Statistics and bioinformatics in nutritional sciences: analysis of complex data in the era of systems biology

Wenjiang J. Fu, Arnold J. Stromberg, Kert Viele, Raymond J. Carroll, and Guoyao Wu,*

aDepartment of Epidemiology, Michigan State University, East Lansing, MI 48824, USA;
bDepartment of Statistics, University of Kentucky, Lexington, KY 40536, USA;
cDepartment of Statistics, Texas A&M University, College Station, TX 77843 USA;
dFaculty of Nutrition and Department of Animal Science, Texas A&M University and Department of Systems Biology and Translational Medicine, Texas A&M College of Medicine, College Station, TX 77843, USA

* This work was supported, in part, by grants from National Institutes of Health (P20RR16481, 2P42 ES007380-09, P20RR020145-01, 1R21 HD049449, and CA57030), King Abdullah University of Science and Technology (KUS-CI-016-04), National Research Initiative Competitive Grants (2008-35206-18764 and 2008-35203-19120) from the USDA Cooperative State Research, Education, and Extension Service, American Heart Association (#0755024Y), and Texas AgriLife Research (H-8200).

* Corresponding author: Tel. (979) 845-1817; Fax (979) 845-6057.

E-mail address: g-wu@tamu.edu (G. Wu).
Abstract

Over the past two decades, there have been revolutionary developments in life science technologies characterized by high throughput, high efficiency, and rapid computation. Nutritionists now have the advanced methodologies for the analysis of DNA, RNA, protein, low-molecular-weight metabolites, as well as access to bioinformatics databases. Statistics, which can be defined as the process of making scientific inferences from data that contain variability, has historically played an integral role in advancing nutritional sciences. Currently, in the era of systems biology, this analytical tool has gained an increasingly important role to quantitatively analyze information about biological macromolecules with the aid of computers. This article describes general terms used in statistical analysis of large, complex experimental data. These terms include experimental design, power analysis, sample size calculation, and experimental errors (type I and II errors) for nutritional studies at population, tissue, cellular, and molecular levels. In addition, we highlighted various sources of experimental variations in studies involving microarray gene expression, real-time polymerase chain reaction, proteomics, and other bioinformatics technologies. Moreover, we provided guidelines for nutritionists and other biomedical scientists to plan and conduct studies and to analyze the complex data. Appropriate statistical analyses of complex data are expected to make an important contribution to solving major nutrition-associated problems in humans and animals (including obesity, diabetes, cardiovascular disease, cancer, ageing, and intrauterine fetal retardation).

Key words: Nutrition research; Statistical analysis; Bioinformatics; Systems biology
1. **Introduction**

1.1. **Nutrition, gene expression and health**

Nutrition plays a vital role in health, genetic predisposition, and disease [1-3]. Many diseases in humans and animals result from a chronic deficiency or excess of certain nutrients in the diet. These disorders include obesity, cancer, type-II diabetes mellitus, osteoporosis, inflammatory disease, stroke, hypertension, and atherosclerosis. Effective prevention and treatment of the diseases require adequate knowledge about the molecular mechanisms responsible for the actions of nutrients and other dietary components on cell metabolism and function [4]. Extensive studies have shown that dietary substances can affect gene expression, protein turnover, and the availability of co-factors for enzymes, thereby altering metabolic fluxes as well as cellular and whole-body homeostasis [5,6].

The genome is organized in chromosomes within the nucleus of the cell, and is transcriptionally heterogenous [7]. Traditional research in molecular nutrition involves the analysis of expression of one or a very few genes at one time. While this approach has led to important findings on the discoveries of key regulatory pathways for nutrient utilization, it generally requires prior knowledge of genes of interest. However, there is increasing evidence that most genes do not function in isolation and that dietary nutrients interact to modulate expression of a set of genes and their biological functions [5]. Thus, nutritionists face a challenging task of defining the complex cellular and molecular mechanisms that control the digestion, absorption and metabolism of dietary nutrients. With the recent completion of sequencing of the genomes of many species, including the human [8,9], mouse [10,11], rat [12] and yeast [13], we now have useful tools to explore and establish the complex interactions between genes and the diet as an environmental factor.
1.2. Bioinformatics and “-omics”

Bioinformatics can be defined as a new discipline of quantitative analysis of information about biological macromolecules with the aid of computers [14]. This includes the sequence, structure and function of genes and their products (e.g., DNA, RNA, and proteins). Sequence analysis includes genome comparison, phylogeny, gene and promoter prediction, motif and pattern discovery, sequence database search, and sequence alignment. Functional analysis includes gene expression profiling, prediction of protein subcellular localization and interactions, and modeling of metabolic pathways. Advanced technologies with the Greek suffix “-ome” (meaning “complete” or “all”) have been developed to analyze large amounts of complex data [15]. Thus, we now have the terminologies of transcriptome (a complete set of mRNA molecules produced by a cell), proteome (the complement of proteins in cells and physiological fluids, as well as their interactions), and metabolome (a complete set of low-molecular-weight metabolites and substances in cells, tissues, organs and physiological fluids). Transcriptome profiling is routinely performed, using microarray analysis, subtractive hybridization, serial analysis of gene expression, and differential display. The analysis of the proteome by highly specific, sensitive, and accurate mass spectrometry (MS) has been made possible with matrix-assisted laser desorption ionization (MALDI) or electrospray ionization (ESI) of proteins and large peptides [5,6]. Further, MS, coupled with liquid chromatography and gas chromatography, is greatly facilitating the analysis of amino acids, fatty acids, carbohydrates, water- and fat-soluble vitamins, micro- and macro-minerals, and the numerous metabolites of nutrients [16].

Robust bioinformatics has unified genomics with proteomics and metabolomics, and this integration plays an essential role in understanding the complex interactions between nutrition and genetics [15]. Therefore, bioinformatics offer powerful means to study nutrigenomics (the
influence of nutrients and other dietary components on the genome, including its expression and stability) and nutrigenetics (concerning how the genome of an individual affects its responses to diet; namely genetic polymorphisms). Single nucleotide polymorphism (SNP), a single-base substitution within DNA, provides a molecular basis for genetic and biological variability in nutrient requirements among individuals [14]. Of particular interest, more than 2.8 million SNPs have been discovered and linked to discrete locations in the human genome [8,9].

1.3. Statistics in the era of systems biology

Statistical analysis can be generally defined as the process of making scientific inferences from data that contain variability. Statistics has historically played an integral role in advancing nutritional sciences. This tool has gained an increasingly important role in the systems biology era to analyze large, complex data sets generated from genomics, proteomics and metabolomics studies [17]. Particularly, analyses of data from real-time polymerase chain reaction as well as microarray, proteomics and other bioinformatics studies require statistical models to account for various sources of experimental variations [18,19]. This aids in identifying which genes in a study sample are differentially expressed in the face of substantial biological and technical variation. Also, collaboration between nutritionists and statisticians can minimize systemic errors, improve the precision of statistical tests, control false positives (type I error) and false negatives (type II error), and optimize data analysis [18-20].

The major objective of this review is to provide guidelines for nutritionists and other biomedical scientists to adequately plan and conduct experiments and analyze complex data particularly from microarray, RT-PCR, proteomics, and other bioinformatics studies.

2. Definitions of general statistical terms

2.1. General considerations
Several considerations are generally needed to plan and conduct a nutrition study. They include: a) study designs; b) sample size calculation or power analysis; c) data collection and experiments; d) statistical modeling and data analysis; and e) interpretation and report of findings. Study design plays a major role in nutrition research and varies with the aim of the investigation. A good study design should ensure the achievement of project aims.

2.2. Hypothesis testing and errors

2.2.1. Null hypothesis and alternative hypothesis

Assume that the study subjects under treatment and controls are drawn randomly from their own populations that follow their probability distributions characterized by defined parameters. For example, the parameters may be two normal distributions for body mass index (BMI) with different means and variances, or two Poisson distributions of the number of people having heart attacks with different mean numbers of heart attacks. If the study aim is to determine whether dietary supplementation with a nutrient (e.g., arginine) will reduce BMI in obese subjects, the hypothesis is to test if the mean BMI of the population receiving the treatment differs from that of the population not receiving the treatment. The null hypothesis ($H_0$) is that two populations have equal mean BMI, while the alternative hypothesis ($H_a$) is that they have unequal mean BMI.

2.2.2. Type I and type II errors

In hypothesis testing, two types of errors may be made when experimental data are analyzed statistically. Type I error (false positive) occurs when a conclusion is drawn in favor of a significant difference while there is no true difference between populations; namely, $H_a$ is claimed to be true while $H_0$ is true. Type II error (false negative) occurs when the null hypothesis is not rejected while the alternative hypothesis is true for the populations; namely, $H_0$ is claimed to be true while $H_a$ is true. Power of a test is the probability of claiming the alternative
hypothesis is true when it is true, i.e. claiming $H_a$ is true when $H_a$ is true. Type I error is usually controlled at a very low probability level, such as 5% or 1%. It is usually called the significance level of hypothesis testing. While the power of a test is the complement probability of type II error, and is usually expected around 80%, but may vary from study to study.

2.3. Sample size

A large sample size yields a powerful test to detect difference between populations. Therefore, sample size calculation is needed to ensure desirable power in statistical hypothesis testing. For this purpose, a difference in the parameter of distributions between study populations needs to be specified, such as a difference of 10 mmHg of mean systolic blood pressure or 5 kg/cm² mean BMI. Based on the desired parameter of biological or clinical significance, a sample size can be calculated on the basis of probability distribution of the measured values with a given significance level (e.g. 5%) and the power of test (e.g. 80%).

2.4. Data collection

Data collection depends on study designs. A cross-sectional study may require a survey to collect data with response variables that may reflect or represent the study outcome for assessing study aims. A case – control study first identifies subjects: cases and controls, where controls may or may not match the cases with other clinical variables. A randomized controlled trial recruits subjects first and then randomly assigns the subjects to the treatment or control group. A longitudinal study assigns subjects randomly to control and treatment groups, monitors the subjects over time, and collects multiple observations. It is important to note that data collected from case-control studies may be subjected to large recollection measurement errors and large bias. This is because nutrient intakes of subjects are usually not based on food consumption records, but rely on subjects’ memory, which can result in large measurement errors and can also
be severely biased toward study aims. Additionally, the subjects’ recollection of past food consumption may be influenced by their knowledge of the outcome.

To achieve high accuracy in data collection and ensure high quality of study findings, samples may be repeatedly collected from the same subject. These repeated samples may improve the study if properly analyzed. However, two kinds of mistakes are often made by investigators. First, repeated samples are treated as independent samples and the correlation between them is ignored in data analysis. Second, the repeated measurements from the repeated samples are averaged out and the averaged measurements are analyzed. The first approach may mistakenly yield false positives due to inflated sample size and power, while the second approach may generate false negatives due to the loss of power.

2.5. Statistical modeling and data analysis

Depending on the study design and the type of response variables, data will be analyzed with different statistical models to reflect the data structure and potential correlation between observations. Categorical response variables are usually analyzed through contingency tables, logistic regressions or generalized estimation equations (GEE) models, while continuous response variables can be analyzed through the t-test, ANOVA, correlation, and regression. Additionally, contingency table can be used for testing the homogeneity of distributions for categorical response or explanatory variables.

Statistical modeling is the data processing step to sort out information from the collected data. This can be achieved by building a quantitative relationship between the outcome or response variables and the explanatory or independent variables through a mathematical model or equation that characterizes the dependence of the former on the latter. In modeling the response variables, correlation between response variables should receive special attention,
because correlated responses (e.g., body weight of the same subject at different time points) are highly correlated and thus the correlation structure should be incorporated into the data analysis. Therefore, longitudinal studies should be carefully analyzed because subjects are monitored with multiple observations at different time points and the correlation structure between observations affects estimation accuracy and subsequent inference.

Interpretation of statistical analysis results is critical for making inference and valid conclusion. Particularly, p-values have been in an irreplaceable position in classical biological studies. However, recent advances in high throughput technologies, such as microarrays, have made it possible to simultaneously analyze thousands of genes and identify some genes that are potentially responsible for the observed differences in the outcome or phenotype of study subjects. This raises a multiple comparison issue in hypothesis testing of multiple genes and thus leads to different criteria for the claim of statistical significance through correction for multiple comparison, such as family-wise error rate (FWER) or false discovery rate (FDR) [21].

2.6. Special concerns over genetic or “omic” data

Genetic or “omic” data are those obtained from genetic, genomic or proteomic studies, such as gene expression data, DNA genetic polymorphism data or protein profiles. Current technologies for conducting genetic and “omic” studies provide measurements of the intensities of genes or proteins, which to a certain degree represent levels of gene or protein expression. Because of the relatively large noise in the microarray gene expression data, significant findings usually need to be confirmed experimentally through quantitative real-time polymerase chain reaction (qRT-PCR), which itself also is subject to variability of other sources. Recent advancement has led to the study of copy numbers that is more intrinsic to the molecular activities and less relies on the technology.
Through high-throughput technologies, a large amount of molecules with known actions (e.g., proteins at the binding site for DNA transcription) can be monitored so that their functions can be studied through measurement of their intensities that may be associated with phenotypes of interest. This approach may provide clues for further study design for causal relationship between risk factors and outcomes. The exceedingly large amount of data can be obtained from the genetic or “omic” studies (e.g., thousands or tens of thousands of genes or proteins, as well as millions of single nucleotide polymorphisms in genome-wide association studies), which are distinct from the traditional biological or clinical studies. These huge amounts of data provide opportunities to biologists to conduct fine-tune studies with great details of genes or proteins, but also present challenges to quantitative scientists (e.g., statisticians and bioinformaticians) to correctly decipher the data and make meaningful conclusions. In terms of statistical modeling and analysis, the “omic” data are characterized as high dimensional (thousands) or highly correlated (genes cooperate to achieve biological functions). However, such studies often involve small sample size, giving rise to the so-called small $n$ – large $p$ problem.

3. Sample size and power calculation

3.1. General considerations

Sample size determination is a major issue in planning quantitative nutrition research. Accurate estimation of sample size will not only ensure the planned study to achieve sufficient power to detect significant differences, but also save time and money by recruiting no more subjects than needed. Many factors affect sample size calculation, including type I error rate, power of test, and expected significance of detection. Sample size calculation for studies not involving microarray or other high-throughput technologies can be found in many biostatistics books, including Rosner [22] and Fleiss et al. [23]. In this section, we summarize methods for
sample size determination for microarray studies. They can be applied similarly to studies involving other high-throughput technologies.

In a microarray study, experiments of 20000 to 30000 genes or features are conducted at the same time and the variability in gene expression differs from each other. Thus, a traditional power analysis would result in 20000 to 30000 different sample sizes. Genes that have similar, but not the same expression in two groups, would require very large sample sizes to detect the minor difference, while genes with dramatic differences can be detected with very small sample sizes. Three per group is usually the minimum sample size for publication, and is reasonable for cell cultures and mice, where between-subject variability is small. Human or other studies, where variability is larger, benefit from at least five to ten subjects per group. Obviously, a larger sample size is always better, but the question is how large is sufficient.

Because the aims of microarray gene expression studies include identification of differentially expressed genes between cases and controls, as well as profiling of subjects based on gene expression levels, the main objective of microarray studies is to discriminate cases from controls. Two major classes of statistical models have been studied so far. One class of models focus on gene expressions, including the ANOVA method [24], and t-test-like method, such as significance analysis of microarrays or SAM [25]. The other class of models focus on the subject label (case versus control, or receiving study treatment versus standard treatment or control), including logistic regression model, or classification models, such as Bayesian hierarchical model [26]. In the first class of models, accurate sample size estimation ensures having enough power to detect genes that are differentially expressed, while in the second class of models, it allows the selection of genes that will be capable to discriminate cases from controls.
Many methods have been proposed to determine sample size when pilot data are available [e.g., 27-35]. In general, two broad approaches have been employed, the model-based approach and the direct control of error rate approach. The model-based approach relies rigorously on the models for microarray data analysis, such as the ANOVA model, which may provide accurate estimation of sample size if the proposed model fits the data properly. This approach is more or less similar to the classical approach to sample size determination based on conventional statistical parametric models with specific assumption on distributions of response variables and experimental risk factors, but differs in the special characteristics of the high throughput data that require error rate adjustment through either FWER or FDR for multiple comparisons. However, it is often difficult to identify a single model that fits the data well. In such cases, the direct control of error rate approach is more appropriate and yields accurate estimation without assuming the models for analysis. Several authors also provided free software to calculate sample size [32, 35]. In addition, free software and public databases of microarray data are also available to support sample size determination for the investigators who have no previous pilot study data for sample size calculation [34]. Their approach does not require a specific class of model, but rather focuses on the distribution of the p-values of single gene expression analysis. This tool makes the sample size determination user-friendly and easily accessible.

3.2. The ANOVA model-based approach

This approach rigorously depends on the statistical model for data analysis, i.e. the ANOVA model, where the individual gene expression or its transformation (usually a log transformation to ensure the normality of intensity data) is assumed to be normally distributed and analyzed using ANOVA model. Popular models are one-way ANOVA or two-way ANOVA, incorporating design factors for the experimental design. See Kerr and Churchill [36] for the
global ANOVA model and Wolfinger et al. [37] for a generalization with random effects. Among the sample size determination methods, Lee and Whitmore [27] described detailed modeling and sample size calculation based on the classical approach to sample size determination for linear models with the adjustment for multiple comparisons through controlling type I error rate, FWER and FDR. They then considered detailed sample size for several standard microarray study designs, including matched-pair data, completely randomized design, and isolated effect design. A sample size table was also provided for each design considered. Their sample size determination is assisted with a software package *sizepower* in R (see [35] for details). Similarly, Dobbin and Simon [33] derived sample size calculation based on a model similar to ANOVA taking into consideration more technical details of microarray technology, such as single-label or due-label, dye-swamp microarrays, etc.

Keep in mind that microarray experiments typically have very small sample sizes. Thus, it is essential to keep variability as small as possible, particularly in the simple one-way or two-way ANOVA analysis. As an example, an experiment that looks at gene expression in liver and heart tissue of mice injected with PCBs or a control. There are two factors (tissue and PCB), each at two levels. A 2x2 ANOVA with five chips per group would use 20 mice, randomly assigned to each of the four treatment combinations. An alternative design would use 10 mice, randomly assigned to PCB or control and RNA from liver and heart of each of the ten mice would be extracted and measured for expression. The second design, because it uses liver and heart from the same mouse, will have far less variability than the first design, and thus have much more power to detect true differences in gene expression.

3.3. *The direct control of error rate approaches*
When there is no *a priori* knowledge of appropriate model for statistical analysis of the proposed data, the required sample size may be calculated based on the direct control of error rate. Muller et al. [28] provided detailed theoretical study on sample size determination of microarray studies based on FDR and FNR (false negative rate), whereas Tsai et al. [29] calculated sample size based on the expected number of false positives using individual comparison-wise error rate. Based on the FDR-control, Jung [31] derived the sample size for a specific number of true rejections while controlling the FDR at a pre-specified level, and provided an exact formula based on a t-distribution assumption. Alternatively, Pounds and Cheng [32] proposed an anticipated FDR (aFDR) method for sample size determination, controlling the FDR, pFDR or positive FDR, and cFDR or conditional FDR. Their method can be easily implemented with R code available on the web at [http://www.stjude research.org/depts/biostats/documents/fdr-library.R](http://www.stjuderesearch.org/depts/biostats/documents/fdr-library.R). Finally, sample size can be estimated using t-statistic, FDR, and large fold change, and other methods [30].

Of note, The PowerAtlas [34] is a power analysis and sample size calculation software package that provides not only sample size calculation, but also needed pilot study data based on publicly available data from previous microarray studies. Their sample size calculation method is based on studies of the distribution of p-values from single gene expression analysis in microarray studies controlling for expected discovery rate (EDR). It allows the use of either investigator's own data or publicly available microarray database already incorporated into the software for sample size calculation. The free software is available on the web site [http://www.poweratlas.org/](http://www.poweratlas.org/).

Sample size calculation is a critical step in designing microarray studies. Accurate estimation of sample size will not only allow optimal design and budgeting of the planned
research but also ensure the desired power to detect significant findings. Since microarray studies
have presented challenges in many different aspects, various methods for sample size calculation
make it difficult to choose an appropriate method for investigators. We suggest selection of a
method based on the design of the study. If the statistical model for data analysis is known from
a pilot study with a desirable aim to achieve, a more specific method of sample size calculation
may be chosen that best fits the analysis plan. If not, an FDR-based approach may be used. If no
pilot study data is available, it may be possible to take the advantage of the PowerAtlas software
to borrow the strength from publicly available microarray study database.

After microarray data have been collected and analyzed, investigators may find it useful to
conduct a power analysis if significant findings were not achieved for the target genes yet some
tendency toward significance was observed. To conduct power analysis, one way is to calculate
the power of a test based on the data and statistical model, which has been provided in many
papers mentioned above. The other is to follow a procedure studied by Fu et al. [38] to examine
if enough subjects have been recruited. This approach assumes that the subjects were
independently recruited in a serial procedure. The classification model will be updated each time
when a subject is recruited and also be tested on a newly recruited subject. It provides a stopping
rule with a pre-specified probability to ensure that, at stopping, the probability of misclassifying
the next subject will be less than certain pre-determined threshold. A bootstrap approach may be
taken to bootstrap the collected samples so that the needed sample size will be able to obtain for
a stopping time based on the target threshold. Random sampling can also be employed to explore
how many more samples will be needed to achieve the pre-determined misclassification level if
stopping is not achieved based on the study data.

4. Statistical analysis of microarray data
4.1. Platform selection

The first decision in a microarray experiment is to pick a platform. The two basic options are one color or two. One color means that only one sample of RNA goes on a chip, while with two colors, two RNA samples go on each chip. While more economical, two color chips are often custom made and require extensive effort in establishing quality control and statistical analysis methods. One color chips are typically professionally produced and reliability of the chips has already been established. The most common one color (or oligonucleotide) arrays are from Affymetrix or Illumina [39] and will be the focus of our discussion. Most of what is said here applies to two color, as well as oligonucleotide arrays. The flow chart of statistical analysis of microarray data is illustrated in Fig. 1.

4.2. The use of replicates and pooling in microarray analysis

The use of replicates in microarray experiments is under constant debate. Technical replicates are using the same mRNA sample on multiple chips. They are useful for establishing the reliability of the platform, but they cannot be used to increase the sample size for statistical calculations. Biological replicates are where different mRNA samples go on each chip, and thus they contribute to the overall statistical sample size for the experiment. In general, for professionally produced microarray chips, technical replicates are not useful. The reliability of the platform has already been well established.

A related issue is the use of pooling, which means putting more than one mRNA sample on each microarray chip. This reduces individual variability, and thus increases power, but at the price of not being able to use individual covariates in the statistical model. In the 2x2 ANOVA example discussed previously, it is possible that the weight of the mouse might impact gene expression. If pooling were used, then weight could not be used as a covariate in the model.
When pooling, it is essential to extract RNA from every sample and then combine equal amounts of RNA from each sample to go on each chip [40].

4.3. Normalization of gene microarray chips

Once the RNA has been appropriately extracted, hybridized to chips and scanned, it is time to normalize the chips so that data between chips can be “fairly” compared. Although plenty of options are available for chip normalization, the most common are MAS 5 from Affymetrix, and gcRMA from Bioconductor (www.bioconductor.org). MAS 5 is the easiest to use and will be the focus of this discussion. gcRMA will typically yield similar results with the exception of the situation where gene expression is very low, in which case, gcRMA is likely better than MAS 5.

Using MAS 5 results in an output file for each chip that contains the probe set ID, the probe set expression, the presence/absence call and the presence/absence p-value. The presence/absence p-value is used to declare each probe set either “P” for present (p<0.05), “A” for absent (p>0.065), or “M” for marginal (0.05 < p < 0.065). The p-value cutoffs for each label are adjustable. Technically, the assumptions of independence that the statistical test makes are not satisfied, but the P/A call is still useful as we will see below. MAS 5 can also output what Affymetrix calls a “change p-value” for each probe set on a pair of chips. Change p-values are statistically wrong and misleading and should never be used in practice.

4.4. Data reduction in microarray analysis

The next step in the statistical analysis is data reduction. If there are probe sets that are not of interest, then statistical calculations should not be done on those probe sets. On most, if not all, Affymetrix chips, the first approximately fifty probe sets are quality control probe sets used by the MAS 5 software. Typically, there is no need to do statistics on these probe sets. Similarly, many researchers are not interested in Expressed Sequence Tags (ESTs) which are genes that
have not been annotated. These should also be removed from the data set if they are not of interest. The final group of probe sets that are typically removed are probe sets that are labeled as absent (A) on all the chips in the experiment. If the P/A call determines that the probe set is not expressed on any chip in the experiment, there is no reason to do statistical analysis on that probe set. If a fairly large number of chips are involved in the experiment, say a total of 20 or more, then the condition of all absent calls on all chips could be relaxed to allow a few marginal or even “present” calls would still result in being removed prior to statistical analysis. At this stage, researchers should also identify any subsets of the probe sets that are of particular interest. For example, a particular pathway or annotation feature. These subgroups can be statistically analyzed separately as well as together with the rest of the probe sets.

4.5. Log transformation of microarray data

The final step prior to statistical analysis of the data is to decide whether or not to take the log transformation of the data. For most microarray data sets, the probe sets with larger expression levels benefit from a log transformation, but the smaller expression levels should not be logged. Most researchers choose to log their data, but many do not. Typically for one color microarrays, there is not much difference in gene lists with or without logging the data.

4.6. Methods of statistical analysis of microarray experiments

The statistical analysis of microarray data is typically done row by row using the analysis appropriate for the experimental design. The most common designs are two sample t-tests, one and two way ANOVA. Almost always, it’s assumed that experimental errors are normally distributed. Obviously, this assumption is not true for all probe sets, but the small sample sizes for most microarray data sets make the normality assumption a good choice. The most common error at this point is likely failure to treat dependencies between the chips properly. If RNA is
taken from the same subject more than once, then a repeated measures statistical model is needed. It is helpful to consult a statistician before proceeding.

The end result of the statistical analysis will be one or more p-values for each probe set. The overall p-value for each row tests whether there are any statistical differences between the rows. A histogram of these p-values provides useful information. If that histogram looks like that of a uniform distribution (a rectangle), then there may be little if any differences between the treatment groups. On the other hand, a histogram with a large peak for low p-values indicates that large differences exist between the treatment groups. Histograms with a low or moderate peak for small p-values indicate that more chips would likely result in smaller p-values for probe sets that are actually differentially expressed.

The next decision is how to determine the list of probe sets that have changed. Traditionally, a p-value less than 0.05 rejects the null hypothesis of no change. In a microarray experiment in which say, 10,000 tests are done, using a p-value cut off of 0.05 could mean as many as 0.05x10,000 = 500 false positives. The false discovery rate (FDR) of Benjamini and Hochberg [21] chooses the cut off by a user specified expected proportion of false positives on this list. Ten or twenty percent are common choices. For experimental conditions which cause differential expression, but not large changes, the FDR method may not find any genes that change. As an alternative to using p=0.05 or the FDR method, many researchers simply use p=0.01 as the cutoff.

Once the overall p-values are used to determine the list of genes that changed, many researchers attempt to use cluster analysis to determine genes that are responding similarly to the experimental conditions. To avoid excess noise in the gene clusters, be sure to cluster only genes that are determined by statistical methods to be differentially expressed. Many different types of
cluster analysis are possible, and they often yield results that are hard to interpret. Statistical
pattern matching (e.g., Liu et al. [41]) is an alternative that can be used to divide that list into
sublists of genes that changed similarly. For example, if two sample t-tests were used to generate
the overall p-values, then the list should be sorted into up regulated and down regulated genes.
For more complicated experimental designs, the patterns will be more complicated. Consult with
a statistician to determine appropriate patterns.

The biological interpretation of the resulting list(s) is done by first annotating the gene lists
using the manufacturer’s web site. There may be obvious biological conclusions that can be
made at this point. A more statistical approach is to provide the list plus a larger list, say the
entire chip, to a statistical software package that statistically determines gene ontology categories
that are over represented on the smaller list compared to the larger list.

5. Statistical analysis of quantitative RT-PCR (qRT-PCR) data

5.1. General considerations

Due to large variability in gene intensity data inherent in the microarray technology,
analysis of microarray gene expression data is subject to wild noise and thus significant findings
should be confirmed by a more reliable method. Potential sources of the wild noise in the
microarray technology include fluorescent scanning, uneven spray of reagents within arrays,
control of environmental factors, and varying experimental conditions for different arrays. These
factors lead to large variability, reflect artifacts that may not be due to the biological variability
in controls and cases or in subjects receiving different treatments, and may not be resolved by the
within-array and between-array normalization in data pre-processing. The huge number of genes
or probes in microarray studies, usually around tens of thousands, may also lead to the
identification of a large number of false positives. To further confirm the findings in the
microarray studies, quantitative (real-time) reverse-transcription (RT) polymerase chain reaction (PCR) or qRT-PCR experiment is often carried out on the DEGs identified in the microarray gene expression analysis.

5.2. Threshold cycle in quantitative RT-PCR analysis

qRT-PCR quantifies the amplification of genes and records the real time (a threshold cycle of each gene to achieve a pre-set intensity level). This threshold will be used to calculate the initial expression of the gene in the tissue sample and to further compare gene expression between cases and controls in terms of fold or percent change. For this purpose, a reference gene (endogenous) is usually pre-specified to be amplified together with the selected genes. To confirm the microarray analysis findings on the DEGs, tissues from the selected cases and controls will be collected for the qRT-PCR experiment on the DEGs. Usually, the number of subjects in each group of treatments or controls is smaller than the corresponding microarray study. Since qRT-PCR experiments generate threshold cycle data for each selected genes, data analysis need to be conducted with proper statistical modeling to ensure correct analysis results. Although mathematical models of PCR experiments have been studied in the literature, statistical modeling has not received much attention so far. Only two statistical methods for qRT-PCR data analysis have been proposed. We will focus on these two methods to provide a detailed discussion on the models and methods below.

5.3. Mathematical models for RT-PCR analysis

DNA sequences are amplified in RT-PCR through DNA polymerase. During the exponential amplification phase of RT-PCR, a copy of target gene doubles in one cycle, and then quadruples in the next cycle. Therefore, the amplification is in the power of 2, or exponential amplification, and can be described by an equation \( Y_n = Y_0 \ 2^n \) with \( Y_0 \) being the initial
expression level of a target gene, and \( Y_n \) the expression level after \( n \) cycles. Because the amplification in an experiment is subject to variations in experimental conditions and may not be 100% efficient and the amplification process may not end with an exact number of cycles, the above equation may be written generally as \( Y_t = Y_0 (1+ e)^t \), where \( t \) is the duration of the amplification process in continuous number of cycles, and \( e \) is the amplification efficiency, which may depend on many experimental conditions and sequence properties (e.g. GC content), which usually ranges between 0 (completely inefficient) and 1 (fully efficient). Thus, it is important to have an endogenous gene to serve as an internal reference to ensure the validity of RT-PCR results. The above equation applies generally to both target genes and the reference gene in cases and controls. Also, the target genes may have different amplification efficiency from the reference gene. Because the samples from both cases and controls are processed simultaneously in one RT-PCR experiment along with the reference gene, the efficiency for samples from cases and controls can be assumed identical [20].

5.4. Statistical models for RT-PCR analysis

The above mathematical equations provide the principle of how to analyze qRT-PCR data to calculate the fold change of genes. However, statistical models are needed to implement the above procedure to ensure that the principles are followed for correct data analysis. This is particularly necessary if multiple samples are drawn from each subject. So far, two statistical methods have been proposed, the GEE model – a generalization of the ANOVA model incorporating the correlation between samples within subject [20], and an ANOVA / ANCOVA model that treats within-subject samples using a random effect model [42]. These two models were based on the same mathematical principle as stated above, and both recognize the within-subject correlation and incorporate it in the modeling. In addition, both papers provide readers
with a user-friendly SAS program code. However, they employ two different methods to conduct
data analysis. The GEE model was proposed originally to deal with response variables, either
continuous or categorical, to conduct analysis in longitudinal studies with multiple observations
from each subject. The qRT-PCR data with repeated samples from each subject, which is used
by biologists to emphasize the importance of measurement accuracy, make the GEE model a
perfect approach to qRT-PCR data analysis. The ANOVA / ANCOVA model with a random
effect is also a good approach to this special type of qRT-PCR data.

However, we found that the SAS program for the ANOVA model and ANCOVA model
has two major flaws. First, it does not recognize the correlation between the target gene and
reference gene from the same subject. Because estimating the fold change of each gene from the
initial quantity of the cDNA to the termination of the PCR amplification is the goal of the
analysis, the initial quantity of genes from the same sample are highly correlated. Indeed, these
two initial quantities are perfectly matched in the statistical sense. Thus, the statistical methods
that fail to address this correlation will surely lose its power in the detection of significance.
Second, it assumes constant variance in the ANOVA model for target gene and reference gene
with potentially different concentrations of samples and, therefore, fails to recognize the
heteroscedasticity, i.e. the different variance of expression for different genes. Because gene
expression is known to vary greatly from gene to gene, some of them are expressed at high levels
and may have large variance while some others are expressed at low levels and may have small
variance. The statistical models that fail to address this heteroscedasticity will surely lead to
incorrect inference. Therefore, current methods using the ANOVA/ANCOVA model needs to be
revised or otherwise may generate misleading results. In the GEE model, these two issues can be
simply resolved by taking the difference of the gene expression between each target gene and the
reference gene, which is the recommended approach in biostatistics to work with paired observations and also resolves the varying variance issue automatically. In addition, the GEE model will yield the correlation coefficient between within-subject repeated samples. The SAS program provided for the GEE model is easy to implement.

Finally, it should be noted that the above two methods assume the full efficiency of the PCR amplification. Although there is discussion provided in the ANOVA/ANCOVA model, statistical methods have not been proposed to analyze data with varying efficiency. Interestingly, it has been observed that the amplification efficiency is mostly determined by the experimental conditions, and less influenced by the different gene structure [43]. This makes the above assumption of equal efficiency valid in single RT-PCR experiment.

6. Statistical analysis of proteomics data

6.1. General considerations

The central advantage of proteomic experiments over microarray analysis is the ability to see farther along the “system pipeline” in terms of the function of the underlying genes [6]. In other words, we may investigate not what the genes code for, but whether those products actually appear in the cellular mechanics. However, in many respects, proteomics lags far being microarray experiments due to several challenges in the collection of mass spectroscopy data. Among these challenges are that identification of peptides is a central feature/difficulty of the analysis (as opposed to microarrays where each gene occupies a “dot” on the array), the quantitation information for peptides requires careful manipulation of possibly overlapping peptide peaks, and missing data is much more of a problem (e.g. in many cases peptides are not identified in an experiment even when they are present) than in microarray experiments. See the
recent review by Bantscheff et al. [44] on the experimental side of proteomics studies, including equipment differences and different mechanisms for preparing biological samples.

6.2. Mass spectrometry data

Central to all mass spectroscopy experiments is the preparation of a sample containing a large amount of peptides, which is then broken down through some chemical reaction (most commonly trypsin digestion). This sample is then placed in a tandem MS/MS machine. Ions from the sample then pass through the machine until they reach a detector. At prespecified regular intervals, this first layer of the MS calculates a mass spectrum (measurements of ion abundance on a mass to charge, or m/z, scale) for the ions arriving at that time. This mass spectrum will contain peaks corresponding to the m/z values of abundant peptides in the original sample. After each scan performed by the first MS layer, the machine selects an m/z region and allows the ions in this region to pass through to the second layer of the tandem MS/MS equipment. These ions are then further broken down and move through the machine to a second detector, which computes a mass spectrum as well. Typically the region of m/z values chosen for “pass through” is one of the largest peaks in the MS scan, but most equipment will also allow the experimenter to program specific rules for this operation. This selection process is important, as always choosing high abundance peaks will typically result in never identifying low abundance, but important, peptides in the original sample [45].

Thus, the main data structure of a tandem MS/MS run is alternating scans, beginning with ion concentrations from the original sample in one scan (MS) and then the ion concentrations from the tandem MS/MS scan. The m/z value is selected to pass ions through to the tandem MS/MS layer. This second layer (the tandem MS/MS scans) is used to identify peptides, while the first layer (the MS scans) is used for quantitation.
6.3. Identification of peptides

The data from the second tandem MS/MS scan is chosen for identification of peptides. They provide information about the mass of the original peptide (because this mass range was selected to pass through) and the masses of the pieces where it was broken down. Peptide identification is based on the principle that the breakdown mechanism for each possible peptide is known (e.g., trypsin predominantly cleaves peptide chains at the carboxyl side of lysine and arginine, except when either is followed by proline). Therefore, the original mass and the masses of the broken down component result in a “signature” allowing for reconstructing and identifying the original peptide in the first layer (MS) scan. Doing this from scratch without a database of peptide, simply from known masses of amino acids, is called de novo sequencing [46]. Available software for this task includes PEAKS [47], PepNovo [48], AUDENS [49], and NovoHMM [50]. All of these algorithms rely on some heuristic search techniques to search through a set of possible amino acid sequences to reconstruct the original peptide.

More commonly, a peptide is identified through the search of a database of known peptides [51]. Each peptide in the database, combined with the chemical method for protein hydrolysis, results in a “signature” set of peaks in the tandem MS/MS scan. The observed spectrum is compared to each possible peptide in the database and a score is assigned based on the agreement between the observed spectrum and the expected spectrum. If a peptide in the database receives a sufficiently high score, then the observed peak is considered to be identified as that peptide from the database. Most software for this purpose reports not only the tentatively identified peptide but a numerical “confidence score” reflecting the uncertainty in that identification. The earliest method, and still a standard, is SEQUEST [52]. SEQUEST works by taking each peptide in the database and determining its expected spectrum (using knowledge of B, A, and Y ions). Then the
correlation between the observed spectrum and the expected spectrum is observed. In addition,
the correlation between the expected spectrum and shifted versions of the observed spectrum are
taken (e.g. all the peaks in the observed spectrum are moved to the left or right in the graph, and
the correlation recalculated). If the correlation is exceptionally high for the observed spectrum,
but moderately low for the shifted spectrum, this indicates the observed spectrum aligned well
with the expected spectrum. This is because high correlations will be expected when the peaks
are in the same locations. This calculation is feasible for the number of peptides in a typical
database (fortunately processing power has been increasing while the size of the databases have
also been increasing). If SEQUEST finds one particular peptide in the database aligns well, it can
be concluded that the observed spectrum came from that peptide.

Unfortunately, as databases grow large, it is possible that multiple peptides produce
reasonably high correlations. Note that the spectra are observed with noise, which often peaks in
the expected spectrum and may not be represented in the observed spectrum, lowering the
accuracy of this method. If so, one may be interested in relative fit, not absolute fit. In other
words, a peptide is identified only if the match to the model spectrum is significantly better than
others in the database. The program X!Tandem [53] attempts to do this quickly. X! Tandem
works by only summing the peaks in the tandem MS/MS spectra which match the model spectra
(thus no shifting, which is computationally costly). A scaled version of this quantity is called the
Hyperscore. The hyperscores are then calculated for each peptide in the database. Histograms of
these hyperscores follow a distribution one would expect from random matching, which is
reasonable considering most peptides in the database should NOT be a match for the observed
spectrum. X! Tandem then assigns an “E-value” that indicates how good the observed match is
compared to what would be expected by chance matching. If the E-value is sufficiently low (e.g.
we would almost never expect this good a match by chance), then the protein is declared to be
identified. This calculation is quite similar to a statistical p-value. A similar, competing idea to
X! Tandem is MASCOT [51]. See Brosch et al. [54] for a comparison of the two. It is possible to
combine results from several of these algorithms for even better identification. The MASCOT
method has been successfully used in nutritional proteomics research [55,56].

Peptide Prophet provides a method for converting scores (from any identification method)
into a probabilistic identification (e.g. Peptide Prophet reports a probability that the peptide is
correctly identified) [57]. Peptide Prophet begins by producing a discriminant score for each
spectra in the sample. These discriminant scores fortunately fall into two groups – a correctly
identified group and an incorrectly identified group. These groups slightly overlap, so Peptide
Prophet uses the EM algorithm to fit a mixture distribution and then assigns a probability to each
identification [58]. If the discriminant score is firmly within the “identified” group, this
probability is near 1, while if the discriminant score is firmly within the non-identified group this
probability is near 0. For discriminant scores in the middle where the overlap occurs, one can get
probabilities of identification anywhere between 0 and 1. Note the user may then choose which
peptides to pursue. Only looking at peptides with high identification probabilities results in being
surer of the results, but one may miss some correctly identified peptides. Using lower probability
scores may be more useful in an exploratory setting where one is looking for interesting peptides.

Note that post-translational modifications of the peptides, such as glycosylation or
phosphorylation [59], can cause difficulties with these methods. The reason is that such post-
translational modifications can alter the chemical structure of the protein and thus change the
way it breaks down. This can lead to misidentification of peptides because the peptide is not
hydrolyzed according to the “expected signature”.
6.4. Isotope labeled quantification

Quantification experiments in proteomics attempt to determine differences between protein expression across treatment groups (e.g. Alzheimer's versus normal patients). These experiments typically take the form of “shotgun proteomics”, where a sample containing a large number of peptides is analyzed with the hope of identifying several differentially expressed proteins for further study. In an isotope labeled experiment, two treatment groups are treated with different isotopes, one heavy and one light. There are a variety of abbreviation heavy methods for this (ICAT, SILAC, iTRAQ, see Bantscheff et al. [44]). The samples are then mixed together and treated as a single sample for the remainder of the mass spectroscopy run. Note that statistically this is important because any variation that can be attributed to handling after the samples are mixed occur equally to all treatment groups. This is in contrast to "label-free" methods, which combine results of multiple mass spectroscopy runs and thus may have differences attributed to experimental variation not present in the labeled experiment.

Quantification is done through the first set of MS runs. Thus, recall the second set of MS runs provides identification, the first set provides quantification information. For identified peptides, we not only have an m/z region (the region allowed to pass through for the identification to take place), but also the time this occurred. We then collect the “first MS” scans for those times in a neighborhood around the time of the “identifying” scan (Fig. 2). The x-axis of figure one contains the scan number while the y-axis contains the observed intensities for the light isotope (red) and the heavy isotope (blue). The dark red part of the bars indicates overlapping red and blue portions. To acquire these observed intensities, we have to examine the MS scan where the peptide was identified. The m/z value that was passed through to the second, tandem MS scan show us where to look in the MS scans in that neighborhood for that peptide. In
addition, we also have to look for m/z values representing the “shift” in mass due to isotope labeling. This is expected to succeed because each isotope should have the same chemical properties, thus the labeling results in different masses between the same peptide across the treatment groups. Furthermore, if the peptide is known, we can determine the expected difference in mass between the light and heavy isotope. Thus, in addition to assessing the expression for the identified peptide, we also look in the mass area where we expect the “companion” peptide from the alternative treatment group to be located. Similarly, to the original peptide, we also acquire a peak and compute the area within that peak to determine the expression of the companion peptide. The ratio of these expressions is then used to quantify the relative expression of the peptide across the treatment groups. Fundamental to this process is the fact that the ion concentrations are measured with noise, and peptides which have relatively low expression (relative to the background variation) are particularly difficult to quantify.

A straightforward but somewhat thorny issue is simply to determine how long before and after the identified peak to add together (in Fig. 2). For example, we have to determine how far out to “color” the bars red or blue for inclusion into the calculation. At the center of a good peak, the expression is far above the baseline noise; however, at the extremes the peak begins to slide back into the baseline noise. If a very small window is chosen, fewer values are summed, yielding higher variation in the resulting expression estimates (namely, a smaller sample of points is used). On the other hand, if a very broad window is selected, pure noise is added up into the expression level, resulting in poor estimates of protein expression.

One common method for calculating expression ratios is XPRESS [60], which performs this task using a low pass Butterworth filter and simply taking a sum of the smoothed values. Another is ASAPRatio [61], which combines a Savitzky-Golay filter with a normal (Gaussian)
fit of the peak to determine expression. Finally, in RelEx, a Savitzky-Golay filter is applied for 100 scans before and after the identified peak [62]. Then, a linear regression is performed (Fig. 3) on the expressions at each scan. Namely, a data point in the linear regression is the expression for the heavy isotope at that scan plotted against the expression for the light isotope at that scan. Ideally (if the expression has similar shapes), these points should form a line, whose slope is the expression ratio. In practice, the points are measured with errors, and thus the points fall in a linear trend, but not exactly on a line. RelEx is convenient in that the quality of the regression provides a quick measure of the confidence in the expression ratio. If the regression diagnostics (e.g., $R^2$) are poor, this may indicate a poorly identified peptide or other problem.

In these algorithms, expression ratios for peptides, not proteins, are quantified. If a researcher has confidence in a certain list of peptides form a particular protein, protein expression ratios may be obtained by combining the estimated expression ratios for each component peptide. Typically this is done on the log scale as estimate of the log expression ratio are more stable than estimates of the expression ratios themselves, particularly when one of the isotopes is minimally expressed in the denominator of the ratio. Averaging ratios on the log scale is equivalent to taking the geometric mean of the ratios themselves. As averages are normally distributed, a standard error for the protein ratio can be computed as well.

Unfortunately, for various experimental reasons, one also runs into the problem that a peptide is identified for one of the treatment groups but no corresponding peak may be found for the other treatment group. This can occur for reasons other than non-expression in the other group (e.g. it cannot simply assumed that the peptide was not expressed in the other group). Thus, missing peptides will limit the ability to perform quantification.

6.5. Label-free quantitation
Isotope labeling is a difficult process which is not possible in many areas, in particular human proteomics. Thus, an active area of research is in “label-free” methods. In a labeled experiment, one mass spectroscopy run is completed with both the heavy and light isotopes together in the input. In a label-free experiment, multiple MS runs are performed. For example, in an experiment attempting to determine protein expression differences between Alzheimer’s patients from normal controls, two samples would be prepared, one from the Alzheimer’s patients and another from the controls, and each would have a separate MS run. Experimentally, this immediately creates the difficulty that any variation from run to run will now also be included in the estimate of the protein expression ratio, something controlled for in isotope labeled experiments. Among these difficulties is that the peaks are contained in different scans during the experiment. Specifically, the scan numbers containing a particular peptide in one MS run may not be sufficiently close to the scan numbers containing the peptide in another MS run.

Thus, the central statistical difficulty in the analysis of label free methods is finding corresponding peaks in multiple MS runs. Specifically, one must be able to identify something along the lines of “the peak at such and such m/z value in scan X of the first MS run corresponds to the same peptide as the peak at such and such m/z value in scan Y of the second MS run.” One popular method for handling this problem is PEPPrER (Platform for Experimental Proteomic Pattern Recognition, [45]), a method which heavily utilizes normal mixture models [63] and bootstrapping [64] to identify common peaks which are then used for quantitation. Another possibility is ProtQuant [65] which utilizes the XCorr scores from the SEQUEST identification for quantitation. Label free quantitation methods is one of the more active research areas, with multiple software packages being developed rapidly.

6.6. Software platforms
A wide variety of software is available for the analysis of proteomic data. Some of this software is open source and freely downloadable, while other software is propriety. We will focus on the open source software. A fundamental problem in proteomics is the reasonable large amount of data generated in a single experiment (many MS and tandem MS scans, each containing ion frequencies at many m/z values results in large arrays of data). Thus, data storage formats are an initial problem. At present there are several standard formats available, see Droit et al. [66] who discusses variants based on extensible markup language (XML) including mzXML, PepXML, and ProtXML. These data formats are used in a variety of software packages, and thus provide a convenient format for analyzing results with different software packages in addition to a standard format for posting datasets to the internet.

A popular pipeline for the analysis of proteomic database is the Trans-Proteomic Pipeline (TPP), available for free from the Seattle Proteome Center (http://tools.proteomecenter.org/software.php). This software includes a large number of tools for data handling (including the conversion of data from a variety of formats), identification of peptides and proteins (PeptideProphet) and quantification tools such as XPRESS and ASAPRatio, in addition to many others. A key advantage of any complete open source pipeline is that it may be amended by users, and thus new analysis methods may be easily compared to older analysis methods (see Nelson et al. [66]).

On fundamental issue where proteomics could be improved is a fuller understanding of the error processes underlying proteomics data. This is a particularly thorny problem due to the detailed interactions present in the chemistry. However, such an understanding is fundamental to proper statistical procedures. Many papers in the literature are light on descriptions of the statistical methods used (in terms of the exact numerical procedure and their error properties) in
favor of experimental comparisons. This is compounded by several pieces of software only being available in closed source, propriety forms. Thus, it is essentially impossible to determine exactly how the data is being processed.

A “mundane” statistical procedure like linear regression is not used because it has performed well only in a few experimental datasets. It can be shown that, given certain assumptions (e.g., normally distributed independent errors), linear regression is the optimal way of estimating parameters. These assumptions can then be verified through the use of residual plots.

7. Statistical analysis of other bioinformatics data

7.1. Linkage studies

In addition to studies of microarray gene expressions and proteomics for protein functions, bioinformatics research also include linkage studies through analysis of family pedigree, genetic association studies through analysis of SNP genotype, transcription regulatory region studies, copy number variation studies, and the most recently developed genome-wide association studies of human genome.

Linkage studies have been developed extensively in biological research for genetically inheritable diseases. The requirement of data collection through multiple generations makes it difficult to implement, especially with late on-site diseases (e.g., Alzheimer disease), and less attractive and less powerful compared with case-control studies of unrelated individuals. On the other hand, genetic association studies among independent cases and controls offer promises to enhance the detection of the association between specific genes or gene sets and diseases or phenotypes, where the presence of specific allele of genes may alter the risk significantly. Furthermore, human genetic variation studies offer great promise in deciphering the genetics of
complex diseases, such as hypertension, diabetes, through genome-wide association studies [68] and copy number variation (CNV) studies [69,70]. Recent development of copy number estimation methods and SNP genotyping techniques through high density SNP array technologies, such as Affymetrix 5.0 and 6.0 SNP arrays [71], further expedites the process of genome-wide association studies and copy number variation studies, and thus makes it possible to conduct large scale research with millions of SNPs and tens of thousands of subjects [68].

Genetic association studies focus on the identification of genes or specific alleles of genes, gene – gene interaction or haplotypes (combination of specific alleles of different genes on the same copy of chromosome) that is highly associated with the disease or phenotype, or their interaction with environmental risk factors. The case-control study is the most popular design for genetic association investigations, where a logistic regression model is often employed. While a single gene association model may be straightforward, gene - gene and gene - environment interaction may pose greater challenges and also lead to more significant findings. For example, haplotypes may present special gene-gene interaction, but are not observable with double heterozygous SNPs because of an ambiguous phase that needs a special treatment with the statistical expectation-maximization (EM) algorithm for missing data [72-74].

Genetic association studies are based on the genotyping of SNPs, where each SNP generally has two alleles, either the same (homozygous SNP of ‘AA’ or ‘BB’ type) or different (heterozygous SNP of ‘AB’ type), that vary with individuals. Genetic association studies and the most recently developed genome-wide association studies are based on genotype data of a large number of SNPs annotated with high density SNP microarrays, such as Affymetrix 100K SNP arrays, 500K SNP arrays (or 5.0 arrays) and 6.0 SNP arrays. Hence, it is crucial to accurately annotate each SNP from the microarray probe intensity data. So far, a few genotype calling
methods have been developed on the basis of Affymetrix high density SNP arrays. In contrast to the single array-based genotype calling methods including the whole genome sample assay (WGSA) and dynamic modeling (DM), the following machine learning-based methods including the RLMM [75], BRLMM [71], CRLMM [76], MAMS [77] and a single array approach GEL [78] have improved genotyping accuracy. Such techniques usually require multi-array training with the gold-standard annotation of SNP genotype of the HapMap samples. A novel approach based on a robust model of DNA copy numbers requires only one array, which further improves genotyping with high accuracy consistently for different prototypes of arrays and resolves the missing data problem in genotype data [79]. These methods provide a broad spectrum in SNP annotation for genetic association studies, particularly for genome-wide association studies.

7.2. Estimation of DNA copy numbers

DNA copy number variation has been reported to play a critical role in cancer research [80]. Nutritional studies have also been conducted to provide important clues to the development of cancers, including breast cancer, colon cancer, and prostate cancer [81,82]. Several methods have been proposed to estimate copy numbers, including a high-resolution method [83], the CNAG method [84], the CARAT method [85], and the PICR method [79]. Of note, the PICR method provides accurate estimation of copy numbers at each SNP site and thus provides high resolution of copy number detection. This allows for both detection of copy number alteration and SNP genotype calling with high resolution at high accuracy consistently for cross-laboratory studies, even cross-array prototypes.

7.3. Genome-wide association studies

Large scale genome-wide association studies have become increasingly popular in recent years to obtain significant findings for many diseases or phenotypes at an unprecedented speed.
Often, multiple diseases or phenotypes are studied together and millions of SNPs are annotated to discover major associations between diseases and genetic risk factors, including gene-gene and gene-environment interactions. The quality of data from such big projects is no doubt of great importance, because it will lay a solid foundation for powerful detection of associations and accurate assessment of scientific findings.

Because one of the major goals of nutritional studies is to identify how expression of genes is regulated through dietary intervention [1-3], the identification of transcription regulator regions (e.g., binding sites) or transcription factors (e.g., promoters and enhancers) plays a major role in advancing the field. Although methods have been developed in this research area, many of them (e.g., hidden Markov models) have technical limitations and may not be suitable for data from different cell types or animal species. Most recently, a new word-counting method has been shown to be robust for studies of different eukaryotes [86]. Because the word-counting method does not have much technical limitations, it is believed to be applicable to a wide arrange of studies as demonstrated [86].

8. Conclusion and perspective

Statistical analysis is a necessary means to test hypotheses in nutritional and other biomedical sciences. In the post-genome era, there is increasing interest in quantifying the effects of nutrients on simultaneous expression of thousands of genes and proteins in cells or tissues [87-92]. This has offered new exciting opportunities for nutritionists but also presented technical challenges in experimental designs and valid statistical methods for data analysis. Sample size calculation is a critical step in designing microarray and other high throughput studies. Accurate estimation of sample size will not only allow optimal design and budgeting of the planned research but also ensure the desired power to detect significant findings. It is also
crucial that data obtained from microarray gene expression, RT-PCR, proteomics, and other bioinformatics studies are subjected to appropriate statistical analysis. Particularly, in microarray analysis, statistical significance in levels of differentially expressed genes among treatment groups is commonly determined by a combination of p-value and FDR. Likewise, GEE models that properly reflect the structure of data is often employed in statistical models for assessing fold change of gene expression in real-time RT-PCR experiments. Moreover, a number of software platforms (e.g., MASCOT, Peptide Prophet, Sequence, and X!Tandem) are available for identification of proteins in biological samples. Additionally, levels of protein expression can be determined using isotope-labeled and isotope-free quantification methods. Finally, bioinformatics tools have been developed for SNP genotyping, genetic linkage or genetic association studies in nutrition research. We anticipate that this article will provide useful guidelines for nutritionists and other biomedical scientists to plan and conduct sound studies at molecular, cellular, tissue, and whole-body levels and to employ appropriate statistical methods for analysis of experimental data in the era of systems biology.

Acknowledgments: The authors would like to thank Bill Nelson for help with the Trans-Proteomics Pipeline and Frances Mustcher for office support.
References


Figure legends

**Fig. 1.** Flow chart for microarray experiment and data analysis. A microarray experiment involves platform selection, sample size calculation, adequate design, data collection and processing, and normalization of gene chips. Statistical significance in levels of differentially expressed among treatment groups is commonly determined by a combination of p-value and the false discovery rate. Results of microarray studies are normally verified by real-time RT-PCR analysis.

**Fig. 2.** A well identified peak. The x-axis contains the scan number while the y-axis contains the observed intensities for the light isotope (red) and the heavy isotope (blue). Dark red areas indicate overlapping bars. The estimated protein expression ratio is simply the sum of the red bars divided by the sum of the blue bars.

**Fig. 3.** Idealized version of RelEx. The bottom panes of the figure show the observed light (red) and heavy (blue) ion intensities. The intensities used are colored. The points in the main pane are the light and heavy intensities observed for each scan plotted against each other. The slope of the regression line (determined without an intercept) estimates the expression ratio while the correlation of the points (how well the peaks align) is a measure of how confident we should be in the estimated expression ratio. The full implementation of RelEx adds various smoothers and other enhancements.
Figure 1
Figure 2
Figure 3