and the response as **Y**, but in some situations other options may be more useful. For example, referring to Fig. 8, relating transcriptome data ($\mathbf{X}_3$) as input to phenotypic data ($\mathbf{X}_6$).

The data blocks may be either continuous or categorical. When the output of a model is continuous, the model defines a regression model or a prediction model, whereas when the outputs are categorical the model defines a discriminant analysis or classification model.

The typical characteristic of omics data is the dimensionality of the data matrix, with an abundance of features and usually only a few samples. For example, a typical transcriptome data may consist of more than 20,000 features, one for each gene transcript, whereas the number of samples is often limited due to high cost. The classical statistical regression methods usually require more samples than features, which triggered massive statistical research on how to deal with this situation.

For an input matrix with large number of features compared with samples, the main problem is the so-called rank problem. This means that some of the features may be replaced (at least approximately) by a linear combination of some other features. Hence, there is a situation concerning data redundancy. By traditional statistical approaches this is considered as a problem. Classical methods in linear regression and discriminant analysis, such as ordinary least squares (OLS) and linear discriminant analysis (LDA), cannot be fitted to such data in their regular form.

4.22.4.4.1 Supervised bilinear methodologies

With projection methods the observed variation is projected onto a smaller subspace spanning the underlying variation in the data as was also described for PCA, illustrated in Fig. 11. Similar two-block methods are developed as illustrated in Fig. 16. Rather than finding the components that best describe the variation within one data block, it maximize the covariance between the data blocks. The arrow in Fig. 16 shows that the information in one response data block (the block here called Y) is utilized to define the best component in X to predict the response.

One approach is to maximize the covariance between the two data blocks, which is the principle of the methods called Partial least squares regression (PLSR) and Partial least squares discriminant analysis (PLS-DA).[22,42,47,58] By these approaches the "problem" of collinearity is converted to a strength as several features together may better represent an underlying variation in the data than one variable at the time. Most often, correlated variables together describe a phenomenon of interest.

To validate the different PLS factors, cross-validation may be used where the data are split into a number of segments of one or more samples. One segment at a time is left out of the regression analysis and used for validation. This is repeated until all segments have been left out once ("cross validation"). Another strategy is to split the data into one calibration set and one test set, or better, use a completely new data set for validation as a real test set.

The Norwegian cohort of the human study (Data set 2, Box 1) is here used to demonstrate two-block regression. The transcriptome data were used as input and the clinical data as response, using as response parameters the body mass index (BMI), high-density lipoprotein cholesterol (HDL) and triglyceride (TG). PLS regression model of all transcripts are displayed in Fig. 17A–C. In the score plot (Fig. 17A), all samples prior bariatric surgery are located towards the right along PLS factor 1, and all samples after bariatric surgery are located towards the left. In the corresponding loading plot of the responses (Fig. 17C), BMI is located most strongly towards the right, which shows that the main pattern of variation in this data was reduced BMI after the bariatric surgery. The loading plots of the gene expression data (Fig. 17B) reflect relative changes in the gene expression related to the changes in the response (each

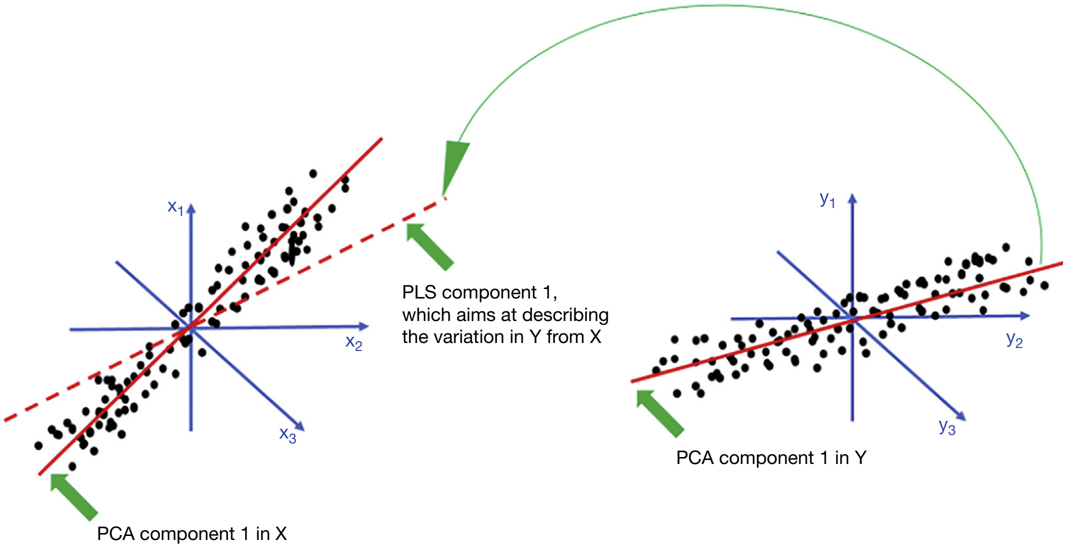


Fig. 16 Graphical illustration of the supervised multivariate method Partial Least Squares Regression (PLSR). By PLS, one data block (Y to the right), regarded as response, is used to guide the component (PLS factor) estimated in the the input variables (X to the left).

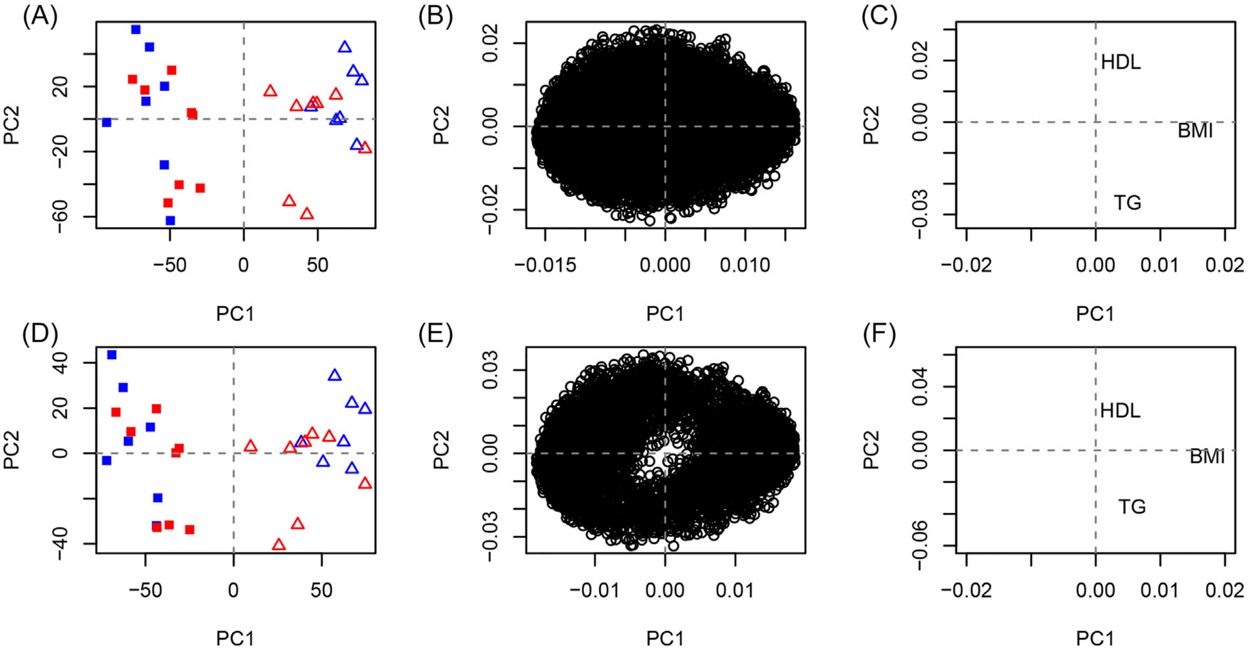


Fig. 17 Human data on obesity, type 2 diabetes and bariatric surgery (Data set 2 **Box 1**). PLS regression of the Human data with the transcriptome data as input and the clinical data body mass index (BMI), high-density lipoprotein cholesterol (HDL) and triglyceride (TG) as response. (A, D) 2D scatter plot of the scores of the samples for the first two PLS factors color coded by diabetes (in *red*) vs. non-diabetes (in *blue*) of samples taken before bariatric surgery (open *triangles*) and after the bariatric surgery (filled squares). (B, E) The corresponding loading plots of the transcriptome input data for the same PCs. (C, F) the corresponding loading plots of the clinical response data for the same PCs. (A–C) PLS regression performed on all transcripts, (D–F) The same PLS regression where only gene transcripts selected as significant by Jackknife for any of the response parameters were included. The data were mean centered and scaled to unit variance prior to PCA to let all transcripts have the same impact on the analysis.

gene is here represented as a point in the graph). As the data are mean centered, this must be viewed in relative terms; genes with positive loading, located towards the right along the horizontal axis, PLS factor 1, is higher expressed when BMI is high, and genes located towards the left are lower expressed when BMI is high, relative to the means of the data. To select the most relevant genes, feature selection was applied, where we applied the test Jackknife adapted to bilinear models.[60] This method performs a test of the stability of the regression coefficients across all cross-validation segments. For each loop in the cross-validation routine, regression coefficients are calculated. A *t*-test is subsequently performed to validate the stability of the regression coefficients or other relevant parameters when leaving out one segment at a time from the calibration to be used for validation. This validation procedure selected 5828 gene transcripts as having stable regression coefficients for BMI, 1100 transcripts were selected for HDL, and 1402 for TG. As will be illustrated below for the *Lactobacillus* data, the selected genes may be plugged into some bioinformatic tools to study the biological meaning they represent. A new analysis was performed on the transcripts selected for at least one of the three response parameters (Fig. 17D–F). The analysis before feature selection as well as the analysis after feature selection were dominated by the effect of bariatric surgery. In the loading plot of the responses (Fig. 17C and F), BMI is located towards right along the first PLS factor, and in the corresponding score plots (Fig. 17A and D) samples prior bariatric surgery are located towards the right and samples post-surgery towards the left, negatively related to the response features. The second PLS factor separated the patients according to HDL and TG, although this was not related to bariatric surgery or diabetes vs. non-diabetes. Notably, differences related to diabetes did not reveal any pattern of variation related to any of the clinical parameters by the two first PLS factors, nor when including more PLS factors (results not shown). It will be seen at the end of this book chapter that supervised multivariate analysis enables detection of critical diabetes related genes in this data when the transcriptome was related to the experimental design factors.

The direction of the regression may not always go from the left to the right in Fig. 8. For the *Campylobacter jejuni* data (Data set 1, Box 1) one might for example be interested in exploring how the FTIR pattern relates to the transcriptome data using the FTIR pattern as input in the model and the transcriptome as output, or vice versa. As an example, a model was fitted to the *Campylobacter jejuni* data with the FTIR data as a block of input features. Instead of using the whole gene transcript block as a response matrix, PCA was conducted on the mean-centered transcriptome data in advance, and the PCs from the PCA were used as response, one at a time. Thus, one omics data block is related to a compressed representation of another omics data block. The FTIR data were centered and standardized to allow all features to have the same impact on the analysis. We also performed preprocessing by Extended Multiplicative SCatter correction (EMSC), which is treated in other chapters of the present book. The FTIR spectra showed poor relation to the first PC (not shown). Thus, the behavior of the few samples that dominated the first PC of the transcriptome data, did not affect the FTIR spectra (see also Fig.12. C,D). The second PC, on the other hand, could be well predicted from the FTIR spectra, although this pattern accounted for only a small proportion of the gene expression.

Separate models can be trained based on smaller, but interesting parts of the FTIR spectra as regressor features. It is known that the fat have signals in the wavelength 3000–2800. Results of a model including only a fat region from wavelength 3000–2800 are presented in Fig. 18, where PC2 from the PCA analysis of the transcriptome data is used as response. The correlation coefficient between predicted and measured value was $r = 0.90$ and the model could account for 82% of the variation in the response parameter using 11 PLS factors. The first and most important PLS factor described 33% and the second PLS factor 10% of the total variance. The score plot reveals a temperature axis along the first PLS factor (Fig. 18A), whereas the second PLS factor separates biological replicate 1 from the other replicates (Fig. 18B). The loadings of features in this model (Fig. 18C), are located towards the same direction as the temperature of 25 °C (seen in the score plot, Fig. 18A). The interpretation of this is that, after omitting the pattern

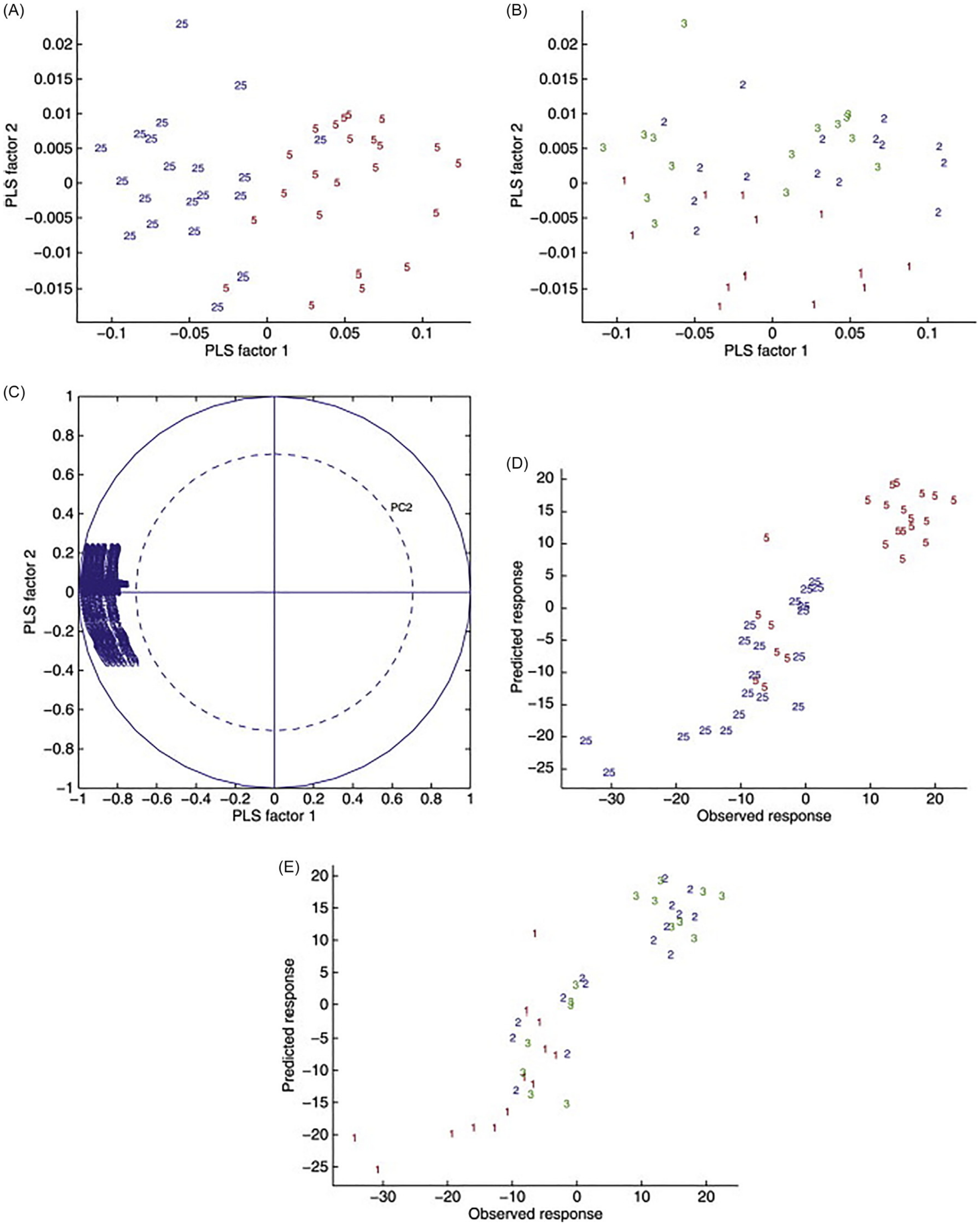


Fig. 18 *Campylobacter jejuni* (Data set 1, **Box 1**). Results of PLS regression analysis using the fat region of FTIR (wavelength 3000–2800) as regressor features and PC2 from PCA of the transcriptome as response for the *Campylobacter* data (**Box 1**). Scores (A) and (B) of PC1 vs. PC2, (C) Correlation loadings plot for PC1 and PC2; (D) and (E) predicted vs. observed. The samples were labeled and colored by temperature in panels (A) and (D), where 25 °C is in blue, and 5 °C is in red, and in panels (B) and (E) by replicate number.

of variation described by the first PC, the second PC reflects a pattern where the bacteria exposed to 25 °C contain more of fatty acids as reflected by the FTIR measurement, compared with those exposed to 5 °C.

The observations that the main pattern of variation in the transcriptome data, as reflected by the first PC of the transcriptome data, did not correlate to the spectroscopic data, whereas a clear connection between the two data blocks is seen when considering the second PC of the transcriptome data, illustrate the usefulness of projection methods. By working on the level of latent structure, we were able to detect this relation.

On this data we will also demonstrate the soft-threshold PLS method of Sæbø et al.[61] Soft-threshold PLS includes a threshold parameter controlling a shrinkage of the PLS weights towards zero with the aim of selecting the smallest possible number of features that can discrimimate the response parameter. All weights smaller than the threshold in absolute value will have a zero contribution to the PLS component. For large threshold values, the number of contributing features may become very small. The soft-threshold PLS found that a model with only two contributing genes was sufficient to obtain a near-perfect classification with only one misclassified sample grown at the two different temperatures. The two genes that were identified are plotted in Fig. 19. The gene Cj0408 is frdC Fumarate reductase cytochrome B subunit and the gene Cj0420 is a putative periplasmic protein.

Fig. 19 illustrates a very important aspect of multivariate data. Unless the two genes are considered simultaneously, it would not be possible to detect the relevance of these genes. In particular along the vertical axis, the two temperatures completely overlap, and by any univariate methods this gene would never be detected as relevant to separate the temperature effects. But seen along with the other protein, a separation is achieved. In general, the combination of features may reflect variation that is hidden when a univariate approach is taken.

In the *Lactobacillus* data, the phenome was dominated by interacting pattern where one strain responded more than the other on reduced glucose availability, whereas explorative multivariate analysis by PCA did not reflect this pattern of variation in the proteome and transcriptome (Fig. 14). Even by the supervised method PLSR, applied with the proteome as input and all four end-products as response, where the end-products are used to guide the estimated components, we still do not see a clear interacting pattern where one strain change more upon changed growth condition by the first PLS factors (results not shown), which was so clearly seen by PCA of the end-products (Fig. 14). But after feature selection (Fig. 20), there is a larger change for the strain LS25 than for the strain 23K along the first PLS factor. In a regression analysis of the data of the *Lactobacillus sakei* the transcriptome or the proteome was used as input and the end-products as response. One approach for feature selection, called "shaving", repeatedly removes a certain percent of the features (or "shaved off"). This can be repeated until the optimal prediction is achieved. Different criteria can be chosen for the shaving process. Selectivity Ratio (SR) performs variable selection by calculating the ratio of explained to residual variance of the **X** features on the projection vector.[62] Another approach, Loading weight (LW), uses the loading weight at the optimum number of PLS factors as selection criteria. When applying shaving, a plot displays the reduction in the error as the number of features included in the model decreases (not shown here).

To visualize the pattern of variation of the selected features, we here apply PCA of the selected features. For the proteome, proteins with this interacting pattern could be selected when using SR, and for the transcriptome LW was applied (Fig. 20).

For the *Lactobacillus sakei* data, under restricted nutrient availability, the end-products are changed in a direction that gives more ethanol, formate, and acetate and less lactate. This implies more energy per glucose unit, as ethanol, formate and acetate yields more ATP per glucose unit than lactate. The analysis of the end-products revealed that this was seen more for LS25 than for 23K. In the PCA plot of the selected features (Fig. 20) this was reflected by the first PC. By the two-block analysis where the proteomes are related to the end-products we identified proteins that may shed light on the underlying molecular mechanism. For interpretation of the selected features we applied graph analysis.

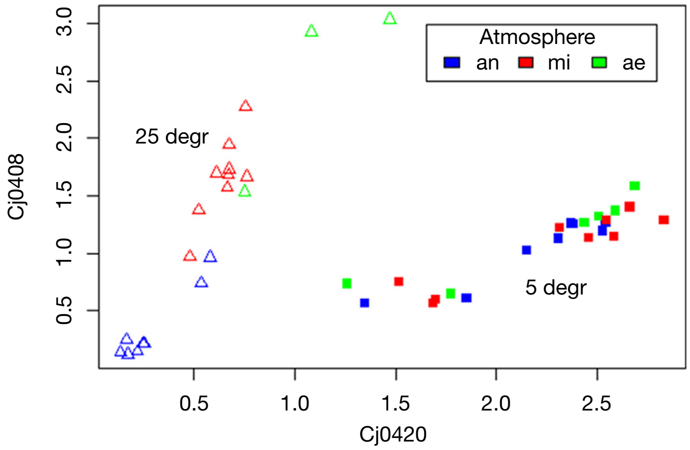


Fig. 19 *Campylobacter jejuni* (Data set 1, Box 1). The gene expression of the two transcripts identified by soft-threshold PLS for the discrimination between temperatures at 5 °C (open triangles) and 25 °C (filled squares), color-coded by the atmosphere: anaerobe (*blue*), microanaerobe (minor low O_2 availability) (*red*), and aerobe condition (*green*).

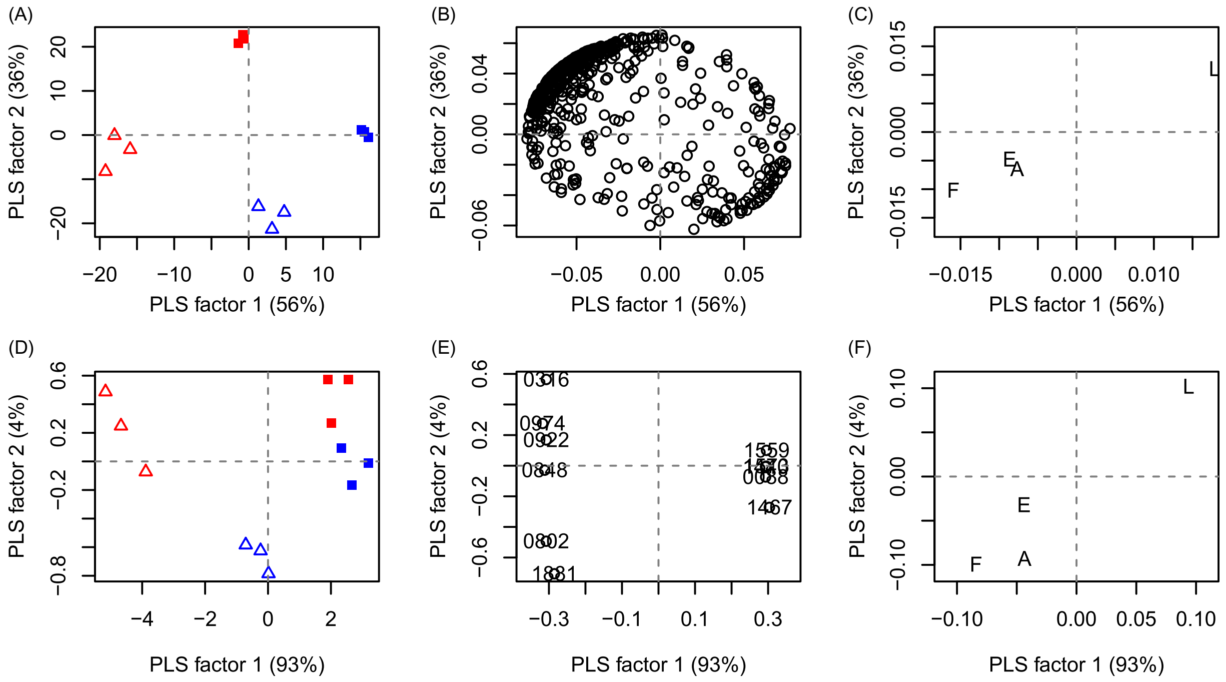


Fig. 20 Lactobacillus sakei (Data set 3, **Box 1**). PLS regression of all proteins in the proteome as input (A-C) and proteome after feature selection as input (D-F) with end-products as response, after variable selection by the approach shaving using the criteria Selectivity Ration (SR). Scores of the samples (A,D) and loadings for the selected features (B,E) and for the response (C,F) for the two first PLS factors are shown. The samples in the score plots are L. sakei 23K (blue) and LS25 (red) at high (filled squares) and low (open triangles) glucose availability.

4.22.4.4.2 Graph analysis

By using available bioinformatic software where the selected genes are plugged into the software, a graph can visualize the meaningful biological information. In Fig. 21 we have here selected proteins according to their position in the PCA plot (Fig. 20) so that the graph mirrors the pattern of variation in the loadings of the first PC in the PCA. Located towards the right side of the graph are proteins which increased in their expression when the glucose availability was reduced, a change more predominant for the strain LS25 than for 23K as seen in the score plot in Fig. 20. Accompanied with the metabolic change, there was a change in the expression of the proteins formate C-acetyltransferase (pyruvate formate lyase), NADH oxidase, and redox sensing transcriptional repressor Rex (all located towards the right in Fig. 21). The interpretation of this is altered pyruvate metabolism to benefit the bacteria by

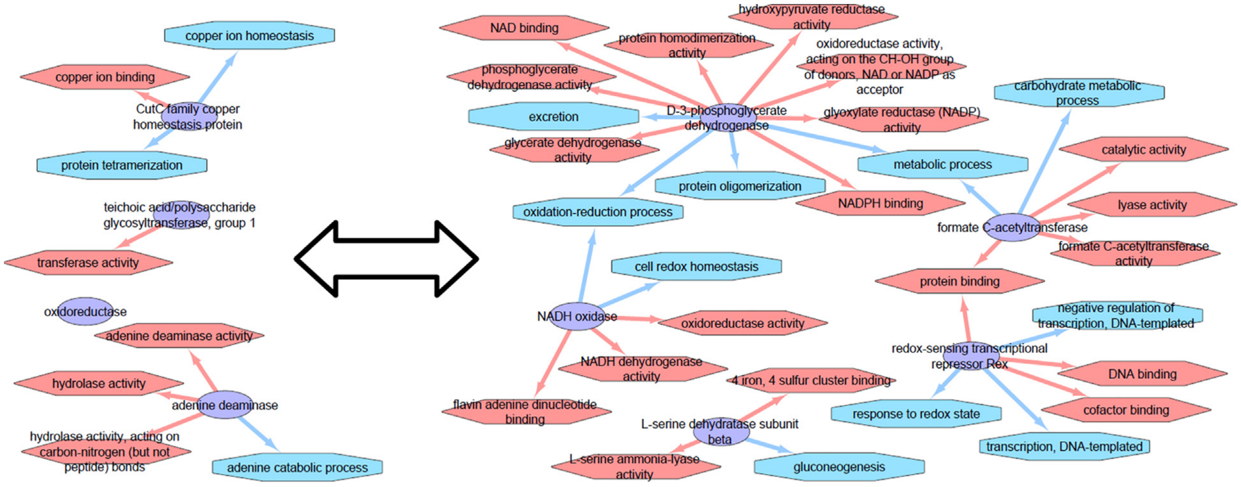

copper ion homeostasis
copper ion binding
CutC family copper homeostasis protein
protein tetramerization
teichoic acid/polysaccharide glycosyltransferase, group 1
transferase activity
oxidoreductase
adenine deaminase activity
hydrolase activity
adenine deaminase
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
adenine catabolic process
NAD binding
protein homodimerization activity
hydroxypyruvate reductase activity
oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
phosphoglycerate dehydrogenase activity
excretion
D-3-phosphoglycerate dehydrogenase
glyoxylate reductase (NADP) activity
carbohydrate metabolic process
glycerate dehydrogenase activity
metabolic process
catalytic activity
protein oligomerization
NADPH binding
oxidation-reduction process
lyase activity
formate C-acetyltransferase
formate C-acetyltransferase activity
cell redox homeostasis
protein binding
NADH oxidase
oxidoreductase activity
negative regulation of transcription, DNA-templated
redox-sensing transcriptional repressor Rex
NADH dehydrogenase activity
4 iron, 4 sulfur cluster binding
DNA binding
flavin adenine dinucleotide binding
response to redox state
cofactor binding
L-serine dehydratase subunit beta
transcription, DNA-templated
L-serine ammonia-lyase activity
gluconeogenesis

Fig. 21 *Lactobacillus sakei* (Data set 3, **Box 1**). A parallel plot to PC1 in the loading plot in **Fig. 20**, where proteins selected by PLS regression using the approach "Shaving" with "Selectivity Ratio" on proteome data regressed towards the end-product ethanol. The horizontal axis represents PC1 whereas the position within each side is randomly made for the best visualization. Proteins (purple circles), biological processes (blue octagon) and molecular processes (pink hexagon) are indicated., obtained by Cytoscape version 3.7.0.

generating more ATP, or by gaining NAD^+ which facilitates other end-products than lactate. Furthermore, L-serine dehydrogenase subunit beta, which is involved in the conversion of serine to pyruvate, is also closely linked to these proteins (also located towards the right in Fig. 21), with elevated expression under strongly restricted glucose availability, more so for the strain LS25 than for the strain 23K. These findings are described more in details in our original publication on this material.[54] Located towards the left is adenine deaminase involved in purine metabolism, and copper ion homeostasis and putative oxidoreductase involved in redox reactions in the cell.

An important aspect of the interpretation of multivariate data is to consider features that are closely correlated in the view of their common pattern. A relevant question for the biologist/molecular biologist/chemists to rise is therefore what are the biological link between features that display a common pattern of variation from a biological/chemical point of view?

4.22.4.4.3 Regularization methods

Ridge regression is a frequently applied approach to reduce the number of features in multivariate data. By this method, a small tuning parameter λ, called a penalty term, is added to the algorithm to make the estimated regression coefficients more stable. By adding the penalty term, large regression coefficients will be reduced, and this in turn may reduce overfitting and improve the prediction error. In Ridge regression none of the input terms are reduced all the way to zero. Therefore, it does not perform variable selection. Least absolute Shrinkage and Selection Operator (LASSO) performs both variable selection and regularization in order to enhance the prediction accuracy and improve the interpretability of the statistical model. Elastic net is the combination of the two, where a tuning parameter α defines the balance between them. Thereby elastic net allows to strike a balance between ridge regression and LASSO.

4.22.4.4.4 Support vector machines (SVM)

Support vector machine (SVM) is a multivariate classification method. In its simplest form, SVM ends up in a linear classifier in that a linear combination of variables are compared to some threshold. The way it does this is however somewhat different from e.g. LDA in that it attempts to maximize the margin between the class boundaries and training samples rather than placing the boundary at midpoints between class prototypes.

SVM may be transformed to a non-linear one in the original variables by adding functions of these as new variables. Such transformations may be justified either empirically (with better performance), or sometimes based on an expected relation between variables and classes. As a hypothetical example, classifying someone as obese requires calculating the body-mass index (BMI): $weight/height^2$, and comparing it to the threshold of BMI, here denoted b, equals to 30. Imagine a linear classifier using variables weight and height: it would invariably get the classes wrong. However, if the inputs were transformed to weight and squared height, the solution is simple and can be expresses as a linear classifier in these variables: when $1 \cdot weigth\ - b \cdot (height^2)$ is greater than zero, then the BMI is greater than the level defining obese.

SVMs with extensions, such as the mechanism described here, or using other mechanisms such as "the kernel trick," are very powerful classifiers. They do however tend to get very "large" in the sense that they remember all (or most) training samples in order to be applied. This may pose memory issues with large training sets.

4.22.4.4.5 Neural networks (NN)

In recent years, classifiers based on neural networks have gained much traction. Such classifiers were originally inspired by actual neural networks by constructing "neurons" that are sensitive to a number of inputs and "fire" when combinations of these increase past some threshold. Such neurons are connected through several "layers" in a network and for the case of a binary classifier end up with a single output which is a nonlinear function of the network's inputs. NNs are hence nonlinear classifiers by construction.

Training of neural networks is an iterative process where errors in predictions are used to update the network. This is analogous to a child learning to do a task with a teacher correcting the answers until the child gets all (or most) of the answers right. Key in this analogy is that the teacher does not explain *why* the child is wrong in any given problem, only that the answer *is* wrong. This may be a poor teacher, but this is the teacher typically used in training simple neural networks.

The non-linear property of neural network classifiers constitutes both their strength and complexity. Assuming a sufficiently complex network and enough data to train on, one strength is the claimed ability to approximate the non-linear relationship between inputs and classes automatically. For instance, in the above example with obesity the inputs could still be weight and height, and the network should be able to deduce the relationship to squared height automatically. An important complexity is relating given classifications to the inputs: it may become very difficult to learn any general rule between inputs and classes from a trained network—even though the network is apparently working flawlessly. This is related to the above learning process where the teacher does not explain the relationship – and in many cases, even for a human performing a complicated task perfectly such as recognizing the face of Jennifer Aniston, this person may not even be capable of explaining *what* makes her face different from all others. The choice of person is not entirely chance: a study led by Quiroga found a cell in a subject that fired precisely on pictures of Jennifer Aniston, see Quiroga and coworkers[63] for more information.

4.22.4.4.6 Genetic algorithm (GA)

The genetic algorithm (GA)[64,65] is a method for variable subset search that can be used prior to model fitting. A genetic algorithm mimics the process of natural evolution to generate useful solutions to optimization and search problems. A genetic algorithm is a search heuristic that is inspired by Charles Darwin's theory of natural evolution. This algorithm reflects the process of natural selection where the fittest individuals are selected for reproduction in order to produce offspring of the next generation.

The algorithm has successfully been applied to various megavariate data sets both in chemometrics[64,65] and bioinformatics.[66] The GA is an example of a random sampling-based method for variable selection though modified through evolutionary steps. Another method, which may succeed in finding favorable variable combinations through random variable sampling, is the random forests method of Breiman.[67]

The purpose of this section is not to list every possible approach for analyzing data from megavariate omics. There are numerous approaches presented in the literature, which are not mentioned here. Ideally, some guidance as to what would be the better approach should be given, but unfortunately there is no always-optimal choice of method. The best choice of predictor and method for variable selection varies from one data set to the other.

In a practical situation, it is probably a good strategy to explore a range of methods with different properties to find a good approach. However, a consequence of screening several approaches for a given data set, possibly in combination with variable selection, is the danger of overfitting. This necessitates careful validation of the final conclusion, preferably using independent test data.

4.22.4.5 Modeling the Effects of Experimental Design Factors and Pseudofactors

In the following section, we describe methods to study the effects of the design parameters on the responses based on a traditional statistical framework that is further developed into multivariate data.

4.22.4.5.1 Single-response analysis

The classical approach to analyze experimental design studies[18,19] is based on the fundament laid by R. Fisher and his co-workers,[68,69] who published the first use of ANOVA for experimental data approximately 100 years ago. This method builds on a linear model where the design parameters are the inputs in the model and the response are the outputs.

There is always a certain level of random variation in any experiments. To test if there is a difference between the performances of, e.g., a set of genotypes, the differences between the genotypes should be of a certain magnitude compared with the random variation of the experiment. ANOVA is, as the name indicates, analysis of the variances. To illustrate the methodology, we consider situations where we only have fixed effects. The analysis can then be described as given in **Table 2**. Fixed effects means that we are interested in the classes or levels of each factor that is tested. This contrasts to a situation where we select a set of samples that should represent a larger population. For example, when we select a set of locations in Norway for the experiment, where we are not particularly interested in the sites that are chosen. The sites may have been choosen to represent growth conditions in different regions in Norway. Sticking to the situation with only fixed effects we can describe the ANOVA as presented in **Table 2**. The effects of all factors are tested towards a common estimate of the random error in the data, which is based on the residual of the model, i.e. the variation that cannot be described by the experimental factors.

The mathematical formula for this analysis is a general linear model as a sum of the effects (Eff) of each factor:

$$\text{Response} = \text{Intercept} + \text{Eff}_{\text{A}} + \text{Eff}_{\text{B}} + \text{Eff}_{\text{AB}} + \text{Residual}$$

Any statistical model assumes that the samples are conducted in random order. If the analysis is performed in systematic order, e.g. all replicates of sample 1 first followed by all replicates of sample 2, etc., then it is not possible to distinguish between the real differences between the samples, and random variation that my occur during the analysis procedure. Sometimes, however, it is not possible to completely randomize all samples. For example, when the samples are run in batches. In gel electrophoresis one may for example run only 20 samples at the time as one batch. If the experiment consists of 20 samples and 3 biological replicates of each sample, then each batch should have all 20 samples from one replicate. Often the number of samples is not exactly the same as the capacity for each batch, and this should then be considered in the statistical analysis. There are different strategies to take this into account into the statistical analysis covered in statistical literature.

Table 2 Two-way ANOVA. A, B are fixed effects and there are several biological replicates of each combination of A and B.

	Sources of variation	*Means of Squares (MS)*	*f-values*
	Total		
A	Genotype	SD^2 of means of A	$\text{SD}^2_{\text{A}}/\text{SD}^2_{\text{residual}}$
B	Environment	SD^2 of means of B	$\text{SD}^2_{\text{B}}/\text{SD}^2_{\text{residual}}$
A*B	Genotype*Environment	SD^2 of means of the combination A and B	$\text{SD}^2_{\text{A*B}}/\text{SD}^2_{\text{residual}}$
Residuals and interactions with residuals		common SD^2 for the experiment, estimated from the replicates and all interaction terms with the residuals	

The f-value obtained from the data are compared with tabulated F-valued for the normal distribution. The f-value of the normal distribution is the distribution obtained if there were no significant effects of the input factors.

4.22.4.5.2 Multiple testing for multiple responses

With several response features, one could analyze each response separately by ordinary univariate *F*-tests. This will typically lead to multiple testing problems as many tests are performed. A number of different approaches are described to account for the multiple testing problem. When a test is performed on a large number of features, the risk of stating significant effects when there are truly no effects may be large. Adjustment of the *p*-values by False Discovery Rate (FDR) ensure that only a given percentage (e.g. 5%) of the responses reported as significant are false. The FDR criterion has become very popular in functional genomics. There exist several methods to calculate FDR adjusted *p*-values. Most do not take dependence among the responses into account.[70] An FDR variant that allows any kind of dependence among the responses is described in Moen et al.[50] The calculations are then based on rotation testing or permutation testing. It is important to realize that these tests are nevertheless univariate, and do not consider the combined actions of the features as illustrated in Fig. 19.

4.22.4.5.3 Multivariate analyses of data with multiple input factors and multivariate responses

Several methods are developed as a combination of a linear model as in ANOVA and a multivariate response. Some methods combine PCA with the linear model in ANOVA. By PCA, the data are projected onto new orthogonal PCs found as linear combinations of the original features. One approach is to apply PCA or simultaneous component analysis to the effects of each term in the linear model. This gives ANOVA-PCA when using effects plus the residuals. The method ANOVA and simultaneous component analysis (SCA) (ASCA) performs the bilinear multivariate method directly on the effects and utilizes differences between the replicates thereafter. A supervised alternative based on bilinear methodologies is obtained by applying target projection or partial least squares discriminant analysis which leads to ANOVA-TP[71] or AoV-PLS.[72]

In ASCA the original samples are projected onto the score space of SCA or PCA. Explained variance of the features given the design factor in focus can be calculated, and further sub-divided into contributions from the principal components.

Traditionally, significance in factor level differences was estimated using permutation testing. A strategy which produces exact significances for balanced designs, both for main effects and interactions, was introduced by Liland and coworkers.[73] It is based on classical multivariate ANOVA and can be used to produce informative significance ellipsoids in the score plots or calculate tables of pair-wise comparison statistics, possibly with compact letter displays of factor level groups.

ASCA performed on the *Campylobacter* data (data set 1, Box 1) is displayed in Fig. 22. The ellipsoids surrounding each factor levels show that there are significant differences between all levels of the main effects as well as on the interacting effects which is seen as the ellipsoids are not overlapping.

The corresponding loadings revealed interacting pattern between atmosphere and temperature reflecting deviating pattern for a few dominating genes. A more detailed plot along with the gene names would unravel more detailed information on which genes are elevated/reduced under the different conditions. As described by Moen et al.,[50] the cells survived better under 25 °C anaerobic conditions than under either the microaerobic or aerobic conditions tested (as determined by plate counts). The general trends were that the cells lost the ability to be recovered on plates before the membrane integrity was disrupted. The expression of the genes Cj1356c, putative integral membrane protein and Cj0200c, putative periplasmic protein, were among the genes that contributed to the interacting effects between atmosphere and temperature. Another gene observed by a detailed view of the loading plot of the interaction pattern (not shown) highlighted the gene Cj1385, Catalase, katA, which decomposes hydrogen peroxide into water and oxygen. Hydrogen peroxide is a major redox metabolite involved in redox sensing, signaling, and redox regulation.

ASCA performed on the FITR data (Fig. 23) show that the interacting pattern was not as strong for the FTIR data as for the gene expression data. This corresponds to the strong interacting pattern revealed by the first PC of the PCA of the gene expression data that was not related to the FITR data as revealed by the PLS regression analysis in the section of the two-block relation.

We here present a generic approach, or a platform, that we call Effect plus Residual (ER) modeling, which covers different methods that combines a linear model as in ANOVA with multivariate analysis of the effects of one term at the time while utilizing the residual of the complete model (Additional file A1). This approach is not only meant to address investigations with experimental design, but also any situation where the data may be influenced by several known or unknown confounding factors that can be organized as orthogonal factors. The method is flexible with respect to the downstream analysis, and it may perform analysis of each factor within each level of the other factors.

The method is here first illustrated in the *Lactobacillus sakei* study (Data set 1, Box 1), where the design factors are defined as factor_A (strain LS25 vs. 23K) and factor_B (growth condition, High and Low), and the interaction term $\text{factor}_{\text{A}*\text{B}}$, which describe different responses of the strains to changed growth condition. The linear model on the data with two experimental factors are:

$$\text{Observed data} = \text{Eff}_\text{A} + \text{Eff}_\text{B} + \text{Eff}_\text{AB} + \text{Residuals}$$

As described for ASCA multivariate intervention may be applied directly on the effects of each factor. Another approach is to add the residuals to each effect, which gives what is called ER values of each factor.

$$\text{E}_\text{A} + \text{R} = \text{Eff}_\text{A} + \text{Res} \tag{1}$$

$$\text{E}_\text{B} + \text{R} = \text{Eff}_\text{B} + \text{Res} \tag{2}$$

$$\text{E}_\text{AB} + \text{R} = \text{Eff}_\text{AB} + \text{Res} \tag{3}$$

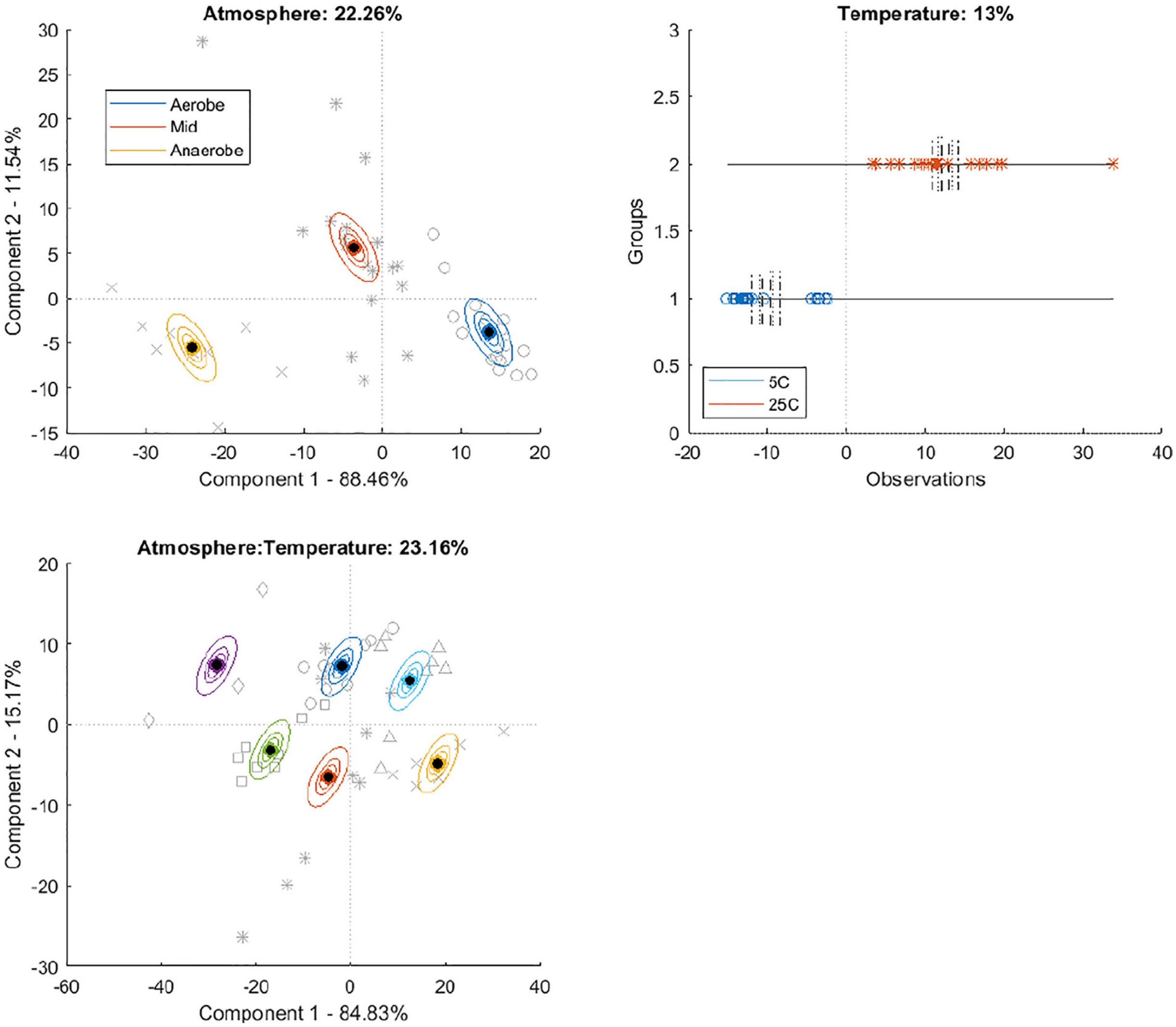


Fig. 22 *Campylobacter jejuni* (Data set 1, **Box 1**). Score plots of the samples of ASCA performed on the gene expression data where the main effects of atmosphere, the main effects of temperature and their interaction term are displayed. From the inner to the outer ellipsoid, these represent 40%, 68% and 95% of the variation of the data. The two upper plots display main effects of atmosphere (to the left) and main effects of (temperature) to the right, and the interacting effects, which display all combinations of atmosphere and temperature (in the lower figure). Although we here visualize a two-dimensional plot of two components (for atmosphere and the interaction plot) and one-dimensional plot (for the temperature), and hence ellipses or just colored points are displayed, the term ellipsoid is used as the number of components in the data is more than two.

The addition of the residuals to each effect is illustrated for one feature (ethanol) in the *Lactobacillus sakei* data (Data set 3, Box 1) in Fig. 24. A benefit of adding the residual of the complete model directly to the effects is that new quantities are obtained that reflect only the effects of one factor at the time, and have the residual for validation, and this quantity can be utilized for any univariate or multivariate intervention.

The effects of each factor plus the residual of the complete model were for the *Lactobacillus sakei* data subjected to elastic net for feature selection (Fig. 25).[74] Only a few features were needed to discriminate between the two strains (in the plot to the *left*), between the two growth conditions (in the *middle*) and the differential response of the two strains (to the *right*).

ER modeling with PLS-DA of *Campylobacter jejuni* is displayed in Fig. 26 focusing on the effects of atmosphere, temperature and their interaction. By ER modeling a separate data table is made of each of these effects plus the residual of the complete model, and the supervised multivariate method PLS-DA with Jackknife for features selection are performed with the design parameters as response. The analysis of the atmosphere separated the samples well in the score plot according to the atmosphere condition (Fig. 26A) and the corresponding pattern of variation in selected genes are displayed in the corresponding loading plot (Fig. 26B). Likewise, the effects of temperature is well identified with genes spanning this variation in the loading plot (Fig. 26C and D). For the interaction term (Fig. 26E and F), the first PLS factor separate different atmospheric conditions: anaerobe, microaerobe and aerobe, and the second PLS factor separating low (upper) and high (lower) temperature, where the lowest temperature, do not span any variation along the first PLS factor. Thus, the interacting pattern is clearly visible in the score plot. To the right in the

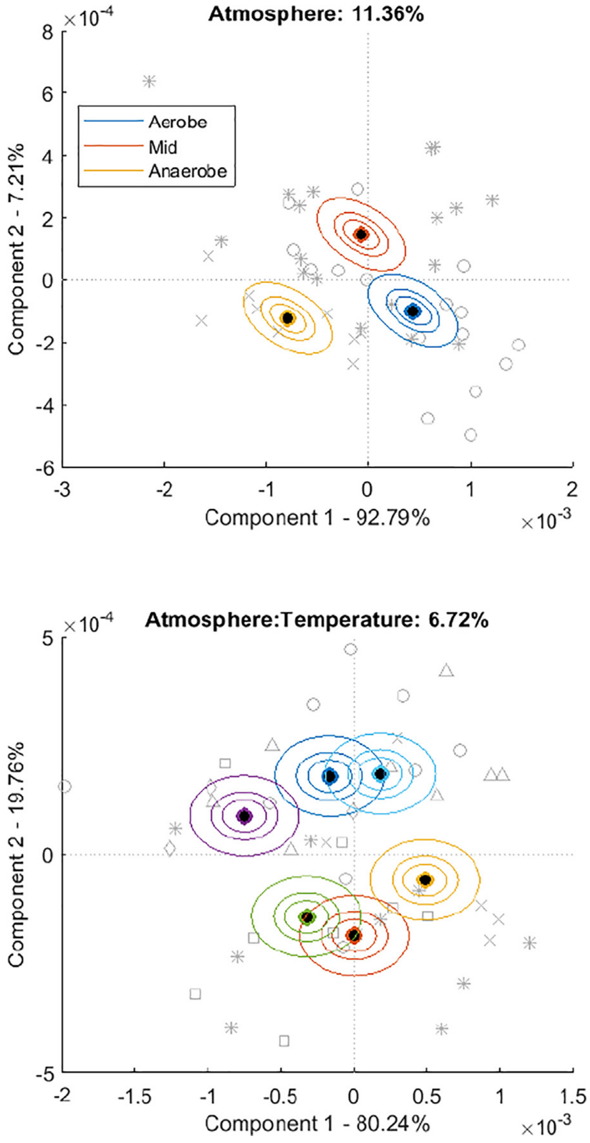


Fig. 23 *Campylobacter jejuni* (Data set 1 **Box 1**). Score plots of the features of ASCA performed on the FTIR data at selected wavelength where the main effects of atmosphere, the main effects of temperature and their interaction term are displayed.

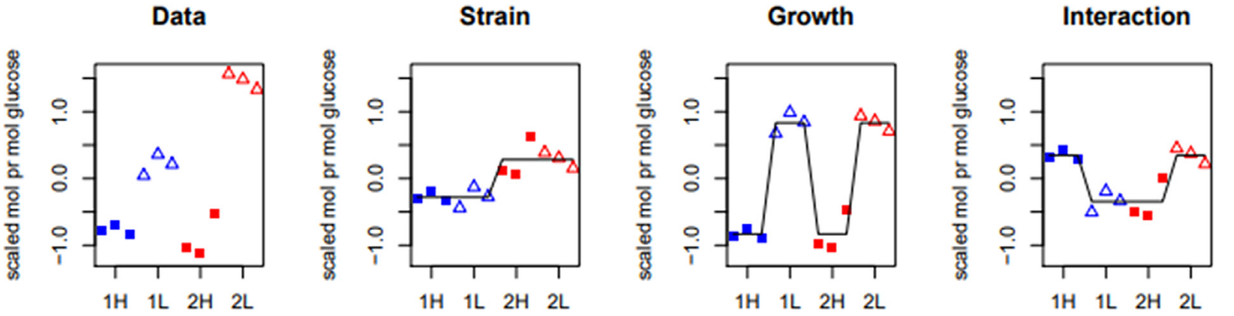


Fig. 24 *Lactobacillus sakei* (Data set 3, **Box 1**). Illustration of ER modeling for the protein redox-sensing transcriptional repressor Rex (Ass. number LCA_0848). From the left: the data, ER values of strain, ER values of growth condition, ER values of the interaction between strain and growth condition where the effects are in the three latter figures displayed as lines and the effects plus residuals are displayed as points for the strain 23K (*blue*) and LS25 (*red*) at high (*filled squares*) and low (*open triangles*) glucose availability.

loading plot are genes involved in the interaction. One of the genes located towards the right is Cj1359 which is the gene PPK. The aerobe condition at high temperature, which had the poorest survival, was located to the left in the score plot, opposite to this gene and the other selected genes along this PLS factor. The PPK gene, encoding polyphosphate kinase, plays an important role in stress tolerance. It has been found that this gene is involved in the loss of stress resistance, biofilm formation, colonization capacities as well as being important in the regulation of the production of other proteins.[75]

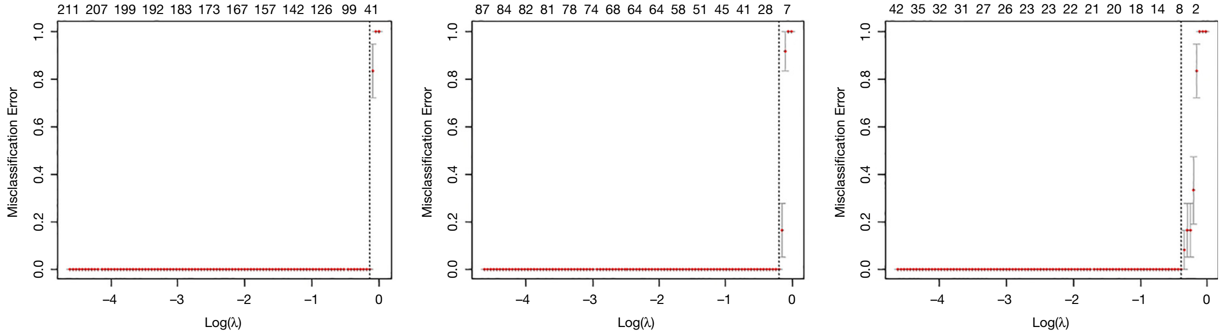


Fig. 25 *Lactobacillus sakei* (Data set 3, **Box 1**). Elastic net of the transcriptome with the design as response run separately using Effect+Residual values for each of the factors: strain (to the *left*), growth condition (in the *middle*) and their interaction (to the *right*) as input, with the belonging design parameter as response. Cross-validation (*red*) with upper and lower standard deviation (*gray*) with increasing number of features in the model as the regularization parameter changes. The figures show the misclassification as the number of features in the model increases, as red from the right to the left. Thus, these figures show that in all models, just a few features were needed to obtain good prediction.

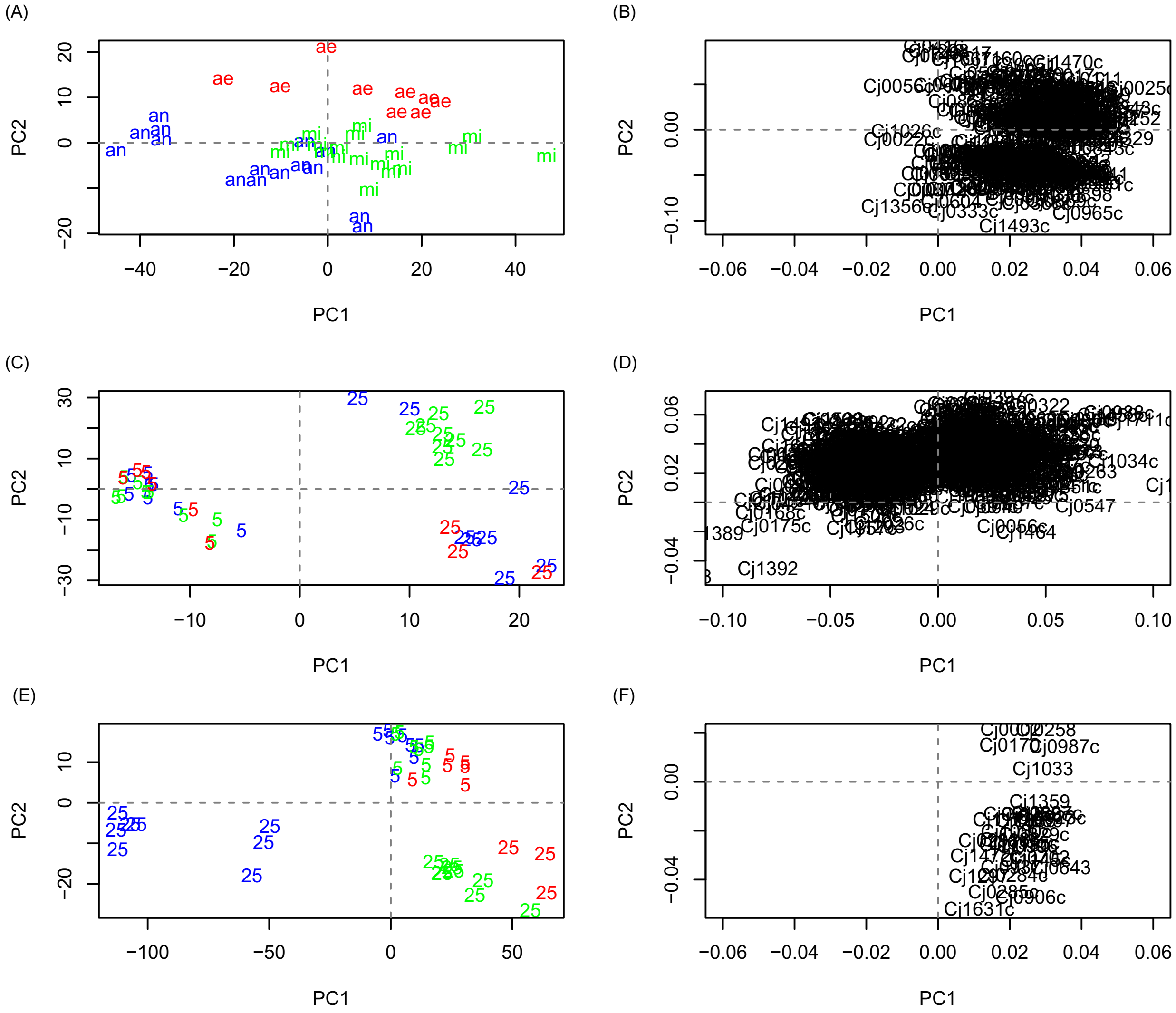


Fig. 26 *Campylobacter jejuni* (Data set 1 **Box 1**). Plots of PLS discriminant model of the gene expression after E+R modeling which isolate the effects of each factor. (A,B) PLS discriminant analysis of the effects of atmosphere, (B,D) PLS discriminant analysis of the effects of atmosphere, and (E,F) PLS discriminant analysis of the effects of the interaction between temperature and atmosphere. (A,C,E) scores of the samples and (B,D,F) corresponding loadings of the genes. (A) aerobic condition (ae) (in red), microaerobic condition (mi) (in green) and anaerobic condition (an) (in blue), (B) effects of temperatures where the color symbols are as in (A) with temperature (5 and 25) given in the plots (C) the interaction between atmosphere condition and temperature with symbols and colors as in (B).

For the human study of diabetes and bariatric surgery (Data set 3, **Box 1**) we will here present the use of ER modeling to combine data from two independent cohorts for validation across the cohorts. The two studies used the same transcriptome platform and the same patient groups: persons with and without diabetes with high BMI subjected to bariatric surgery.

Within each data set (the Norwegian cohort and the Mexican cohort) we applied ER modeling to separate the effects of diabetes vs. non-diabetes from other influencing factors that were internal for each data set. This resulted in two data tables, one from each cohort that reflected only the effects of diabetes vs. non-diabetes plus the residuals of the complete model within each data set. Details of the study is to be presented elsewhere (in prep). Here we show that we were able to use ER modeling to combine the two data cohorts to give one large set of data. For the combined data set we applied a ER modeling with country as factor A, diabetes as factor B, and their interaction term factor as A*B, to isolate the effects of diabetes vs. non-diabetes across all patients in the two studies, and to study differential effects of diabetes vs. non-diabetes between the two countries (Fig. 27). As the samples were taken at the bariatric surgery in the Mexican cohort, these samples are most comparable with samples taken prior bariatric surgery from the Norwegian cohort. However, as we by ER modeling could omit effects of bariatric surgery, we can utilize all data, to make one large data table of all patient samples.

A data model was made using one cohort as training set and the other as test set to predict diabetes vs. non-diabetes from the transcriptome record. This resulted in nearly perfect correct classification of diabetes vs. non-diabetes for the completely independent validation cohort (Fig. 27). The blue curve displays the observed data where −1 is non-diabetes and +1 is diabetes, and the red curve displays the predicted classifications, which shows that all except one sample are correctly classified. The scores and loading of the calibration models are displayed in **Fig. 28**.

Thus, in each study, PLS factors were obtained that identified gene transcripts related to diabetes that were consistent across the two cohorts.

In addition, specific expression pattern that differed between the two studies will be presented elsewhere. The observation that we were able to connect data from two completely independent studies by using ER modeling opens new opportunities for extension of data and validation, to explore common and different regulation mechanisms revealed by different cohorts. This is of particular relevance in omics studies considering the high cost and often limited number of samples within each study.

To identify the most relevant genes that were affected by diabetes, we applied Jackknife in the PLS regression, which selected a number of gene transcripts in both cohorts, that will be presented in detailed elsewhere. In the PLS plot of the loadings the selected genes are marked with filled green circles in the PLS loading plot (**Fig. 28**). A PCA plot of only the selected genes is presented in **Fig. 29**. In this cloud of the selected transcripts was the pathway renin-angiotensin system and renin secretion. As described by Ramalingam and coworkers,[76] the renin-angiotensin system may serve as a close link between obesity and insulin resistance. The adipose tissue secretes bioactive molecules such as the inflammatory hormone angiotensin II, generated in the Renin Angiotensin System. Obesity is a complex disease characterized by excessive expansion of adipose tissue and is an important risk factor for chronic diseases such as type 2 diabetes. Moreover, obesity is a major contributor to inflammation and oxidative stress, which are key underlying causes for insulin resistance and type 2 diabetes.

4.22.5 Regression Algorithms for Megavariate Data

One of the strategies one can use to handle very wide data sets, is to multiply the data table by itself. If the number of rows (samples) is lower than the number of columns (features) which is often the situation in omics data, then matrix multiplication may result in a data table which is limited to the size of the number of rows both as rows and also as columns. Thus the original data is collapsed in the largest dimension. This matrix is called a kernel matrix, and the corresponding methods kernel methods. An efficient approach is to do most of the data analysis in this kernel matrix. One kernel method is the Parsimonious Kernel PLS (PKPLS).[77] As soon the kernel matrix has been computed, PKPLS can compute scores and cross-validated predictions without any

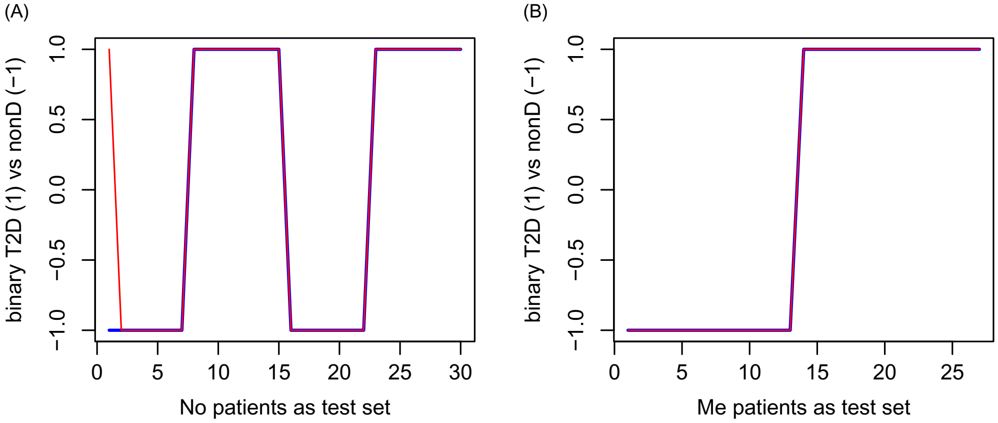


Fig. 27 Human data on obesity, type 2 diabetes (T2D) and bariatric surgery (Data set 2 Box 1). Classification results in PLS-DA analysis after ER modeling of the two cohorts (Mexican and Norwegian) using one cohort for training and the other for testing. The response are the class values for T2D (+1) and non-diabetes (−1). (A) Classification using the Norwegian cohort for training (*blue curve*) and the Mexican cohort for testing (*red curve*). (B) Classification using the Mexican cohort for training (*blue curve*) and the Norwegian cohort for testing (*red curve*).

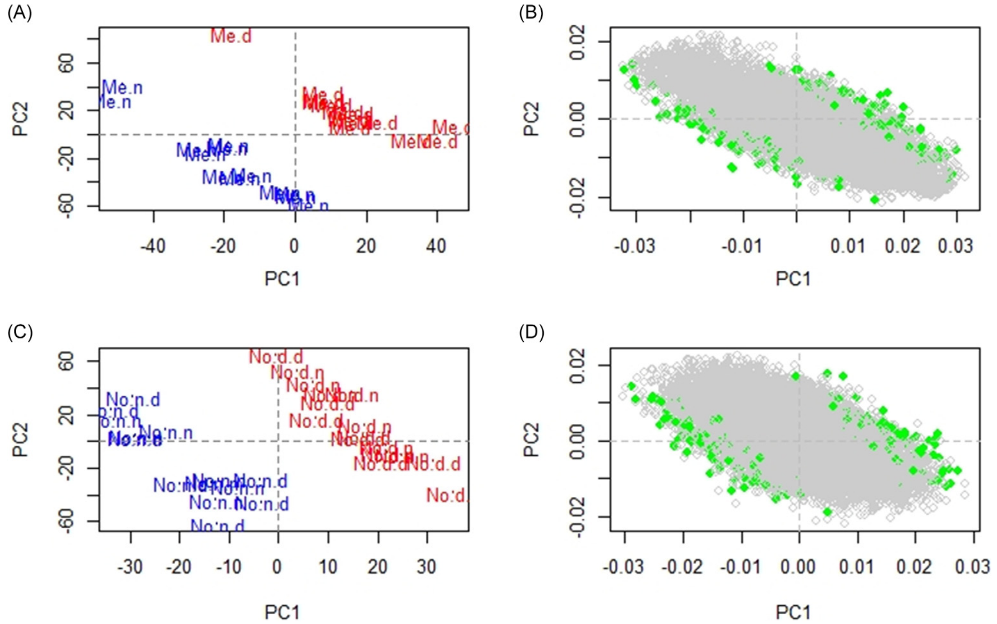


Fig. 28 Human data on obesity, type 2 diabetes (Data set 1 and 2, **Box 1**). Plots of PLS discriminant analysis of the combined data of two cohorts after E+R modeling. In the PLS discriminant analysis, the transcriptome data is input and diabetes vs. non-diabetes response (A, B) Model with Mexican data as training set, and (C,D) model with the Norwegian data as training set. (A,C) scores of the samples for patients with diabetes (in *red*) vs. non-diabetes (in *blue*). (B, D) The corresponding loading plots displaying transcripts selected for diabetes in both cohorts (filled *green* circles) and the remaining transcripts (*gray* open circles).

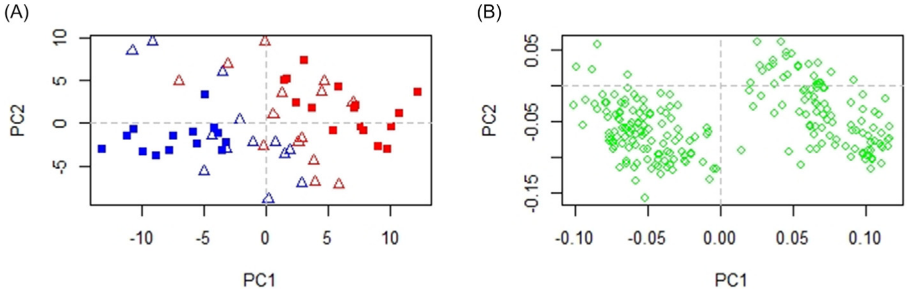


Fig. 29 Human data on obesity, type 2 diabetes and bariatric surgery (Data set 2, **Box 1**). PCA of selected gene transcripts from the discrimination of patients with vs. without diabetes using the transcriptome data as input and diabetes vs. non-diabetes as response after omitting the effect of country and the interaction between country and diabetes by Effect+Residual modeling. (A) 2D scatter plot of the score plots of the samples for the first two PCs color-coded by diabetes (in *red*) vs. non-diabetes (in blue). (B) The corresponding loading plot of displays transcripts selected for diabetes in both cohorts (filled *green circles*).

calculations outside the dimensions of this small data table. The kernel matrix is reused across cross-validation segments (possibly with nested/double cross-validation) and across multiple responses or multiple classes. In cases with more than 1000s features and many underlying components, the calculations can be speeded up by several hundred-fold, without any loss of precision.

It is important to realize that with a limited number of samples compared with the number of columns, the dimensionality of unrelated phenomena in the data, as described by linear models like PCA, is limited to the number of rows. That is why calculations on the kernel matrix does not loose precision.

The practical consequence of such a method is that there is no upper limit to the number of features that can be analyzed, if focus is on predictions or subspace complexity estimations (deciding the number of PLS components). This opens up possibilities for the simultaneous analysis of more genes, transcripts, proteins, metabolites etc. It also enables extending data sets with interactions of variables, transformations or combinations of data sources that together explain more of the response being predicted.

4.22.6 Concluding Remarks

This chapter gives a presentation of some data analytical tools useful for analyzing functional genomic data or other comprehensive data. A number of different methods are illustrated using the same data set throughout the chapter. Emphasis is to give considerations about the complexity and possible pitfalls when analyzing such data.

A Appendix

A.1 Additional file 1

A.1.1 Effect plus with Residual modeling (ER)

A linear model of an experiment with two input factors, f1 and f2 and their interaction term f1*f2, results in estimates of the effects of each input variable. The original data in a data block (a data table) is thereby split in a set of data tables of the same size as the original data table, where each data table reflects the effects of only one input factor (here illustrated in green, pink, and yellow). Plus the residual table (Res). The input factor may be experimental design factors or any other orthogonal factors relevant for the data at hand. The effects of each input factor (Eff$_{f1}$, Eff$_{f2}$ and Eff$_{f1*f2}$) are estimated from the data. By considering the effects of only one factor at the time plus the residuals of the complete model, values are created that isolate the effects of one factor at the time, omitting other influencing factors. We call this "ER for each factor". These ER can be used for any univariate and multivariate intervention. In a discriminant analysis, class identity of one factor at the time is used as response and the corresponding corrected values of the same factor is the input (Fig. A1A). A regression analysis may

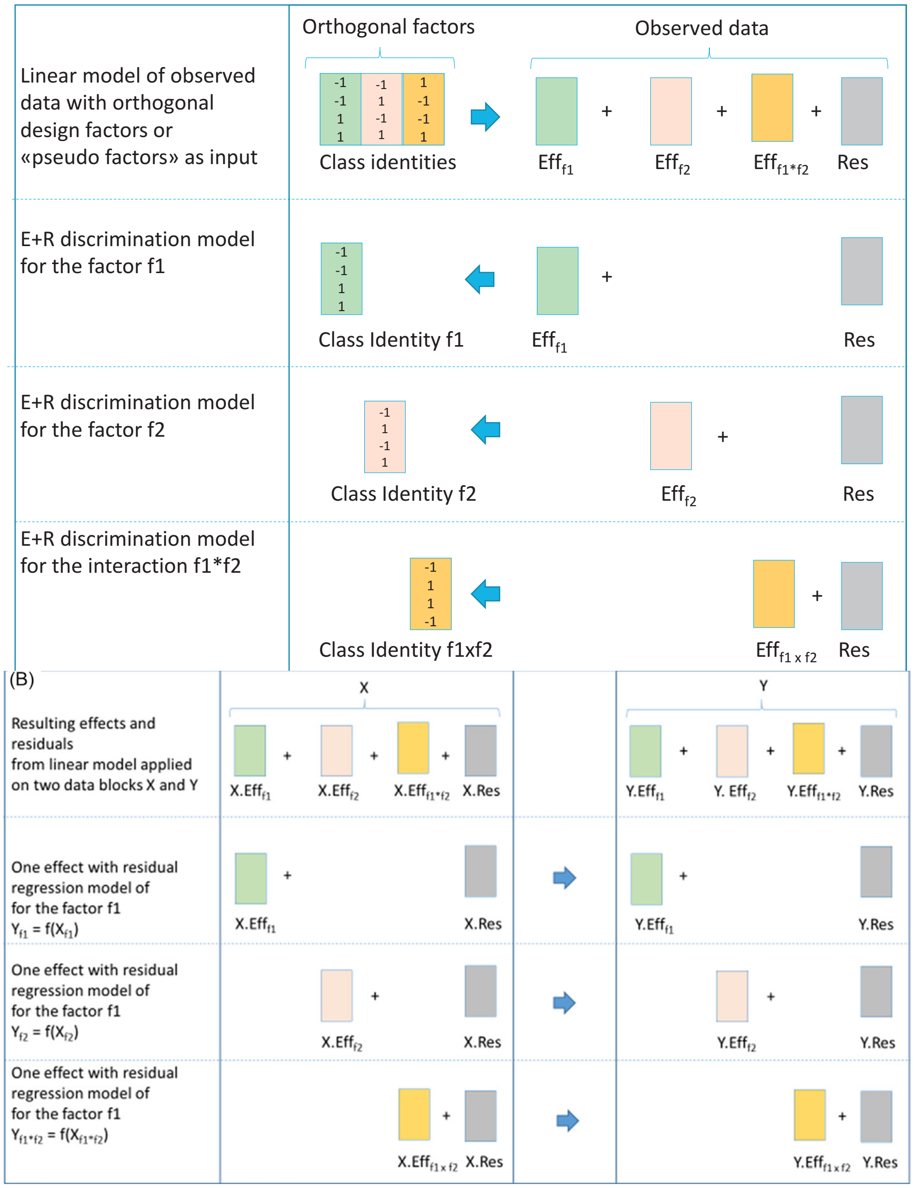
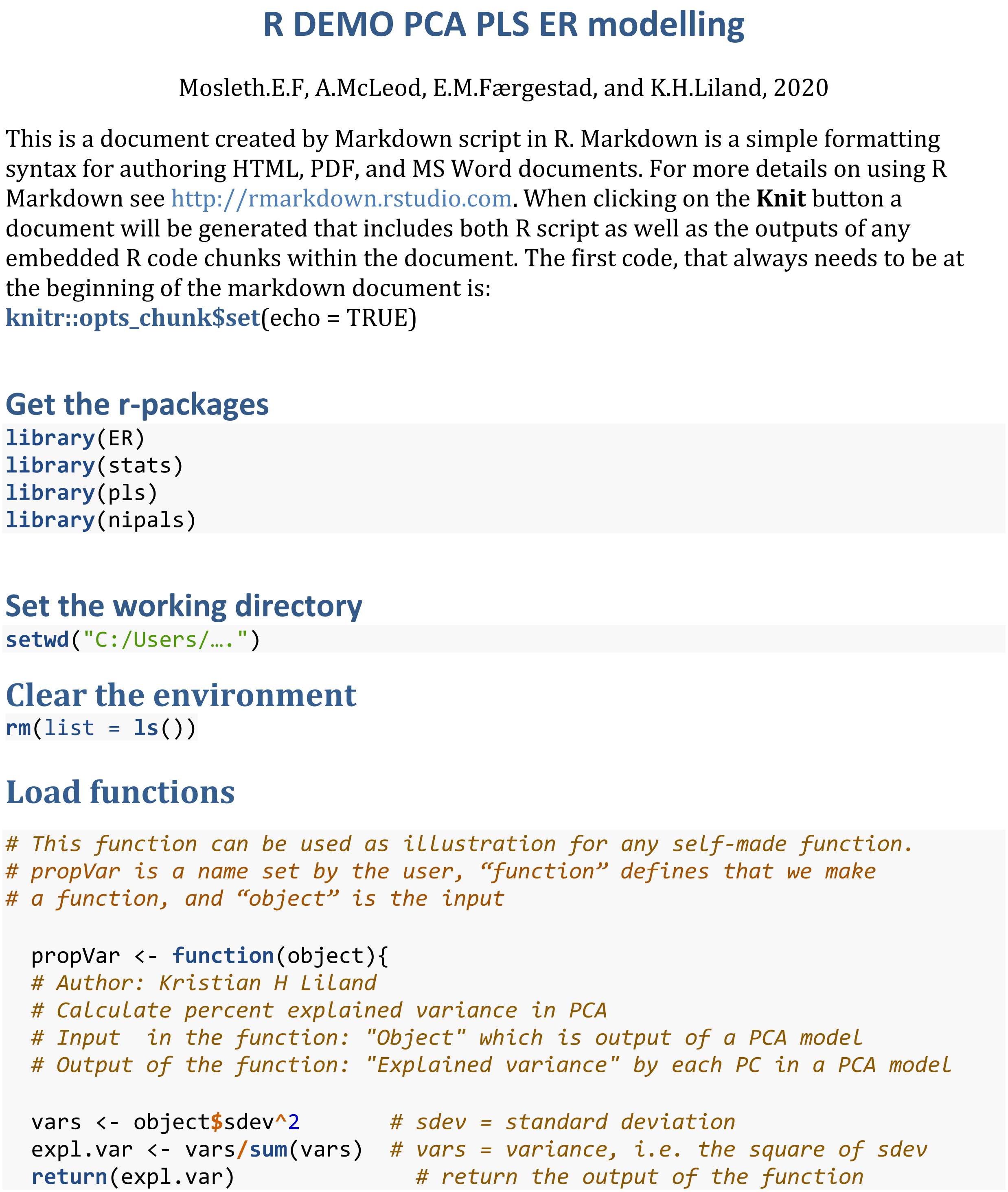

Orthogonal factors
Observed data
Linear model of observed data with orthogonal design factors or «pseudo factors» as input
-1 -1 1 1
-1 1 -1 1
1 -1 -1 1
Class identities
Eff$_{f1}$ + Eff$_{f2}$ + Eff$_{f1*f2}$ + Res
E+R discrimination model for the factor f1
-1 -1 1 1
Class Identity f1
Eff$_{f1}$ + Res
E+R discrimination model for the factor f2
-1 1 -1 1
Class Identity f2
Eff$_{f2}$ + Res
E+R discrimination model for the interaction f1*f2
-1 1 1 -1
Class Identity f1xf2
Eff$_{f1 x f2}$ + Res
(B)
X
Y
Resulting effects and residuals from linear model applied on two data blocks X and Y
X.Eff$_{f1}$ + X.Eff$_{f2}$ + X.Eff$_{f1*f2}$ + X.Res
Y.Eff$_{f1}$ + Y. Eff$_{f2}$ + Y.Eff$_{f1*f2}$ + Y.Res
One effect with residual regression model of for the factor f1 $Y_{f1} = f(X_{f1})$
X.Eff$_{f1}$ + X.Res
Y.Eff$_{f1}$ + Y.Res
One effect with residual regression model of for the factor f1 $Y_{f2} = f(X_{f2})$
X.Eff$_{f2}$ + X.Res
Y.Eff$_{f2}$ + Y.Res
One effect with residual regression model of for the factor f1 $Y_{f1*f2} = f(X_{f1*f2})$
X.Eff$_{f1 x f2}$ + X.Res
Y.Eff$_{f1 x f2}$ + Y.Res

Fig. A1 (A) Discriminant analysis defined for each factor in a linear model where the effects of other factors are omitted. The class identities are used as response and the effects of the corresponding factor plus the residual of the complete model are used as input. (B) Regression analysis defined for each factor in a linear model where the effects of other factors are omitted. The corrected value for each factor from one data block is used as input and the corrected value of the same factor is used as response.

also be defined using the ER of a factor from one data block as input and the corrected values of the same factor from another data block as response (Fig. A1B).

A.2 Additional file 2



R DEMO PCA PLS ER modelling

Mosleth.E.F, A.McLeod, E.M.Færgestad, and K.H.Liland, 2020

This is a document created by Markdown script in R. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com. When clicking on the **Knit** button a document will be generated that includes both R script as well as the outputs of any embedded R code chunks within the document. The first code, that always needs to be at the beginning of the markdown document is:
knitr::opts_chunk$set(echo = TRUE)

Get the r-packages

```
library(ER)
library(stats)
library(pls)
library(nipals)
```

Set the working directory

```
setwd("C:/Users/…. ")
```

Clear the environment

```
rm(list = ls())
```

Load functions

```
# This function can be used as illustration for any self-made function.
# propVar is a name set by the user, “function” defines that we make
# a function, and “object” is the input

  propVar <- function(object){
  # Author: Kristian H Liland
  # Calculate percent explained variance in PCA
  # Input  in the function: "Object" which is output of a PCA model
  # Output of the function: "Explained variance" by each PC in a PCA model

  vars <- object$sdev^2        # sdev = standard deviation
  expl.var <- vars/sum(vars)  # vars = variance, i.e. the square of sdev
  return(expl.var)              # return the output of the function
```

Load Data

```
data("Lactobacillus")
```

Get data from the array

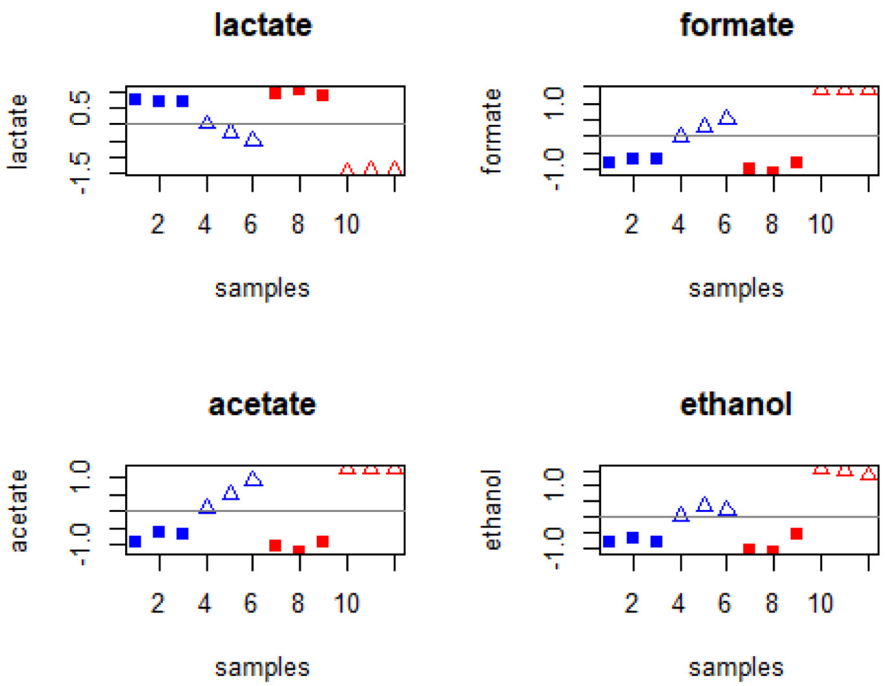
```
An array is a collection of data tables
names(my.array) # gives the names of the data tables in the array
str(my.array)   # gives details of the content in the array

The objects in R are defined by a class category.
When extracting data tables from an array, the tables must first be
unclassified, then classified as data.frame as shown here.

This experiment has two design factors A and B, and several data tables
of observed features

Design       <- as.data.frame(unclass(my.array$Design))
factorA      <- as.data.frame(unclass(my.array$factorA))
factorB      <- as.data.frame(unclass(my.array$factorB))
features.End <- as.data.frame(unclass(my.array$features.End))
features.P   <- as.data.frame(unclass(my.array$features.P))
features.T   <- as.data.frame(unclass(my.array$features.T))
```

Define colors and points

```
Define colors and points for plotting, here two columns in Design are used
cls        <- c("blue","red")      # colour to be used in plots
points     <- c(2,15)              # points to be used in plots
my.cls.s   <- cls[unlist(Design$A.12)]    # colors according to factorA
my.pch.s   <- points[unlist(Design$B.12)] # points according to factorB
```

Plots of scaled data

```
Inn <- scale(features.End)# scale: subtracts mean and scale to unit variance
k   <- dim(Inn)[2]        # k is thereby the number of columns
par(mfrow=c(2,2))
for (i in 1:k){
plot(Inn[,i],ylab=colnames(Inn)[i],xlab='samples',
     col=my.cls.s,pch=my.pch.s,main=colnames(Inn)[i])
abline(h=0,v=0,col='gray50')
}
```



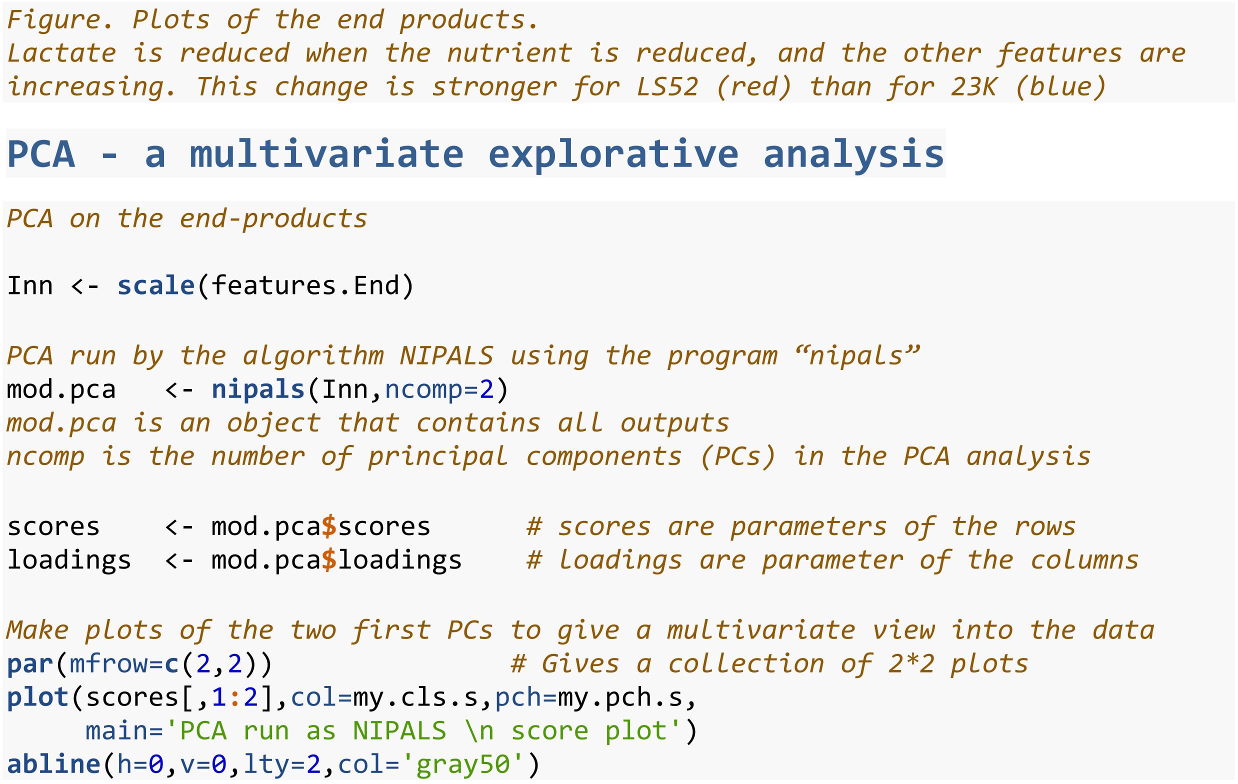


Figure. Plots of the end products.
Lactate is reduced when the nutrient is reduced, and the other features are increasing. This change is stronger for LS52 (red) than for 23K (blue)

PCA - a multivariate explorative analysis

```
PCA on the end-products

Inn <- scale(features.End)

PCA run by the algorithm NIPALS using the program “nipals”
mod.pca   <- nipals(Inn,ncomp=2)
mod.pca is an object that contains all outputs
ncomp is the number of principal components (PCs) in the PCA analysis

scores    <- mod.pca$scores      # scores are parameters of the rows
loadings  <- mod.pca$loadings    # loadings are parameter of the columns

Make plots of the two first PCs to give a multivariate view into the data
par(mfrow=c(2,2))                # Gives a collection of 2*2 plots
plot(scores[,1:2],col=my.cls.s,pch=my.pch.s,
     main='PCA run as NIPALS \n score plot')
abline(h=0,v=0,lty=2,col='gray50')
```

```
my.xlim   <- c(-0.8,0.8);my.ylim <- c(-0.8,0.8)
plot(loadings[,1:2],cex=0.0001,
     xlim=my.xlim,ylim=my.ylim,main='PCA run as NIPALS \n loading plot')
text(loadings[,1:2],labels=rownames(loadings),xlim=my.xlim,ylim=my.ylim)
abline(h=0,v=0,lty=2,col='gray50')

PCA run by the algorithm SVD. This can be performed by the r-program “prcomp”
mod.pca   <- prcomp(Inn, scale = TRUE)
prop.var  <- round(propVar(mod.pca),digits = 2)
scores    <- scores(mod.pca)
loadings  <- loadings(mod.pca)

plot(scores[,1:2],col=my.cls.s,pch=my.pch.s,
     main='PCA run as SVD \n score plot',
     xlab=paste0('PC1 (',prop.var[1],'%)'),
     ylab=paste0('PC2 (',prop.var[2],'%)'))
abline(h=0,v=0,lty=2,col='gray50')

plot(loadings[,1:2],cex=0.0001,
     xlim=my.xlim,ylim=my.ylim,main='PCA run as SVD \n loading plot',
     xlab=paste0('PC1 (',100*prop.var[1],'%)'),
     ylab=paste0('PC2 (',100*prop.var[2],'%)'))
text(loadings[,1:2],labels=rownames(loadings),xlim=my.xlim,ylim=my.ylim)
abline(h=0,v=0,lty=2,col='gray50')
```

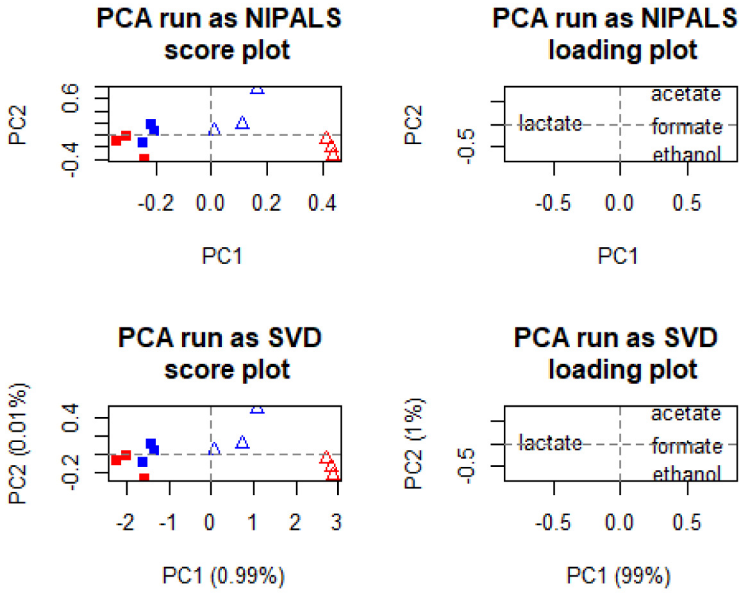
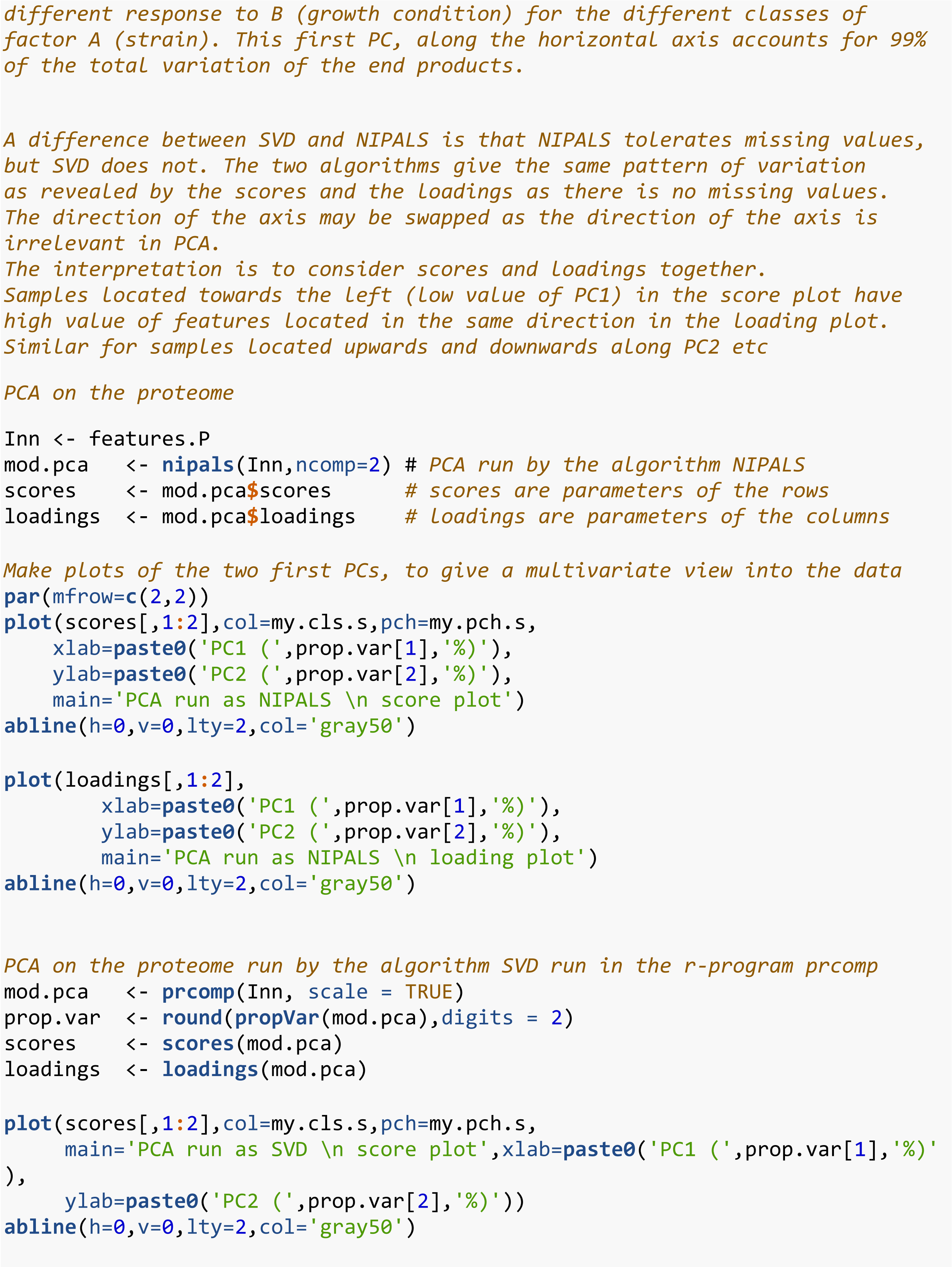


Figure. PCA of the end products.
The interpretation is seen by considering simultaneously each principal component (PC) both in the score plot of the samples, and in the X loading plot of the features. The first PC is here dominated by interacting effects:



```
different response to B (growth condition) for the different classes of
factor A (strain). This first PC, along the horizontal axis accounts for 99%
of the total variation of the end products.


A difference between SVD and NIPALS is that NIPALS tolerates missing values,
but SVD does not. The two algorithms give the same pattern of variation
as revealed by the scores and the loadings as there is no missing values.
The direction of the axis may be swapped as the direction of the axis is
irrelevant in PCA.
The interpretation is to consider scores and loadings together.
Samples located towards the left (low value of PC1) in the score plot have
high value of features located in the same direction in the loading plot.
Similar for samples located upwards and downwards along PC2 etc

PCA on the proteome

Inn <- features.P
mod.pca   <- nipals(Inn,ncomp=2) # PCA run by the algorithm NIPALS
scores    <- mod.pca$scores      # scores are parameters of the rows
loadings  <- mod.pca$loadings    # loadings are parameters of the columns

Make plots of the two first PCs, to give a multivariate view into the data
par(mfrow=c(2,2))
plot(scores[,1:2],col=my.cls.s,pch=my.pch.s,
    xlab=paste0('PC1 (',prop.var[1],'%)'),
    ylab=paste0('PC2 (',prop.var[2],'%)'),
    main='PCA run as NIPALS \n score plot')
abline(h=0,v=0,lty=2,col='gray50')

plot(loadings[,1:2],
        xlab=paste0('PC1 (',prop.var[1],'%)'),
        ylab=paste0('PC2 (',prop.var[2],'%)'),
        main='PCA run as NIPALS \n loading plot')
abline(h=0,v=0,lty=2,col='gray50')



PCA on the proteome run by the algorithm SVD run in the r-program prcomp
mod.pca   <- prcomp(Inn, scale = TRUE)
prop.var  <- round(propVar(mod.pca),digits = 2)
scores    <- scores(mod.pca)
loadings  <- loadings(mod.pca)

plot(scores[,1:2],col=my.cls.s,pch=my.pch.s,
     main='PCA run as SVD \n score plot',xlab=paste0('PC1 (',prop.var[1],'%)'
),
     ylab=paste0('PC2 (',prop.var[2],'%)'))
abline(h=0,v=0,lty=2,col='gray50')
```

```
plot(loadings[,1:2],
     main='PCA run as SVD \n loading plot',
     xlab=paste0('PC1 (',prop.var[1],'%)'),ylab=paste0('PC2 (',prop.var[2],'%
)'))
abline(h=0,v=0,lty=2,col='gray50')
```

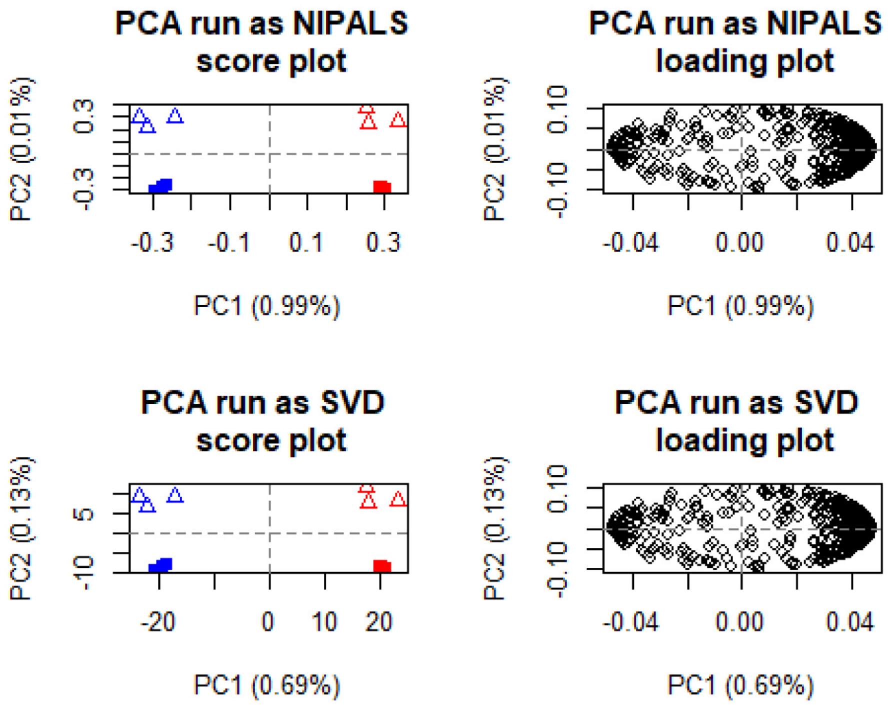


Figure. PCA of the proteome. The two first PC describe main effects of the design factor A (strain) and B (growth condition).

Multivariate supervised analysis

PLS is a supervised multivariate analysis to study how one data block is related to other data blocks.
Here: PLSR of X (Protome) => Y (Phenome).
PLS is run by the program plsr in the package "plsr"

```
my.X        <- scale(as.matrix(features.P))
my.Y        <- scale(as.matrix(features.End));

par(mfrow=c(3,3))
my.ncomp    <- 4   # define the number of PLS factors used in the model
my.data     <- data.frame(my.Y,my.X)
mod.plsr    <- plsr(my.Y ~ my.X, ncomp= my.ncomp, data = my.data,
                    validation = "LOO", jackknife = TRUE)

scores    <- mod.plsr$scores    # the parameters of the rows for both X and Y
Xloadings <- mod.plsr$loadings  # the parameters of the input columns in X
rownames(Xloadings)<- colnames(my.X)
Yloadings <- mod.plsr$Yloadings # the parameters of the input columns in Y
jkn       <- jack.test(mod.plsr, ncomp = my.ncomp, use.mean = TRUE)
p.jkn     <- jkn$pvalues;
l         <- 0.001
N.sel        <- names(which(p.jkn[,1,]<l))
par(mfrow=c(2,3))
plot(scores[,1:2],col=my.cls.s,pch=my.pch.s,
     xlab='PLS factor 1',ylab='PLS factor 2',
     main='Score plot')
abline(h=0,v=0,lty=2,col='gray50')

plot(Xloadings[,1:2],
     xlab='PLS factor 1',ylab='PLS factor 1',
     main= 'Xloading plot')
abline(h=0,v=0,lty=2,col='gray50')

plot(Yloadings[,1:2],
     xlab='PLS factor 1',ylab='PLS factor 2',
     main= 'Yloading plot',
     col='gray')
abline(h=0,v=0,lty=2,col='gray50')
text(Yloadings[,1:2],labels=colnames(my.Y))

plot(scores[,3:4],col=my.cls.s,pch=my.pch.s,
     main='Score plot',
     xlab='PLS factor 3',ylab='PLS factor 4')
abline(h=0,v=0,lty=2,col='gray50')

plot(Xloadings[,3:4],
     xlab='PLS factor 3',ylab='PLS factor 4',
     main= 'Xloading plot')
abline(h=0,v=0,lty=2,col='gray50')

plot(Yloadings[,3:4],
     xlab='PLS factor 3',ylab='PLS factor 4',
     main= 'Yloading plot',
     col='gray')
abline(h=0,v=0,lty=2,col='gray50')
text(Yloadings[,3:4],labels=colnames(my.Y))
```

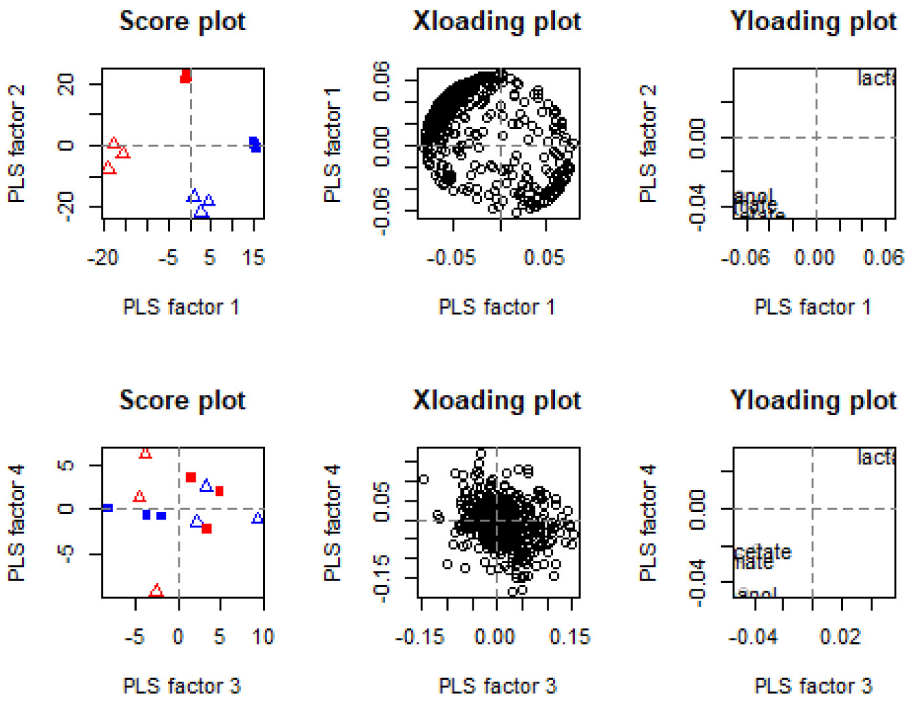


Figure. PLS regression of proteome (input) towards end-products (response). The two first PLS factors (the upper plots) describe main effects of the design factor A (strain) and B (growth condition).
The third PLS factor (horizontal along the lower plots) reflects interacting pattern describing different responses for the two strains

Effect + Residual (ER) modelling

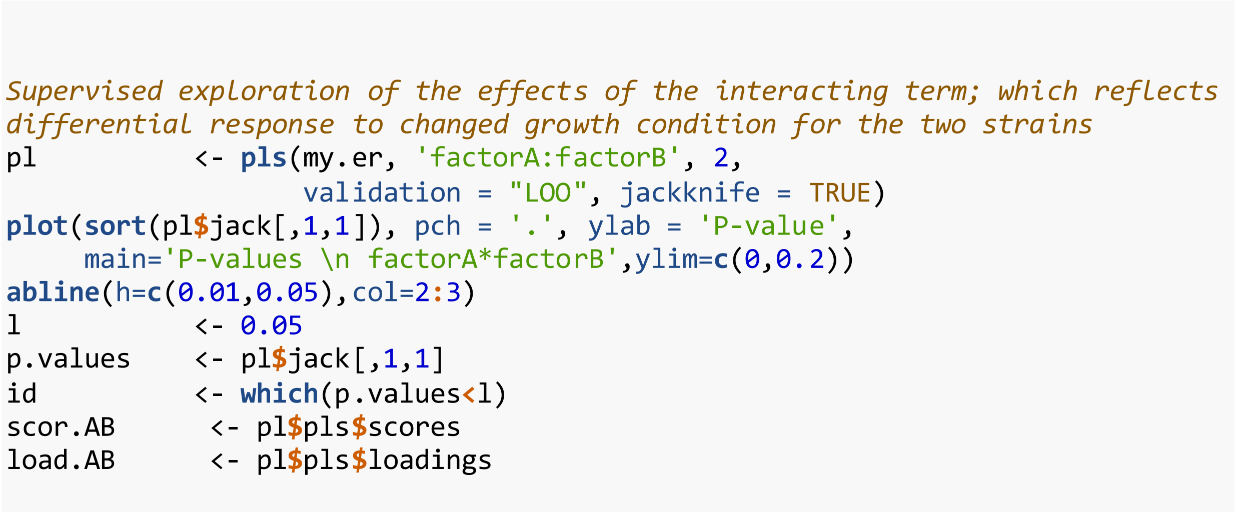
Performs a linear model that first estimates effects of each input factor on each feature. The residual of the complete model is thereafter added to each effect. For an experiment of 2 factors and their interaction term this results in three data tables; one for each main effect and one for the interaction term.

```
my.er     <- ER(scale(features.P) ~ factorA*factorB, data = my.array)
```

PLS-DA after Effect + Residual modelling

Any suitable algorithms can be performed for feature selection. Here we apply Jackknife

```
my.ncomp    <- 2
par(mfrow=c(2,3))

# Supervised exploration of the effects of strain on the proteome
pl          <- pls(my.er, 'factorA', 2, validation = "LOO", jackknife = TRUE)
plot(sort(pl$jack[,1,1]), pch = '.', ylab = 'P-value',
     main='P-values \n factorA',ylim=c(0,0.2))
abline(h=c(0.01,0.05),col=2:3)
l           <- 0.05
choose a rejection limit l, p=0.05 often used, but should be considered accor
ding to the data
p.values    <- pl$jack[,1,1]
extract p-value from the jackknife test, which validates the stability of the
regression coefficients
id          <- which(p.values<l)
identify the features with p-values below the chosen level of l
scor.A      <- pl$pls$scores
extract scores of the samples to be used for plotting and visualisation
load.A      <- pl$pls$loadings
extract loadings of the features to be used for plotting and visualisation
n.s.jkn.A   <- rownames(load.A)[id]
features selected for factorA


Supervised exploration of the effects of diabetes on the transcriptome
pl          <- pls(my.er, 'factorB', 2, validation = "LOO", jackknife = TRUE)
plot(sort(pl$jack[,1,1]), pch = '.', ylab = 'P-value',
     main='P-values \n factorB',ylim=c(0,0.2))
abline(h=c(0.01,0.05),col=2:3)
l           <- 0.05
p.values    <- pl$jack[,1,1]
id          <- which(p.values<l)
scor.B      <- pl$pls$scores
load.B      <- pl$pls$loadings
n.s.jkn.B   <- rownames(load.B)[id] # features selected for factorB
```



```
Supervised exploration of the effects of the interacting term; which reflects
differential response to changed growth condition for the two strains
pl          <- pls(my.er, 'factorA:factorB', 2,
                   validation = "LOO", jackknife = TRUE)
plot(sort(pl$jack[,1,1]), pch = '.', ylab = 'P-value',
     main='P-values \n factorA*factorB',ylim=c(0,0.2))
abline(h=c(0.01,0.05),col=2:3)
l           <- 0.05
p.values    <- pl$jack[,1,1]
id          <- which(p.values<l)
scor.AB      <- pl$pls$scores
load.AB      <- pl$pls$loadings
```

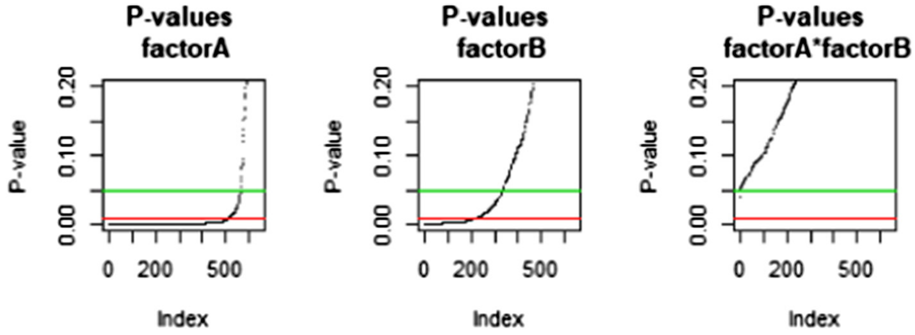
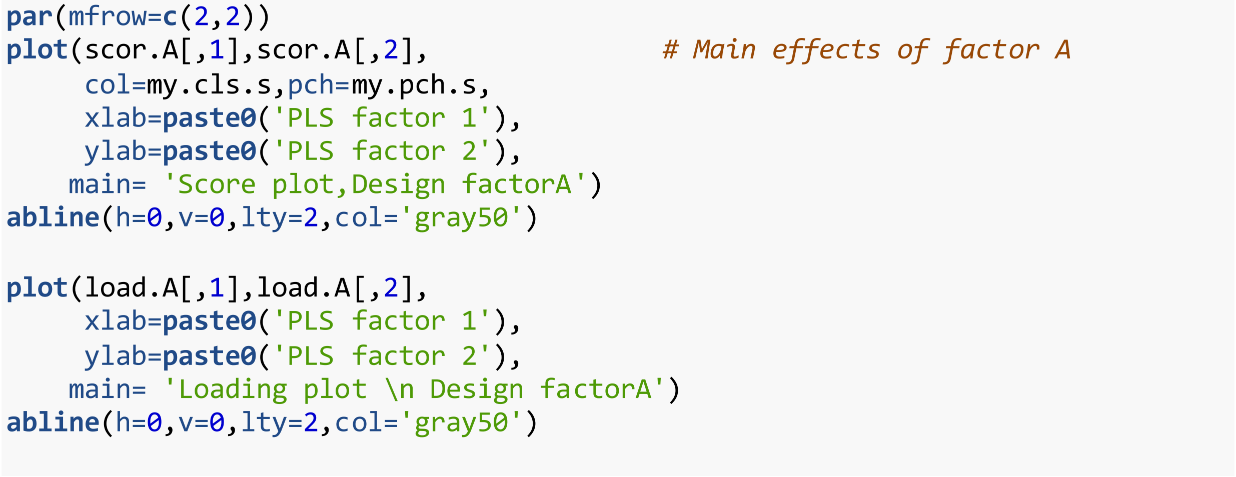


Figure. P-values obtained when using Jackknife for feature selection. Two lines are displayed to visualize the number of features selected by two different rejection limits

Plot PLS-DA results



```
par(mfrow=c(2,2))
plot(scor.A[,1],scor.A[,2],                  # Main effects of factor A
     col=my.cls.s,pch=my.pch.s,
     xlab=paste0('PLS factor 1'),
     ylab=paste0('PLS factor 2'),
    main= 'Score plot,Design factorA')
abline(h=0,v=0,lty=2,col='gray50')

plot(load.A[,1],load.A[,2],
     xlab=paste0('PLS factor 1'),
     ylab=paste0('PLS factor 2'),
    main= 'Loading plot \n Design factorA')
abline(h=0,v=0,lty=2,col='gray50')
```

```
plot(scor.B[,1],scor.B[,2],                    # Main effects of factor B
     col=my.cls.s,pch=my.pch.s,
     xlab=paste0('PLS factor 1'),
     ylab=paste0('PLS factor 2'),
     main= 'Score plot \n Design factorB')
abline(h=0,v=0,lty=2,col='gray50')

plot(load.B[,1],load.B[,2],
     xlab=paste0('PLS factor 1'),
     ylab=paste0('PLS factor 2'),
     main= 'Loading plot \n Design factorB')
abline(h=0,v=0,lty=2,col='gray50')
```

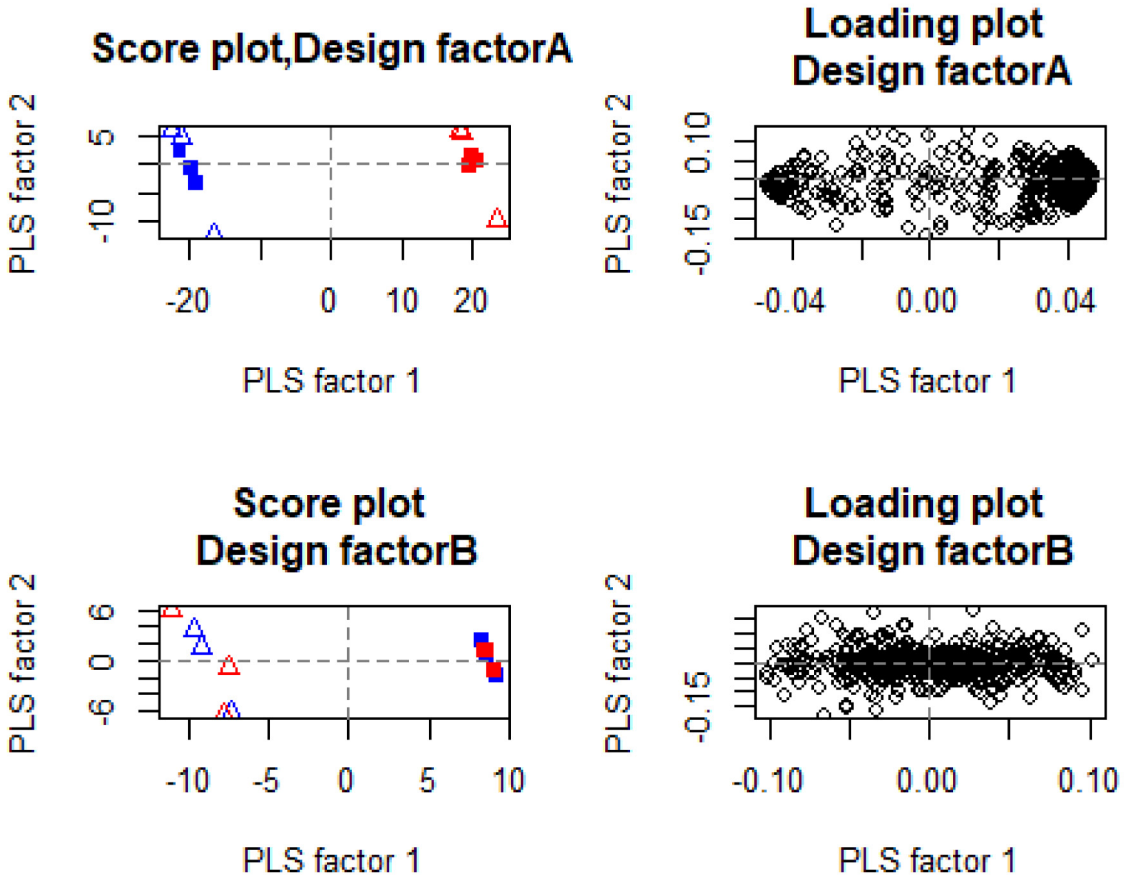
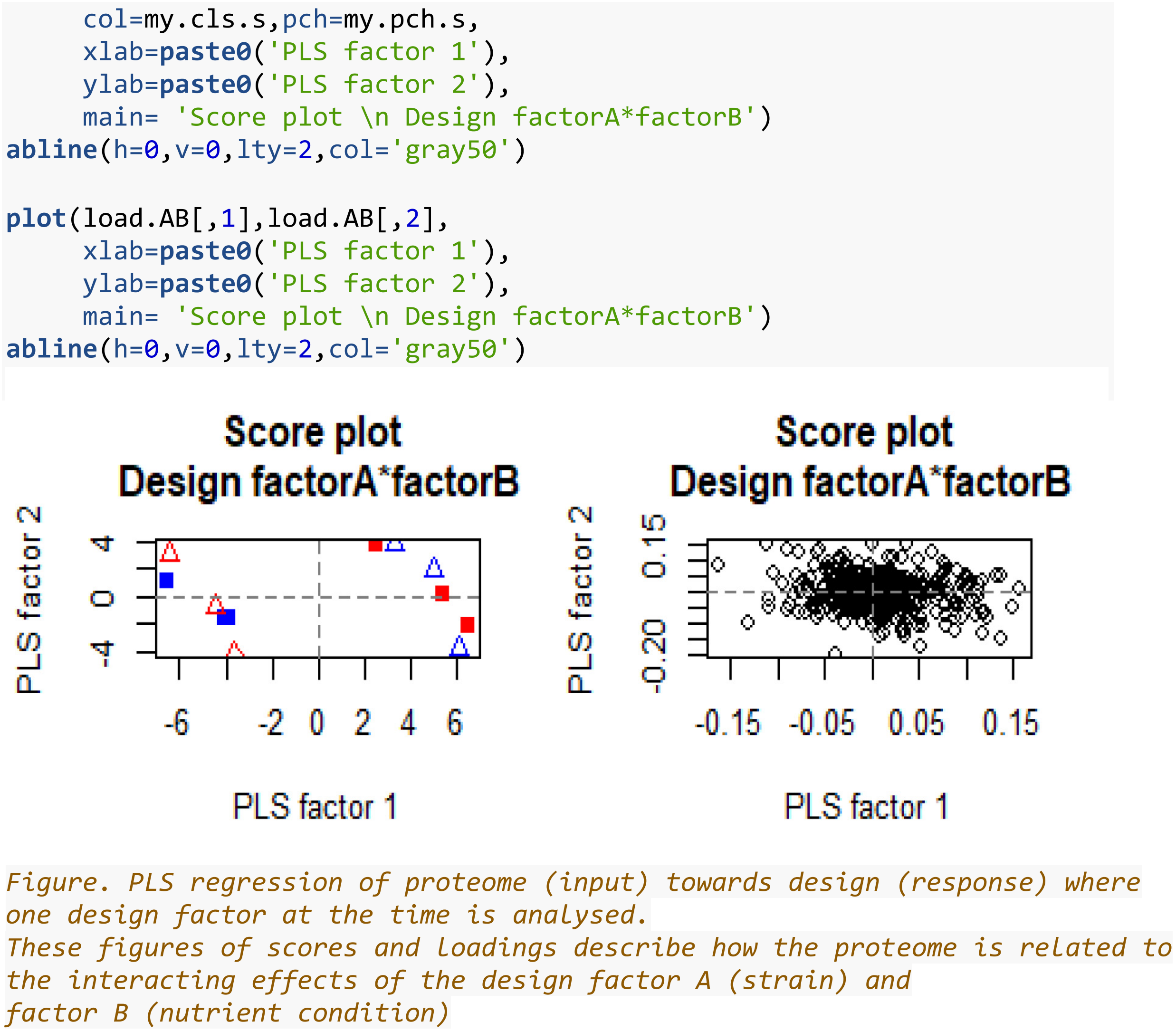


Figure. PLS regression of proteome (input) towards design (response) where one design factor at the time is analysed.
The two first figures of scores and loadings describe how the proteome is related to the design factor A (strain) and the two next figures of scores and loadings describe how the proteome relates to factor B (growth condition)

```
par(mfrow=c(2,2))
plot(scor.AB[,1],scor.AB[,2],                      # The interacting effects
```



```
     col=my.cls.s,pch=my.pch.s,
     xlab=paste0('PLS factor 1'),
     ylab=paste0('PLS factor 2'),
     main= 'Score plot \n Design factorA*factorB')
abline(h=0,v=0,lty=2,col='gray50')

plot(load.AB[,1],load.AB[,2],
     xlab=paste0('PLS factor 1'),
     ylab=paste0('PLS factor 2'),
     main= 'Score plot \n Design factorA*factorB')
abline(h=0,v=0,lty=2,col='gray50')
```

Figure. PLS regression of proteome (input) towards design (response) where one design factor at the time is analysed.

These figures of scores and loadings describe how the proteome is related to the interacting effects of the design factor A (strain) and factor B (nutrient condition)

A.3 Additional file 3

Preprocessing

This section will include how to use Python to implement some of the steps described in the preprocessing section of this book. The package scikit-learn will be used throughout the section. More information about scikit-learn can be found at https://scikit-learn.org/

The data can be loaded with pandas as a data frame:

```
[1]: import pandas as pd

     dataframe = pd.read_excel('Example Data.xlsx', index_col=0)
     dataframe
```

```
[1]:            Gene 1  Gene 2  Gene 3  Gene 4  Gene 5  Gene 6  Gene 7  Class Label
     Sample 1     1097     145     132     127     168     132     133            0
     Sample 2      694     142     128     130     161     127     132            0
     Sample 3      959     156     127     129     183     126     128            0
     Sample 4      887     140     129     127     208     118     140            0
     Sample 5      946     151     119     129     180     123     135            0
     Sample 6      877     157     126     129     180     122     147            1
     Sample 7      784     174     128     131     185     123     135            1
     Sample 8      737     148     128     121     166     124     143            1
     Sample 9      814     140     128     133     178     116     136            1
     Sample 10     617     143     136     131     205     138     130            1
```

The data can then be separated into data and response variable, X and y:

```
[2]: X = dataframe.drop(columns='Class Label')
     y = dataframe['Class Label']
```

```
[3]: X.head() # the first 5 rows of X
```

```
[3]:           Gene 1  Gene 2  Gene 3  Gene 4  Gene 5  Gene 6  Gene 7
     Sample 1    1097     145     132     127     168     132     133
     Sample 2     694     142     128     130     161     127     132
     Sample 3     959     156     127     129     183     126     128
     Sample 4     887     140     129     127     208     118     140
     Sample 5     946     151     119     129     180     123     135
```

```
[4]: y
```

```
[4]: Sample 1     0
     Sample 2     0
     Sample 3     0
     Sample 4     0
     Sample 5     0
     Sample 6     1
     Sample 7     1
     Sample 8     1
Sample 9      1
Sample 10     1
Name: Class Label, dtype: int64
```

The data can then be split into training data and test data to evaluate the performance of the model.

In the example below, test_size = 0.3 will split 30% of the data into the dataset X_test and 70% into X_train. To ensure reproducibility, random_state is used and can be set to any integer. For classification tasks, stratify = y will ensure the data is split with the same class proportions as the original data.

```
[5]: from sklearn.model_selection import train_test_split
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,␣
      ↪random_state=1, stratify=y)
```

Scaling

As discussed in the preprocessing section of this chapter, it is common to scale features so that features with larger values do not have more influence than features with smaller values. In machine learning, it is common to scale and center the data at zero since machine learning models typically preform best under these conditions.

Traditionally, logarithmic and square root transformations have been applied to omics data to make the data more normally distributed. These techniques, however, will not center the data at 0 or ensure the features are on the same scale. Both logarithmic and square root transformations are types of Box-Cox transformations and can be implemented in scikit-learn through the PowerTransformer class. PowerTransformer uses log-likelihood to optimize the parameter λ in the formula below to make the data as close to normally distributed as possible.

$$x_i^{\lambda} = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(x_i) & \text{if } \lambda = 0 \end{cases} \tag{1}$$

PowerTransformer will by default also standardize the data, but this can be changed by setting standardize=False.

In machine learning, standardization and normalization are most commonly used and can be implemented using the classes StandardScaler and MinMaxScaler, respectively. StandardScaler will give a distribution with a mean of 0 and a standard deviation of 1 but is susceptible to outliers. MinMaxScaler will scale each feature to a range of 0 to 1, however it is also very susceptible to outliers.

Normalization:

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}$$

Standardization:

$$\frac{x_i - mean(x)}{stdev(x)} \tag{3}$$

Using the X_train and X_test datasets, Box-Cox and scaling can be applied using the following code:

```
[6]: from sklearn.preprocessing import PowerTransformer
     pt = PowerTransformer(method='box-cox', standardize=True)
     X_train_pt = pt.fit_transform(X_train)
     X_test_pt = pt.transform(X_test)

     from sklearn.preprocessing import StandardScaler
     ss = StandardScaler()
     X_train_ss = ss.fit_transform(X_train)
     X_test_ss = ss.transform(X_test)

     from sklearn.preprocessing import MinMaxScaler
     mms = MinMaxScaler()
     X_train_mms = mms.fit_transform(X_train)
     X_test_mms = mms.transform(X_test)
```

References

1. Pollard, T. D.; Earnshaw, W. C.; Lippincott-Schwartz, J.; Johnson, G. *Cell Bilogy*, Elsevier Health Sciences, 2016; p 908.
2. Brown, T. A. *Genomes 4*, Garland Science: New York, 2018.
3. Perkins, R. C. In *Functional Proteomics. Methods in Molecular Biology;* Wang, X., Matthew, K., Eds., Humana Press, Springer Protocols: New York, NY, 2018.
4. Pawar, A. K.; Samhitha, T.; Prasanna, B.; Asif, M. L. Metabolomics: Current Technologies and Future Trends. *Indo Am. J. Pharm. Sci.* **2018,** *5,* 5114–5121.
5. Pinu, F. R.; et al. Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. *Metabolites* **2019,** *9.*
6. Campbell, R. F.; McGrath, P. T.; Paaby, A. B. Analysis of Epistasis in Natural Traits Using Model Organisms. *Trends Genet.* **2018,** *34,* 883–898.
7. Kogenaru, M.; de Vos, M. G. J.; Tans, S. J. Revealing Evolutionary Pathways by Fitness Landscape Reconstruction. *Crit. Rev. Biochem. Mol. Biol.* **2009,** *44,* 169–174.
8. Nghe, P.; Kogenaru, M.; Tans, S. J. Sign Epistasis Caused by Hierarchy Within Signalling Cascades. *Nat. Commun.* **2018,** *9.*
9. Deery, M. J.; et al. Proteomic Analysis Reveals the Role of Synaptic Vesicle Cycling in Sustaining the Suprachiasmatic Circadian Clock. *Curr. Biol.* **2009,** *19,* 2031–2036.
10. Ruben, M. D.; et al. A Database of Tissue-specific Rhythmically Expressed Human Genes Has Potential Applications in Circadian Medicine. *Sci. Transl. Med.* **2018,** *10.*
11. Harrigan, G. G.; Goodacre, R. *Metabolic Profileing: Its Role in Biomarker Discovery and Gene Function Analysis*, Springer Nature, 2003.
12. Yan, M.; Xu, G. W. Current and Future Perspectives of Functional Metabolomics in Disease Studies-a Review. *Anal. Chim. Acta* **2018,** *1037,* 41–54.
13. Tabery, J. R.; Fisher, A. Lancelot Hogben, and the Origin(S) of Genotype-Environment Interaction. *J. Hist. Biol.* **2008,** *41,* 717–761.
14. Falconer, D. S.; Mackey, T. F. C. *Introduction to Quantitative Genetics*, Addison-Wesley Longman: Harlow, UK, 1996.
15. Rice, J. C.; Allis, C. D. Review Histone Methylation Versus Histone Acetylation: New Insights Into Epigenetic Regulation. *Curr. Opin. Cell Biol.* **2001,** *13(3),* 263–273.
16. Greally, J. M. A User's Guide to the Ambiguous Word 'Epigenetics'. *Nat. Rev. Mol. Cell Biol.* **2018,** *19,* 207–208.
17. Hawe, J. S.; Theis, F. J.; Heinig, M. Inferring Interaction Networks From Multi-Omics Data. *Front. Genet.* **2019,** *10.*
18. Box, G.; Hunter, J. S.; Hunter, W. G. *Statistics for Experimenters. Design, Innovation, and Discovery,* 2nd edition;, Wiley, 2005.
19. Montgomery, D. C. *Design and Analysis of Experiments,* 8th edition;, Wiley, 2013.
20. Jackson, O. L.; Liyanage, R.; Borgmann, S.; Wilkins, C. L. Problems with the "Omics". *TrAC, Trends Anal. Chem.* **2006,** *25,* 1046–1056.
21. Wold, H. Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach. *J. Appl. Probab.* **1975,** *12,* 117–142.
22. Wold, S.; Martens, H.; Wold, H. In *Proc. Conf. Matrix Pencils, Lecture Notes in Mathematics;* Ruhe, B. K. A., Ed., Springer: Heidelberg, 1983; pp 286–293.
23. Martens, H. Quantitative Big Data: Where Chemometrics Can Contribute. *J. Chemom.* **2015,** *29,* 563–581.
24. Tauler, R.; Parastar, H. *Angew. Chem. Int.* **2018**.
25. Wright, S. Correlation and Causality. *J. Agric. Res.* **1921,** *557.*
26. Schwammle, V.; Verano-Braga, T.; Roepstorff, P. Computational and Statistical Methods for High-Throughput Analysis of Post-Translational Modifications of Proteins. *J. Proteomics* **2015,** *129,* 3–15.
27. Tyanova, S.; et al. The Perseus Computational Platform for Comprehensive Analysis of (Prote)Omics Data. *Nat. Methods* **2016,** *13,* 731–740.
28. He, J. M.; et al. MassImager: A Software for Interactive and in-Depth Analysis of Mass Spectrometry Imaging Data. *Anal. Chim. Acta* **2018,** *1015,* 50–57.
29. Chen, K.; Park, J.; Li, F.; Patil, S. M.; Keire, D. A. Chemometric Methods to Quantify 1D and 2D NMR Spectral Differences among Similar Protein Therapeutics. *AAPS PharmSciTech* **2018,** *19,* 1011–1019.
30. Zhang, Z. M.; et al. Chemometrics in Instrumental Analysis of Complex Systems-in Honor and Memory of Yi-Zeng Liang. *J. Chemom.* **2018,** *32,* 22.
31. Richards, S. E.; Holmes, E. *Chemometrics Methods for the Analysis of Genomics, Transcriptomics, Proteomics, Metabolomics, and Metagenomics Datasets*, Woodhead Publishing Series in Food Science, Technology and Nutrition, 2015.
32. Gidskehaug, L.; Anderssen, E.; Flatberg, A.; Alsberg, B. K. A Framework for Significance Analysis of gEne Expression Data Using Dimension Reduction Methods. *BMC Bioinformatics* **2007,** *8.*
33. Acar, E.; Bro, R.; Smilde, A. K. Data Fusion in Metabolomics Using Coupled Matrix and Tensor Factorizations. *Proc. IEEE* **2015,** *103,* 1602–1620.
34. Gardlo, A.; et al. Normalization Techniques for PARAFAC Modeling of Urine Metabolomic Data. *Metabolomics* **2016,** *12.*
35. Timmerman, M. E.; Hoefsloot, H. C. J.; Smilde, A. K.; Ceulemans, E. Scaling in ANOVA-Simultaneous Component Analysis. *Metabolomics* **2015,** *11,* 1265–1276.
36. Vis, D. J.; et al. Analyzing Metabolomics-Based Challenge Tests. *Metabolomics* **2015,** *11,* 50–63.
37. Hasdemir, D.; Smits, G. J.; Westerhuis, J. A.; Smilde, A. K. Topology of Transcriptional Regulatory Networks: Testing and Improving. *Plos One* **2012,** *7.*
38. Jansen, J. J.; et al. Between Metabolite Relationships: An Essential Aspect of Metabolic Change. *Metabolomics* **2012,** *8,* 422–432.

39. Jansen, J. J.; Szymanska, E.; Hoefsloot, H. C. J.; Smilde, A. K. Individual Differences in Metabolomics: Individualised Responses and between-Metabolite Relationships. *Metabolomics* **2012,** *8,* S94–S104.
40. Szymanska, E.; Saccenti, E.; Smilde, A. K.; Westerhuis, J. A. Double-Check: Validation of Diagnostic Statistics for PLS-DA Models in Metabolomics Studies. *Metabolomics* **2012,** *8,* S3–S16.
41. Smilde, A. K.; et al. ANOVA-Simultaneous Component Analysis (ASCA): A New Tool for Analyzing Designed Metabolomics Data. *Bioinformatics* **2005,** *21,* 3043–3048.
42. Deleted in review.
43. Farag, Y.; Berven, F. S.; Jonassen, I.; Petersen, K.; Barsnes, H. Distributed and Interactive Visual Analysis of Omics Data. *J. Proteomics* **2015,** *129,* 78–82.
44. Oveland, E.; et al. Viewing the Proteome: How to Visualize Proteomics Data? *Proteomics* **2015,** *15,* 1341–1355.
45. Csala, A.; Hof, M. H.; Zwinderman, A. H. Multiset Sparse Redundancy Analysis for High-Dimensional Omics Data. *Biom. J.* **2019,** *61,* 406–423.
46. Csala, A.; Zwubdernan, A. H. In *Computational Biology;* Husi, H., Ed., Codon Publications: Brisbane, 2019.
47. Esbensen, K.H., Swarbrick, B., Westad, F., Whitcombe, P. & Anderson, M.J. Multivariate Data Analysis: An introduction to Multivariate Analysis, Process Analytical Technology and Quality by Design. Camo: Oslo; 2018, p. 462.
48. Bianconi, F.; Antonini, C.; Tomassoni, L.; Valigi, P. Robust Calibration of High Dimension Nonlinear Dynamical Models for Omics Data: An Application in Cancer Systems Biology. *IEEE Trans. Contr. Syst. Technol.* **2020,** *28,* 196–207.
49. Smilde, A.; Bro, R.; Geladi, P. *Multi-Way Analysis With Applications in the Chemometrical Sciences*, John Wiley & Sons: Chichester, U.K., 2004.
50. Moen, B.; et al. Explorative Multifactor Approach for Investigating Global Survival Mechanisms of Campylobacter Jejuni under Environmental Conditions. *Appl. Environ. Microbiol.* **2005,** *71,* 2086.
51. Oust, A.; et al. Analysis of Covariance Patterns in Gene Expression Data and FT-IR Spectra. *J. Microbiol. Methods* **2006,** *65,* 573–584.
52. Dankel, S. N.; et al. Switch from Stress Response to Homeobox Transcription Factors in Adipose Tissue After Profound Fat Loss. *Plos One* **2010,** *5.*
53. Barajas-Olmos, F.; et al. Altered DNA Methylation in Liver and Adipose Tissues Derived from Individuals With Obesity and Type 2 Diabetes. *BMC Med. Genet.* **2018,** *19.*
54. McLeod, A.; et al. Effects of Glucose Availability in *Lactobacillus sakei*; Metabolic Change and Regulation of the Proteome and Transcriptome. *PLoS One* **2017,** *12.*
55. Robinson, N. A.; et al. Response of the Salmon Heart Transcriptome to Pancreas Disease: Differences between High- and Low-Ranking Families for Resistance. *Sci. Rep.* **2020,** *10,* 868.
56. Babicki, S.; et al. Heatmapper: Web-Enabled Heat Mapping for all. *Nucleic Acids Res.* **2016,** *44,* W147–W153.
57. Hotelling, H. Analysis of a Complex of Statistical Variables into Principal Components. *J. Educ. Psychol.* **1933,** *24* (417–441), 498–520.
58. Martens, H.; Næs, T. *Mulivariate Calibration*, Chichester, U.K: John Wiley & Sons Ltd, 1989.
59. Kern, L.; et al. Obesity-Induced TNF alpha and IL-6 Signaling: The Missing Link between Obesity and Inflammation-Driven Liver and Colorectal Cancers. *Cancer* **2019,** *11.*
60. Martens, H.; Martens, M. Modified Jack-Knife Estimation of Parameter Uncertainty in Bilinear Modelling by Partial Least Squares Regression (PLSR). *Food Qual. Prefer.* **2000,** *11,* 5–16.
61. Saebo, S.; Almoy, T.; Aaroe, J.; Aastveit, A. H. ST-PLS: A Multi-Directional Nearest Shrunken Centroid Type Classi Er Via PLS. *J. Chemom.* **2008,** *22,* 54–62.
62. Rajalahti, T.; et al. Biomarker Discovery in Mass Spectral Profiles by Means of Selectivity Ratio Plot. *Chemom. Intell. Lab. Syst.* **2009,** *95,* 35–48.
63. Quiroga, R. Q.; Reddy, L.; Kreiman, G.; Koch, C.; Fried, I. Invariant Visual Representation by Single Neurons in the Human Brain. *Nature* **2005,** *435,* 1102–1107.
64. Leardi, R. Application of Genetic Algorithm-PLS for Feature Selection in Spectral Data Sets. *J. Chemom.* **2000,** *14,* 643–655.
65. Leardi, R.; Gonzalez, A. L. Genetic Algorithms Applied to Feature Selection in PLS Regression: How and When to Use Them. *Chemom. Intell. Lab. Syst.* **1998,** *41,* 195–207.
66. Li, H.; Gui, J. Partial Cox Regression Analysis for High-Dimensional Microarray Gene Expression Data. *Bioinformatics* **2004,** *20,* 208–215.
67. Breiman, L. Random Forests. *Mach. Learn.* **2001,** *45,* 5–32.
68. Fisher, R. A.; Mackenzie, W. A. Studies in Crop Variation. II. The manurial response of different potato varieties. *J. Agric. Sci.* **1923,** *13,* 311–320.
69. Fisher, R. A.; Wishart, J. The Arrangement of Field Experiments and the Statistical Reduction of the Results. *Imp. Bur. Soil Sci. Tech. Commun.* **1930,** *10,* 23.
70. Storey, J. D.; Tibshirani, R. Statistical Significance for Genomewide Studies. *Proc. Natl. Acad. Sci. U. S. A.* **2003,** *100,* 9440–9445.
71. Marini, F.; de Beer, D.; Joubert, E.; Walczak, B. Analysis of Variance of Designed Chromatographic Data Sets: The Analysis of Variance-Target Projection Approach. *J. Chromatogr. A* **2015,** *1405,* 94–102.
72. Angelina, E.; Qannari, E.; Moyon, T.; Alexandre-Gouabau, M. C. AoV-PLS: A New Method for the Analysis of Multivariate Data Depending on Several Factors. *Electron. J. Appl. Stat. Anal.* **2015,** *8,* 214–235.
73. Liland, K. H.; Smilde, A.; Marini, F.; Naes, T. Confidence Ellipsoids for ASCA Models Based on Multivariate Regression Theory. *J. Chemometr.* **2018,** *32.*
74. Zou, H.; Hastie, T. Regularization and Variable Selection Via the Elastic Net. *J. R.l Stat. Soc. B* **2005,** *67,* 301–320.
75. Peng, L.; Jiang, Q.; Pan, J. Y.; Deng, C.; Yu, J. Y.; Wu, X. M.; Huang, S. H.; Deng, X. Y. Involvement of Polyphosphate Kinase in Virulence and Stress Tolerance of Uropathogenic Proteus mirabilis. *Med. Microbiol. Immunol.* **2016,** *205,* 97–109.
76. Ramalingam, L.; et al. The Renin Angiotensin System, Oxidative Stress and Mitochondrial Function in Obesity and Insulin Resistance. *BBA-Mol. Basis Dis.* **2017,** *1863,* 1106–1114.
77. Liland, K. H.; Stefansson, P.; Indahl, U. G. Much Faster Cross-validation in PLSR-modElling by Avoiding Redundant Calculations. *J. Chemometr.* **2020**.

Further Reading

Pray, L. A. Discovery of DNA Structure and Function: Watson and crick. *Nature Education* **2008,** *1.*

Singer, M.; Berg, P. *Genes and Genomes: A Changing Perspective*, University Science Books: Mill Valley, CA, 1991.