# Part One
# General Overview



*Statistical Diagnostics for Cancer: Analyzing High-Dimensional Data*, First Edition.
Edited by Frank Emmert-Streib and Matthias Dehmer.

# 1
# Control of Type I Error Rates for Oncology Biomarker Discovery with High-Throughput Platforms

*Jeffrey Miecznikowski, Dan Wang, and Song Liu*

## 1.1
## Brief Summary

This chapter provides an overview of the genetic and proteomic high-throughput platforms and the statistical methods used to evaluate molecular biomarkers for cancer diagnosis. Commonly, these experimental platforms are used in cancer diagnosis where the biomarkers can be used to determine cancer subtypes and thus potential treatments. Because of the large amount of data from these platforms, accurate testing methods are necessary. In this chapter, we highlight the statistical methods used to evaluate each potential biomarker and limit the number of false positives under a specific error rate.

## 1.2
## Introduction

Since the invention of microarray technology and related high-throughput technologies, researchers have been able to compile large amount of information. This amount of information enables researchers to uncover potentially new targets for therapies or to enhance our knowledge of biological systems. These high-throughput platforms have become commonly used experimental platforms in the biological realm [1]. A high-throughput platform is designed to measure large numbers (thousands or millions) of signatures in a biological organism at a given time point. These platforms are a function of the postgenomic era and are often used to determine how genomic expression is regulated or involved in biological processes. These platforms often use hybridization and sequence-based technologies such as gene expression microarrays and RNA-Seq platforms.

Specifically, these platforms and technologies have revolutionized the way researchers study cancer, especially with regard to diagnosis and prognosis. Current cancer classification consists of more than 200 subtypes of cancer [2]. In order to receive the most appropriate therapy, the clinician must identify as accurately as possible the cancer subtype, stage, and/or grade. Clinicians commonly use

*Statistical Diagnostics for Cancer: Analyzing High-Dimensional Data*, First Edition.
Edited by Frank Emmert-Streib and Matthias Dehmer.

morphologic characteristics of biopsy specimens but "it gives very limited information and clearly misses much important tumor aspects such as rate of proliferation, capacity for invasion and metastases, and development of resistance mechanisms to certain treatment agents" [3]. Therefore, in order to improve these classification methods, new molecular diagnostic methods are needed. Thus, the huge amount of molecular information that can be extracted and integrated to find common patterns is a major advantage of these high-throughput platforms. These new technologies will allow researchers to enhance cancer diagnostics by (1) classifying tumor samples into known and new taxonomic categories, (2) discovering new diagnostic and therapeutic markers, and (3) identifying new subtypes that correlate with treatment outcome.

The design of high-throughput platforms, the cost of high-throughput platforms, and the amount of information received from these platforms necessitate the need for statisticians to be involved in the analysis of these experiments. Often the statistician's primary task determines the genomic/proteomic regions of interest for further interrogation, verification, or validation. These regions of interest should be regions of the genome or proteome that are statistically significantly correlated with the outcome of interest, for example, survival, drug response, cancer subtype, and so on. With these large numbers of tests, reporting significance based on univariate $p$-values less than 0.05 leads to a large number of false positives. Besides limiting the number of false positives, another challenge in developing valid high-throughput-derived biomarkers is obtaining large enough datasets with sufficient patient follow-up time [4, 5]. Hence, in light of these concerns the concept of statistical significance has been re-evaluated over the last 20 years, most notably with the study of the false discovery rate (FDR) in [6]. Namely, multiple testing procedures have been greatly studied and refined in order to control a suitable Type I error in these experiments. The goal of these modern statistical procedures is to limit the number of false positive probes or genes that proceed to the validation phase of these experiments.

This chapter is designed to study some of the high-throughput technologies that are employed in these experiments and the Type I error methods to control the results. In the remaining sections, this chapter outlines the high-throughput platforms for cancer diagnosis and the statistical methods to obtain a univariate measure of significance for each gene/protein/probe. Each chapter subsection also contains a hypothetical cancer experiment that would employ the described statistical technique. The chapter concludes by outlining methods that use the univariate $p$-values to control multiple testing based Type I errors for these experiments. Finally, the chapter concludes with a conclusion and perspective for future work.

## 1.3 High-Throughput Platforms

In the following subsections, we outline several of the common high-throughput platforms used in experiments designed for cancer diagnosis. These platforms

were chosen to illustrate the diversity of platforms available for interrogating deoxyribonucleic acid (DNA), ribonucleic acid (RNA), or proteins.

### 1.3.1 Gene Expression Arrays

The human genome consists of DNA sequences located within the nucleus of each cell. Specific DNA sequences are copied (transcribed) into messenger RNA (mRNA). These mRNA copies transition from the nucleus to the cytoplasm of the cell in order for the corresponding sequence to be used in manufacturing various protein molecules.

Genetic microarray technology makes use of this process [7–9]. Generally speaking, there are two types of microarray technology: two-dye spotted pin technology and Affymetrix technology. In the two-dye spotted pin technology, target complementary DNA (cDNA) elements are laid out on a microscopic glass slide and are probed with dye-labeled samples. The target cDNA elements generated in advance are physically arrayed in a two-dimensional grid on a chemically modified glass slide. Then for the two-dye spotted method, equal amounts of two purified mRNA samples are separately reverse transcribed using primer sets labeled with two different fluorescent dyes. The two resulting dye-labeled samples are used as probes in a competitive hybridization reaction with the target elements on the chip. After hybridization, a laser scanner generates two images of the chip at the wavelengths of light corresponding to each sample for each spot on the chip. References [10, 11] discuss these chips further and the preprocessing and analysis methods used on the images that create the microarray data.

Affymetrix gene expression microarrays represent the other major type of gene expression microarray technologies. The Affymetrix DNA microarrays, called "GeneChips" according to the Affymetrix trademark, are generated using semiconductor and photolithography manufacturing techniques. The major distinction between Affymetrix and spotting techniques is that multiple short probes (20–40 base pairs) are used to measure gene expression levels. For this reason, preprocessing methods are critical for this chip. A thorough outline of these methods are available in [10] with a comparison of the various techniques provided in [12]. Regardless of the type of gene expression microarray employed after preprocessing, the array experiment will result in an intensity level for a given probe/subject that reflects the amount of expression for that given probe/subject.

Both types of gene expression arrays are commonly employed to study cancer diagnosis such as in [13–17] and cancer prognosis such as in [18–22].

### 1.3.2 RNA-Seq

Recently, RNA-Seq has emerged as a powerful new technology for transcriptome analysis [23]. A typical RNA-Seq experiment takes a sample of purified RNA, has it sheared, converted to cDNA and sequenced on a massively parallel sequencer, such

as the Genome Analyzer (or HiSeq) from Illumina Inc, SOLiD from Life Technologies Inc, or 454 from Roche Inc. This process generates short (e.g., 75 bp) reads taken from either one end of both ends of each cDNA fragment. Depending on the sequencing depth, the number of sequenced short reads per sample could range from around 10 to 100 millions. By mapping millions of RNA-Seq reads to individual genes' transcripts, one can estimate the overall mRNA abundance and detect differentially expressed genes. Unlike gene expression microarrays that rely on prior probe design and existing transcript annotations, RNA-Seq can be used to analyze any gene and any transcriptome. Applications to cancer studies can be found in [24, 25]. The development of analytic methods to process and analyze the RNA-Seq data is an active area of ongoing research [26].

### 1.3.3 DNA Methylation Arrays

As a major epigenetic modification, DNA methylation plays a vital role in transcriptional regulation and chromatin remodeling. The aberration of DNA methylation profile has been associated with many human diseases including cancer [27]. Use of DNA methylation microarray is a popular approach in studies to characterize the epigenetic landscape of human cells [28].

Three widely used commercial platforms to perform methylation profiling are the GoldenGate Methylation Beadarray, Infinium HumanMethylation27 BeadChip, and Infinium HumanMethylation450 BeadChip provided by Illumina Inc. The first two arrays quantitatively target 1505 cytosine-phosphate-guanine (CpG) loci covering around 800 genes and 27,578 CpG sites targeting around 14,000 genes, respectively, while the last one covers 99% of RefSeq genes and 96% of CpG islands within the human genome. For each targeted locus, the raw fluorescent signals from both methylated (Cy5) and unmethylated (Cy3) alleles are extracted to create the average methylation (β) value derived from multiple replicate methylation measurements. The resulted methylation level (β value) for each locus ranges between zero and one. Zero indicates absent methylation and one indicates complete methylation. Since their release, many analytic methods have been developed to process and analyze the Illumina DNA methylation array data [29].

### 1.3.4 Mass Spectrometry Platforms

Mass spectrometry is an analytic tool used to identify proteins, where the associated instrument (a mass spectrometer) measures the masses of molecules converted into ions via the mass-to-charge ($m/z$) ratio. This technology can be used to profile protein markers from tissue or bodily fluids, such as serum or plasma in order to compare biological samples from different patients or different conditions. Matrix-assisted laser desorption and ionization – time-of-flight (MALDI-TOF) is a popular tool used by scientists, where a metal plate with the matrix containing the sample is placed into a vacuum chamber that is excited by a laser, causing the protein

molecules to travel (or "fly") through the tube until they strike a detector that records the time-of-flight for the various proteins under study; surface-enhanced laser desorption and ionization – time-of-flight (SELDI-TOF) is an analog of MALDI-TOF. The interested reader is referred to [30] for discussion regarding the experimental design that creates the data, and elaboration on the MALDI and SELDI constructs. The resulting data are spectral functions containing the $m/z$ ratio and associated intensity, where the peaks in the spectral plots correspond to proteins (or peptides) present in the sample. These procedures generate large amounts of spectral data and can detect protein differential expression and modification in different treatment groups. Noisy data, however, can lead to a high rate of false positive peak identification. This is particularly an issue when working to establish an unbiased, automated approach to detect protein changes, particularly in low abundance proteins. Nevertheless, various mass spectrometry platforms have been used in experiments describing cancer diagnosis such as in [31–33] where pitfalls and concerns with these platforms in cancer diagnosis are noted in [34, 35].

### 1.3.5
### aCGH Arrays

Array-based Comparative Genomic Hybridization (aCGH) technology is similar to cDNA arrays and is an extension from conventional CGH that is used to identify and quantify DNA copy number changes across the genome in a single experiment [36]. The advantages of aCGH include high-resolution and high-throughput measurement capability allowing for more quantitative analysis of the genomic aberrations.

In BAC aCGH arrays, the probes corresponding to locations on a genome are cloned (grown) in a bacterial culture and then arrayed to a glass slide. BAC aCGH technology can be employed to discover markers in diseases as in [37–40] and for detecting genomic imbalances in cancers as described in [41–51]. In BAC aCGH studies, the markers for cancer are often discovered by comparing the signal at a given chromosome loci between the tumor sample and a control sample. Specifically, researchers often examine the logarithm (base 2) of the ratio of the tumor sample to the control sample ($\log T/C$). Some of the normalization methods for this logarithmic ratio are described in [53]. This normalized ratio will allow researchers to determine the presence of an imbalance in copy number for a given marker between the tumor sample ($T$) and the control sample ($C$).

### 1.3.6
### Preprocessing HT Platforms

In short, preprocessing algorithms are required in nearly all high-throughput experiments (see, for example, [52]). This is due to the fact that high-throughput platforms measure both biological signal and technical signal. Therefore, the goal of preprocessing algorithms is to remove the technical signal. This technical signal can be considered in terms of background correction and normalization to adjust

across experiments. Often these preprocessing techniques are specific to the platform employed (see, for example, [53]). For these reasons, we will not cover all of the preprocessing methods available.

However, to give the reader a feel for preprocessing methods, we discuss quantile normalization, a technique that has been applied and adopted in several different high-throughput platforms [54]. A nice feature of quantile normalization is that it does not require the construction of (non) linear models to describe the experimental system. As each experimental unit (e.g., mouse, patient, cell line, or sample) will be measured via the proposed high-throughput platform, a (genetic) profile for this experimental unit will be obtained. In quantile normalization, we impose the same empirical distribution of the high-throughput intensity for each profile (e.g., the profile for each experimental unit will have the same quartiles, median, etc.). The algorithm proposed in [54] is designed so that all profiles are matched (aligned) with the empirical distribution of the averaged sample profiles.

## 1.4 Analysis of Experiments

After preprocessing the experiment, we ultimately obtain a $N \times M$ summary matrix, $\mathbf{X} = (x_{nm})$, where $x_{nm}$ denotes the normalized measure of probe (gene/protein) $m$ in sample $n$. This data matrix will be used for subsequent statistical analysis. This data matrix can be interpreted as a collection of $M$ explanatory vectors each of length $n$. In our setting, we assume that the researcher is interested in examining which of the $M$ vectors are correlated with the outcome vector of interest $Y$. Since each of the $M$ vectors represents a gene/protein or, generally speaking a "probe" we can consider this analysis as a "probe by probe" analysis where each probe represents a potential biomarker.

For each of the following subsections, we assume that each column in our $X$ matrix corresponds to a biomarker under consideration. Our goal in the following subsections is to assign a $p$-value measuring the correlation between the biomarker and the outcome of interest. The outcome of interest is denoted by $Y$ and contains a value for each sample in the matrix $X$. The outcome of interest can be of several forms, (1) continuous (or nearly continuous) variable, for example, size of tumor, (2) categorical for example, healthy versus disease, or (3) censored continuous variable, for example, survival times, or time to recurrence. In the following subsections, we outline the analysis for each outcome variable setting and provide a cancer-related hypothetical experiment suitable for statistical analysis via the proposed methods.

### 1.4.1 Linear Regression

In a linear regression setting for discovering high-throughput biomarkers, our goal is to determine which biomarkers are significantly correlated with our outcome of interest which, for this section, is assumed to be suitably continuous. Examples of

continuous outcomes in biomarker discovery may include drug level concentrations, white blood cell count, marker staining percentage, and tumor size. The remainder of this section first introduces the simple linear regression model, and later addresses the multivariate regression model designed to assess the correlation between our markers and the continuous outcome of interest.

#### 1.4.1.1 Simple Linear Regression

**Example 1.1**

An experiment is conducted to study the correlation between gene expression and tumor size (a surrogate measure for the extent of disease) in breast cancer patients at the time of diagnosis. To that end, we obtain breast cancer tumor samples from a random cohort of patients recently diagnosed with breast cancer. These tumor samples are processed to obtain mRNA and are interrogated with a gene expression array to obtain the expression level for set of genes. The outcome of interest is the tumor size. We would like to know which genes are significantly associated with the tumor size and which are not.

In a simple linear regression, we consider only a single biomarker, which is considered a predictor or explanatory variable for the outcome or response variable $Y$. In a simple linear regression with $N$ observations, the model is stated as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, 2, 3, \ldots, N \tag{1.1}$$

where $Y_i$ is the outcome for the $i$th sample, $\beta_0, \beta_1$ are (unknown) parameters and $X_i$ is the value of the biomarker (probe) for the $i$th sample. In this model, we assume that the error terms $\varepsilon_i$ are independent with a constant (unknown) variance $\sigma^2$. In a simple linear regression, we can estimate our unknown parameters, $\beta_0, \beta_1$, $\sigma^2$, using least squares estimators. In a least squares estimation, our goal is to determine values for the parameters that minimize our error in the fitted model. For the pairs of observations $(X_i, Y_i)$, we consider the deviation of $Y_i$ from its fitted value from the linear regression by examining the deviation (DEV) defined as

$$\text{DEV}_i = Y_i - (\beta_0 + \beta_1 X_i). \tag{1.2}$$

With the definition of deviation capturing our concept of "error," the goal in the least squares estimation is to minimize the sum of the squared deviations:

$$Q = \sum_{i=1}^{N} (Y_i - (\beta_0 + \beta_1 X_i))^2 = \sum_{i=1}^{N} \text{DEV}_i^2. \tag{1.3}$$

As shown in [55], the following formulas yield the point estimators $b_0$ and $b_1$ for $\beta_0$ and $\beta_1$, respectively, that minimize $Q$,

$$b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}, b_0 = \overline{Y} - b_1 \overline{X}. \tag{1.4}$$

Note that $\overline{X}$ and $\overline{Y}$ are the sample means of the $X_i$ and the $Y_i$ observations, respectively. A commonly used estimator for $\sigma^2$ is given by the mean squared error (MSE):

$$\text{MSE} = \frac{\sum (Y_i - (b_0 + b_1 X_i))^2}{N-2} = \frac{\sum \text{DEV}_i^2}{N-2}. \tag{1.5}$$

In order to measure the significance of the correlation between the predictor and response, we need to make an assumption about the form of the distribution of $\varepsilon_i$. We assume that the error terms $\varepsilon_i$ are independently normally distributed with mean 0 and variance $\sigma^2$ (denoted by $N(0, \sigma^2)$). With this assumption, we have the ability to assess the significance of $\beta_1$, or, in other words, ask the question, "Is $\beta_1$ significantly different from 0?" Specifically, the hypothesis test of interest is stated as

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}. \tag{1.6}$$

In short, hypothesis testing allows researchers to make decisions between two hypotheses, based on observed data. An introduction to statistical hypothesis testing is provided [56, 57] with more advanced treatments in [58, 59].

In order to evaluate our test in (1.6), we need to derive a test statistic and its distribution under the null hypothesis. For our point estimator $b_1$ in (1.4), it can be shown (as in [55]) that, under the null hypothesis, $b_1$ is normally distributed with mean 0 and variance given by

$$\sigma^2(b_1) = \frac{\sigma^2}{\sum (X_i - \overline{X})^2}, \tag{1.7}$$

where $\sigma^2(b_1)$ denotes the variance of $b_1$. As mentioned earlier, $\sigma^2$ is unknown, but MSE is a commonly used estimator for $\sigma^2$. Hence, our estimator for $\sigma^2(b_1)$ in (1.7) can be expressed as

$$s^2(b_1) = \frac{\text{MSE}}{\sum (X_i - \bar{X})^2}, \tag{1.8}$$

where MSE is given in (1.5). Since $b_1$ is normally distributed, we have that the standard statistic $(b_1 - \beta_1)/\sigma(b_1)$ is a standard normal variable. Using our estimator in (1.8) for $\sigma^2(b_1)$, our test statistic $Z$ can be given as

$$Z = \frac{b_1}{s(b_1)}. \tag{1.9}$$

Note that under the null distribution $Z$ follows a $t$ distribution with $N-2$ degrees of freedom (denoted by $Z \sim t_{N-2}$).

Once obtaining the test statistic and its distribution under the null hypothesis, we can obtain the *p*-value: a value indicating the probability of obtaining a test statistic at least as extreme as the observed statistic under the assumption that the null hypothesis is true. See [56, 58, 60] for a more thorough discussion of *p*-values. For our biomarker $X$, using the statistic in (1.9), we can

calculate a (univariate) $p$-value for the test in (1.6) by the following:

$$p\text{-value} = 2K(-|Z|) \tag{1.10}$$

where $K$ denotes the cumulative distribution function (CDF) for the $t$ distribution with $N-2$ degrees of freedom. This $p$-value is univariate in the sense that it refers to the level of significance for a single biomarker. The univariate $p$-value in (1.10) does not address the significance in light of testing $M$ possible biomarkers (see Section 1.5). Nevertheless by using (1.10), we can compute a $M$ length vector consisting of $p$-values for each of the biomarkers under consideration.

#### 1.4.1.2 Multiple Regression

**Example 1.2: Continued from Example 1.1**

To further our analysis of the genes associated with tumor size, we would like to adjust for patient age. That is, the experimenters are interested in which genes are significantly associated with tumor size after adjusting for the patient's age. In this experiment patient age acts as another explanatory variable.

Multiple regression represents an extension of the ideas developed in Section 1.4.1.1. In a multiple regression, we include multiple predictor variables in the model to explain the response variable. This setting is useful to evaluate potential biomarkers in light of other variables, for example patient age or patient race. For example with two predictor variables, for example, two biomarkers, $X_1$ and $X_2$ the first-order multiple regression model is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i. \tag{1.11}$$

The model in (1.11) is first order in the sense that each variable is included in the model, but there is no interaction variable $X_1 \times X_2$ included in the model. Following the methodology in (1.1) and (1.11), we can generalize our regression model for $m$ variables $X_1, X_2, X_3, \ldots, X_m$ as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_m X_{im} + \varepsilon_i. \tag{1.12}$$

Following similar methodology outlined in Section 1.4.1.1, we can formally test the significance of variable $X_j$ using likelihood or least squares methods to estimate our parameters in (1.11) and (1.12), see [55, 61, 62].

### 1.4.2 Logistic Regression ($Y$ Discrete)

**Example 1.3**

Bladder cancer clinicians are interested in proteomic biomarkers associated with the two major subtypes of bladder cancer. This information will further

> improve the ability of clinicians to diagnose and classify bladder cancer patients. Hence, a mass spectrometry experiment is performed to analyze the proteome in a set of bladder cancer tumors from a cohort of the papillary transitional cell carcinoma subtype and a cohort of the nonpapillary transitional cell carcinoma subtype. The experimenters are interested in proteins that are differentially expressed between the two bladder cancer subtypes.

In this setting, we assume that $Y$ is a binary (two categories) random variable. For example, $Y$ may denote healthy or disease subjects. Note, it is outside the scope of this chapter to fully explore regression settings where $Y$ consists of more than two categories. Hence, for the remainder of this section, we code our bivariate outcome variable $Y$ as 0 or 1. Using a logistic regression model, and a single predictor variable $X$ we can model our outcome as follows:

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}, \tag{1.13}$$

where $E(\,)$ represents the expected value function and $\exp()$ represents the exponential function. Similar to the regression models, we can use likelihood methods or least squares methods to obtain $b_0, b_1$ estimators of $\beta_0, \beta_1$. Note unfortunately, closed form solutions for $b_0, b_1$ do not exist and so computer intensive numerical search procedures such as those employed in R (see [63]) and SAS® software are necessary.

Once we have determined estimates for $\beta_0$ and $\beta_1$ as $b_0$ and $b_1$, respectively, our goal is to test the same hypothesis as in (1.6). Unfortunately, our statistic and ultimately, calculating the $p$-value in this setting is not as straightforward as in linear regression. A common test for $\beta_1$ in the logistic regression setting is the likelihood ratio test [56–58]. In short, we compute the partial deviance representing the deviance between the model containing $\beta_1$ and the model where $\beta_1 = 0$. Before defining partial deviance, we define deviance (DEV) for the logistic regression model in (1.13) as

$$\text{DEV} = -2\sum_{i=1}^{N} \left( Y_i \log(\hat{Y}_i + (1 - Y_i)\log(1 - \hat{Y}_i) \right), \tag{1.14}$$

where $\hat{Y}_i$ is the fitted value for sample $i$ in the logistic regression model. The fitted value $\hat{Y}_i$ is obtained by using $b_0$ and $b_1$ in place of $\beta_0$ and $\beta_1$ in (1.13). Thus, we can define the partial deviance (PD) as the difference between the deviance (calculated in (1.14)) for a model containing $\beta_1$ (as in (1.13)) and the deviance for a model where $\beta_1 = 0$. Under the null hypothesis in (1.6), we have (asymptotically) that PD follows a chi-squared distribution with 1 degree of freedom. Thus, we can obtain a $p$-value measuring the significance of $\beta_1$ as

$$p\text{-value} = 1 - G(\text{PD}), \tag{1.15}$$

where $G$ represents the CDF for a chi-squared distribution with 1 degree of freedom. Thus, for a bivariate outcome, we can use (1.15) to obtain a $p$-value for each biomarker under consideration.

#### 1.4.2.1 Multiple Logistic Regression

**Example 1.4: Continued from Example 1.3**

Researchers have determined that cigarette smoking plays a role in bladder cancer. Hence, the researchers would like to know the proteins associated with bladder cancer subtype after adjusting for smoking pack years – a measure quantifying the amount of cigarette smoking for each patient. In this example, smoking pack years acts as an additional explanatory variable.

Similar to the multiple linear regression model, we can generalize our model in (1.13) for $m$ biomarkers as follows:

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m)}. \tag{1.16}$$

Similarly, we can test for the significance of $\beta_j$ using analogs of (1.14) and (1.15) where asymptotically the statistic follows a chi-squared distribution with 1 degree of freedom under the null hypothesis.

The simple and multiple logistic regression models can be extended to situations where the outcome variable has more than two groups or levels (polychotomous). Treatment for these situations can be found in [64–66].

### 1.4.3 Survival Modeling

Survival analysis is commonly performed in biomarker testing within oncology research. Prognostic or predictive biomarkers, by design, are meant to explain the patients overall cancer outcome or the effect of a therapeutic intervention. Commonly the outcome or effect studied in these situations is the survival time, time to recurrence, or time to disease progression. In all three situations, this variable is considered right censored where the event is observed only if it occurs prior to some prespecified time. For example, patients may be followed with events recorded for up to five years. The amount of follow up time should be a balance based on the number of expected events and the resources required to follow the patients over that time frame.

The following texts provide thorough treatments for survival analysis [67–69]. Our goal in this section will be to introduce the Kaplan–Meier estimator as a method to, ultimately, test and obtain a $p$-values representing the significance of a biomarker in assessing time to event data.

#### 1.4.3.1 Kaplan–Meier Analysis

**Example 1.5**

Researchers are interested in what DNA copy number changes are associated with shorter survival in ovarian cancer patients. To study this question,

researchers analyze a set of ovarian cancer tumors using aCGH technology. Each patient in this study has been followed for at least 5 years with their survival times documented. The goal is to determine copy number imbalances that are significantly associated with shorter survival. To this end, in each sample, the aCGH-derived data for each probe or location on the genome is dichotomized into either normal copy number or copy number imbalance. This dichotomized data is examined to determine what regions are significantly correlated with patient survival.

In a survival analysis setting, we define $S(t)$ to be the probability that an experimental unit from a given population will have a lifetime exceeding $t$. That is, for a random variable $T$ representing the lifetime of the experimental unit, we have

$$S(t) = \Pr(T > t), \tag{1.17}$$

where $Pr(\ )$ denotes probability. Related to the survival function, we define the hazard function, denoted by $h(t)$, as the event rate at time $t$, conditional on survival until time $t$ or later. Mathematically, when $T$ is a continuous random variable, we have

$$h(t) = -\mathrm{d}\log S(t)/\mathrm{d}t. \tag{1.18}$$

For a sample from this population of size $N$ let the observed times until an event of $N$ sample members be given as follows:

$$t_1 \leq t_2 \leq t_3 \cdots \leq t_N. \tag{1.19}$$

Corresponding to each $t_i$ is $n_i$ – the number of patients at risk just prior to time $t_i$ and $d_i$ – the number of events at time $t_i$. With this notation we define the Kaplan–Meier estimator designed to estimate the survival function $S(t)$ for a random variable $T$as

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}, \tag{1.20}$$

where $\Pi$ represents the product operator [70]. In a simple bivariate biomarker setting, we can use (1.20) to estimate the survival function for each population of the biomarker under consideration. For the remainder of this section, we assume that our biomarkers under consideration have two levels (e.g., low/high or expressed/unexpressed) or two groups.

To assess the correlation of our biomarker with survival, we can use a logrank test to compare the survival distribution of two sample populations corresponding to the two levels of our biomarker. This test was first proposed by Mantel [71].

The logrank test statistic is designed to compare estimates of the hazard functions of the two groups at each observed event time. The statistic computes the observed and expected number of events in one of the groups at each observed event time and then adds these to obtain an overall summary across all time points

where there is an event. In this way we test the following hypotheses,

$$\begin{aligned} H_0 &: S_1(t) = S_2(t) \\ H_1 &: S_1(t) \neq S_2(t) \end{aligned}, \tag{1.21}$$

where $S_1(t)$ and $S_2(t)$ denote the survival functions for group 1 and group 2, respectively. Let $j = 1, \ldots, J$ be the distinct times of observed events in the either group. For each time $j$, let $N_{1j}$ and $N_{2j}$ be the number of subjects at risk (have not yet had an event or been censored) at the start of period $j$ in the groups, respectively. Let $N_j = N_{1j} + N_{2j}$. Let $O_{1j}$ and $O_{2j}$ be the observed number of events in the groups respectively at time $j$, and define $O_j = O_{1j} + O_{2j}$.

Given that $O_j$ events happened across both groups at time $j$, under the null hypothesis (of the two groups having identical survival and hazard functions) $O_{1j}$ has the hypergeometric distribution with parameters $N_j$, $N_{1j}$, and $O_j$. This distribution has expected value $E_{1j} = O_j \frac{N_{1j}}{N_j}$ and variance $V_j = \frac{O_j(N_{1j}/N_j)(1 - N_{1j}/N_j)(N_j - O_j)}{N_j - 1}$. The logrank statistic compares each $O_{1j}$ to its expectation $E_{1j}$ under the null hypothesis and is defined as

$$Z = \frac{\sum_{j=1}^{J}(O_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^{J} V_j}}. \tag{1.22}$$

Under the null hypothesis in (1.21), the logrank statistic in (1.22) follows (asymptotically) a standard Normal distribution. Thus we can obtain a (two-sided) $p$-value as follows:

$$p\,\text{value} = 2\Phi(-|Z|), \tag{1.23}$$

where $\Phi$ represents the CDF for a standard normal distribution. Thus, using (1.23), we can obtain a $p$-value representing the level of significance for each biomarker with survival.

As an aside, when our biomarker is not bivariate, there are alternative methods such as Cox proportional hazards modeling that can be employed to assess the correlation between the biomarker and time to event (survival) data. The interested reader is encouraged to see [67, 72].

## 1.5 Multiple Testing Type I Errors

The experiments performed in the high-throughput platforms in Sections 1.3.1–1.3.5 often have a goal of narrowing down the genome or proteome to a subset of interesting or significant genes/loci/proteins/regions of the genome. In this sense, the scientists are performing a data reduction where the goal is to choose a subset from the high-throughput scope that are related or associated with the outcome. This relationship with the outcome is assessed using a hypothesis test and can be summarized statistically with the $p$-value (see Section 1.4). When performing the

**Table 1.1** A summary of results from analyzing multiple hypothesis tests[a)].

| | $H_0$ Retained | $H_0$ Rejected | Total |
|---|---|---|---|
| $H_0$ True | $U$ | $V$ | $M_0$ |
| $H_0$ False | $T$ | $Q$ | $M_1$ |
| | $M - R$ | $R$ | M |

a) We consider $M_0$ and $M_1$ as fixed (unknown) parameters representing the number of true nulls and the number of true alternatives, respectively. The random variables $U$ and $Q$ represent the number of the correct decisions, while the random variables $T$ and $V$ represent the number of incorrect decisions.

tests there are four possible scenarios that can occur (see Table 1.1). Two scenarios represent correct decisions, while the other two are incorrect decisions or "errors": (1) when one rejects the null hypothesis when it is actually true, and (2) when one does not reject the null hypothesis when it is actually false.

The probability associated with the first scenario is referred to as Type I error and the second scenario is termed as Type II error. In a multiple testing scenario, we would like to control a function of the number of Type I errors committed (see Table 1.2). A related quantity is statistical power. For a single test, power is defined as the probability of correctly rejecting a true alternative hypothesis. With multiple testing, this is commonly generalized by examining the average power. In other words, let $M_1 = M - M_0$ be the number of true alternatives. With this notation, researchers commonly use $E(Q/M_1)$, representing the average power, to compare the error rates in Table 1.2.

Using the notation in Table 1.1, our goal in the remainder of this section is to control the Type I error rate. In Table 1.1, $V$ denotes the number of false positives and Table 1.2 represents Type I errors that can be controlled. Note that all of the quantities in Table 1.1 are a function of $V$. We focus on the generalized family wise error rate ($k$-FWER) as this is a common error rate to control in high-throughput experiments [73]. Hence, the remainder of this chapter will focus on methods that control $k$-FWER.

**Table 1.2** Table summarizing the Type I errors using random variables defined in Table 1.1 (see similar tables in [74] and Table 15.1 in [75])[a)].

| Abbr. | Name | Quantity |
|---|---|---|
| FWER | Family wise error rate | $Pr(V \geq 1)$ |
| $k$-FWER | Generalized family wise error rate | $Pr(V \geq k)$ |
| FDR | False discovery rate | $E[V/R]$ |
| $k$-FDR | Generalized false discovery rate | $E\left[\frac{VI(V \geq k)}{\max(R,1)}\right]$ |
| PCER | Per comparison error rate | $E[V]/M$ |
| TPPFP | Tail probabilities for the proportion of false positives | $Pr(V/R > q)$ |

a) Note $I(\,)$ in the equation for $k$-FDR denotes the indicator function and $max(\,)$ denotes the maximum operator. Further note that q in TPPFP should be determined prior to testing.

### 1.5.1 FWER, *k*-FWER Methods

The *k*-FWER error rate is a generalized version of the family wise error rate (FWER). Control of FWER refers to controlling the probability of committing one or more false discoveries. If we let $V$ denote the number of false positives from $M$ hypothesis tests (biomarkers), then notationally, (according to [76]) control of FWER at the level of $\alpha$ can be expressed as,

$$Pr(V \geq 1) \leq \alpha \tag{1.24}$$

or equivalently,

$$Pr(V = 0) \geq 1 - \alpha \tag{1.25}$$

Note that $\alpha$ is usually chosen to be small, for example, 0.05. Often (1.24) is abbreviated as FWER $\leq \alpha$. In *k*-FWER the equation becomes

$$\Pr(V \geq k) \leq \alpha \tag{1.26}$$

where $k$ and $\alpha$ are usually determined prior to the analysis. Similar to FWER, control of *k*-FWER at level $\alpha$ can be expressed as *k*-FWER $\leq \alpha$. Practically speaking, controlling *k*-FWER allows researchers to claim that with high probability there are no more than $k$ false positives in their list of significant biomarkers. Naturally the choice of $k$ is critical when controlling *k*-FWER. The choice should be made prior to the analysis and it should be based on the resources available to validate the biomarkers in the significance list. If there are relatively limited resources available to validate the biomarkers, then $k$ could be rather small (conservative), otherwise $k$ should be larger (liberal). The following subsections discuss the variety of methods available to control FWER and *k*-FWER.

#### 1.5.1.1 Adjusted Bonferroni Method

The adjusted Bonferroni method to control *k*-FWER is a generalized version of the Bonferroni correction designed to control FWER [76]. The Bonferroni correction is designed to control the FWER at level $\alpha$ by doing each individual test at significance level $\alpha/M$ where $M$ is the number of tests. The adjustment given in [76] to control *k*-FWER at $\alpha$ is done by performing each test at level $k\alpha/M$. That is, a biomarker and the corresponding hypothesis test is considered significantly associated (reject the null) if the *p*-value is less than $k\alpha/M$. Under this scheme, the probability against $k$ or more false positives is no larger than $\alpha$, that is, *k*-FWER $\leq \alpha$. The proof is supplied in [76] and is a generalization of the proof for the original Bonferroni method designed to control FWER.

#### 1.5.1.2 Holm Procedure

A method to control *k*-FWER using the Holm procedure is given in [76]. This method is an adjustment to the Holm method designed to control the FWER [77]. The Holm method is considered a "step-down" procedure [58] which, essentially, means the *p*-value cut point for significance is based on considering the ranked

vector of $p$-values starting with the most significant $p$-values. The following procedure describes the Holm method to control FWER at level $\alpha$ for $M$ tests. Let

$$\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_M, \tag{1.27}$$

be constants defined by $\alpha_i = \alpha/(M - i + 1)$. For each of the $M$ biomarkers under consideration, we denote their corresponding null hypotheses by $H_1, H_2, \ldots, H_M$. We let the ordered $p$-values (smallest to largest) be denoted by $p_{(1)} \leq \cdots \leq p_{(M)}$ corresponding to the ordered null hypotheses, $H_{(1)}, \ldots, H_{(M)}$. If $p_{(1)} > \alpha_1$, then reject no null hypothesis. Otherwise, if

$$p_{(1)} \leq \alpha_1, \ldots, p_{(r)} \leq \alpha_r, \tag{1.28}$$

then reject hypothesis $H_{(1)}, \ldots, H_{(r)}$ where the largest $r$ satisfying (1.28) is used. With this framework to control FWER at level $\alpha$, the Holm method to control $k$-FWER at level $\alpha$ as stated in [76] is done by redefining $\alpha_i$ as

$$\alpha_i = \begin{cases} k\alpha/M, & i < k \\ k\alpha/(M + k - i), & i \geq k \end{cases}. \tag{1.29}$$

Note, as stated in [76], when the $p$-values are independent (with $1 \leq k \leq 1/\alpha$), a more powerful version of the Holm method is obtained by redefining $a_i$ in (1.29) as

$$\alpha_i = \begin{cases} \left(\alpha \prod_{j=1}^{k} \frac{j}{M - k + j}\right)^{1/k}, & i < k \\ \left(\alpha \prod_{j=1}^{k} \frac{j}{M - i + j}\right)^{1/k}, & i \geq k \end{cases}. \tag{1.30}$$

#### 1.5.1.3 Generalized Hochberg Procedure

The generalized Hochberg procedure is originally presented in [78] as a method to control FWER. It is expanded in [76] and is shown to be closely related to the generalized Holm procedure. In fact it is stated in [76] that "Hochberg's procedure is the step-up version of Holm's step-down procedure." Recall that Holm's procedure is considered a step-down procedure because it starts by considering the most significant $p$-values and once a $p$-value is larger than a threshold the process stops and all smaller $p$-values (hypotheses) are considered significant (reject the null). In an analogous way, a step-up procedure starts with the least significant $p$-values and once a $p$-value is smaller than a threshold, the process stops and all smaller $p$-values (null hypotheses) are rejected. With this notion is mind and assuming independent $p$-values, we can state Hochberg's procedure as given in [76] as follows: if $p_{(M)} \leq \alpha_M$, then reject all the null hypotheses, that is, accept all alternative hypotheses. Otherwise, reject null hypothesis $H_{(1)}, \ldots, H_{(r)}$ where $r$ is the largest integer satisfying $p_{(r)} \leq \alpha_r$ with the $\alpha_i$ defined in (1.30).

#### 1.5.1.4 Generalized Šidàk Procedure

A thorough treatment of the generalized Šidàk method is presented in [73]. Note, that the notation and technical details required for their presentation of this method are outside the scope of this chapter. However, using reasonable assumptions, we can simplify the generalized Šidàk method presented in [73]. We consider

using a beta-uniform model (BUM) as the distribution of the $p$-values [79]. The BUM model represents a mixture model for generating $p$-values. With a BUM model, we assume that the $p$-values are independently distributed according to a mixture model where the $p$-value observations are either from a uniform distribution (true null hypotheses), or a beta distribution (true alternative hypotheses). We expect that true alternatives will yield, on average, small $p$-values and hence the Beta distribution with a mean near zero is a reasonable model for the alternative $p$-values. With this setting, the generalized Šidàk procedure works by rejecting all hypotheses with a $p$-value less than $p_{\text{cut}}$ where $p_{\text{cut}}$ is such that

$$F(k-1|M, p_{\text{cut}}) = 1 - \alpha \tag{1.31}$$

where $F$ is the CDF of a Binomial random variable of size $M$ and probability of success $p_{\text{cut}}$. Notationally, we have $W \sim \text{bin}(M, p_{\text{cut}})$. In short, the Šidàk method can be interpreted in light of mixture models that interpret the $M$ hypotheses as a mixture of alternative hypotheses (discoveries) and null hypotheses. For a collection of $M$ tests, the Šidàk method is designed to select a success probability parameter in a binomial distribution where a success means the test follows the alternative hypothesis, while a failure means the test follows the null hypothesis. Under this assumption, the proof that the Šidàk method controls $k$-FWER can be found in [73, 80].

#### 1.5.1.5 minP and maxT procedures

Recently two data driven methods, minP and maxT, have been proposed to control $k$-FWER [81–83]. The methods require a bootstrap step or permutation step to estimate the null distribution [84]. This is in contrast to the adjusted Bonferroni method, the Holm method, the Hochberg method, and the generalized Šidàk method that only require the $M$-length vector of $p$-values. Due to the complexity of these algorithms, we feel these methods are outside the scope of this chapter.

## 1.6 Discussion

It cannot be understated that there are numerous assumptions that must be verified for the regression and survival analysis methods to be valid. The field of regression/survival diagnostics refers to the general class of techniques for detecting whether the assumptions are valid with these methods. We encourage the reader to explore the references in each of the sections for more thorough coverage of the assumptions and diagnostic techniques for each method.

Accurate Type I error control in high-throughput experiments is crucial in order to avoid costly downstream experiments attempting to validate false positives. Further, it is important to understand the assumptions and implications involved in choosing a Type I error to control (see Table 1.2). For example, in our work of pathway-based microarray analysis [85], we showed that $k$-FWER methods are more robust than the other error rate control methods.

The dependence structure in our tests is a key aspect of these $k$-FWER methods. For example, independent test statistics ($p$-values) are required for the presented versions of the Šidàk, Holm, and Hochberg methods. Due to the similarity of probes and their genomic inter-relatedness, this assumption is most likely unreasonable in high-throughput experiments. Recently there has been several works that discuss the dependence structure assumptions for these methods [86–89]. Future work for these $k$-FWER methods will continue to explore the robustness of these methods to violations in the dependence structure.

In this chapter, we have highlighted several methods designed to control $k$-FWER, where $k$-FWER is designed to control a probability statement about the distribution of $V$, the number of false positives. We can generalize our treatment of Type I error rates by considering Type I error, generically, as a functional of a Type I error (e.g., $V$ or $V/R$). That is, the Type I error can be characterized in terms of a general functional $\theta(F)$, where $F$ represents the distribution corresponding to the (error) random variable of interest, for example, $F$ is the distribution of $V$, the number of false positives. Future work in this area explores the possibility of unifying the assumptions required for generic Type I control and the possibility of formulating general expressions for power.

In addition to the probe-by-probe testing we discussed in this chapter, there are alternative methods to analyze this data including principal components analysis (PCA) and gene set enrichment analysis (GSEA). Within principal components, the analysis can be supervised: outcome taken into consideration, or unsupervised: outcome information ignored. Both approaches have their merit and can be used in prediction and classification [90, 91]. Meanwhile, GSEA methods are designed to assess the significance of a cohort or group of probes/genes. These methods test a hypothesis of significance for each cohort or group of genes and thus a $p$-value can be assigned to each group of genes rather than an individual gene. Commonly used algorithms include the original GSEA algorithm [92] and more recently the gene set analysis (GSA) algorithm [93]. Most commonly these methods used predetermined gene sets compiled from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [94] database or the human protein reference database (HPRD) [95]. These databases include pathways for metabolism, genetic information processing, environmental information processing, cellular processes, human diseases, and drug development.

## 1.7 Perspective

In addition to common clinical–pathological variables used in cancer diagnosis, with the success of the human genome project researchers are using molecular variables to aid in the diagnosis and subtype classification of cancer. Genetic markers such as estrogen receptor gene and breast cancer susceptibility gene mutations have been commonly used for years. However, researchers continue to search for novel putative biomarkers derived from interrogating the entire genome or

proteome. These high-throughput experiments to find novel biomarkers yield high-dimensional datasets. In general with these high-dimensional datasets, the task of the statistician is to reduce the dimension of the data. This dimension reduction, sometimes called feature extraction, should be performed in a way that removes noise while retaining biological signal. In a biological high-throughput experiment, this reduction can be performed by selecting a subset of biological probes that are significantly associated with the outcome of interest. Statistical significance of association is assessed in light of controlling a Type I error designed to control the number of false positives when simultaneously testing all of the biological probes with the outcome of interest. These biological probes can be genes, genetic regions, proteins, peptides, or microRNAs, while the outcome of interest may be continuous, discrete, or censored, and the Type I error might be controlling the rate of false positives or the probability of committing a certain number of false positives. In this chapter, we describe the high-throughput platforms that generate this type of high-dimensional data and the statistical methods employed to assess overall statistical significance with the various outcomes of interest. These statistical methods can be used in large-dimensional datasets obtained from high-throughput platforms designed to discover potentially novel biomarkers in the diagnosis of cancer.

Future work in these areas will include further development and validation techniques for the putative markers obtained from these types of experiments. Statisticians continue to advance statistical methods to control Type I errors and are keenly interested in designing methods to control Type I error in light of correlation. The strategy for choosing a Type I error method/scheme based on the type of data under consideration as well as on the validation methods (and their error rates) that will be used for the markers is an active area of ongoing research. Also, recently scientists have started to explore combining datasets from multiple high-throughput experiments. This field of integrative analysis will require a new set of statistical methods to integrate DNA, RNA, and proteomic data all gathered on the same set of patients. These integrative approaches hold promise for researchers looking to gain insights on complex interactions involving multiple biological systems.

## References


1 Rajan, S., Djambazian, H., Dang, H., Sladek, R., and Hudson, T. (2011) The living microarray: a high-throughput platform for measuring transcription dynamics in single cells. *BMC Genomics*, **12** (1), 115.

2 National Cancer Institute (2011) The cancer genome atlas. http://cancergenome.nih.gov/newsevents/forthemedia/backgrounder.

3 Perez-Diez, A., Morgun, A., and Shulzhenko, N. (2007) Microarrays for cancer diagnosis and classification. *Microarray Technol. Cancer Gene Prof.*, **593**, 74–85.

4 Pepe, M. (2005) Evaluating technologies for classification and prediction in medicine. *Stat. Med.*, **24**, 3687–3696.

5 Pepe, M. and Longton, G. (2005) Standardizing diagnostic markers to evaluate and compare their performance. *Epidemiology*, **16** (5), 598–603.

6 Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to

multiple testing. *J. Roy. Stat. Soc. B Met.*, **57** (1), 289–300.

**7** Brown, P. and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nat. Genet.*, **21** (Suppl 1), 33–37.

**8** Duggan, D., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. (1999) Expression profiling using cDNA microarrays. *Nat. Genet.*, **21** (Suppl 1), 10–14.

**9** Lockhart, D. and Winzeler, E. (2000) Genomics, gene expression and DNA arrays. *Nat. London*, **405**, 827–836.

**10** Dudoit, S., Fridlyand, J., and Speed, T. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97** (457), 77–87.

**11** Yang, Y., Dudoit, S., Luu, P., Lin, D., Peng, V., Ngai, J., and Speed, T. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30** (4), e15–e15.

**12** Zhu, Q., Miecznikowski, J., and Halfon, M. (2010) Preferred analysis methods for Affymetrix GeneChips: II. an expanded, balanced, wholly-defined spike-in dataset. *BMC Bioinform.*, **11** (1), 285.

**13** Glas, A., Floore, A., Delahaye, L., Witteveen, A., Pover, R., Bakx, N., Lahti-Domenici, J., Bruinsma, T., Warmoes, M., Bernards, R. *et al.* (2006) Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics*, **7** (1), 278.

**14** Gordon, G., Jensen, R., Hsiao, L., Gullans, S., Blumenstock, J., Ramaswamy, S., Richards, W., Sugarbaker, D., and Bueno, R. (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res.*, **62** (17), 4963.

**15** Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci.*, **98** (26), 15149.

**16** Statnikov, A., Aliferis, C., Tsamardinos, I., Hardin, D., and Levy, S. (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, **21** (5), 631–643.

**17** Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci.*, **99** (10), 6567.

**18** Barrier, A., Boelle, P., Roser, F., Gregg, J., Tse, C., Brault, D., Lacaine, F., Houry, S., Huguier, M., Franc, B. *et al.* (2006) Stage ii colon cancer prognosis prediction by tumor gene expression profiling. *J. Clin. Oncol.*, **24** (29), 4685–4691.

**19** Michiels, S., Koscielny, S., and Hill, C. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365** (9458), 488–492.

**20** Miecznikowski, J., Wang, D., Liu, S., Sucheston, L., and Gold, D. (2010) Comparative survival analysis of breast cancer microarray studies identifies important prognostic genetic pathways. *BMC Cancer*, **10** (1), 573.

**21** Sotiriou, C., Neo, S., McShane, L., Korn, E., Long, P., Jazaeri, A., Martiat, P., Fox, S., Harris, A., and Liu, E. (2003) Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. USA*, **100** (18), 10393.

**22** Van't Veer, L., Dai, H., van de Vijver, M., He, Y., Hart, A., Mao, M., Peterse, H., Van der Kooy, K., Marton, M., Witteveen, A. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415** (6871), 530–536.

**23** Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

**24** Levin, J., Berger, M., Adiconis, X., Rogov, P., Melnikov, A., Fennell, T., Nusbaum, C., Garraway, L., and Gnirke, A. (2009) Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol.*, **10** (10), R115.

**25** Pflueger, D., Rickman, D., Sboner, A., Perner, S., LaFargue, C., Svensson, M., Moss, B., Kitabayashi, N., Pan, Y., De La Taille, A. *et al.* (2009) N-myc downstream

regulated gene 1 (NDRG1) is fused to ERG in prostate cancer. *Neoplasia*, **11** (8), 804.
**26** Oshlack, A., Robinson, M., and Young, M. (2010) From RNA-seq reads to differential expression results. *Genome Biol.*, **11**, 220.
**27** Portela, A. and Esteller, M. (2010) Epigenetic modifications and human disease. *Nat. Biotechnol.*, **28**, 1057–1068.
**28** Laird, P. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
**29** Siegmund, K. (2011) Statistical approaches for the analysis of DNA methylation microarray data. *Hum. Genet.*, **129**, 585–595.
**30** Levner, I. (2005) Feature selection and nearest centroid classification for protein mass spectrometry. *BMC Bioinform.*, **6**.
**31** Kolch, W., Neusüß, C., Pelzing, M., and Mischak, H. (2005) Capillary electrophoresis.mass spectrometry as a powerful tool in clinical diagnosis and biomarker discovery. *Mass Spectrom. Rev.*, **24** (6), 959–977.
**32** Koopmann, J., Zhang, Z., White, N., Rosenzweig, J., Fedarko, N., Jagannath, S., Canto, M., Yeo, C., Chan, D., and Goggins, M. (2004) Serum diagnosis of pancreatic adenocarcinoma using surface-enhanced laser desorption and ionization mass spectrometry. *Clin. Cancer Res.*, **10** (3), 860–868.
**33** Paweletz, C., Trock, B., Pennanen, M., Tsangaris, T., Magnant, C., Liotta, L., and Petricoin, E. III (2001) Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer. *Dis. Markers*, **17** (4), 301.
**34** Diamandis, E. (2004) Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. *J. Natl. Cancer Inst.*, **96** (5), 353–356.
**35** Diamandis, E. (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool. *Mol. Cell. Proteomics*, **3** (4), 367–378.
**36** Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A. N., Pinkel, D., and Albertson, D.G. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.*, **29** (3), 263–264.
**37** Lugtenberg, D., de Brouwer, A., Kleefstra, T., Oudakker, A., Frints, S., Schrander-Stumpel, C., Fryns, J., Jensen, L., Chelly, J., Moraine, C. *et al.* (2006) Chromosomal copy number changes in patients with non-syndromic X-linked mental retardation detected by array CGH. *J. Med. Genet.*, **43** (4), 362.
**38** Miyake, N., Shimokawa, O., Harada, N., Sosonkina, N., Okubo, A., Kawara, H., Okamoto, N., Kurosawa, K., Kawame, H., Iwakoshi, M. *et al.* (2006) BAC array CGH reveals genomic aberrations in idiopathic mental retardation. *Am. J. Med. Genet. Part A*, **140** (3), 205–211.
**39** Stankiewicz, P. and Beaudet, A. (2007) Use of array CGH in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation. *Curr. Opin. Genet. Dev.*, **17** (3), 182–192.
**40** Ullmann, R., Turner, G., Kirchhoff, M., Chen, W., Tonge, B., Rosenberg, C., Field, M., Vianna-Morgante, A., Christie, L., Krepischi-Santos, A. *et al.* (2007) Array CGH identifies reciprocal 16p13.1 duplications and deletions that predispose to autism and/or mental retardation. *Hum. Mutat.*, **28** (7), 674–682.
**41** Albertson, D. (2003) Profiling breast cancer by array CGH. *Breast Cancer Res. Treat.*, **78** (3), 289–298.
**42** Albertson, D., Ylstra, B., Segraves, R., Collins, C., Dairkee, S., Kowbel, D., Kuo, W., Gray, J., and Pinkel, D. (2000) Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nat. Genet.*, **25** (2), 144–146.
**43** Albertson, D.G., Collins, C., McCormick, F., and Gray, J.W. (2003) Chromosome aberrations in solid tumors. *Nat. Genet.*, **34** (4), 369–376.
**44** Garnis, C., Coe, B., Zhang, L., Rosin, M., and Lam, W. (2003) Overexpression of LRP12, a gene contained within an 8q22 amplicon identified by high-resolution

array CGH analysis of oral squamous cell carcinomas. *Oncogene*, **23** (14), 2582–2586.

**45** Hackett, C.S., Hodgson, J.G., Law, M.E., Fridlyand, J., Osoegawa, K., de Jong, P.J., Nowak, N.J., Pinkel, D., Albertson, D.G., Jain, A., Jenkins, R., Gray, J.W., and Weiss, W.A. (2003) Genome-wide array CGH analysis of murine neuroblastoma reveals distinct genomic aberrations which parallel those in human tumors. *Cancer Res.*, **63** (17), 5266–5273.

**46** Hodgson, G., Hager, J., Volik, S., Hariono, S., Wernick, M., Moore, D., Albertson, D., Pinkel, D., Collins, C., Hanahan, D. *et al.* (2001) Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nat. Genet.*, **29** (4), 459–464.

**47** Idbaih, A., Marie, Y., Lucchesi, C., Pierron, G., Manié, E., Raynal, V., Mosseri, V., Hoang-Xuan, K., Kujas, M., Brito, I. *et al.* (2008) BAC array CGH distinguishes mutually exclusive alterations that define clinicogenetic subtypes of gliomas. *Int. J. Cancer*, **122** (8), 1778–1786.

**48** Pinkel, D. and Albertson, D. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, **37**, S11–S17

**49** Pollack, J., Sorlie, T., Perou, C., Rees, C., Jeffrey, S., Lonning, P., Tibshirani, R., Botstein, D., Borresen-Dale, A., and Brown, P. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. USA*, **99**, 12963–12968

**50** Rossi, M., Conroy, J., McQuaid, D., Nowak, N., Rutka, J., and Cowell, J. (2006) Array CGH analysis of pediatric medulloblastomas. *Genes Chromosomes Cancer*, **45** (3), 290–303.

**51** Veltman, J.A., Fridlyand, J., Pejavar, S., Olshen, A.B., Korkola, J.E., DeVries, S., Carroll, P., Kuo, W.-L., Pinkel, D., Albertson, D., Cordon-Cardo, C., Jain, A. N., and andWaldman, F.M. (2003) Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors. *Cancer Res.*, **63** (11), 2872–2880.

**52** Leek, J., Scharpf, R., Bravo, H., Simcha, D., Langmead, B., Johnson, W., Geman, D., Baggerly, K., and Irizarry, R. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11** (10), 733–739.

**53** Miecznikowski, J., Gaile, D., Liu, S., Shepherd, L., and Nowak, N. (2011) A new normalizing algorithm for BAC CGH arrays with quality control metrics. *J. Biomed. Biotechnol.*, **2011**.

**54** Bolstad, B., Irizarry, R., Åstrand, M., and Speed, T. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19** (2), 185–193.

**55** Neter, J., Wasserman, W., and Kutner, M. (1989) Applied Linear Regression Models, Richard D. Irwin, Homewood, IL.

**56** Casella, G. and Berger, R. (2001) Statistical inference.

**57** Wasserman, L. (2004) *All of Statistics: A Concise Course in Statistical Inference*, Springer, Berlin

**58** Lehmann, E. (1997) *Testing Statistical Hypotheses*, Springer, Berlin

**59** Schervish, M.J. (1995) *Theory of Statistics*, Springer, Berlin

**60** Schervish, M.J. (1996) *P* values: what they are and what they are not. *Am. Stat.*, **50**, 203–206.

**61** Draper, N. and Smith, H. (1998) Applied regression analysis (Wiley series in probability and statistics).

**62** Rawlings, J., Pantula, S., and Dickey, D. (1998) *Applied Regression Analysis: A Research Tool*, Springer, Berlin

**63** R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

**64** Hosmer, D. and Lemeshow, S. (2000) *Applied Logistic Regression*, vol. **354**, Wiley-Interscience, New York

**65** Lesaffre, E. and Albert, A. (1989) Multiple-group logistic regression diagnostics. *Appl. Stat.*, **38**, 425–440.

**66** Marshall, R. and Chisholm, E. (1985) Hypothesis testing in the polychotomous logistic model with an application to

detecting gastrointestinal cancer. *Stat. Med.*, **4** (3), 337–344.

**67** Klein, J. and Moeschberger, M. (2003) *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, Berlin

**68** Lee, E. and Wang, J. (2003) *Statistical Methods for Survival Data Analysis*, vol. **364**, Wiley-Interscience, New York

**69** Prentice, R. and Kalbfleisch, J. (1980) *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, New York.

**70** Kaplan, E. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, **53**, 457–481.

**71** Mantel, N. *et al*. (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemoth. Rep. 1*, **50** (3), 163.

**72** Hosmer, D. and Lemeshow, S. (1999) *Applied Survival Analysis: Regression Modeling of Time to Event data*, Wiley Online Library, New York.

**73** Guo, W. and Romano, J. (2007) A generalized Sidak–Holm procedure and control of generalized error rates under independence. *Stat. Appl. Genet. Mol. Biol.*, **6** (1), Article 3.

**74** Nichols, T. and Hayasaka, S. (2003) Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat. Method Med. Res.*, **12** (5), 419–446.

**75** Gentleman, R., Carey, V., Huber, W., Dudoit, S., and Irizarry, R. (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer, Berlin

**76** Lehmann, E. and Romano, J. (2005) Generalizations of the familywise error rate. *Ann. Stat.*, **33**, 1138–1154.

**77** Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.

**78** Hochberg, Y. (1988) A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, **75** (4), 800.

**79** Pounds, S. and Morris, S. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of *p*-values. *Bioinformatics*, **19** (10), 1236–1242.

**80** Miecznikowski, J., Gold, D., Shepherd, L., and Liu, S. (2011) Deriving and comparing the distribution for the number of false positives in single step methods to control k-FWER. *Stat. Probabil. Lett.* **81** (11), 1695–1705.

**81** Dudoit, S., van der Laan, M., and Pollard, K. (2004) Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Stat. Appl. Genet. Mol. Biol.*, **3** (1), 1040.

**82** van der Laan, M., Dudoit, S., and Pollard, K. (2004) Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Stat. Appl. Genet. Mol. Biol.*, **3** (1), 1041.

**83** Westfall, P. and Young, S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*, Wiley-Interscience, New York.

**84** Efron, B. and Tibshirani, R. (1997) *An Introduction to the Bootstrap*, Chapman & Hall, New York.

**85** Gold, D., Miecznikowski, J., and Liu, S. (2009) Error control variability in pathway-based microarray analysis. *Bioinformatics*, **25** (17), 2216–2221.

**86** Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.

**87** Bhattacharjee, M., Dhar, S., and Subramanian, S. (2011) *Recent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics*, vol. **4**, World Scientific Publishing, Singapore.

**88** Sarkar, S. (2008) Generalizing Simes' test and Hochberg's step-up procedure. *Ann. Stat.*, **36** (1), 337–363.

**89** Sarkar, S., Guo, W., and Finner, H. (2011) On adaptive procedures controlling the familywise error rate. *J. Stat. Plan. Infer.*, **142**, 65–78.

**90** Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006) Prediction by supervised principal components. *J. Am. Stat. Assoc.*, **101** (473), 119–137.

**91** Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D., and Brown, P. (2000)

Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, **1** (2), 0003–1.

**92** Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.*, **102** (43), 15545–15550.

**93** Efron, B. and Tibshirani, R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1** (1), 107–129.

**94** Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28** (1), 27–30.

**95** Mishra, G., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T. *et al.* (2006) Human protein reference database – 2006 update. *Nucleic Acids Res.*, **34** (Database Issue), D411.