Statistical Modelling 2017; **17(4-5)**: 305–315

Discussant paper on 'Statistical contributions to bioinformatics: Design, modelling, structure learning and integration'

Tianzhou Ma[1], Chi Song[2] and George C. Tseng[1]
[1]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh Pittsburgh, PA, USA.
[2]Division of Biostatistics, College of Public Health, Ohio State University, Columbus, OH, USA.

This article by Morris and Baladandayuthapani (abbreviated as MB paper hereafter) has nicely covered many important contributions from statisticians to the bioinformatics field. Being fortunate to witness the rising and maturing of bioinformatics and computational biology in the past two decades, statisticians have undoubtedly made significant contributions in the field and have gradually grown from traditional consulting role to leadership role in many areas of bioinformatics, genomics and extended biomedical fields. Table 1A shows the number of National Institutes of Health (NIH) R01 grants with statisticians as principal investigator (PI) or multiple PI compared to the total number. The ratio has grown from 0.56% during 2001–2005 to 0.94% during 2011–2015 (odds ratio = 1.7, Fisher's exact p-value = 5.5E–32). If restricted to National Human Genome Research Institute (NHGRI) and National Library of Medicine (NLM) (two institutes likely to fund bioinformatics methodological grants), the growth was less or non-significant (8.92% to 13.16% for NHGRI and 8.79% to 6.58% for NLM). The trend is similar if we extend the numbers to all NIH grants in Table 1B. Although this analysis covers the entire biomedical field and is not restricted to bioinformatics, it suggests that statisticians are taking more leadership role in scientific grants funded by disease-related NIH institutes (e.g., through collaborations using multiple PI mechanism) or institutes are funding more disease-relevant methodological grants led by statisticians. For example, our lab received an R01 grant from National Cancer Institute (NCI) to develop power calculation and study design methodologies for RNA-seq, methyl-seq and fusion transcript detection.

The term 'bioinformatics' covers a wide array of topics and biological applications. For example, the 'scope guideline' of *Bioinformatics* journal states 10 research categories: genome analysis, sequence analysis, phylogenetics, structural bioinformatics, gene expression, genetics and population analysis, systems biology, data and text mining, databases and ontologies, and bioimage informatics. Table 2 shows the number of papers published in the journal by category with any co-author affiliated

Address for correspondence: George C. Tseng, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA, USA.
E-mail: ctseng@pitt.edu

 10.1177/1471082X17705992

Table 1A Number of NIH R01 grants in RePORTER

Year	Number of grants statisticians as PI or multiple PI	Number of other grants	Total number of grants	Proportion and odds ratio
All				
2011–2015	1 234	129 666	130 900	0.94%
2001–2005	849	150 282	151 131	0.56%
				OR = 1.7 (p = 5.5e–32)
NHGRI				
2011–2015	97	640	737	13.16%
2001–2005	59	602	661	8.92%
				OR = 1.5 (p = 0.013)
NLM				
2011–2015	20	284	304	6.58%
2001–2005	24	249	273	8.79%
				OR = 0.73 (p = 0.35)

Source: Retrieved 10 April 2017, from https://projectreporter.nih.gov/reporter.cfm.

Table 1B Number of all NIH grants in RePORTER

Year	Number of grants statisticians as PI or multiple PI	Number of other grants	Total number of grants	Proportion and odds ratio
All				
2011–2015	2 787	386 434	389 221	0.72%
2001–2005	1 719	359 862	361 581	0.48%
				OR = 1.51 (p = 6.8e–42)
NHGRI				
2011–2015	178	2 751	2 929	6.08%
2001–2005	112	2 098	2 210	5.07%
				OR = 1.21 (p = 0.13)
NLM				
2011–2015	53	1 007	1 060	5%
2001–2005	68	1 218	1 286	5.29%
				OR = 0.94 (p = 0.78)

Source: Retrieved 10 April 2017, from https://projectreporter.nih.gov/reporter.cfm.

to a statistics or biostatistics department in 2006, 2011 and 2016. Although this analysis is preliminary (e.g., it misses statisticians housed in medical, bioinformatics or genetics departments), it clearly demonstrates strength of statisticians in conducting certain bioinformatics topics such as 'genetics and population analysis' (20–60%), 'gene expression' (31–38%) and 'genome analysis' (14–24%). The MB paper has covered issues mostly in the analysis and modelling of high-throughput experimental data, an area we statisticians usually take leadership role. Although the review is impossible to be complete, the authors have nicely focused on four key areas and have illustrated with many examples from their own works and perspectives. The story telling writing style is smooth to read, educational and insightful for statisticians to reflect in depth what we have done, what we could have done better and how we can do more in the future for the broader quantitative biology and medicine field. Below, we expand with some topics that we feel statisticians have made essential

and impactful contributions in Section 1, and it is followed by several key points that we statisticians could pay more attention to enhance our leadership in Section 2.

1 Selected contributions by statisticians in bioinformatics

Statisticians have made numerous contributions to bioinformatics. Below we present some additional work by statisticians, including some relevant to our experiences. The coverage cannot be comprehensive; for example, many excellent works in statistical genomics (e.g., genome-wide association and disease loci mapping), gene regulation (e.g., analysis of ChIP-chip and ChIP-seq data, Hi-C data and eQTL analysis) or more recently single-cell data analysis are covered neither by us nor by the MB paper.

1.1 Methods for differential expression analysis

The MB paper discussed methods for finding differentially expressed proteins and differentially methylated regions but the most cited papers from statisticians are probably mRNA differential expression (DE) detection methods for microarray and RNA-seq data. Detection of such differentially behaved features in disease progression or treatment response is usually a beginning step to understand disease mechanism and identify candidate markers. Tusher et al. (2001) developed the Significance Analysis of Microarray (SAM; 11 233 Google Scholar citations) software using a fudge parameter to stabilize standard deviation in the denominator of gene-specific *t*-statistics and followed by permutation analysis and false discovery rate (FDR) control. Later, Smyth (2004) proposed a Linear Models for Microarray (LIMMA; 8 650 citations) using an empirical Bayes approach. These two software packages are probably the most cited microarray DE analysis tools so far. For DE analysis in RNA-seq, edgeR by Robinson et al. (2010) (4686 citations) and DEseq by Anders and Huber (2010) (5 065 citations) using over-dispersed negative binomial model for count data are the most popular methods. To account for abundant number of zero counts in RNA-seq, Van De Wiel et al. (2012) developed a ShrinkSeq method with zero-inflated negative binomial for better model fitting.

1.2 Study design and power calculation

Complexity of high-throughput genomic experiments has created difficulties for biologists to properly design and utilize the tool for biological investigation. Batch effect and reproducibility mentioned in the MB paper are important aspects in this regard. Yang and Speed (2002) presented design issues in two-color cDNA microarray experiments (945 citations). Several influential papers in microarray by Richard Simon have discussed wider study design issues (Simon et al., 2002; 206 citations), pitfalls and strategies for diagnostic and prognostic classification (Simon et al., 2003; 1 009 citations; Simon, 2005; 354 citations) and sample size determination (Dobbin and Simon, 2005; 152 citations). Allison et al. (2006) later

Table 2 Papers involving statisticians in the journal of 'Bioinformatics' by category

	2016			2011			2006		
Category	Total number of articles	Number of articles involving statisticians	Proportion	Total number of articles	Number of articles involving statisticians	Proportion	Total number of articles	Number of articles involving statisticians	Proportion
Genome analysis	86	21	0.24	104	15	0.14	32	5	0.16
Sequence analysis	96	5	0.05	105	5	0.05	72	9	0.13
Phylogenetics	10	1	0.1	12	1	0.08	11	2	0.18
Structural Bioinformatics	33	5	0.15	75	1	0.01	56	2	0.04
Gene Expression	32	12	0.38	77	24	0.31	95	30	0.32
Gene and Population Analysis	25	15	0.6	31	11	0.35	10	2	0.2
System Biology	78	11	0.14	127	7	0.05	54	2	0.04
Data and Text Mining	28	0	0	66	12	0.18	30	1	0.03
Databases and Ontologies	32	2	0.06	39	3	0.08	14	0	0
Total	420	72	0.17	636	79	0.12	374	53	0.14

Source: Retrieved 10 April 2017, from https://academic.oup.com/bioinformatics.

Note: The section of 'Bioimage Informatics' was introduced into the journal in 2012, thus is not listed here for comparison.

provided a nice summary of well-recognized microarray data analysis strategies for practitioners to follow (1 297 citations). It has become a consensus that statisticians should be involved in early-stage grant proposal writing and study design discussion to avoid potential pitfalls that are impossible to correct in later data analysis stage. With the increasing complexity and amount of data generated, we expect statisticians to continue an essential role in study design of omics-type experiments.

1.3 Machine learning

Supervised machine learning (aka classification) and unsupervised machine learning (aka clustering) are two commonly encountered analytical purposes in high-throughput experimental data analysis. For supervised machine learning, class labels of samples are given. The purpose is to construct an accurate classification model from training data that can predict well in future general population or test data. Statisticians have applied and developed multiple methods that are commonly used and highly cited in biological applications. Tibshirani et al. (2002) (2 432 citations) developed a shrunken nearest centroid method to allow simultaneous feature selection and model construction for microarray classification. Geman et al. (2004) (240 citations) and Tan et al. (2005) (283 citations) proposed a non-parametric top-scoring pair method by utilizing mRNA expression ranking of gene pairs for robust prediction in different platforms or study design. Díaz-Uriarte and De Andres (2006) (1 360 citations) and Furey et al. (2000) (2 214 citations) applied popular random forest (Breiman, 2001) and support vector machines methods for microarray classification. Integrative methods using multi-omics profiles have been developed for drug sensitivity prediction in cancer patients and many of which have been assessed and compared via a collaborative effort between NCI and the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project (Costello et al., 2014). Methods based on multiple kernel learning algorithm (Gönen and Alpaydn, 2011) and random forest-based methods are among the top-performing approaches.

For unsupervised machine learning of microarray, statisticians also contributed significantly in relevant applications. The purpose of clustering microarray data may be two-folds. Clustering samples allows discovery of novel disease subtypes to form basis of personalized medicine. Clustering genes generates highly correlated gene modules that are co-expressed and potentially co-regulated. McLachlan et al. (2002) (573 citations) and Medvedovic et al. (2004) (186 citations) applied frequentist and Bayesian Gaussian mixture model clustering on microarray gene clustering. Weighted correlation network analysis (WGCNA; Langfelder and Horvath, 2008; 1 676 citations) is a popular microarray gene clustering software for module detection, network construction and calculating topological properties. Monti et al. (2003) (976 citations) and Tseng and Wong (2005) (181 citations) developed resampling-based approach for stable clustering result. Tibshirani et al. (2001) (2 438 citations) and Tibshirani and Walther (2005) (368 citations) proposed gap statistics and prediction strength approaches for estimating the number of clusters in cluster analysis. In the MB paper, clustering using multi-level omics data has been discussed

(see also Tseng et al., 2015). Recently, Kim et al. (2016b) and Huo and Tseng (2017) have proposed overlapping group lasso approaches to utilize inter-omics feature relations as prior information to enhance sample cluster analysis. Although statisticians face competition from computer scientists and applied mathematicians in the large machine learning and pattern recognition field, the boundary across these disciplines has become less clear and statisticians trained to master uncertainty in data should continue to play an important role in this area.

1.4 Horizontal meta-analysis of high-throughput experimental data

In high-throughput omics studies, each individual study usually has small or moderate sample size. Combining multiple studies can improve statistical power and estimation accuracy. Such genomic information integration of multiple transcriptomic studies is also termed 'horizontal meta-analysis', to distinguish from vertical integration of multi-omics data. Rhodes et al. (2002) (681 citations) was probably the first to apply microarray meta-analysis for differentially expressed gene detection in cancer research and many classical and newly developed methods have been used in biological applications since then (Choi et al., 2003; Li & Tseng, 2011; Wang et al., 2012). To advance from microarray to next generation sequencing (NGS) technology, Ma et al. (2016) and Ma et al. (2017) proposed Bayesian hierarchical model for meta-analysis to combine multiple RNA-seq studies or to combine microarray and RNA-seq cross-platform studies. In addition to these methods for detecting differentially expressed genes, omics meta-analysis for other biological purposes have been developed, including quality control (Kang et al., 2012), pathway analysis (Shen and Tseng, 2010), cluster analysis (Huo et al., 2016), classification analysis (Kim et al., 2016a), dimension reduction (Kim et al., 2017), gene regulatory association (Wang et al., 2017) and differential co-expression network analysis (Zhu et al., 2016).

1.5 Evaluation and comparative studies

Statisticians are experts in understanding model assumptions, model fitness and pros and cons of different methods for a given analytical purpose. As a result, they contribute well to perform comprehensive and unbiased comparative studies that are instrumental to guide practitioners when they select a proper method to use. In early microarray stage, Dudoit et al. (2002) (2 627 citations) performed comparison of multiple classical machine learning tools for microarray classification. Datta and Datta (2003) (362 citations) and Thalamuthu et al. (2006) (216 citations) performed comprehensive comparison of microarray gene clustering methods, and Brock et al. (2008) (100 citations) compared different missing value imputation methods. Tseng et al. (2012) (138 citations), Evangelou and Ioannidis (2013) (202 citations) and Begum et al. (2012) (72 citations) provided comprehensive review of methods for microarray and GWAS meta-analysis. For NGS data analysis, Soneson and Delorenzi (2013) (366 citations) performed a comparative analysis of 11 RNA-seq DE methods. Liu et al. (2016) recently performed a large-scale evaluation for fusion

transcript detection in RNA-seq data and proposed a meta-caller by combining three top-performing tools for better detection accuracy.

2 Looking ahead: How can we better contribute and participate to lead?

Looking ahead, statisticians should continue to contribute and participate in the bioinformatics field to take the lead in areas we excel. Below are a few observations and suggestions that we could focus to achieve the goal.

1. **Software packages and computational considerations:** Many good papers in top statistical journals receive slow and few citations. One main reason, in our opinion, is the lack of software packages and practical consideration of computing. In our discipline, there has been little motivation to provide user friendly software. With the fast development of R and Bioconductor, there is no more excuse not to provide an easy package for each methodological paper, especially in the bioinformatics field. For biological labs, it has been found that open sharing of high-throughput experimental data is associated with increased citation (Piwowar et al., 2007; 460 citations). We argue that methodological papers with convenient software packages will also find greater impact and higher citations.

2. **Reproducibility of methodological papers:** In the MB paper, reproducibility for bioinformatics data analysis has been comprehensively discussed for biological studies. Here, we emphasize that reproducibility for methodological papers is equally important. It has been a common frustration for statisticians to identify an exciting methodological paper and end up spending weeks or longer to implement and reproduce results in the paper, if it can be successful. We suggest that, in addition to a software package for implementation, programming code for simulation, data preprocessing and down stream analyses should be provided publicly. For example, our research group has made it a routine to provide all data and code used in each paper on our research website or github http://github.comso that anyone can reproduce the results easily. Since 1 September 2016, *JASA Applications and Case Studies* section has required all submissions to include data, programming code and instructions for reproducing results in the submitted papers. As we statisticians tend to be slow in providing research solutions compared to computer scientists or data analysts, the transparency will allow improved turnaround in method development.

3. **Early involvement in project planning and study design:** As shown in Tables 1A and 1B, statisticians are taking more leadership in early project planning, particularly in grant writing stage. Statisticians in genomic field should particularly be collegial and proactive in a team science setting to increase our impact and contribution.

4. **Publish in subject journals:** From publications cited by this article and the MB paper, one can easily find the trend that statistical methods published in subject journals (usually with much higher impact factor) tend to receive higher citations.

There are still reasons to publish solid method papers in *JASA Application and Case Studies*, *JRSS-C* or *Annals of Applied Statistics* to maintain mathematical rigour and communication of technical novelty among statisticians. But 'If your application of statistical methods is truly addressing an important scientific question, then shouldn't the scientists in the relevant field want to hear about it?', quoted by the 'Do we really need applied statistics journals?' article in the simplystatstics.org blog. The blog also contains many nice tips of how to utilize preprint servers and other issues in publication. One major hurdle of publishing in subject journals relates to the evaluation of junior faculty or PhD students when they publish in many non-statistical journals. Compared to 10–15 years ago, our sense is that tenure review for junior faculty promotion now recognizes better for publications in subject journals. The constraint remained is mostly on PhD students or post-docs looking for tenure-track faculty positions in statistics or biostatistics departments. Without so-called 'solid' statistical publications in *JASA*, *JRSS* series or *Biometrics*, publications in subject journals are often still discounted in job interview since members in the faculty search committee may not have the expertise to judge paper quality in biological journals. This is an area our field can work to improve.

5. **Close collaboration with scientists and data access issues:** We statisticians analyze data but normally do not generate data. New scientific questions and new data types have been a key driving force in our discipline. Although remote collaboration using teleconference technology has become easy, it is clear that close collaboration with scientists locally in the same campus or even in the same building often provides better efficiency in genomic research. When first-hand data are not available, public data are indispensable for methodological development. In microarray era, public databases such as Gene Expression Omnibus (GEO) were paradise for statisticians that provided tons of real data to explore, motivated many scientific questions, and allowed development and validation of many methodologies. Many journals have encouraged or even forced biological publications to deposit expression profiles in GEO. When we transited to next generation sequencing stage, National Center for Biotechnology Information (NCBI) established a similar data repository space call Sequencing Read Archive (SRA). However, the culture of data sharing has largely disappeared now because raw sequencing data can potentially infringe privacy of patients by exposing their identities and genetic risks to diseases. With this excuse in mind, journals generally demand less public sharing of sequencing data now. Even if a lab graciously shares their sequencing data, they have to be stored in protected databases such as dbGaP. Access to such protected databases can be administratively tedious and time consuming, normally waiting for half to one year and with complicated administrative paperwork to maintain. One potential solution may be similar to the minimum information about a microarray experiment (MIAME) protocol to describe useful and minimal information a data generator should submit without affecting patient privacy (e.g., submitting read counts, fragments per kilobase of transcript per million mapped reads (FPKM) or transcripts per kilobase million mapped reads (TPM) values without actual sequencing read information

in FASTQ or BAM files). Statisticians could make important contributions in such an issue.

References

Allison DB Cui X, Page GP and Sabripour M (2006) Microarray data analysis: From disarray to consolidation and consensus. *Nature Reviews Genetics*, **7**, 55–65.

Anders S and Huber W (2010) Differential expression analysis for sequence count data. *Genome Biology*, **11**, R106.

Begum F, Ghosh D, Tseng GC and Feingold E (2012) Comprehensive literature review and statistical considerations for gwas meta-analysis. *Nucleic Acids Research*, **40**, 3777–84.

Breiman L (2001) Random forests. *Machine Learning*, **45**, 5–32.

Brock GN, Shaffer JR, Blakesley RE, Lotz MJ and Tseng GC (2008) Which missing value imputation method to use in expression profiles: A comparative study and two selection schemes. *BMC Bioinformatics*, **9**, 12.

Choi JK, Yu U, Kim S and Yoo OJ (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**, i84–i90.

Costello JC, Heiser LM, Georgii E et al. (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, **32**, 1202–12.

Datta S and Datta S (2003) Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, **19**, 459–66.

Díaz-Uriarte R and De Andres SA (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.

Dobbin K and Simon R (2005) Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, **6**, 27–38.

Dudoit S, Fridlyand J and Speed TP (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77–87.

Evangelou E and Ioannidis JP (2013) Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, **14**, 379–89.

Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M and Haussler D (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–14.

Geman D, d'Avignon C, Naiman DQ, Winslow RL et al. (2004) Classifying gene expression profiles from pairwise mrna comparisons. *Statistical Applications in Genetics and Molecular Biology*, **3**, 1071.

Gönen M and Alpaydn E (2011) Multiple kernel learning algorithms. *Journal of Machine Learning Research*, **12**, 2211–68.

Huo Z and Tseng GC (2017) Integrative sparse k-means for disease subtype discovery using multi-level omics data. *Annals of Applied Statistics*, in press.

Huo Z, Ding Y, Liu S, Oesterreich S and Tseng G (2016) Meta-analytic framework for sparse k-means to identify disease subtypes in multiple transcriptomic studies. *Journal of the American Statistical Association*, **111**, 27–42.

Kang DD, Sibille E, Kaminski N and Tseng GC (2012) Metaqc: Objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic Acids Research*, **40**, e15–e15.

Kim S, Lin C-W and Tseng GC (2016a) Metaktsp: A meta-analytic top scoring pair method for robust cross-study validation of omics prediction analysis. *Bioinformatics*, **32**, 1966–73.

Kim S, Oesterreich S, Kim S, Park Y and Tseng GC (2016b) Integrative clustering of multi-level omics data for disease subtype discovery

using sequential double regularization. *Biostatistics*, **18**, 165–79.

Kim S, Kang D, Huo Z, Park Y and Tseng GC (2017) Meta-analytic principal component analysis in integrative omics application. *Bioinformatics*, under review.

Langfelder P and Horvath S (2008) Wgcna: An r package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Li J and Tseng GC (2011) An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics*, **5**, 994–1019.

Liu S, Tsai W-H, Ding Y et al. (2016) Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end rna-seq data. *Nucleic Acids Research*, **44**, e47–e47.

Ma T, Liang F and Tseng GC (2016) Biomarker detection and categorization in ribonucleic acid sequencing meta-analysis using bayesian hierarchical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, in press.

Ma T, Liang F, Oesterreich S and Tseng GC (2017) A joint bayesian modeling for integrating microarray and rna-seq transcriptomic data. *Journal of Computational Biology*, in press.

McLachlan GJ, Bean R and Peel D (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–22.

Medvedovic M, Yeung KY and Bumgarner RE (2004) Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, **20**, 1222–32.

Monti S, Tamayo P, Mesirov J and Golub T (2003) Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, **52**, 91–118.

Piwowar HA, Day RS and Fridsma DB (2007) Sharing detailed research data is associated with increased citation rate. *PloS One*, **2**, e308.

Rhodes DR, Barrette TR, Rubin MA, Ghosh D and Chinnaiyan AM (2002) Meta-analysis of microarrays. *Cancer Research*, **62**, 4427–33.

Robinson MD, McCarthy DJ and Smyth GK (2010) Edger: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–40.

Shen K and Tseng GC (2010) Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, **26**, 1316–23.

Simon R (2005) Roadmap for developing and validating therapeutically relevant genomic classifiers. *Journal of Clinical Oncology*, **23**, 7332–41.

Simon R, Radmacher MD and Dobbin K (2002) Design of studies using dna microarrays. *Genetic Epidemiology*, **23**, 21–36.

Simon R, Radmacher MD, Dobbin K and McShane LM (2003) Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, **95**, 14–8.

Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, 3.

Soneson C and Delorenzi M (2013) A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, **14**, 91.

Tan AC, Naiman DQ, Xu L, Winslow RL and Geman D (2005) Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, **21**, 3896–3904.

Thalamuthu A, Mukhopadhyay I, Zheng X and Tseng GC (2006) Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**, 2405–12.

Tibshirani R and Walther G (2005) Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, **14**, 511–28.

Tibshirani R, Walther G and Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal*

Statistical Society: Series B (Statistical Methodology), **63**, 411–23.

Tibshirani R, Hastie T, Narasimhan B and Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, **99**, 6567–72.

Tseng GC and Wong WH (2005) Tight clustering: A resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, **61**, 10–16.

Tseng GC, Ghosh D and Feingold E (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research*, **40**, 3785–99.

Tseng GC, Ghosh D and Zhou XJ (2015) *Integrating Omics Data*. Cambridge, England: Cambridge University Press.

Tusher VG, Tibshirani R and Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, **98**, 5116–21.

Van De Wiel MA, Leday GG, Pardo L, Rue H, Van Der Vaart AW and Van Wieringen WN (2012) Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, **14**, 113–28.

Wang L, Liu S, Ding Y, Yuan S-s, Ho Y-Y, and Tseng GC (2017) Meta-analytic framework for liquid association. *Bioinformatics*, in press.

Wang X, Kang DD, Shen K et al. (2012) An r package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. *Bioinformatics*, **28**, 2534–36.

Yang YH and Speed T (2002) Design issues for cdna microarray experiments. *Nature Reviews Genetics*, **3**, 579–88.

Zhu L, Ding Y, Chen C-Y et al. (2016) Metadcn: Meta-analysis framework for differential co-expression network detection with an application in breast cancer. *Bioinformatics*, **33**, 1121–29.