Macrothink Institute™

Journal of Biology and Life Science
ISSN 2157-6076
2014, Vol. 5, No. 1

A Brief Overview of Multivariate Data Analysis in Biological Sciences

Mohammad OHID ULLAH (Corresponding author)

Associate Professor, Department of Statistics

Shahjalal University of Science and Technology, Sylhet-3114, Bangladesh

E-mail: ohidullah@gmail.com

Received: December 25, 2013 Accepted: January 8, 2014

doi:10.5296/jbls.v5i1.4829 URL: http://dx.doi.org/10.5296/jbls.v5i1.4829

Abstract

Microarrays have been used to quantify the mRNA expression for different genes in different living beings. Recently RNA-seq be the major techniques for genome-wide assessment of transcriptomics data as it provides more accurate estimates of transcriptomics data than microarrays for either known or unknown transcripts in a larger dynamic range. To analyse the high-throughputs transcriptomics data as well as proteomics and metabolomics data, it's very essential to know the proper statistical tools and software that are suitable for the respective design and data. Therefore, in this review an attempt has been made to discuss briefly some useful multivariate techniques to analyze and integrate multi-groups datasets as well as to choose appropriate models based on design and data with short description.

Keywords: mRNA, RNA-seq, Multivariate data, PCA, MFA, Models.

Abbreviations: mRNA, messenger RNA; RNS-seq, RNA sequence; PCA, Principal component analysis; MFA, Multi factor analysis.

1. Introduction

1.1 Microarray and RNA-Seq Study

Last decades, DNA microarrays have been used extensively to quantify the abundance of mRNA corresponding to different genes. Alternate of microarrays, more recently high-throughput sequencing (RNA-seq) be the major technique for genome-wide assessment of transcriptomics data (Mortazavi et al., 2008; Nagalakshmi et al., 2008; Charlotte and Mauro, 2013). Compared with microarrays, RNA-seq provides more accurate estimates of transcriptomics data than microarrays for either known or unknown transcripts in a larger

dynamic range (Wang et al., 2009 and it's protocols were develops in 2008 (Wei and Yijuan, 2013). As the cost of arrays and sequencing decreases, it is possible that the use of microarray and RNA-seq tools will increase rapidly and become popular tools in developed as well as developing countries for trancriptomics analysis (David et al., 2013). Recently researchers would like to study using whole genome. The main concern in such kind of studies is the interest in a large number of parameters, while deal with a small number of biological sample size. This is, for example, the case in a microarray experiment and recently RNA-seq experiment.

Microarrays instrument can be used to measure the relative quantities of specific mRNAs in two or more samples of thousands of genes simultaneously. Microarray experiments are subject to a quite special limitation, mainly the cost of a slide. As a result, the biological replications as well as the technical replications are usually not enough for the application of conventional statistical tools. Churchill (2002) reports that technical replicates on a single slide are highly reproducible, but it is the reproducibility of biological replicates that is important for inference. Sometimes it is suggested that with three replicates it is possible to detect one anomalous observation (Wit and McClure, 2004). Despite the limitation, microarrays remain a popular instrument to measure gene expression compare to other measurements. Microarray study as well as RNA-seq includes the experimental design, data normalization, detect differentially expressed genes, prediction of gene expression, class discovery, gene set analysis, systems biology etc. Microarrays and RNA-seq are an essential technology for studying gene expression. It can be handled the level of expression of thousands genes, even an entire genome, can be estimated for a sample of cells.

The most important question in molecular biology is: how living organisms function in different levels. To find out this answer, a numerous of research has been done in the world. In 1990, the Human Genome Project was founded to identify the function of almost all of the genes in human DNA (Collins et al., 2003).

As we know the location of gene, afterwards other questions arise how it's regulated, what's the function. Gene expression is the regular process in cell and the information is carried by a gene and then transformed into a protein. This process can be influenced by a number of internal (disease) or external (environmental) factors. This is also important to know the association between genes and other internal or external factors. This kind of association and research not only in individual gene level but also can be done in group of genes and organ level.

Basically microarray and RNA-seq studies are based on the multivariate analysis. Both of these techniques produce enormous data of gene expression. To analyze this kind of data we have to use data mining technique as well as multivariate data analysis techniques. Many statistical tools were developed to handle this kind of data. In this review we are going to focus the application of some useful R packages.

For instance, Limma (Linear model for microarray) was developed by Gordon Smyth (Smyth, 2004) using moderated t statistic as well as empirical bayes estimate to find out the differential expressed gene and this gives better power than SAM (Significance analysis of microarray) package (Tusher et al., 2001). To analyze the microarray data at first one have to normalized

her data and then develop the design matrix of her data. To see the normalizing technique see limma package in R.

In biological and social sciences researchers collect lots of data to explore their study. For instance, in biology, microarray and RNA-seq studies contain huge amount of data. In social sciences, researchers want to find out socio-economic parameters. Now-a-days multivariate technique is going to be more popular to analyze such kind of study. For example, Multivariate analysis of variance (MANOVA), Principal component analysis (PCA), Multifactor analysis (MFA), Partial least square (PLS), partial least square path model (PLSPM), and Cluster/hierarchical analysis (CA). In this review only a brief illustration of PCA and MFA will be mentioned using R packages.

Though a number of software packages have been developed for the analysis of microarray data, software itself is insufficient. One needs knowledge about the different aspects of data analysis in order to select and utilize software successfully. Many methods have already being published and it is very difficult for biologist to conclude which methods are valid and suitable for their study. Therefore in this review a short description will be mentioned about the some software packages with their application as well as when a biologist will choose appropriate models/statistical tools to analyze their data.

1.2 Principal Component and Multifactor Analysis

Many multivariate analytical techniques are available now-a-days. The most frequent useful tool named PCA (Principal component Analysis) approach to reduce your dimension (Johnson and Wichern, 2006). Here we would like to give some hypothetical examples of multivariate data from Biological field for PCA and MFA by and *FactoMineR* (de Tayrac et al., 2009) package. Integration between two datasets (trancriptomics and proteomics) may also be analyzed by integrOmics package (Le Cao et al, 2009) in R. In Social sciences (Sociology, Economics, Political sciences) and Life Sciences (Plant Sciences, Forestry, Ecology, Nutrition, Epidemiology, Medical study, Genetics, Bioinformatics and recently Systems and Synthetic Biology), the researchers collect many information in their research. To handle these kinds of data multivariate analytical techniques need to be applied (Ullah, 2010). Here some examples of multivariate analysis were mentioned, however different kind of multivariate data can also be analyzed by using these tools.

1.2.1 PCA (Principal Component Analysis)

PCA is a multivariate statistical analysis technique that determines linearly independent sets of variables in large, multi-dimensional data sets. It is useful for analyzing and visualizing complicated and possibly redundant data sets and it converts correlated variables to uncorrelated new latent variables. Let's assume that you have factor (diet) with 5 levels (different quantities) and every level you used 5 mice and measured the gene expression from heart tissue from the 20 mice. Assume that you collected data using microarray instruments and LIMMA (Smyth, 2004) package in R and got normalized data. To see the effect of treatment you can use LIMMA package in R or else. Many related packages are available in http://cran.r-project.org/web/packages/.

Here I would like to find out the important genes (variables) respect to treatment using multivariate techniques; therefore, I just use a subset (4 genes) of the hypothetical microarray data.

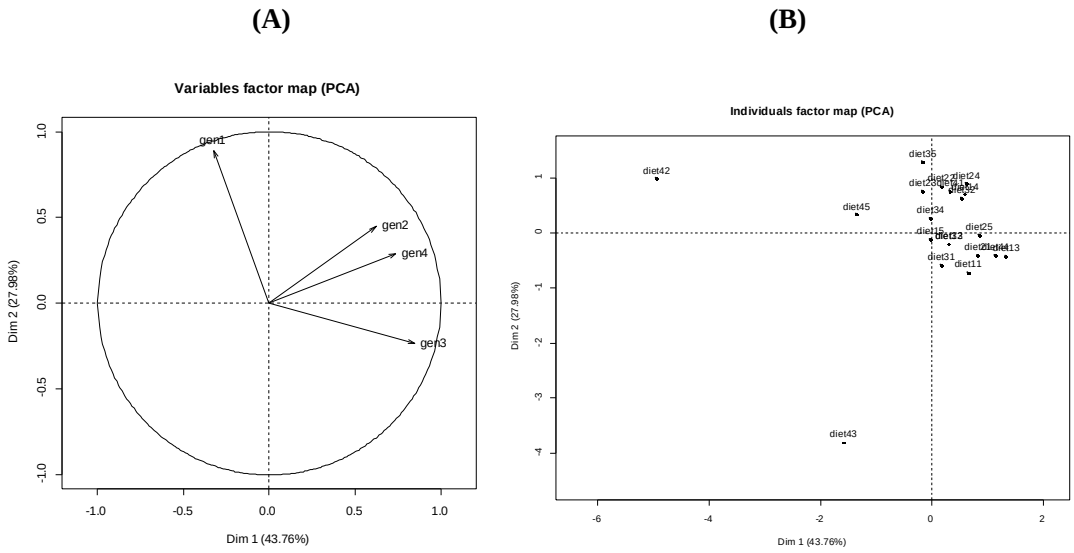

Figure 1. (A) Score plot/individual factor map of PCA. (B) Loading plot/ Variables factor map of PCA

Observing the score plot and loading plot it is clear that gene1 is influenced by diet4 compare to other diets because in both plot gene1 and diet4 belongs to the same compartment. Similarly gene2 and gene3 were influenced by diet2. From the score plot it was also observed that the gene expressions by diet4 are distinct from the other diets. In the loading plot the same directions of vector genes indicate they are positively associated and opposite directions indicate genes are negatively associated. From the above loading we can conclude that gene2 and gene4 are positively associated.

1.2.2 MFA (Multifactor Analysis)

If you have responses in different groups/blocks then you can use MFA (Multi factor analysis) option in FactoMineR package to integrate multi-datasets. Assume that you have data from gene expression, protein and metabolomics and you want to find out the relationship among them. Now to use MFA you can observe how the groups differ of each other and to find out the importance of genes/proteins/metabolomics respect to the factor/ treatment. The main difference between PCA and MFA is: MFA able to handle multi datasets on the other hand PCA able to handle single dataset.

Following hypothetical example was used to explore the multifactor analysis, which contains 4 genes, 3 proteins and 3 metabolomics data from the 4 different diets and each diet has 5 replications.

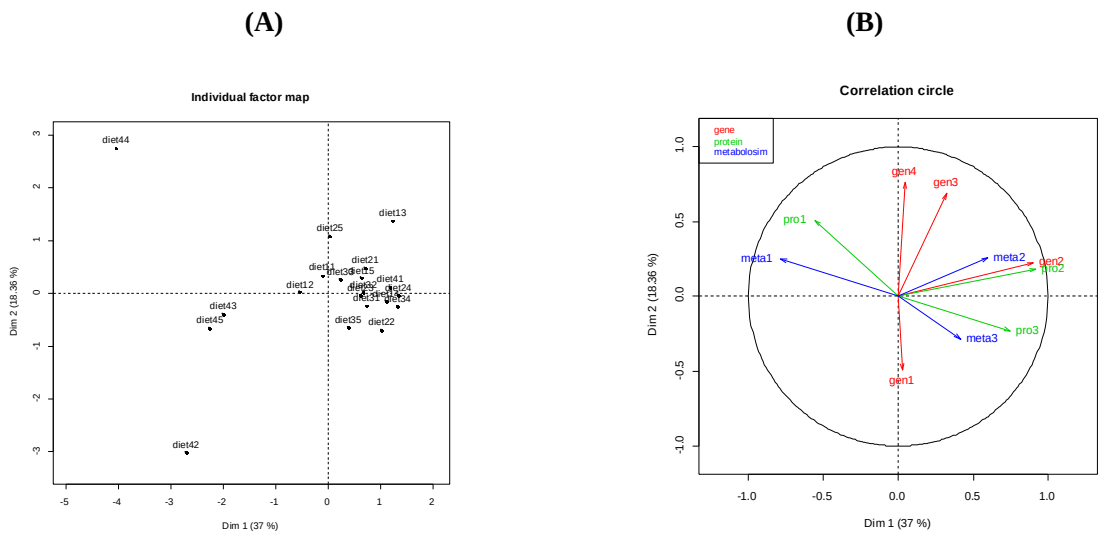

Figure 2. (A) Score plot/individual factor map of MFA. (B) Loading plot/ correlation circle of MFA

The score plot/individual factor map shows diet4 are different from other diets and diet1, 2, and 3 are belonged to one cluster. Loading plot or correlation circle shows how genes, proteins and metabolites are associated to each other. Here we observed that meta2, gene2 and pro2 are highly positively related because they are going same direction with adjacently. Parallel observing the score plot and loading plot we can find out which variables are influenced by which diets. The results show that pro1 and meta1 are influenced by diet4 because these are located in the same compartment of the Figure 2A and 2B.

Some useful packages with their function are presented in the following table.

Table 1. Some useful software tools to analyse and integrate biological datasets.

Software	Function	Reference
FactoMineR (R package)	Able to handle multi-data as well as to integrate between them using multivariate statistical tools	(de Tayrac et al., 2009)
integrOmics /mixOmics (R package)	Integrate two different datasets.	(Le Cao et al, 2009)
Simca-p	Able to handle multivariate normal and non-normal data by using PCA and PLS techniques	(Wold, 1966)
Unscrambler	Able to handle multivariate data as well as integrate between them using different kinds of statistical tools	(Martens, 1989)
Plspm (R package)	Able to handle several groups of multivariate data with causal relationships by path models instead of SEM (Structural Equation Model).	(Wold, 1982)
lmbc	Analyzing RNA-seq data	(David et al., 2013)
limma	Normalization and finding differential expressed genes	(Smyth, 2004, 2005)

1.3 Choosing Appropriate Models

As biological data contains enormous amount of information, therefore, we need to know which variables are discrete/categorical or continuous as well as which they are dependent or independent before fitting model(s). Independent variable is one that is not influence or predicted by other variable(s). That is it causes to change the response or outcome. Dependent variable is one that is influenced by other variables (independent variables)-which is also known as response or outcome. Some variables are dependent to each other is known as interdependence variables (Joshep et al., 2009). We have also to aware about the nature (discrete or continuous) of the data to choose appropriate model. A discrete variable is one that measures specific values (for instance- 'yes', 'no'), or several categories or count. A continuous variable is one that measures any value of real number.

The following table shows the basic overview to choose appropriate models:

Table 2. An overview to choose appropriate models for analysing different kinds of datasets

Independent variable(s) (X)	Dependent variable(s) or outcomes (Y)	Design of the study	Model
Discrete: 2 categories	Continuous	Observational or intervention	Not necessary to fit model, just use parametric test: t-test if normality and homoscedasticy assumptions fulfilled. Otherwise use Non-parametric test: Mann-Whitney (Agresti, 2002; Kutner et al., 2005)
Discrete: more than 2 categories	Continuous	Intervention	Fit ANOVA (Analysis of variance) if all the assumptions of random error terms are fulfilled. Especially normality and homoscedasticity. If not, then use non-parametric test: kruskall-Wallis Whitney (Agresti, 2002; Kutner et al., 2005)
Discrete: more than 2 categories	Continuous	Observational	Fit multiple regression model after creating dummy (1, 0) variables of independent variable. If the number of category k then number of dummy variables should be k-1 (Kutner et al., 2005)
Discrete as well as Continuous	Continuous	Intervention	Fit ANCOVA (Analysis of covariance). For discrete independent variable, better to create dummy variable for less than 5 categorical discrete one (Kutner et al., 2005)
Continuous	Continuous	Observational	Fit linear regression model. For one independent variable, simple linear regression model. For more than one independent variables multiple linear regression model (Kutner et al., 2005)
Discrete or continuous	Discrete (binary)	Observational Or case-control study	Fit Logistic regression model (Agresti, 2002; Bonita et al., 2006)
Discrete or continuous	Discrete (more than	Observational Or cohort study	Multinomial logistic regression or proportional odds model (Agresti, 2002)

	two category)		
Discrete or continuous	Continuous	Longitudinal	Linear Mixed models (Verbeke and Molenberghs, 2000)
Discrete or continuous	Discrete (categorical or count)	Longitudinal	Generalizing Estimating Equations (GEE), Alternating Logistic regression (ALR) (Molenberghs and Lesaffre, 1994; Molenberghs and Verbeke, 2005)
Discrete or continuous	Discrete or continuous (survival time)	Cohort study	Proportional hazard model (Cox model) (Joshep et al., 2009; Bonita et al., 2006; John and Melvin, 1997)
Discrete or continuous	More than one Discrete or continuous (survival times)	Cohort study	Frailty model (Joshep et al., 2009)
Discrete (one or more)	Continuous (several)	Observation or intervention	MANOVA (multivariate analysis of variance) (Johnson and Wichern, 2006)
Discrete or continuous	Continuous (several)	Observation or intervention	MANCOVA (Johnson and Wichern, 2006)
-	Continuous (several)	Observation or intervention	Principal component analysis/Factor analysis. Alternatively nonparametric method is PLS (Johnson and Wichern, 2006; Joshep et al., 2009)
-	Discrete (several)	Observation or intervention	Correspondence analysis. Alternatively nonparametric method is PLS (Johnson and Wichern, 2006; Joshep et al., 2009)
continuous (more than one)	Discrete	Observational or intervention	Discriminate analysis (Johnson and Wichern, 2006)
Discrete or continuous (several)	Discrete or continuous (several)	Observational or intervention	Structural Equaltion Model (SEM), if independent and dependent variable are interdependence. It's better if the assumptions: especially normality and misspecification hold and there are no problems about missing values and multicollinerarity. Otherwise non-parametric method such as PLSPM (partial least squares path model) may be applied to integrate multi-blocks datasets (Johnson and Wichern, 2006; Joshep et al., 2009; Ullah, 2012; Lohmoller, 1989)
Discrete or categories (one or more)	Continuous (time)	Observational or intervention	ANOVA mixed-effect model (Ping et al., 2009)
Discrete or continuous (several)	Discrete or continuous (several)	Observational or intervention	Canonical correlation (Johnson and Wichern, 2006)

2. Conclusion

Multivariate statistical tools that described in this review can be applied to analyze microarray/RNA-seq data as well as other biological sciences related data. Recently an emerging field named executable biology or integrative biology (Krawetz, 2009; Chen et al., 2009; Aderem, 2007; Fisher and Henzinger 2007) that integrate different datasets by multivariate statistical tools to understand biological systems. Some methods have already been developed for modelling, network and integration for individual gene specific analysis. However, it's not enough yet to see how systems in the cell/organ working on in the living beings. Grouping homogeneous genes/proteins in a cluster or module or meta gene/protein and then might be possible to apply modeling and network approach to see how these meta genes/proteins evolve over time in a specific organ and how they are related of each other. Genomics and transcriptomics were followed by proteomics and metabolomics. The expansion and use of such omics platforms, particularly transcriptomics, proteomics and metabolomics is discussed in detail by (Kussmann et al., 2008). In future an integrative analysis including different datasets can be done in the biological research in order to see how diets (treatments)-mRNA-protein-metabolism-function is related of each other with phenotypes (like body mass index, disease status etc.) for discover the biological system. The different packages/software and choosing appropriate modeling techniques based on their design and datasets discussed in this short review may helpful for the biological researcher as well as other researchers from different fields to analyze their data.

Acknowledgement

I am thankful to the Nutrition, Metabolism and Genomics group, Division of Human Nutrition, Wageningen UR, NL where I learned about biology during my PhD work.

References

Aderem, A. (2007). Systems biology-editorial. Curr. *Opin. Biotech. 18*, 331. http://dx.doi.org/10.1016/j.copbio.2007.07.010

Agresti, A. (2002). Categorical Data Analysis. *John Wiley & Sons*, New York. http://dx.doi.org/10.1002/0471249688

Bonita, R., Beaglehole, R., & Kjellstrom, T. (2006). *Basic Eoidemiology.* WHO press, Geneva.

Charlotte, S., & Mauro, D. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics. 14*, 91. http://dx.doi.org/10.1186/1471-2105-14-91

Chen, L., Wang, R-S., & Zhang, X-S. (2009). *Biomolecular Networks: Methods and Applications in Systems Biology*. Wiley. http://dx.doi.org/10.1002/9780470488065

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nat. Genet. 32* Suppl, 490-495. http://dx.doi.org/10.1038/ng1031

Collins, F.S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. *Science. 300*(5617), 286-290. http://dx.doi.org/10.1126/science.1084564

David, D., Xuming, H., & Ping, M. (2013). Bias Correction in RNA-Seq Short-Read Counts Using Penalized Regression. Stat. *Biosci. 5*, 88–99. http://link.springer.com/article/10.1007%2Fs12561-012-9057-6

de Tayrac, M., Le, S., Aubry, M., Mosser, J., & Husson, F. (2009). Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics. 10*, 32. http://dx.doi.org/10.1186/1471-2164-10-32

Fisher, J., & Henzinger, T. A. (2007). Executable cell biology. *Nat. Biotechnol. 25*(11), 1239-1249. http://dx.doi.org/10.1038/nbt1356

John, P. K., & Melvin, L. M. (1997). Survival analysis-techniques for censored and truncated data. Springer-Verlage, New York.

Johnson, R. A., & Wichern, D. W. (2006). Applied Multivariate Statistical Analysis. Prentice-Hall, New York.

Joshep, F. H. Jr., William, C. B., Barry, J. B., & Roleph, E. A. (2009). Multivariate data analysis. Prentice Hall.

Krawetz, S. (2009). Bioinformatics for Systems Biology. Springer. http://dx.doi.org/10.1007/978-1-59745-440-7

Kussmann, M., Rezzi, S., & Daniel, H. (2008). Profiling techniques in nutrition and health research. Curr. Opin. *Biotechnol. 19*(2), 83-99. http://dx.doi.org/10.1016/j.copbio.2008.02.003

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models. McGraw-Hill, New York.

Le Cao, K. A., Gonzalez, I., & Dejean, S. (2009). integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics. 25*(21), 2855-2856. http://dx.doi.org/10.1093/bioinformatics/btp515

Lohmoller, J. B. (1989). Latent variable path modeling partial least squares. Physica-Verlag, Heidelberg. http://dx.doi.org/10.1007/978-3-642-52512-4

Martens, H., & Naes, T. (1989). *Multivariate Calibration*. Wiley, London

Molenberghs, G., & Lesaffre, E. (1994). Marginal modeling of correlated ordinal data using a multivariate Placket distribution. *Journal of the American statistical Association. 89*, 633- 644. http://dx.doi.org/ 10.1080/01621459.1994.10476788

Molenberghs, G., & Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* Springer-Verlag, New York.

Mortazavi, A, Williams, B.A., McCue, K, Schaeffer, L., & Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods. 5*(7), 621–628. http://dx.doi.org/10.1038/nmeth.1226

Nagalakshmi, U., Wang, Z., Waer,n K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008)

The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. *320*(5881), 1344–1349. http://dx.doi.org/10.1126/science.1158441

Ping, M., Wenxuan, Z., & Jun, S. L. (2009). Identifying Differentially Expressed Genes in Time Course Microarray Data. Stat. *Biosci. 1*, 144–159. http://dx.doi.org/10.1007/s12561-009-9014-1

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. *Appl. Genet. Mol. Biol. 3*, Article3. http://dx.doi.org/10.2202/1544-6115.1027

Smyth, G. K. (2005). *Limma: linear models for microarray data*. Springer, New York.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. 98*(9), 5116-5121. http://dx.doi.org/10.1073/pnas.091062498

Ullah, M. O. (2010). Some statistical advices for the non-statisticians. VDM Verlag, Saarbrucken.

Ullah, M. O. (2012). *Nutritional Systems Biology of Fat- Integration and Modeling of Transciptomics Datasets Related to Lipid Homeostasis*. Thesis Wageningen University, Wageningen.

Verbeke, G., & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet. 10*(1), 57–63. http://dx.doi.org/10.1038/nrg2484

Wei, S., & Yijuan, H. (2013). eQTL Mapping Using RNA-seq Data. *Stat. Biosci. 5*, 198-219. http://dx.doi.org/10.1007/s12561-012-9068-3

Wit, E., & McClure, J. (2004). *Statistics for Microarrays*. John Wiley and Sons, West Sussex. http://dx.doi.org/10.1002/0470011084

Wold, H. (1966). *Estimation of principal components and related models by iterative least squares*. Academic Press, New York.

Wold, H. (1982). *Soft modeling: the basic design and some extensions*. North Holland, Amasterdam.

Copyright Disclaimer