4.08 Analysis of Megavariate Data in Functional Genomics

E. M. Færgestad, Ø. Langsrud, M. Høy, and K. Hollung, Nofima Mat, Norwegian Institute of Food, Fisheries and Aquaculture Research, Ås, Norway

S. Sæbø and K. H. Liland, Norwegian University of Life Sciences, Ås, Norway

A. Kohler, Nofima Mat, Norwegian Institute of Food, Fisheries and Aquaculture Research, Ås, Norway; Norwegian University of Life Sciences, Ås, Norway

L. Gidskehaug, Norwegian University of Life Sciences, Ås, Norway; Centre for Integrative Genetics, Ås, Norway

J. Almergren, Centre for Integrative Genetics, Ås, Norway

E. Anderssen, Norwegian University of Science and Technology, Trondheim, Norway

H. Martens, Nofima Mat, Norwegian Institute of Food, Fisheries and Aquaculture Research, Ås, Norway; Norwegian University of Life Sciences, Ås, Norway; Centre for Integrative Genetics, Ås, Norway

Symbols

x	a scalar
$\mathbf{x}$	a vector (a column of numbers)
$\mathbf{X}$	a matrix (a table of numbers)
$\underline{\mathbf{X}}$	*N*-way array (i.e., a 3-way array is $\mathbf{X}$

The choice of strategy for analyzing data from functional genomics must be based on an understanding of the system under study. We therefore start with a description of crucial aspects of functional genomics and the data typically generated as the empirical basis of functional genomics 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 approaches for data analysis.

4.08.2 Molecular Basis of Functional Genomics

An overview of different types of data and their relations in functional genomics is given in **Figure 1**. The figure describes the flow of information from genetic makeup to the final phenotypic expression. Below we go through each of these sources of information. For more comprehensive information on these topics see, for example, Brown.[1]

4.08.2.1 Genome

The genome is usually composed of deoxyribonucleic acids (DNAs) (**Figure 2**), which are long polymeric molecules of nucleotides. Each nucleotide consists, of (1) a phosphate group, (2) a sugar molecule, and (3) 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 allow 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 may be alternative 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. Most eukaryotic organisms (higher plants and animals) are so-called diploids, which means that each normal cell carries two copies of each chromosome. 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. 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 **Figure 2**). 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) have a single circular DNA, whereas eukaryotic organisms have long linear DNA macromolecules.

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

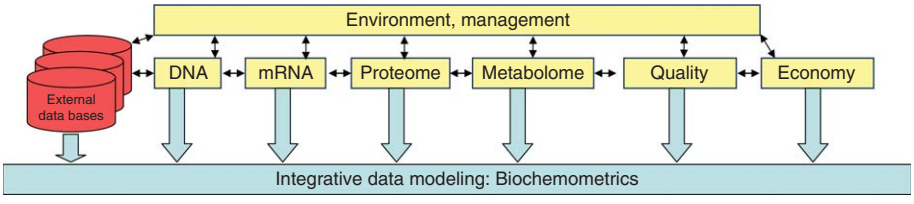

Figure 1 Overview of different function genomics data.

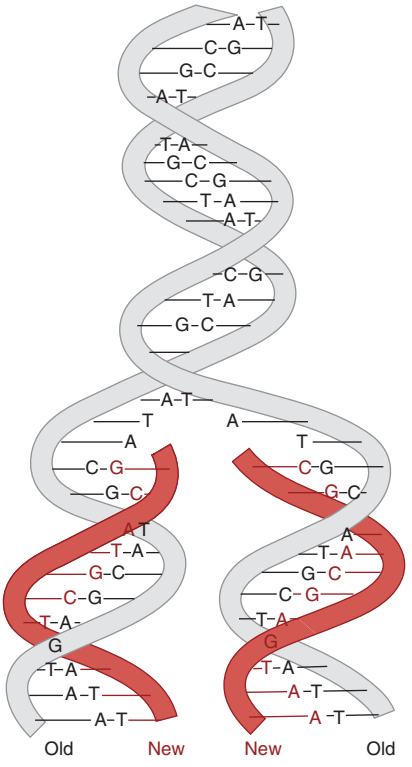

Figure 2 DNA duplication during mitosis. Reproduced with permission from Singer, M.; Berg, P. *Genes and Genomes*: *A Changing Perspective*. University Science Books: Mill Valley, CA, 1991.

constitute the basis for biological diversity. Most mutations are harmful or lethal to the cell or the organism, whereas others do not affect the phenotype at all. In some cases, however, the mutation may be beneficial and will, by the survival of the fittest, be manifested in the following generations.

Prokaryotes mostly reproduce by cell division, whereas eukaryotic organisms usually have sexual reproduction. In the production of germ cells (meiosis), the two chromatid pairs are split, 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, one allele dominates over the other with respect to the resulting phenotype; that is, one allele is dominant and the other is recessive. 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. Further, one gene may affect several phenotypic characteristics. This is called pleiotropy. Another important phenomenon is that of a gene suppressing the action of allelic variation of other genes. Such interaction among genes, called epistasis, is a major challenge to address. For illustration of epistasis we can consider a situation comprised of two genes, each with two alleles where one is dominant over the other, and the dominant allele of one of the genes is epistatic over the allelic variation at the other gene (see **Table 1**). This implies that for the example in **Table 1**, whenever the gene A is present, A will determine the phenotypic expression. Phenotypic differences between genotypes having B versus b will

Table 1 Allelic combinations for two genes, each with two alternative alleles, A vs. a and B vs. b, and the resulting phenotype. The Phenotypes are expressed as cursive of the dominant gene being expressed

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

Note: A is dominant over a, B is dominant over b, and A is epistatic over the allelic variation of the other gene.

only be expressed in the absence of A. Thus, the 16 alternative allele combinations will express phenotypic variation in the ratio 12:3:1. With this situation, it would 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, and the epistatic effect of A over allelic variation at the other gene, is, however, obscure and difficult to detect. Understanding the gene interaction really 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 one or a few genotypes reflects 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.

The genome can be analyzed by so-called DNA sequence mapping. This can be done in various ways. One method is the detection of single nucleotide polymorphisms (SNPs), which may be used to map or characterize different alleles of a gene. SNPs are single nucleotide hot spots for high diversity between individuals. SNP detection technologies are used to scan for new polymorphisms and to determine the allele(s) of a known polymorphism in target sequences. SNP detection technologies have evolved from labor-intensive, time-consuming, and expensive processes to some of the most highly automated, efficient, and relatively inexpensive methods for genome studies. A great deal of time and money is currently being invested in the production of large libraries of SNPs.

4.08.2.2 Transcriptome

When a gene is activated, a transcript of that particular gene is made as a single-stranded RNA molecule. 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. The part of RNA encoding a protein is called messenger RNA (mRNA). The composition of the transcriptome can be analyzed by microarray techniques. As the mRNAs are short-living molecules, microarray techniques give dynamic snap shots into a short period of time for the organism, tissue, or cell from which the transcripts are collected. The snap shot 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 experiments are conducted according to some experimental design, the aim is usually to identify 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 tight time-course study in such experiments searching for changes occurring over a period of time.

4.08.2.3 Proteome

Proteins are chains of amino acids connected by peptide bonds. The order of the base pair on mRNA comprises a code for the synthesis of proteins as a sequence of three, and three base pairs on the mRNA constitute the code for one given amino acid. For example, the base CCA is the code for the amino acid prolamine. The order of amino acids is, thus, directly connected to the order of the base pairs of the gene coding for the given protein. In total there exist more than 20 different amino acids variants, and the structure and functionality of the protein is determined by the order and the properties of the amino acids.

The proteins can be cleaved by specific enzymes to shorter amino acid chains called peptides. The proteome is the collection of the proteins in a cell at a given time. Some proteins are metabolic enzymes involved in energy metabolism, others may act as hormones or signalling 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. 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 posttranslation modifications.

A large number of different techniques are used in proteome research. For a review see, for example, Nelson and Cox[2] and Simpson.[3] These techniques are generally more time consuming than the techniques used for genome/transcriptome studies. To map presence or absence and the amounts of the various proteins in a cell at a given time, proteins are most often separated by electrophoresis or chromatographic techniques combined with mass spectroscopy (MS).

Mapping all proteins in a tissue is an ambitious goal that is seldom achieved in complex biological samples, but a large number of proteins can be separated by these approaches. Irrespective of the techniques used to separate the proteins, the resulting pattern is complex, and suitable data analytical tools are required for identifying the proteins or polypeptides which are of relevance for the biological questions in focus.

4.08.2.4 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.[4] The metabolome is a result of the biochemical reactions being catalyzed by the proteins of the proteome. 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. However, the metabolome is constantly changing due to all the chemical reactions occurring in the cell.

For metabolite profiling, the aim is to identify and quantify a selected number of metabolites. For this purpose, sensitive chromatographic methods such as GC–MS and LC–MS are often used. For more comprehensive information on this topic see Lindon *et al.*[5]

4.08.2.5 Environmental Impact

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. Bazanov[6] defined a morphogenetic field as giving permissions for the environment in which organisms are developed, and a geometrical and mathematic model was proposed.

4.08.2.6 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 situation one single quality parameter has the primary focus as, for example, the presence or absence of one 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.

In many situations, a broad fingerprint of the chemical composition of a tissue of interest might give very useful information. A wide range of analytical methods are available for analyzing chemical composition. Biospectroscopic methods like Fourier-transform infrared spectroscopy (FTIR), near-infrared spectroscopy (NIR), nuclear magnetic resonance (NMR), and direct-injection mass spectroscopy (DIMS) can give fingerprints of the chemical and physical composition, which can be used for screening purposes, as a supplement to other more time-consuming methods, and to obtain an overview of the phenotypic expression. This has been thoroughly discussed by Munck *et al.*[7] and Munck.[8] By using NIR combined with chemometric tools, Munck *et al.*[7] revealed pleiotropic effects of mutant versus wild-type barley on a large number of traits, and they could thereby reveal specific gene expression patterns.

4.08.2.7 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. An example of how this can be done is given in Section 4.08.4.6.3. 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 variables to deal with some of the data redundancy in functional genomics.

4.08.3 Important Considerations in Functional Genomics

General challenges frequently occurring in chemometric studies, such as the need for background correction and nonlinearity, are described in other parts of this book. These topics may also be relevant for data preprocessing and analysis in genomics studies. In this section, however, we call attention to the aspects particularly relevant for functional genomics.

4.08.3.1 Scientific Strategy

The traditional approach for statistical research has been to investigate one-to-one relationships between input variables and some response variables. 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 genomic 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 genomics, 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.

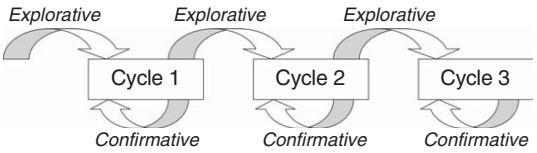

Figure 3 Design strategy.

Alternation between an explorative phase and confirmative phase is a useful strategy to gradually unwrap the secrets of nature (**Figure 3**).

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 its their full strength after sufficient explorative research has been conducted to ensure that 'one is digging where gold is to be found'. In functional genomics, we often face the situation of 'searching for a needle in a hay stack'. A humble attitude to this complexity, with a stepwise cycling 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. A statement of Dr. J. W. Tukey may be useful to keep in mind: "It is better to be approximately correct than precisely wrong".

4.08.3.2 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 both situations, we build a data model where some response parameters (Y) are expressed as a function $f(\cdot)$ of some measurements X. This can be expressed as

$$Y = f(X) + \epsilon \tag{1}$$

where ϵ captures the variability in the response not explained by the functional relationship with $\mathbf{X}$. Usually, the objective of a study is a combination of these two aims, but we can theoretically think of extremes on both sides. In some cases prediction as such may be the major concern. The aim is merely to obtain reliable predictions of future samples with no required understanding of the system leading to the observed effects. In other cases, data modeling for understanding a biological system may be the primary objective of a study. The aim is to capture as much relevant information as possible with no focus on constructing predictors for future observations.

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 variables with best prediction performance. As is well known from classical statistical theory, too complex models may give inflated prediction errors due to overfitting and increased variance of estimated model parameters. Therefore, it is desirable to keep the predictor as simple as possible, thus retaining as a small prediction error 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 variables yielding maximum prediction performance.

In megavariate data sets where there may be multiple, equal sized, disjoint subsets of variables 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 variables 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 will fail when used on future samples. Therefore, if possible, a reliable prediction model should be based on the following:

- variables that are directly linked to the response by causality;
- variables 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);
- variables 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 predictor construction, as the focus is on obtaining insight into all variables showing significant changes across samples. Variables 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 variables.

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 variables. In an early phase, the primary focus might be on ensuring that all variables of possible relevance are identified, whereas the risk of selecting false-positive variables 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 variables 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 toward false positive if the variable set is to be interpreted in the light of background knowledge later.

Even though there is often a close connection between the object of gaining insight versus the object of establishing a prediction model, we treat these topics separately in Section 4.08.4.

4.08.3.3 Considerations on the Experimental Design

Setting up the design of an experiment is critical, and it needs some thoughtful consideration in view of the complexity of functional genomics.

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 some 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 expressions in a bacteria species, and the experiment is conducted under constant conditions on all other factors, such as temperature, for example, the conclusion would then be valid 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 by varying both factors simultaneously according to an experimental plan. One alternative is a full factorial design where all levels of both factors are combined.[9] Then main effects of both factors as well as interaction effects between the two factors can be revealed. Again, there will be limitations to the interpretation of the results as the conclusion, in principle, will be valid only for the particular environmental setting of other factors used, humidity for example, and for the particular genotype investigated. 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.