Front. Biol. 2010, 5(4): 354–360
DOI 10.1007/s11515-010-0053-2

REVIEW

Statistical considerations for high throughput screening data

Xian-Jin XIE

Division of Biostatistics, Department of Clinical Sciences & Simmons Comprehensive Cancer Center, The University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA

Abstract High throughput screening (HTS) is a widely used effective approach in genome-wide association and large scale protein expression studies, drug discovery, and biomedical imaging research. How to accurately identify candidate 'targets' or biologically meaningful features with a high degree of confidence has led to extensive statistical research in an effort to minimize both false-positive and false-negative rates. A large body of literature on this topic with in-depth statistical contents is available. We examine currently available statistical methods on HTS and aim to summarize some selected methods into a concise, easy-to-follow introduction for experimental biologists.

Keywords high throughput screen, false-positive rate, false-negative rate, target discovery, predictive modeling

1 Introduction

Emerging new biotechnologies continue to push back the frontiers of biological sciences, presenting vast opportunities as well as major challenges to biostatisticians and bioinformaticians in proper analysis of the complex data. Though often, such advanced analyses are performed by trained biostatisticians, it is important for biologists to understand the basic biostatistical concepts underlying the statistical and computational methods utilized for their data. Understanding these biostatistical rationales not only enables the biologists to interpret the analytical results more accurately, but also facilitates better design of the experiments.

High throughput screening (HTS) in biomedicine refers to a type of scientific experimentation where thousands of or up to a million biological tests can be examined simultaneously for potential 'target' discovery. This vast number of simultaneous testing has been made possible at an affordable expense by significant advances of modern technology such as nanotechnology and robotic systems for bioassay. Nowadays, pharmaceutical company's compound libraries typically consist of over a million compounds available for lead screening. A normal day of HTS operation may provide more than 100 000 data points, leading the business into the ultra-high-throughput-screening (uHTS) era (Mayr and Bojanic, 2009).

DNA microarrays have single strands of a gene sequence segment (often called probes) immobilized to the quartz wafer surface via hydroxylation (e.g. Affymetrix arrays). Tens of thousands of the probes can be arranged on a single chip. When complementary strands of DNA or RNA are available, these probes hybridize with them and immobilize them at the site of the probes. A widely used method of making the genes visible is to add a dye (e.g. Cy3 or Cy5) to the complementary strand so that the amount of cDNA captured by the microarray chip can be quantified by a proper optical scanner. Obviously, to prepare well labeled cDNA, one has to grow the biological material under controlled conditions, obtain RNA from the sample and build cDNA from the RNA using nucleotides with a dye molecules attached. If a valid experimental protocol is closely followed and proper data normalization is performed, tens of thousands of statistical comparisons can be conducted.

Single nucleotide polymorphism (SNP) array is a type of DNA microarray used to detect a variation at a single site in DNA sequence, which is the most frequent type of variation in the genome. With the completion of the whole human genome sequence and the reduction of costs in SNP genotyping, it is now possible to use an array of one million SNPs for genomewide association studies, resulting in a huge number of hypothesis testing.

Proteomics attempts to identify and evaluate the presence and abundance of all the protein components and their variant isoforms such as post-translational modifications in an organism. It is of great scientific and clinical value for researchers to systematically investigate the different protein patterns among different cell types and tissues and under different developmental, physiological,

Received April 19, 2010; accepted May 11, 2010

E-mail: xian-jin.xie@utsouthwestern.edu

and disease conditions. New insights gained on the molecular basis for fundamental biological processes as well as the course of disease progression ultimately bring the possibilities for improvement in clinical diagnosis and treatment. Because the number of proteins and their variants is considerably large and because proteins often interact with each other to form dynamic multi-subunit macromolecular complexes, the search for biologically relevant proteomic patterns remains a challenging task. Given a single dataset obtained from any one of the most widely used techniques to screen the proteome: 2D gel electrophoresis, or protein and peptide mass spectrometry, it is inevitable that investigators will examine a large number of potential associations (number of protein peaks under examination is substantially larger than the number of biological replicates).

In new drug discovery, high throughput screening is to test a large collection of potentially active chemical compounds (either natural or synthesized) for further pre-clinical evaluations. HTS has become one of the major tools for pharmaceutical industries to look for initial clues in new drug search, complementing the rational drug design where knowledge about the targeted protein and its biological mechanism is the basis for the design of a drug. Although rational drug design has led to certain successes, it may not be always effective since knowledge about the compound's interaction with the target protein and its cellular micro-environment is hardly comprehensive. HTS, however, searches for the potential chemical compounds without a bias and without the need of comprehensive knowledge about the compound. As the collection of the testing population may consist of up to millions of chemical compounds, a large number of statistical tests need to be performed simultaneously.

In biomedical imaging studies, data often consist of information from hundreds of thousands of voxels. The standard analysis of medical images often uses statistics calculated voxel-by-voxel across individuals in each group. This generally requires multiple-comparison adjustments.

It should be noted that multiple comparison issues are not unique with the HTS data analysis. They exist whenever multiple hypotheses are tested. For example, in a clinical trial, if there are more than one outcomes, *P*-values calculated for each of the outcomes need to be adjusted in order for an overall controlled type I error (false positive) rate. The required adjustment also applies when multiple interim analyses are planned for a single outcome measured at different time points during a trial.

Rather than attempting to provide an exhaustive detailed review of all the available statistical methods concerning multiple testing, we build on reviewing these methods and seek to provide a brief introduction and recommend a few selected practically useful guidelines for biologists to consider when designing HTS experiments and analyzing data from them. Section 2 discusses considerations on experimental design and section 3 mainly concerns methods of target identification. Issues about predictive modeling on DNA or protein microarray data and the validation methods are discussed in section 4. Several concluding remarks are made in section 5.

2 Experimental design

The goal of proper design of an experiment is to guard against systematic bias, reduce and control random variation so that signal of interest can be optimally extracted for examination. Therefore, understanding the possible sources of variation is the pivotal point for a better experimental design. This is particularly important for HTS experiments where multiple sources of systematic and random error exist. These errors come directly from the process of technological complexity as well as the preparation of the biological samples. As a general principle, systematic error may often be eliminated or corrected if source of the error can be identified, e.g. appropriate normalization of raw data may remove systematic plate-to-plate variation, making measurements comparable across plates. Random error, however, is more difficult to remove, as every biologist knows that one never gets exactly the same results when repeating the same experiment under the same condition. This type of error comes from a variety of sources including biological, experimental instruments, environment, and human-related factors. It unpredictably influences measurements relative to their true values. As always, the key in designing an HTS experiment is to minimize all extraneous variation not due to biology, such as those due to sample handing and processing.

Since HTS serves as the initial primary step for potential 'hits' identification, designing the experiment with a clear large-scale process of follow-up screen and validation steps is highly recommended. In drug discovery HTS, for instance, 'hits' suggested from the primary HTS will go through a secondary screen for retesting. *Bona fide* hits confirmed by secondary screen will then be evaluated for biological relevance and those with a plausible biological activity according to structure-activity relationship and medicinal chemistry are labeled drug "leads", which can then be developed into drug candidates for pre-clinical and clinical testing.

How many replicates do I need? That is one of the most important questions one has to ask when planning an HTS experiment. Since there may be confusion on biological and technical replication, we emphasize that it is the number of independent biological samples from the population of interest that matters. Though data from the same biological sample vary less, increasing the technical replicates on the same sample only increases the precision of measure on this sample. It does not contribute to estimating the mean measure of the population of samples

out there, which is almost always the researcher's interest. Increasing the number of independent biological replicates, on the other hand, captures more information of the population and estimates the population better. One extreme example is in an experiment to compare differential gene expression profiles between breast cancer patients and women with benign breast tissue. Suppose you have blood sample from one breast cancer patient and one woman with benign tissue. No matter how many technical replicates you can perform, your conclusion on your primary research question regarding the two populations of women will not be reliable. This is because a single independent sample from a population rarely represents the population adequately. Indeed by increasing the number of technical replicates on the two samples, you get better estimates from the two samples, making your conclusions about the samples themselves more reliable. It is the number of biological replicates that enables us to make the leap from samples we have to the population of our interest.

There are two major perspectives concerning the number of biological replicates. The first is feasibility. The number of replicates must be within the budgetary constraints and other resource limits. The second is statistical consideration. The number of replicates depends on the signal to noise ratio and depends on pre-specified acceptable false negative rate and false positive rate. Although the statistical importance of using replicates in HTS has been recognized and many have researched on this topic (Jung, 2005; Jung et al., 2005; Pawitan et al., 2005; Zhang and Heyse, 2009), whether to use replicates (and number of replicates to use) is rarely determined by careful statistical justifications. This is particularly true with the primary drug compound screens where compounds are typically measured once due to cost and time considerations despite the fact that single measurement methods have proven inadequate (Rocke, 2004). In other primary screens such as genome-wide RNAi screens, researchers may choose single measure, duplicates, triplicates, or occasionally four replicates predominantly based on logistic constraints. In almost all the secondary or confirmatory screens, certain number of replicates is used for estimating and controlling for the random error. Measurement precisions gained through the replicates may be valuable and outweigh immediate cost considerations.

It is worth noting that, in addition to considerations on special features of HTS experiments, fundamental principles of general experimental design apply also. These include well thought out strategies on confounding factors, blocking effects, potential crossing effects, balancing and valid randomization, which can be readily found in many statistical textbooks.

2.1 Recommendations on designing HTS experiments

It is the time and cost consideration that often dictates the current HTS experiment design. However, adequate statistical rationale should also be taken seriously. In those primary HTS experiments where a single measure is taken, one has to keep in mind that both false positive and false negative rates can be high and it is extremely hard to estimate these error rates due to the lack of a good variability measure. Because of this reason, it is highly recommended that at least duplicate measures be taken in HTS experiments. A simple measure in statistical gain from replicates is that, for the same variability, increasing the number of replicates from 2 to 6 results in decreased standard error estimates from 71% to 41% (71%, 58%, 50%, 45%, and 41%). For HTS predictive modeling such as those in mRNA and protein expression arrays, widely accepted simple rule to determine the sample size does not exist either. There are proposals and suggestions on estimating required sample size for this type of studies. But they depend on assumptions and specifications that are usually unknown or hard to predict.

3 Target discovery

The definition of a *P*-value is the probability of observing the data or more extreme ones given that the null hypothesis is true and the nominal cutoff *P*-value (e.g. 0.05) we choose is the false positive rate that we can tolerate. When testing a single hypothesis, the false positive rate often is under control (by definition). However, when testing multiple hypotheses, the chance of some true null hypotheses falsely showing "significant" result increases. As the number of hypotheses being tested increases, the number of false positives increases accordingly. That is, with multiple comparison problems, the chance of declaring a false significant test (positive) is higher than the nominal level even if each test is set at that nominal level. For instance, in a genome-wide screening study where 47 400 comparisons are tested, if the nominal cutoff of $P\text{-value} < 0.05$ is not adjusted, under the null hypothesis of no differential expression, one would by chance falsely conclude that some 2370 transcripts are differentially expressed (Xie, 2008). This far exceeds the common tolerable error limit. Note also a widely used definition of error rate for multiple comparisons: the family-wise error rate (FWER), which is the probability of having at least one false significant result among all the tests. It can be written as $\Pr(R > 0)$, where R is the number of rejections of true null hypotheses (false positives). This is a very strict error rate. There are different methods that are used to adjust the single nominal *P*-values so that an FWER can be controlled. The simplest procedure among these methods is the Bonferroni adjustment, where each test is controlled at a lower nominal level: α/m, with m being the total number of comparisons and α being the FWER specified by investigators. Simple and easy to use, however, Bonferroni procedure is overly conservative

when the number of tests becomes large, as is the case in HTS, making its usefulness extremely limited. To improve the power of Bonferroni adjustment, closed testing procedure has been proposed by several groups (Koch and Gansky, 1996; Zhang et al., 1997; Grechanovsky and Hochberg, 1999). Instead of a 'single-step' adjustment in Bonferroni procedure, the closed testing procedure uses critical significance levels larger than α/m, allowing higher power. Different tests can be used in the closed testing procedure, including Hotelling's T^2, Bonferroni-Holm minP test, Westfall-Young Bootstrap minP test, Exact Permutaional minP method, etc. (Westfall and Young, 1993).

A new method for multiple comparisons proposed by Benjamini and Hochberg (1995) takes a different approach. Instead of controlling FWER (the probability of having at least one false positive result), they suggest to control the false discovery rate (FDR), a less restrictive measure but biologically highly relevant and useful. FDR is defined as the expected proportion of false results among all declared significant ones. For example, if we set our target for FDR to be 5%, then having 100 significant results ("discoveries") with only five being false is well under control. This property of FDR is highly desirable and it is often more powerful than traditional statistical methods. To assist biologists with readily bench side access to proficient computing capabilities to analyze for FDR, recently a web portal has been developed for data on differential expressed genes or proteins (Ling et al., 2009).

Because Benjamini's original FDR approach assumes all hypotheses are true null and in certain situations it may be possible to estimate the proportion of true null hypotheses, Storey (2002) extended the Benjamini's method to incorporate this information. As a result, the extended FDR of Storey is more powerful when information about the proportion of true null hypotheses is available.

The original Benjamini-Hochberg FDR method was proposed under the situation where all the tests are assumed to be independent. Under this assumption, controlling FDR is quite straightforward. But in many cases the measurements from HTS are dependent. This is especially true with the DNA microarray or proteomics experiments where dependency may come from co-regulations of sets of genes or measurement errors due to technical limitations (e.g. spatial correlations). During last ten years or so, many have attempted to generalize the original Benjamini-Hochberg FDR method to situations where tests are dependent (Yekutieli and Benjamini, 1999; Benjamini and Yekutieli, 2001; Storey, 2003; Farcomeni, 2007). But based on Farcomeni's simulation studies, a certain degree of dependence is allowed among the tests for the classical procedures to work, provided that the number of tests is large. He argued that under weak dependence, there is no need for any correction to the standard Benjamini-Hochberg FDR method (Farcomeni, 2007). The assumption of weak dependency may be reasonable when considering the whole genome. But strong dependencies do exist within certain small group of genes that share common chromosomal location, regulation pathway, or functional process.

Subramanian et al. (2005) described a powerful knowledge-based method for interpreting genome-wide expression profiles: Gene Set Enrichment Analysis (GSEA). They focus on established gene sets to avoid some of the possible limitations when individual genes are analyzed separately: (1) No individual gene may meet the threshold for statistical significance after multiple comparison correction because the biological differences at single gene level are modest relative to the noise inherent to current microarray technology. (2) Interpretation of a list of statistically significant genes from separate tests without any related biological function may be difficult and *ad hoc*. (3) Single gene analysis may miss important ones on cell pathways since cell functions rely on sets of genes acting collectively. Small change in all genes of a set may be more relevant than a large change of a single gene. If gene set can be defined reliably from prior biological knowledge, GSEA can be powerful in enhancing the signal-to-noise ratio and make it possible to detect modest changes in individual genes that would be hard to detect in single gene analysis. Yan et al. (2005) and Zhou et al. (2005) have also proposed similar approaches, called ontology-based pattern identification (OPI) score, to improve screening results by incorporating both the activity ranking and the number of structurally/functionally related genes.

A deterministic threshold widely used and preferred by many biologists is fold change (FC). It is a simple comparison of the mean measures in experimental group versus control group: $FC = \frac{\overline{x}_i}{\overline{y}_i}$. It is popular not because of its statistical rigor, but because of its simplicity and intuitive appeal. As a matter of fact, a deterministic cut-off of fold change yields unknown statistical properties since it completely ignores the variability of the data. Nevertheless, in some circumstances, it might be necessary to exclude those genes with small mean differences depending on the specific study, associated research questions and plausible biological rationale. When used in combination with a statistical test, fold change may yield excellent results and is very useful in practice (Whitehurst et al., 2007; Xie et al., 2007).

A recent important development in statistical methods for HTS data analysis is the optimal discovery procedure (ODP) proposed by Storey (2007). Not performing a statistical test on each comparison separately and then adjusting for multiple comparisons across the tests, the ODP considers all the comparisons simultaneously and the statistic is calculated based on the information from the whole dataset. The ODP is "optimal" in that it maximizes the number of expected true discoveries for each fixed

number of expected false discovery rate. This is obviously an ideal property suitable for many HTS data analysis. It has been shown that ODP has higher power over a number of current leading methods.

Based on Storey's ODP proposal and James-Stein estimation (James and Stein, 1961), Cao et al. (2009) reported an improved version of the procedure. James-Stein's estimation in decision theory is the phenomenon that when three or more parameters are estimated simultaneously, their combined estimator is more accurate (has lower expected mean-squared error) than any method that handles the parameters separately. This is surprising since the parameters and the measurements might be totally unrelated. Using this important statistical fact, several new statistics have been proposed (Cui et al., 2005; Opgen-Rhein and Strimmer, 2007). Because an introduction of a prior distribution on variance across data naturally implements the James-Stein shrinkage idea, Cao et al. (2009) proposed a Bayesian hierarchical model for simultaneous significance testing by introducing a mixture structure on variance components. They further suggested that estimates from the Bayesian hierarchical model be utilized in Storey's ODP. Their simulation studies show that the Bayesian ODP outperforms many leading methods including the original ODP, especially when the number of replicates is small. Such an improvement is due to the reason that the Bayesian ODP employs the posterior probability of a discovery being true, which has the joint force of shrinkage estimation and borrowing strength across tests. Zhang and Cao (2009) also compared methods based on shrinkage concepts and the double filtering with fold change and t-test and showed that the former may outperform the latter.

3.1 Recommendations for choice of analytical methods

It is critical to choose an appropriate approach for examining the HTS data since the choice may have significant effects on the targets that are identified. Valid biological rationale should always dictate the choice of methods. If changes in mean measures relative to the underlying noise are important, statistics with variance considered (formal statistical testing) should be used. If absolute changes of measures are pivotal to the biological mechanisms, fold change may be preferable regardless of the statistical significance. In situations where both absolute changes and changes relative to noise level are important, statistical tests and fold change may be used together for more reliable and relevant discoveries. When formal statistical testing is performed, overly conservative approaches such as Bonferroni correction for FWER control should always be avoided and Benjamini-Hochberg FDR control should be used for initial screening. Furthermore, Storey's ODP should also be performed for each tolerable number of expected false discovery rate. In most of the HTS data situations where number of sample replicates is small, Bayesian ODP (Cao et al., 2009) is recommended over the original Storey's ODP.

4 Predictive modeling

Section 3 reviews the analytical methods concerning identification of individual targets. This section will briefly discuss necessary statistical considerations when sets of variables are examined together for predictive modeling.

In certain HTS experiments such as genome-wide association and large scale protein expression studies, it is often known what conditions the samples are obtained. For example, in a lung cancer research project, one may have both lung cancer cell lines and normal lung cell lines, and it is of interest to examine whether the gene expression profiles can distinguish the cancer cell lines from the normal cell lines. In a breast cancer study, if samples from patients are available, one may be interested in gene profiles that correlate to different stages of the tumor progression. Such classifications can be viewed as predicting the categories of the observations. If the outcome of interest is not categorical but some ordinal or continuous measures from the observations, then it is of equal importance to discern genes that are correlated to the continuous measure. For example, if the survival times are available for study patients, it would be highly valuable to find out if there is a genetic factor (certain genes) that is associated with the survival time after controlling for treatments and other covariates. These types of correlative explorations are often called risk prediction modeling or predictive modeling. It is one of the most used branches of data mining. The main issue addressed by predictive modeling is to produce approximately accurate predictions for the outcome (response, dependent, or class) variables by a set of predictors (explanatory variables). This is achieved by careful examining noisy data and effectively extracting the associations between the outcome variable and the predictors.

Vast statistical literature on how to build predictive models exists. However, most of these methods that are well accepted deal with data that have a small number of potential predictors (P) relative to a large number of observations (n). In such data situations ($P << n$), the resulting model may capture the underlying mechanism of biology to a degree corresponding to the amount of true information in the data, thus is likely to be reliable in that it predicts well on other independent datasets that are obtained similarly. The main techniques used are classification and regression methods. When a predictive model is built on HTS data where $P >> n$, problem of overfitting arises, resulting in a model that may capture true associations but also peculiar patterns of the specific dataset. Consequently, such a model performs poorly with external validation datasets.

Model building with HTS data is not an easy task. Only

through a series of rigorous statistical investigations can a reliable model be reached. In gene expression array studies for instance, step one is to determine genes that are potentially important predictors in the model. These predictors may be individual genes, sets of genes, or meta-genes (representative gene in a gene cluster), or super-genes resulting from principal component analysis of the gene expression data matrix. Once potentially important predictors are identified, in step two, model fitting can be performed using these predictors by likelihood or penalized likelihood (e.g. LASSO proposed by Hastie et al., 2001) approaches. Note that numeric iterative methods or proper approximations may be required for non-linear model fit where close form solutions do not exist. Due to the currently availability of high computing power, computational search-based non-parametric modeling techniques such as neural networks and decision trees are more and more used in practice (Ripley, 1996). Step three should include goodness-of-fit testing and predictability evaluations on the model reached. For this, one may use Akaike's information criterion or the Bayesian information criterion. Finally, validation of the final model obtained should be performed on independent data for generalizability testing. Obviously, any reliable model should predict well in similar data collected by others. This final validation step using external independent data has been regarded as the 'gold' standard before a model is taken seriously.

Caution of overfitting should always be taken when fitting a predictive model using HTS data. On the other hand, it is also intrinsic that methods of model building on HTS data suffer from very low power, i.e. the chance of detecting the true pattern that is associated with the outcome is very low due to the small number of independent replicates (e.g. 20 000 genes on 50 patients). This low power issue of HTS data analysis is extremely hard to remedy, thus resulting in failure to detect some significant targets.

4.1 Recommendations on HTS correlative analysis

Once the HTS dataset is obtained and secured, it is relatively cost-effective to perform a large number of statistical analyses on it. This is usually the case when potential correlations among factors (e.g. genes) or associations between factors and sample conditions are of interest but none of these potential correlations are pre-specified for formal hypothesis-testing. Due to the reasons discussed above (overfitting as well as very low power), one has to be always cautious of any correlations detected through exploring HTS data. Validation should be part of any experimentation process. It is pivotal in analyzing HTS data. One should test the model on external data before making any definitive conclusions. Internal validation based on an individual study is usually not enough.

5 Conclusions

As with any biological experiments, sound scientific rational and optimal design and high-quality execution of the HTS experiment dictate the following analyses. Besides, valid data preprocessing steps including effective background correction and normalization is necessary before statistical analysis can be performed. Two typical systematic errors in HTS data have to be carefully dealt with in particular. They are spatial patterns of the data from a plate and temporal patterns of the data across experimental time. Analysis results are highly dependent on these data processing and normalization steps. This remains to be an active area of research and is not covered in this article (Owzar et al., 2008). Nonetheless, we should be as diligent on these steps as we are on choosing the most appropriate statistical methods. Although all HTS experiments are different in certain ways, the following suggestions apply in general when planning an HTS experiment:

(1) Establish clearly your primary goal of the experiment, i.e. what exactly would you like to compare or examine?

(2) Think through all possible sources of bias (systematic error) and determine effective ways to eliminate or reduce the bias.

(3) Think through all possible variance components of the data (random error) and determine effective ways to reduce and/or control them.

(4) Within cost constraints, try to use at least three replicates per measurement.

(5) Based on above steps, design experiments that illuminate and magnify the signal, minimize the bias, and control the random error.

(6) Choose the statistical methods that are consistent with the experimental design, in particular, apply sufficient multiple comparison adjustment and guard against overfitting.

(7) Avoid drawing definitive conclusions before external validation and rigorous assessment of new model prediction errors.

Finally and importantly, it is a good idea to get a trained biostatistician involved in your HTS study, from the stage of planning of the HTS experiment, data collection and normalization, to the final statistical analysis and interpretation.

Acknowledgement This work is supported in part by NIH P50-CA70907, NIH U24CA126608, and NASA NNJ05HD36G.

References

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate – a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B Meth, 57: 289–300

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

Cao J, Xie X J, Zhang S, Whitehurst A, White M (2009). Bayesian optimal discovery procedure for simultaneous significance testing. BMC Bioinformatics, 10(1): 5

Cui X, Hwang J T G, Qiu J, Blades N J, Churchill G A (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics, 6:59–75

Farcomeni A (2007). Some results on the control of the false discovery rate under dependence. Scand J Stat, 34(2): 275–297

Grechanovsky E, Hochberg Y (1999). Closed procedures are better and often admit a shortcut. J Statist Plann Inference, 76(1–2): 79–91

Hastie T, Tibshirani R, Friedman J (2001). Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer-Verlag

James W, Stein C (1961). Estimation with quadratic loss. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, Berkeley, 1961. University of California Press, 361–379

Jung S H (2005). Sample size for FDR-control in microarray data analysis. Bioinformatics, 21(14): 3097–3104

Jung S H, Bang H, Young S (2005). Sample size calculation for multiple testing in microarray data analysis. Biostatistics, 6(1): 157–169

Koch G G, Gansky S A (1996). Statistical considerations for multiplicity in confirmatory protocols. Drug Inf J, 30: 523–534

Ling X B, Cohen H, Jin J, Lau I, Schilling J (2009). FDR made easy in differential feature discovery and correlation analyses. Bioinformatics, 25(11): 1461–1462

Mayr L M, Bojanic D (2009). Novel trends in high-throughput screening. Curr Opin Pharmacol, 9(5): 580–588

Opgen-Rhein R, Strimmer K (2007). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. Stat Appl Genet Mol Biol, 6(1): 9

Owzar K O, Barry W T, Jung S H, Sohn I, George S L (2008). Statistical challenges in preprocessing in microarray experiments in cancer. Clin Cancer Res, 14(19): 5959–5966

Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A (2005). False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics, 21(13): 3017–3024

Ripley B (1996). Pattern Recognition and Neural Networks. Cambridge University Press

Rocke D M (2004). Design and analysis of experiments with high throughput biological assay data. Semin Cell Dev Biol, 15(6): 703–713

Storey J D (2002). A direct approach to false discovery rates. J Roy Stat Soc Ser B Meth, 64(3): 479–498

Storey J D (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat, 31(6): 2013–2035

Storey J D (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. J R Stat Soc, B, 69(3): 347–368

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

Westfall P H, Young S S (1993). Resampling-Based Multiple Testing. New York: John Wiley & Sons, Inc.

Whitehurst A W, Bodemann B O, Cardenas J, Ferguson D, Girard L, Peyton M, Minna J D, Michnoff C, Hao W, Roth M G, Xie X J, White M A (2007). Synthetic lethal screen identification of chemosensitizer loci in cancer cells. Nature, 446(7137): 815–819

Xie X J (2008). On multiple testing, validation of gene expression profiling, and translational research. Chin Med J (Engl), 121(13): 1247–1248, author reply 1247–1248

Xie X J, Whitehurst A, White M (2007). A practical efficient approach in high throughput screening: using FDR and fold change. Nat Protoc, doi:10.1038/nprot.2007.188

Yan S F, Asatryan H, Li J, Zhou Y (2005). Novel statistical approach for primary high-throughput screening hit selection. J Chem Inf Model, 45(6): 1784–1790

Yekutieli D, Benjamini Y (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J Statist Plann Inference, 82(1–2): 171–196

Zhang J, Quan H, Ng J, Stepanavage M E (1997). Some statistical methods for multiple endpoints in clinical trials. Control Clin Trials, 18(3): 204–221

Zhang S, Cao J (2009). A close examination of double filtering with fold change and T test in microarray analysis. BMC Bioinformatics, 10 (1): 402

Zhang X D, Heyse J F (2009). Determination of sample size in genome-scale RNAi screens. Bioinformatics, 25(7): 841–844

Zhou Y, Young J A, Santrosyan A, Chen K, Yan S F, Winzeler E A (2005). In silico gene function prediction using ontology-based pattern identification. Bioinformatics, 21(7): 1237–1245