BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 18 2009, pages 2425–2429
doi:10.1093/bioinformatics/btp430

Databases and ontologies

WebArrayDB: cross-platform microarray data analysis and public data repository

Xiao-Qin Xia[1,*], Michael McClelland[1,*], Steffen Porwollik[1], Wenzhi Song[1,2], Xianling Cong[3] and Yipeng Wang[1,4,*]

[1]Vaccine Research Institute San Diego, 10835 Road to the Cure, San Diego, CA 92121, USA, [2]Department of Oralogy, [3]Department of Dermatology, China Japan Union Hospital, Jilin University, Changchun, 130031, China and [4]Department of Pathology & Laboratory Medicine, University of California, Irvine, CA 92697, USA

Received on March 27, 2009; revised on June 9, 2009; accepted on July 5, 2009
Advance Access publication July 14, 2009
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Cross-platform microarray analysis is an increasingly important research tool, but researchers still lack open source tools for storing, integrating and analyzing large amounts of microarray data obtained from different array platforms.

Results: An open source integrated microarray database and analysis suite, WebArrayDB (http://www.webarraydb.org), has been developed that features convenient uploading of data for storage in a MIAME (Minimal Information about a Microarray Experiment) compliant fashion, and allows data to be mined with a large variety of R-based tools, including data analysis across multiple platforms. Different methods for probe alignment, normalization and statistical analysis are included to account for systematic bias. Student's *t*-test, moderated *t*-tests, non-parametric tests and analysis of variance or covariance (ANOVA/ANCOVA) are among the choices of algorithms for differential analysis of data. Users also have the flexibility to define new factors and create new analysis models to fit complex experimental designs. All data can be queried or browsed through a web browser. The computations can be performed in parallel on symmetric multiprocessing (SMP) systems or Linux clusters.

Availability: The software package is available for the use on a public web server (http://www.webarraydb.org) or can be downloaded.

Contact: xqxia70@gmail.com; mcclelland.michael@gmail.com; yipengw@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Large amounts of microarray experimental data are stored in public repositories, making cross-platform analysis of data from different sources (either different laboratories and/or different platforms), an increasingly attractive and important research tool (Moreau *et al.*, 2003). Such analyses are possible because biological treatments usually have a greater impact on measured expression than the noise of a cross-platform analysis (Chen *et al.*, 2008; Larkin *et al.*, 2005; Shippy *et al.*, 2004). Moreover, the combined use of multiple platforms can overcome the inherent biases of individual platforms for identification of the more robust changes in gene expression profiles (Bosotti *et al.*, 2007).

Currently available analysis packages do not provide all the required functions for cross-platform integration, normalization and statistical analysis of data from different sources. Integrative Array Analyzer (iArray; Pan *et al.*, 2006) offers statistical cross-platform analysis functions but does not have probe alignment or data normalization features. MatchMiner (Bussey *et al.*, 2003) is a powerful tool for matching genes and gene products from two platforms, but is not designed for statistical analysis. The Gene Expression Pattern Analysis Suite (GEPAS; Tárraga *et al.*, 2008) integrates many tools for microarray data analysis, but it does not have data storage capability or cross-platform analysis functions. Other online platforms and public repositories are designed mainly for data storage and lack probe matching and cross-platform analysis functions: prominent examples include Expression Profiler (Kapushesky *et al.*, 2004), ArrayExpress (Parkinson *et al.*, 2007), the Stanford Microarray Database (SMD; Demeter *et al.*, 2007), the Longhorn Array Database (LAD; Killion *et al.*, 2003) and the BioArray Software Environment (BASE; Saal *et al.*, 2002; Troein *et al.*, 2006).

An earlier open source online platform for microarray data analysis, WebArray (Xia *et al.*, 2005), did not offer a cross-platform analysis function, but provided an excellent framework for extension to WebArrayDB (http://www.webarraydb.org)—a database system and analysis suite that provides this function. In addition to traditional methods such as median and quantile for between-array normalization, WebArrayDB has integrated median rank scores (MRS), quantile discretization (QD; Warnat *et al.*, 2005), gene quantile (GQ)—a quantile normalization for each individual gene among different platforms, and principal component analysis (PCA; Stoyanova *et al.*, 2004). WebArrayDB provides standard statistical analysis methods, such as Student's *t*-test, eBayes-moderated *t*-test, Significance Analysis of Microarrays (SAM; Tusher *et al.*, 2001), analysis of variance or covariance (ANOVA/ANCOVA) and non-parametric tests, as options for users to explore.

2 DATABASE INFRASTRUCTURE

WebArrayDB includes all fields required for MIAME-compliant microarray data storage (Brazma *et al.*, 2001). Data are classified

*To whom correspondence should be addressed.

into five categories: 'project', 'array', 'platform', 'protocol' and 'sample'. Each record in these tables is given a unique ID ('MPMDB ID'), and all five categories have to be filled for MIAME compliance and subsequent data analysis. All tables in the database have been indexed to speed up queries even when the size of the dataset becomes very large.

The project table serves as the hub of information—most information is linked to a specific project in the database (Fig. 1 and Supplementary Fig. 1). Intrinsic relationships among project, array, platform, protocol and sample are directly linked by references between tables, which permits fast cross-table searching. When defining a platform, users may supply probe information, including user-defined IDs and gene IDs from other public databases, such as RefSeq, UniGene, etc. All of these IDs can serve as references for cross-platform probe alignment. Since there are extensive gene annotations in GO (Gene Ontology database, http://www.geneontology.org/; Ashburner and Lewis, 2002), WebArrayDB is also designed to facilitate the use of GO for probe searching. The GO database in WebArrayDB is updated monthly.

The project table is linked to the 'users' table that contains the user information including user name and password (Fig. 1), enabling data access to be controlled based on user privileges. Every project has an associated release date which determines the public accessibility of the project. By default the project release date will be 2 years from the data deposit date to protect data privacy. The user can change the release date at the time the data is deposited or at any time thereafter.

WebArrayDB is powered by the affy (Gautier *et al.*, 2004) and the Linear Models for Microarray Data (LIMMA, http://bioinf.wehi.edu.au/limma; Smyth, 2005) packages from bioconductor (http://www.bioconductor.org/), which are open source and open development software projects for the analysis and comprehension of genomic data. Thus, many different formats of intensity files are recognized, including data from Affymetrix CEL files, Agilent Feature Extraction, ArrayVision, BlueFuse, GenePix, QuantArray (Version 3 or later), SMD and SPOT. Any formats that affy and LIMMA do not recognize can be accepted when defined by the user in a tab-delimited text file, including data with more than two scanned channels.

WebArrayDB stores parsed data in database tables. The image files, intensity files, probe files, protocol files and other user-supplied raw data files are stored in the file system on servers with indices in the database.

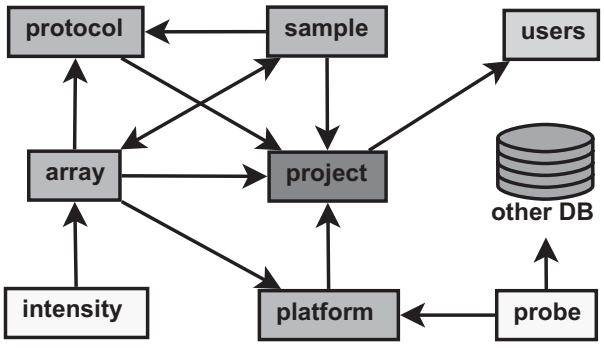

Fig. 1. Information organization in WebArrayDB.

3 DATA ANALYSIS

Data queried from the database can be directly subjected to analysis. WebArrayDB presents a variety of options for data preprocessing, and differential analysis. Conservative default analysis methods and parameters are set so that novice users will be less likely to use flawed analysis strategies.

3.1 Data preprocessing

Data preprocessing includes cross-platform probe alignment, background correction and normalization. For cross-platform analysis, the primary concern is how to match probes from different platforms. Based on the intrinsic relationships between platforms, we offer three approaches to this issue.

- **Direct match**
  Direct match is used when all probes are identical across microarray platforms.
- **Match by reference IDs**
  Probes from two different platforms can be aligned if they share the same reference ID. IDs from well-known public databases, for example, UniGene ID or Ensembl ID, can serve as reference IDs, as can any user-defined category.
- **Match by file**
  Users can align probes by providing a probe-mapping file, in which homologous probes are explicitly mapped.

If multiple platforms are involved, normalization within or between arrays of the same platform can be done directly on the raw data before probe alignment. After alignment, the whole data set can be normalized.

3.2 Differential analysis

Users can analyze data based on either ratio or intensity. The ratio-based model is $R=\mu+\varepsilon$, where R is the ratio, μ represents the intercept of the ratio of the two groups and ε represents the Gaussian random error. We say two samples are different if μ significantly differs from the null hypothesis.

More than one comparison among groups of data can be requested simultaneously. Furthermore, users may apply '+', '−' and parentheses to make more specific comparisons. For instance, given four groups, '$(group1+group2)-(group3+group4)$'computes the global difference between array data supplied in the first two groups compared with array data supplied in the second two groups.

Fold-change analysis, Student's t-test, eBayes-moderated t-test (Smyth, 2004; Smyth *et al.*, 2005), SAM test (Tusher *et al.*, 2001), non-parametric tests (including Wilcoxon rank sum test, Kruskal–Wallis rank sum test and Friedman rank sum test) and ANOVA/ANCOVA are among the choices of algorithms for differential analysis of data in WebArrayDB.

Mixed-effect model ANOVA plays a very important role in microarray data analysis (Churchill, 2002). ANOVA is capable of dealing with multiple factors. The default model in WebArrayDB is

$$E=\mu+G+P+A+D+S+I+\varepsilon$$

where E is the observed log-transformed intensity value, μ is the theoretical 'real' log-transformed intensity value, ε represents the Gaussian random error with 0 as expected value and G is the group factor, which leads to effects of interest, e.g. treatment effects.

P, *A*, *D*, *S* and *I* represent effects of *platform*, *array*, *dye*, *sample* and *individual*, respectively, among which *array* and *individual* are considered random effect factors. Based on the data to be analyzed, more or fewer factors might be used in specific analysis processes.

Experienced users can define new factors and create complicated analysis models. This enables WebArrayDB to analyze data from virtually any experimental design and thereby to retain relevance as methods continue to evolve.

3.3 Other analysis tools

Both raw and differentially analyzed data can be used for further analysis, including hierarchical clustering, correspondence analysis, between group analysis and plotting using genome position. A variety of high-quality charts in PDF and EPS formats can be produced to visualize analysis results.

3.4 Example

3.4.1 Data sources A demonstration of a cross-platform analysis is used as a training example in every WebArray account. This example uses two publicly available prostate cancer microarray datasets. One set was obtained using a custom made cDNA microarray (20K chip, platform MPMDB ID:42) that contains 19 947 sequence verified PCR-amplified human cDNAs representing 15 495 UniGene clusters (Dhanasekaran *et al.*, 2005, project MPMDB ID:76). The other was obtained using a commercially available oligonucleotide microarray (Affymetrix U95A array, platform MPMDB ID:9) that contains 12 626 probe sets consisting of 25-base oligonucleotide probes (Welsh *et al.*, 2001, project MPMDB ID:78). From the two datasets, 49 tumor samples (prostate cancer) and 21 non-tumor samples are analyzed in this example.

3.4.2 Options for analysis Analysis options selected for this demonstration are illustrated in Figure 2. The IDs from the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene) are used to match cDNA clones and Affymetrix probe sets between platforms. Within each study, the median value is used for expression values corresponding to probes of the same UniGene cluster. Genes not mapping to a UniGene cluster present in both microarray platforms are not considered for cross-platform analysis. For the integration and normalization of microarray measurements from different platforms, we apply quantile discretization (Warnat *et al.*, 2005). A common reference sample is used in the two-color cDNA microarray study and the log2 ratios of the intensity values from experimental samples over the common reference sample are calculated for each individual array and used for further analysis. A non-parametric analysis method, the Wilcoxon rank sum test, is used for differential analysis.

3.4.3 Results A total of 4690 probes are identified as common to both datasets, among which 661 are reported to be differentially expressed between tumor and non-tumor samples at $P < 0.01$, with 267 retained after false discovery rate adjustment by the step-up method of Benjamini and Hochberg (1995). Hierarchical clustering is performed for the top 30 most significant differential expressed gene sets (Fig. 3). Clustering results show that the samples were separated into two major groups correlating with their biological origin (tumor versus non-tumor) instead of their platforms. In general, discriminative gene sets found in two datasets on different platforms are likely to be more reliably the characteristic of tumor status than the genes obtained from each individual dataset (Warnat *et al.*, 2005).

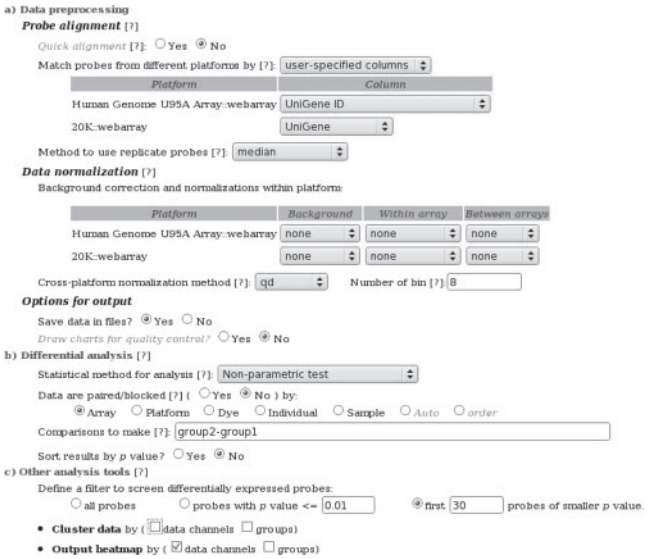

Fig. 2. Options selected in an analysis of two publicly available prostate cancer microarray datasets. See text for details.

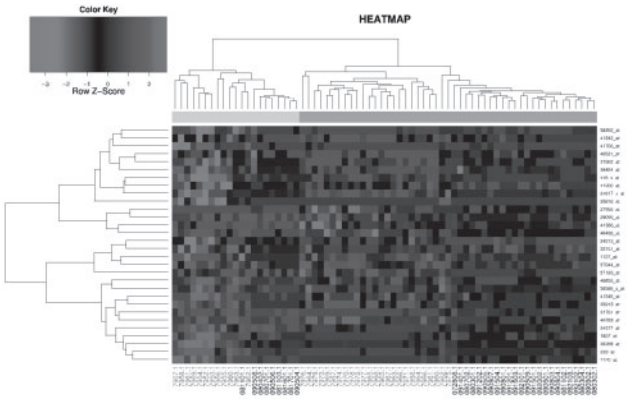

Fig. 3. Heat map of the 30 most significantly differentially expressed probes between tumor and non-tumor samples. The tumor samples are marked at the top of the plot by a brown bar and the non-tumor group by a yellow bar. Arrays of the 20K platform are named in blue font at the bottom of the plot, Affymetrix U95A arrays in black font.

4 IMPLEMENTATION

WebArrayDB has been implemented on a LAMP system (a Linux server with Apache, MySQL and Python) in a typical browser/server model (Fig. 4). In a deployment, the WebArrayDB web server, database server and file server can be located on a single machine or on separate machines. Most modules are written in Python (http://www.python.org), while analysis functions are powered by R language (http://www.r-project.org) (R Development Core Team, 2006) and Bioconductor (Gentleman *et al.*, 2004). Our

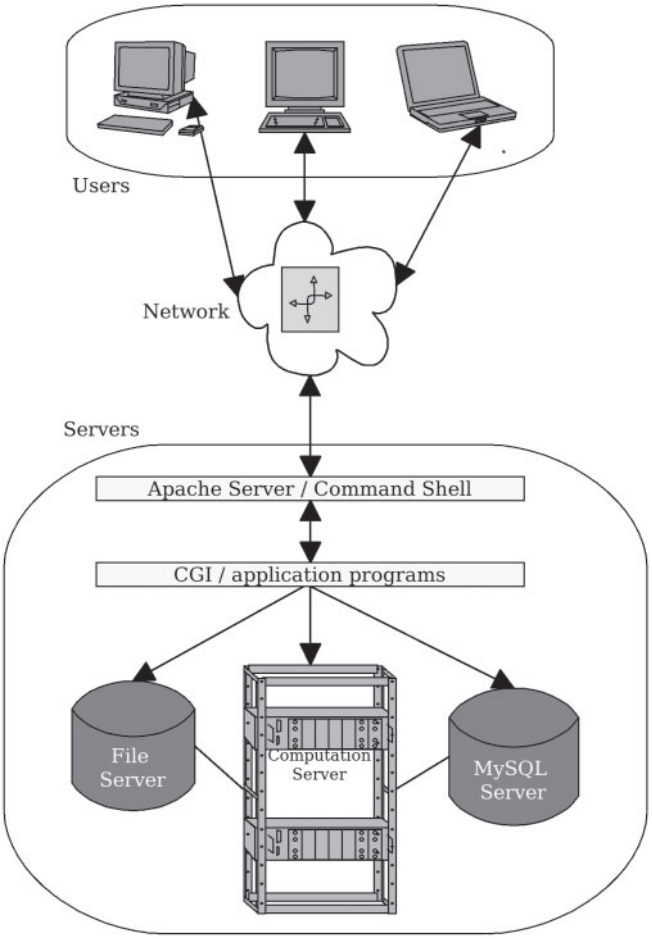

Fig. 4. Architecture of WebArrayDB.

WebArrayDB is hosted on a Dell server with four CPU cores with hyper-threading technology, 24 GB of RAM, 1 TB main hard disk and 1 TB hard disk for backup. The configuration will be upgraded depending on the burdens of computation and increases in the data stored.

Parallel computation can be done at two levels:

- Multiple analysis requests from users can be processed simultaneously. In order to avoid too many active requests, WebArrayDB will automatically determine a maximum number of requests that can be processed simultaneously, limiting both the number per user and the total number, while keeping other requests waiting in the queue. The default values can be adjusted by the administrator.
- Even in a single analysis request, computation can be distributed into many processes that run in parallel. The number of processes can be adjusted by the administrator. The package SNOW (Rossini *et al.*, 2003) was adopted for this purpose, so Message Passing Interface (MPI), Parallel Virtual Machine (PVM) or SOCKET can be used for communication in parallel computation.

Although WebArrayDB is presented as a web server on the internet, a package is downloadable for those who want to build their own dedicated servers with Win32 or POSIX (Portable Operating System Interface) on symmetric multiprocessing (SMP) systems or Linux clusters. WebArrayDB is designed as a lightweight database with a user-friendly web interface facilitating ease of use for bench scientists. Although a curator is always desirable there is no necessity for one. WebArrayDB is an ideal tool for individual researchers, laboratories or small research institutes, to store, share and analyze the microarray data. The installation of the WebArrayDB server and maintenance is likely to require only a few hours of assistance of IT staff.

5 TUTORIAL AND EXAMPLES

A web-based tutorial, presented in English, Chinese and Spanish at the WebArrayDB web site (http://www.webarraydb.org), shows how to upload data and to process a simple example. The input data and analysis results used in the tutorial (simple analysis) and this article (complex cross-platform comparison) are available for viewing by all WebArrayDB users. Analysis methods other than the preselected ones can be chosen for these examples, and results of these changes can be viewed and stored in the user-specific accounts. Thus, all new users have the opportunity to familiarize themselves with the powerful capabilities of WebArrayDB by browsing and editing both the simple and the complex examples in the 'demo' account upon first entry into the system.

ACKNOWLEDGEMENTS

This work was made possible by the generous support of Sidney Kimmel, Ira Lechner, Eileen Haag and Ron Neeley. We also thank Yong Jiang, Krzysztof Studziński, Rocio Canals and Sang-Ho Choi for testing WebArrayDB, and Fred Long for maintaining the server. This work was performed in the laboratory of Michael McClelland.

Funding: Prostate Cancer Foundation and Mary Kay Ash Foundation (in parts); National Institutes of Health (grants R01AI034829, R01AI052237, R01CA68822 and U01CA114810, in parts).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. and Lewis,S. (2002) On ontologies for biologists: the gene ontology–untangling the web. *Novartis Found Symp.*, **247**, 66–80; discussion 80–83, 84–90, 244–52.

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

Bosotti,R. *et al.* (2007) Cross platform microarray analysis for robust identification of differentially expressed genes. *BMC Bioinformatics*, **8** (Suppl. 1), S5.

Brazma,A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.

Bussey,K.J. *et al.* (2003) MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol.*, **4**, R27.

Chen,Q.-R. *et al.* (2008) An integrated cross-platform prognosis study on neuroblastoma patients. *Genomics*, **92**, 195–203.

Churchill,G.A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.*, **32** (Suppl. 2), 490–495.

Demeter,J. *et al.* (2007) The stanford microarray database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res.*, **35**, D766–D770.

Dhanasekaran,S.M. *et al.* (2005) Molecular profiling of human prostate tissues: insights into gene expression patterns of prostate development during puberty. *FASEB J.*, **19**, 243–245.

Gautier,L. *et al.* (2004) affy–analysis of affymetrix genechip data at the probe level. *Bioinformatics*, **20**, 307–315.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Kapushesky,M. *et al.* (2004) Expression profiler: next generation–an online platform for analysis of microarray data. *Nucleic Acids Res.*, **32** , W465–W470.

Killion,P.J. *et al.* (2003) The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD). *BMC Bioinformatics*, **4**, 32.

Larkin,J.E. *et al.* (2005) Independence and reproducibility across microarray platforms. *Nat. Methods*, **2**, 337–344.

Moreau,Y. *et al.* (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet.*, **19**, 570–577.

Pan,F. *et al.* (2006) Integrative array analyzer: a software package for analysis of cross-platform and cross-species microarray data. *Bioinformatics*, **22**, 1665–1667.

Parkinson,H. *et al.* (2007) Arrayexpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*, **35**, D747–D750.

Rossini,A. *et al.* (2003) Simple parallel statistical computing in R. In *UW Biostatistics Working Paper Series, Paper 193*. University of Washington, WA.

R Development Core Team (2006) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Saal,L.H. *et al.* (2002) Bioarray software environment (base): a platform for comprehensive management and analysis of microarray data. *Genome Biol.*, **3**, SOFTWARE0003.

Shippy,R. *et al.* (2004) Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. *BMC Genomics*, **5**, 61.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3.

Smyth,G.K. (2005) LIMMA: linear models for microarray data. In Gentleman,R. *et al.* (eds) *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York, pp. 397–420.

Smyth,G.K. *et al.* (2005) The use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, **21**, 2067–2075.

Stoyanova,R. *et al.* (2004) Normalization of single-channel dna array data by principal component analysis. *Bioinformatics*, **20**, 1772–1784.

Tárraga,J. *et al.* (2008) GEPAS, a web-based tool for microarray data analysis and interpretation. *Nucleic Acids Res.*, **36**, W308–W314.

Troein,C. *et al.* (2006) An introduction to bioarray software environment. *Methods Enzymol.*, **411**, 99–119.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Warnat,P. *et al.* (2005) Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, **6**, 265.

Welsh,J.B. *et al.* (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res.*, **61**, 5974–5978.

Xia,X. *et al.* (2005) WebArray: an online platform for microarray data analysis. *BMC Bioinformatics*, **6**, 306.