Microarray data analysis and mining approaches

Microarray data analysis and mining approaches

Francesca Cordero, Marco Botta and Raffaele A. Calogero

Advance Access publication date 22 January 2008

Abstract

Microarray based transcription profiling is now a consolidated methodology and has widespread use in areas such as pharmacogenomics, diagnostics and drug target identification. Large-scale microarray studies are also becoming crucial to a new way of conceiving experimental biology. A main issue in microarray transcription profiling is data analysis and mining. When microarrays became a methodology of general use, considerable effort was made to produce algorithms and methods for the identification of differentially expressed genes. More recently, the focus has switched to algorithms and database development for microarray data mining. Furthermore, the evolution of microarray technology is allowing researchers to grasp the regulative nature of transcription, integrating basic expression analysis with mRNA characteristics, i.e. exon-based arrays, and with DNA characteristics, i.e. comparative genomic hybridization, single nucleotide polymorphism, tiling and promoter structure. In this article, we will review approaches used to detect differentially expressed genes and to link differential expression to specific biological functions.

Keywords: *transcription; microarray; bioinformatics; statistics; pathways; transcription factors*

INTRODUCTION

Microarray-based transcription profiling is now a consolidated methodology and large-scale microarray studies are becoming a crucial aspect of a new way of conceiving experimental biology. Microarray technology started as two-channel technology [1, 2], i.e. simultaneous hybridization of two different samples performed on the same array. However, single-channel technology, i.e. a RNA sample hybridized on a single array, has more recently become the preferred approach, due to the simpler and flexible experimental design [3, 4].

Within the commercial single channel microarray platforms available on the market, Affymetrix (www.affymetrix.com) is the older, with the largest panel of microarray designed for a variety of different organisms and the higher number of public available data sets (www.ncbi.nlm.nih.gov/geo/; www.ebi.ac.uk/microarray-as/aer/?#ae-main[0]). Affymetrix microarrays are based on chemical synthesis of 25 mer oligonucleotides in 11-5 μm^2 features on glass slides. The high density of oligonucleotides provides adequate space on the chip for use of multiple probes per mRNA transcript. The arrays based on 11 μm^2 features are also called 3′ based expression arrays (3′ IVT arrays) since each transcript is queried by a probe set, made up of 11 probe pairs mapping on 600 bases of the most 3′ end of the transcript. More recently, Affymetrix started the production of arrays based on 5 μm^2 features. The higher density array manufacturing capability enabled the profiling of exon-level expression at the whole-genome scale on a single array (Exon 1.0 ST). In these arrays, each exon is queried by four probes. This technological improvement allowed also the production of arrays where each transcript is queried using 26 probes spread across the full length of the gene (Gene 1.0 ST arrays), providing a more complete and more accurate picture of gene expression than 3′-based expression array designs. A comparison study between 3′ IVT

Corresponding author: Raffaele A. Calogero, Department of Clinical and Biological Sciences, University of Torino, Italy. Tel: +39 0116705417; Fax: +39 0119038639; E-mail: raffaele.calogero@unito.it

Francesca Cordero has a degree in Biological Sciences and she is a Ph.D student in Informatics at Department of Informatics at Torino University, Italy.

Marco Botta has a Ph.D in Computer Science and he an Associate Professor of Computer Science at Department of Informatics at Torino University, Italy.

Raffaele A. Calogero has a degree in Biological Sciences. He is a Professor of Molecular Biology and Bioinformatics at Torino University Medicine School. He is the PI of the Bioinformatics and Genomics Unit at Department of Clinical and Biological Science.

and the new exon platform was recently published [5], indicating that, despite several major technological changes, a high concordance between the two platforms can be observed and the median relative sensitivity is similar in both platforms.

More recently, Illumina (www.illumina.com) has become increasingly popular within the scientific community due to some features of its arrays: long oligonucleotides, probe replication, reduced per hybridization cost, etc. Illumina have created a microarray technology (Bead-Array) based on randomly arranged beads. A specific oligonucleotide sequence is assigned to each *bead type*, which is replicated on the average about 30 times on an array. A series of decoding hybridizations is used to identify every bead [6]. The high degree of replication makes robust measurements for each bead type possible. The BeadChip technology comprises a series of rectangular strips on a slide, each strip containing about 24 000 bead types.

Measured independently by the type of single channel array in use, the main issue in microarray experiment is data analysis and the consequent extraction of biological knowledge. Transcriptome analysis is complicated by multiple factors such as the limited number of possible experiment replications which is always lower than the number of variables, i.e. genes under investigation, or the actual limited knowledge of gene regulation and gene product function. Furthermore, in microarray analysis, it is not possible to identify any specific piece of software that is globally accepted by the scientific community as the gold standard for microarray data analysis. It is clear however that any microarray data analysis can be summarized in four main steps and each step can be completed using different computational tools:

(i) quality control
(ii) data pre-processing:
  - microarray-specific background subtraction
  - experiment-specific background subtraction
  - transcript intensity summary
  - removal of non-significant transcripts
(iii) differential expression detection
(iv) biological knowledge extraction

QUALITY CONTROL

Quality control (QC) is a very important step of microarray analysis. Essentially, qualiy control could be divided in two subareas:

- Detection of array artifacts and outliers.
- Evaluation of the homogeneity of experimental groups.

Furthermore, QC is strongly depended on the microarray platform in use.

In Affymetrix 3′ IVT arrays, each transcript is represented by a probe set made of 11–20 probes pairs. Each probe pair is made of PM and MM probes. One probe designed to perfectly match the target transcript (PM probe) and the other designed to measure the non-specific binding signal of its partner PM probe. The mismatch (MM) probe is identical to its partner PM probe except for the central (13th) nucleotide, which is changed to the complementary base. PMs and MMs are used by the Liu's algorithm [7] to determine whether the transcript of a gene is detected (present) or undetected (absent).

As basic QC, Affymetrix suggests a certain number of checks to be performed at the level of each array (see Affymetrix manual: data_analysis_fundamentals_manual). These checks include:

- Average background and noise, which is a measure of the pixel-to-pixel variation of probe cells on a GeneChip array, (proposed correct range: 20–100).
- The number of probe sets called present relative to the total number of probe sets on the array and replicate samples should have similar percent present values.
- Poly-A RNA spiked-in controls are used to monitor the entire target labelling process and should be all called present.
- Eukaryotic hybridization controls are spiked into the hybridization cocktail, independent of RNA sample preparation, and are thus used to evaluate sample hybridization efficiency on eukaryotic gene expression arrays. They should be called present at least 50% of the time.
- β-actin and GAPDH are used to assess RNA sample and assay quality. Specifically, the signal values of the 3′ probe sets for actin and GAPDH are compared to the signal values of the corresponding 5′ probe sets. The ratio of the 3′ probe set to the 5′ probe set is generally no more than three for the one-cycle assay.

Furthermore, Bioconductor package affyPLM (www.bioconductor.org) allows to perform a Probe Level Model (PLM) fitting. PLM is a model

that is fitted to probe-intensity data. The model is fitted with probe level and chip level parameters on a probe set by probe set basis. In quality control chip level, parameters are a factor variable with a level for each array. The PLM model can be used to plot relative Log expression (RLE) values, which are computed for each probe set by comparing the expression value on each array against the median expression value for that probe set across all arrays. Assuming that most genes are not changing in expression across arrays means ideally most of these RLE values will be near 0. Another QC plot that can be produced using PLM data is normalized unscaled standard errors (NUSE). The standard error estimates obtained for each gene on each array are taken and standardized across arrays so that the median standard error for that gene is 1 across all arrays. This process accounts for differences in variability between genes. An array where there are elevated SE relative to the other arrays is typically of lower quality. Bioconductor packages AffyExpress, affyQCReport also offer the possibility to produce various types of quality controls.

Concerning Illumina arrays BeadStudio allows the generation of a graphical control summary report based on performance of built-in controls (positive/negative hybridization beads, specificity hybridization signals, etc). Bioconductor package beadarray [8] offers some other QC tools. The package has the ability to read the raw data produced from BeadScan. Boxplots, density plots and image plots are generated automatically and summarized in an HTML report and could be used to identify outlier arrays. The limiting issue of this useful package is the large amounts of computer memory required to run these analyses. The lumi package also provides bead level Illumina microarray data analysis. The package covers data input, quality control, variance stabilization, normalization and gene annotation. In particular, the quality control of a LumiBatch object includes a data summary (the mean and standard deviation, sample correlation, detectable probe ratio of each sample) and different quality control plots (boxplots, density plots, pairwise MA or sample correlation).

Furthermore, principal component analysis (PCA) [9] as well as hierarchical clustering [10] can offer a graphical view of the homogeneity of experimental groups and are available in many Bioconductor packages, e.g. oneChannelGUI [11] and lumi.

DATA PRE-PROCESSING

Pre-processing is the process that allows the transformation of the raw fluorescence signal detected by microarray staining into a signal normalized for experimental errors.

The main steps of pre-processing are: background subtraction, experiment normalization, transcript intensity summarization, removal of non-informative and not expressed transcripts.

The first three steps of data pre-processing (i.e. background subtraction, experiment normalization, transcript intensity summarization) are usually combined together in a unique algorithm as in the case of Affymetrix arrays, where the intensities of multiple short probes need to be combined to generate transcript expression level; or in Illumina bead arrays, where the intensities of multiple copies of the same long probe, used to the detect the same transcript, need to be summarized to the average transcript expression level. It should be noted that for the development of summarization algorithms, publicly available dilution experiments as well as spike-in experiments are extremely important as benchmarks in the testing of sensibility and specificity of the algorithms. Both types of benchmark experiments are available for Affymetrix arrays (www.affymetrix.com) whereas for Illumina bead arrays only a few dilution experiments are available [12].

Much has been published on data pre-processing for Affymetrix 3′ IVT array [13], due to the fact that each transcript is described by a group of short 25-mer probes (probe set) and it is necessary to summarize the probe set intensity by taking into consideration various types of noise.

Affymetrix defined an empirical method for summarization of differential expression [14] implemented in the MAS 5.0 software package (www.affymetrix.com/support/technical/technotes/statistical_reference_guide.pdf). The algorithm is based on the specific construction of 3′ IVT Affymetrix arrays (see above) probe set signal is calculated using the One-Step Tukey's Biweight Estimate [14], which yields a robust weighted mean that is relatively insensitive to outliers, even when extreme. The mismatch intensity is used to estimate stray signal. The real signal is estimated by taking the log of the PM intensity after subtracting the stray signal estimate. Stray signal estimate is equal to MM when the MM intensity is lower than the PM intensity. In case of MM values higher than PM values,

an imputed value called change threshold (CT) is used instead of the uninformative MM. If the MM probe cells are generally informative across the probe set except for a few MM, CT is an adjusted MM value based on the bi-weight mean of the PM and MM ratio. If the MM probe cells are generally uninformative, CT is given by a value that is slightly smaller than the PM.

The probe set summary methods currently used widely by the scientific community are mainly model-based methods. These methods model probe set summaries using the information derived from a multi array experiment. The dispersion of the probe set probes in various locations of the array makes the Affymetrix arrays somewhat insensitive to local constructions/hybridization artifacts.

However, an important issue of probe set summarization using 25-mer probes is the definition of the sequence-dependent non-specific hybridization. RMA methodology [15] performs background correction, normalization and summarization in a modular way, but it does not take into account non-specific probe hybridization in probe set background calculation. GCRMA [16] is instead an extension of RMA with a background correction component, which makes use of probe sequence information. More recently, Affymetrix proposed the probe logarithmic error intensity estimate (PLIER) method which produces an improved signal by accounting for experimentally observed patterns in probe behaviour and handling errors at the appropriately low and high-signal values (www.affymetrix.com). Methods such as PLIER and GCRMA, which use model-based background correction, maintain relatively good accuracy without losing much precision. PLIER is also superior to other algorithms in avoiding false positives with poorly performing probe sets [17]. Seo and Hoffmann [17] however highlight the fact that background is a very complex variable and cannot be perfectly estimated. It is therefore not feasible to identify the 'best' probe set algorithm, but this should be defined on the basis of the type of project. A confirmation of the importance to select probe set algorithm on the basis of the experiment type comes from a recent paper [18] shows that MAS5 is the best choice in reverse engineering studies of cellular networks, since a crucial step of GCRMA algorithm is responsible for a systematic overestimation of pairwise correlation, which instead does not affect the detection of differential expression in two and multiple class experiments.

Data pre-processing for Illumina data is relatively simple, mainly because multiple replications of one long oligonucleotide are used to detect the 3′ end of a transcript. The use of long oligonucleotides greatly reduces the non-specific hybridization problem [19] present in Affymetrix arrays. Furthermore, the average 30-fold bead-type redundancy strongly reduces local hybridization artefacts. Pre-processing of Illumina arrays can be performed using Illumina BeadStudio software (www.illumina.com). Bead Studio produces an average value for each bead type on the un-logged scale and provides various normalization and visualization tools. However, loose information is given about replicates of each bead type, data are automatically background corrected and there is no possibility of controlling image processing. Before summarization, BeadStudio detects as outliers all beads of the same type that have an un-logged intensity of more than three median absolute deviations (MAD) and does not include them in intensity summarization. Background is measured for all beads as the mean of the negative controls on an array and is used by BeadStudio software to perform background normalization. Recently, an open source tool [8, 20] allowing the bead-level data handling of Illumina bead arrays improved the flexibility of the Illumina summarization algorithm. In the beadarray package [8] available in Bioconductor [21], the detection of outliers can be done using either un-logged and logged intensities and using a user defined number of MADs. Dunning [20] has also shown that background values for beads are virtually constant within arrays and also across arrays. Local measure of background is equivalent to global value, but background corrected data show much more variability among beads of the same type [20]. This observation therefore suggests the use of the automatic background correction available in BeadStudio be avoided [20]. Log intensity transformation is another part of the pre-processing that can be applied both to Affymetrix and Illumina arrays, and is used to reduce variance and improve precision [20, 22–24]. However, it should be used carefully since the increased precision of log transformation could be at the expense of levels of accuracy [22, 25].

During data generation, numerous factors could alter the outcome through the introduction of systematic biases. Those are mainly linked to the limited

control that the experimenter has on biological objects, i.e. cell cultures, biopsies, reagents, etc., or to the presence of overall disparities of slide surfaces and variation in manufacturing as well as scanner-introduced bias, which influence the RNA quantification process. As a means of identifying and removing systematic biases, data normalization is typically performed.

Global median normalization is usually not recommended since the simple adjustment of the median intensity value within each array does not take into account local intensity bias. Global loess normalization is instead used to address intensity-dependent bias [26, 27].

The loess method was initially proposed by Yang [27]. This approach stems from the M versus A plot, where M is the difference in log expression values and A is the average of those. A normalization curve is fitted to this M versus A plot by using loess, which is a method of local regression. The fits based on the normalization curve are subsequently subtracted from the M values. This method was extended to single channel arrays by Bostald [26]. In his implementation, each array is normalized against all the others for one or two iterations.

A normalization method widely used for single channel arrays is the quantile [26]. The goal of the quantile method is to make the distribution of probe intensities for each of a set of arrays the same. The idea behind the method is that a quantile–quantile plot shows that the distribution of two data vectors is the same if the plot is a straight diagonal line, but not if it is other than a diagonal. This concept can be extended to n dimensions if more than two arrays are available. This suggests that an *n* set of data can be made to have the same distribution by projecting the points of the *n*-dimensional quantile plot onto the diagonal. However, according to very recent results, this normalization method can have an impact on the biological variability and, therefore, appears to be less than optimal from this point of view [28].

A method for the normalization of Illumina bead arrays other than quantile available in the beadStudio is the cubic spline method. This method is similar to the one proposed by Workman [29]. The normalization uses quantiles of sample intensities to fit smoothing B-splines. For each sample, its vector of quantile intensities as well as quantiles for the 'virtual' averaged sample after background subtraction are computed. Cubic B-spline is than computed and used for interpolation. Furthermore, the Bioconductor lumi package supports directly reading of the Illumina Bead Studio toolkit. It contains a variance-stabilizing transformation algorithm that takes advantage of the technical replicates available on every Illumina microarray and a robust spline normalization algorithm, which combines the features of the quantile and loess normalization.

All these methods are based on the assumption that the majority of elements should be not differentially expressed. A recent paper [30] has described a normalization method called orthogonal projections to latent structures (OPLS), which the authors claim to be independent from the previous assumption. This method identifies joint variation within biological samples, allowing the removal of sources of variation that do not correlate with the within-sample variation. This ensures that the structured variation related to the underlying biological samples is separated from the remaining bias-related sources of systematic variation.

A problem in microarray data analysis is the high dimensionality of gene expression space, which prohibits a comprehensive statistical analysis without focusing on particular aspects of the joint distribution of the gene expression levels. A theoretical computation [31] showed that there is an optimal number of hypotheses to be tested which is limited by the number of samples in the experiment. When the proportion of differentially expressed transcripts is small, they tend to get buried among the non-differentially expressed. Possible strategies to overcome this problem are to undertake some kind of biology-driven filtering or to perform signal-driven filtering of genes before the actual statistical analysis [32–34].

Although the integration of biological knowledge is usually associated with the data mining process [35, 36], the use of biology-driven filters could be an ideal choice when the experimenter has a clear idea of which subarea of biology should be investigated at transcription level by microarray analysis. These types of filters are clearly dependent on the availability of biological knowledge and on the robustness of data annotation. The database most used, which links biological information to genes, is Gene Ontology (GO) [37], but many other biology-driven ontologies have become available in recent years [38], increasing the number of biological topics to be used for biology-driven filters.

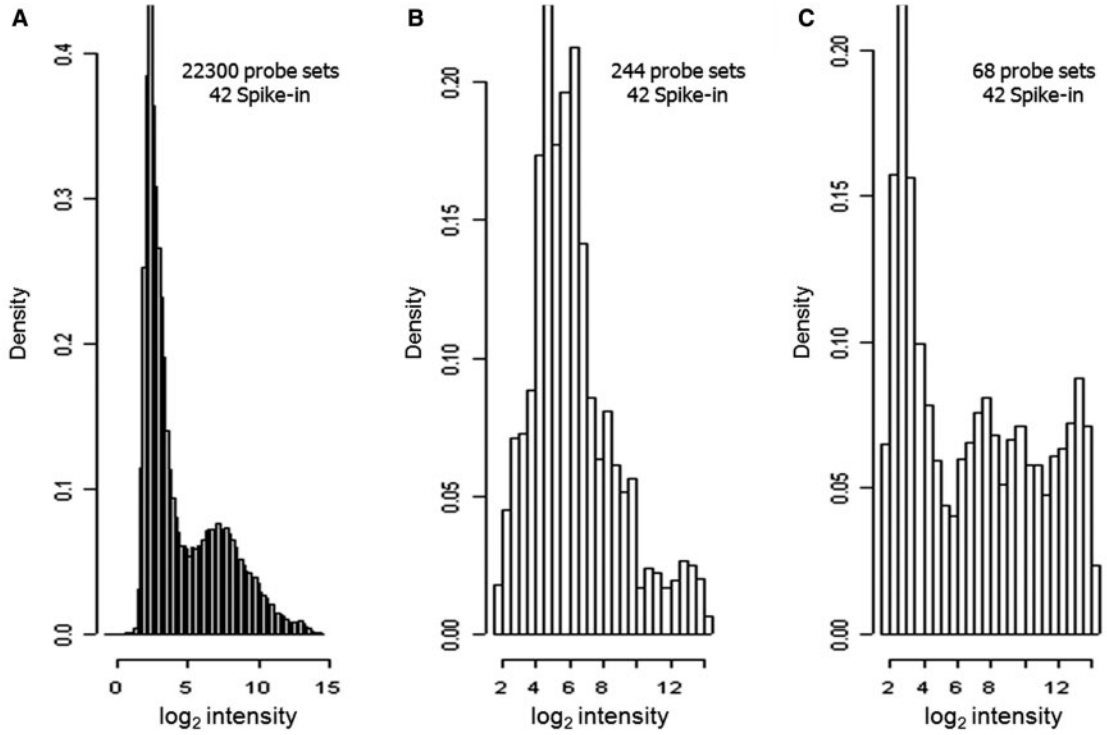

Figure 1: IQR filtering on HGU133A latin square experiments. Forty-two probe sets are spike-in at concentration ranging from 0.125 to 512 pM in a common background. It removes all probe sets which are not characterized by a broad inter-quantile range within the various samples. (**A**) Unfiltered complete set. (**B**) IQR filtering at 0.125. (**C**) IQR filtering at 0.5. This has a tremendous effect in the spike-in experiments, since they are based on a concentration curve from 0 to 512 pM.

Non-informative signals, i.e. those characterized by an expression close to background over all the experimental points, can be detected, in the case of 3′ IVT arrays, by detection (Present/Absent) call algorithm (www.affymetrix.com). This is a p-score that assesses the reliability of each expression level and is produced using a signed rank test to consider the significance of the difference between the PM and MM values for each probe set [7]. This approach could be used to remove data that are not reliably detected, before further analysis [39]. McClintick [39] also observed that the use of a filter based on detection call removes probe sets contributing to a disproportionate number of false positives. Experiment size does however greatly affect the ability to reproducibly detect significant differences, and also impacts on the effect of filtering, i.e. small experiments based on 3–5 samples per treatment group benefit from more restrictive filtering (≥50% present). A similar filtering approach could be applied to Illumina arrays using the detection score, which is given by R/N, where R is the rank of the gene signal relative to negative controls and N is the number of negative controls.

A generally applicable filtering approach called the IQR filter, which eliminates genes that do not show sufficient variation in expression across all samples, as they tend to provide little discriminatory power, was proposed by von Heydebreck [32] and could be used routinely to reduce the number of hypothesis testing [40–43]. This filter is one of the filters implemented in the Bioconductor package genefilter (www.bioconductor.org) and allows the removal of genes that do not show an expression variation over all samples greater than a user defined threshold. The strength of such filtering procedure is palpable when applied to the Affymetrix latin square experiments (Figure 1), where 42 probe sets are spiked-in with concentrations ranging from 0.125 to 512 pM in a common background. This example highlights the fact that invariant transcripts can be easily eradicated, although an important issue of this filtering approach is the homogeneity of the experimental groups. This procedure, if applied to a data set designed to identify cell cycle genes such as Spellman [44], will be inefficient since the vast majority of the genes are characterized by repetitive wave-like fluctuations.

DIFFERENTIAL EXPRESSION DETECTION

Extracting biological information from microarray data requires appropriate statistical methods. Much work has been done to optimize conventional statistical tests to the limited experimental structure usually available in microarray experiments. The main issue in differential expression analysis is the experiment group size, which is always smaller than the number of tests (transcripts) to be investigated. Due to the limited sample size of the majority of experiments involving microarray analysis, statistical tools simply work like a filter that highlights the most significant differentially expressed transcripts, but do not represent the ultimate validation of the differential expression. Transcripts detected by statistical analysis need to get back to the wet laboratory to confirm their differential expression. Furthermore, the integration of different pre-processing steps combined with different statistics does not necessarily detect the same subset of differentially expressed transcripts [45, 46].

Each combination of methods will attain some but not all true signals (Figure 2). At the same time each combination of methods will get some false signals (Figure 2). The trick is to find the best condition to maximize true signals, while minimizing fakes. However, the only way to define the best combination of methods is to know the differential expressed subset of transcripts, which is not known since it is the goal of a differential expression analysis. Using benchmark experiments, however, it is possible to evaluate the performances of different methods [45, 47] in order to identify which method better fits to a specific experimental structure.

The works of Choe [47] and that of Jeffery [45] investigate different aspects of differential expression analysis. Choe compares the performances of two moderated *t*-test statistics, significance analysis of microarrays (SAM) [48] and CyberT [49] in the identification of differentially expressed transcripts in an experiment that resembles the vast majority of microarray experiments designed to highlight the transcriptional events involved in a biological treatment. Jeffery evaluates the performances of SAM, empirical bayes *t*-statistics [50], rank products [51] and other statistics to select meaningful features to be used for classification studies.

SAM is a well-known software within the biological scientific community, in its *t*-statistic a constant value is added to the standard deviation. This constant is called the 'fudge factor' and it is chosen to minimize the dependence of the *t*-statistic variance on standard deviation levels. However, it has been found that SAM does not control well FDR [52–54]. A recent paper of Zhang [55] also indicates that even the most recent improvements in SAM (sam2.20) still produce erroneous and even conflicting results under certain situations.

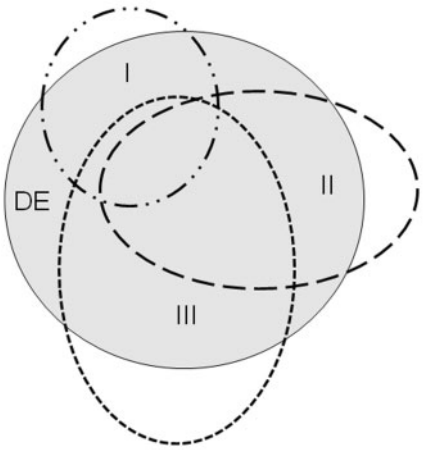

Figure 2: Detection of differential expression integrating different data pre-filtering and statistics. DE: the full set of differentially expressed transcripts associated to a specific biological process. I–III: different integration of pre-filtering and statistical approaches.

CyberT [49] models the standard deviation as a function of signal intensity and its functionality was further enhanced by other linear modelling approaches such as that proposed by Smyth [50] and implemented in the limma Bioconductor package. In particular, CyberT was limited to two-sample control versus treatment designs and its model did not distinguish between differentially and non-differentially expressed genes. Furthermore, CyberT was not characterized by consistent estimators for the hyperparameters [50] and the degrees of freedom associated with the prior distribution of the variances was set to a default value, while the prior variance was simply equated to locally pooled sample variances [50]. Despite these limitations, Choe's paper [47] shows that CyberT performs better than SAM.

The Rank products method [51], implemented in the RankProd Bioconductor package, is based on calculating rank products (RP) from replicate experiments. It is a straightforward and statistically

stringent way to determine the significance level for each gene and allows for the flexible control of the false-detection rate and family-wise error rate in the multiple testing situation of a microarray experiment. RP is more powerful and accurate for sorting genes by differential expression than SAM, in particular with low number of replicates (<10), which are most commonly used in biological experiments [51]. Furthermore, its relative performance is particularly strong when the data are contaminated by non-normal random noise or when the samples are very non-homogenous [56]. RP does however assume equal measurement variance for all genes and tends to give overly optimistic *P*-values when this assumption is violated. It is therefore essential that proper variance stabilizing normalization is performed on the data before calculating the RP values [56]. Where this is impossible, another rank-based variant of RP (average ranks) provides a useful alternative with very similar overall performance [56].

Choe results combined with the description of the limits of SAM and RP suggest that the empirical bayes statistic [50] probably represents the most robust way of identifying differential expression in small experiments designed to have a mechanical view of a biological treatment.

The definition of the best statistical approach could however be different if the task is the features selection for classification purposes. The work of Jeffery [45] highlights the fact that data set characteristics affect the performances of the applied statistics. The empirical bayes statistic represents an accurate way to select features unless datasets have high pooled variance or a low number of samples. In this case, RP has been proved useful.

Although two-sample differential expression analysis is probably the most common experiment, multi-series time-course microarray experiments are useful approaches for exploring biological processes. In these types of experiments, the researcher is frequently interested in studying gene expression changes over time and in evaluating trend differences between the various experimental groups. The large amount of data, multiplicity of experimental conditions and the dynamic nature of the experiments pose great challenges to data analysis. A comprehensive review of research in time series expression data analysis was published by Bar-Joseph in 2004 [57]. Recently, Conesa has published two methods for time-course microarray data analysis [58, 59]. One is maSigPro [59], and is part of Bioconductor packages. This method follows a two-step regression strategy in order to find genes with significant temporal expression changes and significant differences between experimental groups. As a first step, a regression fit for each gene is computed and the *P*-value associated to the *F*-statistic of the model is computed and corrected for multiple comparisons by applying FDR procedure [60]. As a second step, a variable selection procedure [61] to find significant variables for each gene is applied. This will ultimately be used to find what are the profile differences between experimental groups.

The other is ANOVA-SCA [58] and combines ANOVA-modeling and a dimension reduction technique to extract targeted signals from data bypassing structural noise. ANOVA-SCA basically applies PCA to the estimated parameters in each source of variation of an ANOVA model. ANOVA-SCA seems an effective approach for separating the data variability present in a complex time course experiment to extract the signal of interest from noisy data. The selection of significant genes is done by means of two statistics: leverage and squared prediction error (SPE). Leverage is a measure of the importance of a variable (i.e. transcript) in the PCA model and SPE is a measure of the fit of the model for that specific gene. High leverage and low SPE transcripts are transcripts that vary according to the main trend and correspond to major molecular functions affected by the treatment. High SPE transcripts are model diverging data and would correspond to responsive genes with a minority pattern. Low leverage transcripts show low variance and encode functions less specific in the bulk response.

Angelini and coworkers [62] have recently described a fully Bayesian approach to detect differentially expressed genes in time-course experiments. Their approach allows to explicitly use biological prior information and deals with various technical difficulties that arise in microarray time-course experiments such as a small number of observations, non-uniform sampling intervals, missing or multiple data and temporal dependence between observations for each gene. Authors compared their method with that implemented in R-package time course [63] and in the EDGE software [64] claiming that their algorithm provides results which are much closer to a 'biologist's choice' and delivers a lower percentage of false positive and negative answers than other algorithms.

Fischer and coworkers [65] have compared methods for identifying differentially expressed genes on time-series microarray data simulated from artificial gene networks. They suggest the use of ANOVA variants of Cui and Churchill [66] on the bases of simulated data and Efron and Tibshirani's empirical Bayes Wilcoxon rank sum test [67] in the case experimental background cannot be effectively corrected. Shi [68] has instead proposed an approach, based on a probabilistic continuous hidden process model (CHPM), to identify the various biological processes involved in a specific biological experiment. This method integrates time series expression data with GO biological processes, modelling the observed gene expression levels as being generated by a combination of multiple GO biological processes whose activity levels vary over time.

BIOLOGICAL KNOWLEDGE EXTRACTION

Extracting clear and coherent hypotheses from genome-wide expression data remains an important challenge. Much of the initial work has focused on the development of techniques for accurate identification of differentially expressed genes and their statistical significance in a variety of experimental designs. However, the main difficulty in analysis lies not in the identification of differentially expressed genes but in their interpretation. Attempting to understand individual genes on a list of significant genes is demanding and laborious. The problem is compounded when the pathway of interest involves moderate effects that are not captured by the genes near the top of the list. Recent efforts have therefore focused on the discovery of biological pathways rather than individual gene function, with the development of methods that can withstand the inaccuracies of specific gene estimates and provide a more expansive view of the underlying processes.

Pathway analysis

Hosack [46], showed that prevalent biological themes within the set of differentially expressed transcripts derived from the same experiment, but using different transcript selection methods, are a stable representation of the biology underlying the experiment. Therefore, even though differentially expressed transcript lists have only partial overlap [46] they all represent subsets of transcripts associated to a specific biological event (Figure 2).

A much used database for the functional annotation of transcription profiling is the GO [37]. GO is however marked by flaws of certain characteristic types, due to a failure to address basic ontological principles [69, 70]. This problem has been recently at least partially overcome thanks to the availability, as commercial data mining databases, of highly structured knowledge ontologies. Some of the databases are produced by automatic extraction of biological knowledge by means of text mining algorithms, e.g. Ariadne Genomics' PathwayStudio (http://www.ariadnegenomics.com/), others are mainly based on manual curating, e.g. Ingenuity (www.ingenuity.com). The strength of databases such as Ingenuity is not the availability of new statistical methods or proprietary graphical algorithms to depict the relation between functional pathways and differentially expressed transcripts, but the availability of manually curated and fully traceable data derived from primary literature sources.

Routinely, both over- and under-representation of ontology terms can be detected using the standard hypergeometric test [71]. In probability theory and statistics, the hypergeometric distribution is a discrete probability distribution that describes the number of successes in a sequence of n draws from a finite population without replacement. The test based on the hypergeometric distribution is identical to the corresponding one-tailed version of Fisher's exact test. Reciprocally, the *P*-value of a two-sided Fisher's exact test can be calculated as the sum of two appropriate hypergeometric tests. Even though ontology enrichment approaches are widely used, only the most significant portion of the gene list is used to compute their statistic. Furthermore, the order of genes on the significant gene list is not taken into consideration. As a result simply counting the number of gene set members contained in the short list leads to loss of information, especially if the list is long and the difference between the more significant and the less significant is substantial. Finally, the correlation structure of gene sets is not considered at all [72]. More recently, Alexa [73] proposed a conditional hypergeometric test that computes the significance of a GO term based on its neighbourhood. Using the classical approach in which each node is scored independently, only few true significant nodes remain undiscovered. However, the dependencies between top scoring nodes yield a high false-positive rate. Alexa introduced the possibility of weighting genes annotated to a GO

term based on the scores of neighboring GO terms or iteratively removing genes mapped to significant GO terms from more general (higher level) GO terms. The conditional hypergeometric test based on GO terms weightings reduces the false-positive rate, while not missing many true enriched nodes. The other conditional test is more efficient in finding the important areas in the GO graph, it also further reduces the false-positive rate, but with a higher risk of discarding relevant nodes.

A different use of GO is that applied in the SemSim Bioconductor package, which allows the estimation of information content-based similarity scores of GO terms and gene products [74–76]. GO-based semantic similarity scores can be used to perform annotation-based clustering as described by Wolting [77]. Furthermore, the availability of methods like simUI and simLP in the GOstats package [71], which allow the estimation of similarity between lists of differentially expressed genes derived by the induced GO graphs, can be extremely useful to detect the presence of common regulative pathways in meta-analysis experiments made in different laboratories and/or different microarray platforms and biological models.

The integration of transcription profiles with biology knowledge bases (e.g. GO, KEGG, PUBMED, etc.) is another way of mapping differentially expressed transcripts in specific biology knowledge domains. A coordinated change among many gene products can produce potent biological effects, while the effect of each individual transcript can be subtle. The identification of pathways distinctively enriched within a set of differentially expressed transcripts can also be subsequently used to check if more subtle transcriptional variations, not considered in the stringent differential expression analysis, could also be used to strengthen the biological mean of the identified pathway. Another possible application could be the link of alternative splicing events, detected with the new exon-oriented Affymetrix microarray platform, to functional pathways depicted by conventional differential expression analysis.

Two of the most used statistics to evaluate the association between functional pathways and differential expression are the one-tailed Fisher exact test, (FET) [46, 78, 79] and Gene Set Enrichment Analysis (GSEA) [80]. FET is a statistical significance test used in the analysis of categorical data where sample sizes are small. The test is used to examine the significance of the association between two variables in a 2×2 contingency table. GSEA on the other hand evaluates microarray data at the level of *gene sets*. The gene sets are defined based on prior biological knowledge, e.g. published information about biochemical pathways or co-expression in previous experiments. The goal of GSEA is to determine whether members of a gene set S tend to occur toward the top (or bottom) of the list L, in which case the gene set is correlated with the phenotypic class distinction. GSEA acts through three steps:

(i) Calculation of an enrichment score.
(ii) Estimation of significance level of enrichment score.
(iii) Adjustment for multiple hypothesis testing.

Since an accurate and rapid identification of perturbed pathways through the analysis of genome-wide expression profiles facilitates the generation of biological hypotheses, Tian [81] proposed a statistical framework for determining whether a specified group of genes for a pathway has a coordinated association with a phenotype of interest. In this framework, the overall objective of the analysis is to test whether a group of genes has a coordinated association with a phenotype of interest evaluating the following two null hypothesis:

(i) The genes in a gene set show the same pattern of associations with the phenotype compared with the rest of the genes.
(ii) The gene set does not contain any genes whose expression levels are associated with the phenotype of interest.

After the test statistics are computed for testing the two hypotheses gene sets are then ranked in order of their significance and a control for the inflated Type I error due to multiple comparisons of gene sets is also applied. The authors claimed that their approach has more statistical power than currently available methods and can result in the discovery of statistically significant pathways that are not depicted by other methods [81].

Markowetz proposed an algorithm to infer non-transcriptional pathway features based on differential gene expression in silencing assays [82]. The author's idea is that cellular signalling pathways, which are not modulated on a transcriptional level, cannot be directly deduced from expression profiling

experiments. However, when external interventions occur i.e. RNA interference or gene knock-outs, even if the expression of the signalling genes is not changed, secondary effects in downstream genes shed light on the pathway, and allow partial reconstruction of its topology. The core of Markowetz' approach is the definition of a scoring function, which measures how well hypotheses about pathway topology are supported by experimental data.

Promoter analysis

Since microarray data produce a representation of the effect of a specific treatment on the transcriptional machinery, it is extremely important to link the transcriptional output signals, measured by microarray analysis, to promoter elements common to a subset of differentially expressed genes.

In order to perform this association, it is necessary to grasp the hidden structure of eukaryotic promoters. Here, we summarize the characteristics of some of the available methods for transcription-binding site identification.

The computational discovery of regulatory elements is humanly possible because they occur several times in the same genome and because they may be evolutionarily conserved among different species. This means that novel regulatory elements may be discovered by searching for overrepresented motifs across regulatory regions [83]. This apparently simple approach is complicated by the fact that most transcription factor binding sites (TFBSs) are short, and they can have some variation without loss of function. Therefore, most motifs are also found as random hits throughout the genome, and it is a challenging problem to distinguish between false positive hits and true positive binding sites. Motif finding is essentially a signal-to-noise problem. It has been estimated that in human DNA about 3% of inter-genic regions are regulatory elements [84]. For this reason, most algorithms to identify the genomic regulatory elements use orthogonal data. Several algorithms include additional prior knowledge about gene regulation; regulatory elements are not randomly distributed, but tend to form clusters of regulatory modules [85], and the presence of co-occurring motifs can be used to identify putative regulatory modules. Functional sequences are preferentially conserved over the course of evolution by selective pressure. This is another characteristic, along with over-representation, applied by Corà [86] to determine TFBSs in the human genome. The hypothesis that many orthologous genes are expressed similarly in a tissue-specific manner in human and mouse and are likely to be co-regulated by orthologous transcriptional factors (TF) is the base of the *cis*-regulatory regions search [87].

Usually, the TFBSs are represented by a 'consensus sequence'. Consensus sequence has been widely used to represent the specificity of TF. However, the consensus sequence is not flexible enough to account for all variations: in general, it refers to a sequence that matches all of a site closely, but not necessarily exactly [88]. An alternative to consensus sequence is a position weight matrix (PWM) or profile. The PWM summarizes the statistical properties of a collection of TF binding sites and represents the DNA sequences. The PWM is the formalism to represent DNA motifs bound to a particular TF because it contains two kinds of knowledge: the thermodynamic interactions between TF and DNA and the evolutionary selection [89]. The underlying assumptions are that natural selection gave rise to a certain level of sequence specificity for each TF and that sequences that gave rise to the same physically binding affinity are equally likely to be selected [90]. A new algorithm to build PWM was implemented by Foat [91], MatrixREDUCE that uses genome-wide occupancy data for TF (e.g. ChIP-chip). A microarray measurement of TF occupancies and relevant nucleotide sequences for each microarray feature are used as input to MatrixREDUCE. The algorithm performs a least-squares fit to a statistical–mechanical model of TF–DNA interaction, in order to discover the relative contribution to the free energy of binding for each nucleotide at each position in the generalized TF binding site. The measure of significance for the PWM is commonly given by information content of Equation (1), IC, also called relative entropy [92]:

$$I(p) = \sum_{j=1}^{L} \sum_{i=A}^{T} f_{i,j} \log \frac{f_{i,j}}{P_i} \qquad (1)$$

where p is a pattern, L is the pattern length, i is the index of a base at position j of the PWM, $f_{i,j}$ is the frequency of the base i at position j of the PWM, and P_i is the probability of observing that base in the data. The IC is the weighted average for the binding energies from each of the sites represented in the matrix, the lower the IC, the higher the variability in the site [93].

Currently, there are two comprehensive and annotated databases that contain information on

TFs binding site profiles. JASPAR [94] contains a smaller set that is non-redundant (each TF has only one profile), while TRANSFAC [95] contains multiple profile models for some TFs. The discovery of motifs in sequence data was an early problem to be addressed in computational biology. The DNA motif discovery algorithms that have been developed can be divided into three main groups:

- Complete *ab initio* methodologies: parameter-free algorithms for *de novo* identification of potential TFBS. This group contains all methodologies that implement a simple search for the most probable subsequence in a set of sequences. In this case, there are no assumptions about the biological features of the sequences.
- Partial *ab initio* methodologies: algorithms that assume some biological knowledge. There are two categories of algorithms: the first contains algorithms that use 'complementary information' (see below), while the second contains algorithms which assume that the found subsequences are possible TFBS, and describes a sequence motif by means of a position-specific scoring matrix.
- Matrix-based methodologies: algorithms detect potential TFBS by a sliding window search, with one specific PWM, of a match subsequences.

An example of a complete *ab initio* methodology is Weeder [96]. This algorithm extends the exhaustive enumeration of signals without giving as input the exact length of the patterns to be found. Each motif is evaluated according to the number of sequences in which it appears and how well it is conserved in each sequence with respect to expected values derived from the oligo frequency analysis of upstream sequences in the same organism. The algorithm then compares the top-scoring motifs of each run with a clustering method to detect which ones could be more likely to correspond to a TFBS. The consensus for a set of TFBSs can be seen as a perfect form recognized by a TF. The algorithm then enumerates all the possible oligos of the same length of the motif to be found. For each one, it counts how many times it appears in the sequences. The sequences that are overrepresented form a new set of sequences. It then ranks the motifs found according to some statistical measure and gives as output the highest-ranking motifs.

Another algorithm in this category is Yeast Motif Finder (YMF), written by Sinha [97]. YMF uses an exhaustive search algorithm to find motifs with the greatest z-score. The z-score of a motif is the number of standard deviations by which its observed number of instances in the actual input sequences exceeds its expected number of instances.

Both algorithms do not need any input parameter. With many parameters to set, the user explores the parameter space and makes arbitrary judgment calls on which output to trust. Different studies have shown the programs to be quite sensitive to parameters [98].

However, the algorithms that used 'complementary information', as overrepresented in evolutionarily conserved upstream regions or infer about co-regulation (GO and results of a set of microarray experiments), improve the signal/noise ratio by selecting for analysis those portions of the upstream regions that are more likely to be functionally relevant [86]. These methodologies are grouped in the 'Partial *ab initio*' set. An example is the algorithm by Caselle [99], where the genome is grouped in sets based on words that are overrepresented in the upstream region, and their frequencies in the reference sample are then compared to the whole genome. For each of these sets, they compare the average expression in microarray experiments with the genome-wide average. If the difference is statistically significant, the set is a putative TFBS. Other examples in the 'Partial *ab initio*' set are algorithms that used a different type of 'complementary information'. One example is Consensus. This algorithm employs a greedy heuristic [100] and builds up an entire alignment of the sites by adding in a new one at each iteration. The best alignment of a potential site is the one with highest information content. The goal of Consensus is then to determine a sequence alignment that maximizes log-likelihood statistics described in a PWM. An expectation-maximization (EM) method was implemented in the MEME program [101]. MEME method allows for the simultaneous identification of multiple patterns, the starting point derived from each subsequence occurring in the input sequences. For every subsequence, the algorithm evaluated the quality and the accuracy of the statistical significance by a product of the P-value of column information contents. In the latter two algorithms, the basic assumption is that the sequences that are overrepresented in the genome are putative TFBSs; they then consider the alignments for every motif as a starting point on which to build a PWM.

The third group is a set of methodologies which search for the presence of a PWM in all sequence positions using a sliding window approach. One example is MatInspector [102]: this algorithm detects potential sequence matches by automatic searches with a library of pre-compiled matrices. The search method includes position weighting of the matrices based on the information content of individual positions and calculates a relative matrix similarity. Another example is Patser [103]. This algorithm computes the numerical estimation of the *P*-value of the match score between a subsequence and a specific matrix. The *P*-value is the probability of observing a particular score at a particular sequence position. The motif with the highest *P*-value is a putative TFBS.

Tools allowing the integration of microarray data with promoter structure information have been developed [104–109]. The software developed by Kel [104, 106] is commercially available (Explain software, www.biobase.de) and uses a genetic algorithm to predict relevant promoters in a set of given transcripts obtained from microarray analysis, taking advantage of the promoter element matrix database TRANSFAC [110–112]. Werner software [107, 108] is also a commercial tool where promoter elements are identified using MatInspector [113]. Tamada [109] instead has developed a statistical method for estimating gene networks and detecting promoter elements simultaneously. This method integrates microarray gene expression data and the DNA sequence information into a Bayesian network model. The basic idea of the method is that, if a parent gene is a TF, its children may share a consensus motif in their promoter regions of the DNA sequences. The method detects consensus motifs based on the structure of the estimated network and then re-estimates the network using the result of the motif detection.

Although these data mining tools could enable a better comprehension of the complex mechanism of regulation associated to transcription profiling, it should be pointed out that their main limits are related to the quality of the promoter level annotation. A statistic comparing the accuracy of the main tools to discover TFBSs is found in Tompa [114], but it is very difficult to compare the performance of methods, in particular on complex genomes like the human genome.

If sufficient *a priori* knowledge is available, it is possible to reconstruct the gene target network for at least one TF using the method proposed by Barenco [115], which is based on a mathematical technique known as hidden variable dynamic modelling (HVDM). This approach is based on a simple differential equation model that uses hidden information to partially reconstruct, with confidence intervals, the TF target network. The HVDM takes advantage of prior biological knowledge to create a training set of genes, the behaviour of which can be used to derive the activity profile of the controlling TF. The method needs quite a lot of input information, thus rendering its use not easily generally applicable:

(i) Expression time course microarray data, consisting of at least five time points.
(ii) Some prior biological knowledge about the TF under review, e.g. at least three genes in the training set, should be known to be targets of that TF and presumed to be targets of that TF only.
(iii) The transcript degradation rate of one of the known targets, measured in an independent experiment.
(iv) The technical measurement error for each expression value should be known.

CONCLUSIONS

In this review, we have touched upon some of the approaches used for microarray data analysis. The numerical analysis of microarray is now considerably consolidated and when new methods appear they mainly allow for a refinement of the numerical data. However, the true integration of numerical analysis and biological knowledge is still a long way off. The main reason for this lack of integration is the low amount of functional gene annotation and the difficulties of the integration of the massive amount of biological data which are daily published by the scientific community. A further critical issue in high eukaryotes data integration is due to data heterogeneity, which manifests itself in multiple tiers of the biological information base and is a major barrier to progress in the fundamental understanding of biological processes; an example being that biological results produced in *in-vitro* models (e.g. immortalized cell lines) are very useful for investigating specific biological events but they are not representative of the global behavior of a gene in different tissues.

Key Points

- No gold standard software exists for microarray data analysis, but a fixed pipe-line is generally used: quality control, data pre-processing, differential expression detection and biological knowledge extraction.
- Samples size is a critical issue in microarray experiments since it affects the statistical power of the experiment. Researchers should use large number of experimental replicates especially in gene networking studies.
- Functional annotation is still a weak point in microarray analysis. Gene Ontology represent a nice prototype of computer searchable ontology. However, GO has some weak points, e.g. electronic based annotation. Commercial Databases, e.g. Ingenuity, have at least partially overcome GO limitations, but the data mining instruments applicable on these databases are more limited with respect to those available for GO.
- The integration of microarray data and promoter structure is still very rudimentary. Expression data need to be integrated with high-throughput chromatin immune precipitation data to grasp robust functional knowledge.

Acknowledgement

This work was supported by grants from Italian Association for Cancer Research; the Italian Ministero dell'Università e della Ricerca; the University of Torino; and the Regione Piemonte.

References

1. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* 1996;**6**:639–45.
2. Schena M, Shalon D, Davis RW, *et al*. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;**270**:467–70.
3. Vinciotti V, Khanin R, D'Alimonte D, *et al*. An experimental evaluation of a loop versus a reference design for two-channel microarrays. *Bioinformatics* 2005;**21**:492–501.
4. Altman NS, Hua J. Extending the loop design for two-channel microarray experiments. *Genet Res* 2006;**88**:153–63.
5. Abdueva D, Wing MR, Schaub B, *et al*. Experimental comparison and evaluation of the Affymetrix exon and U133Plus2 GeneChip arrays. *PLoS ONE* 2007;**2**:e913.
6. Gunderson KL, Kruglyak S, Graige MS, *et al*. Decoding randomly ordered DNA arrays. *Genome Res* 2004;**14**:870–7.
7. Liu WM, Mei R, Di X, *et al*. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* 2002;**18**:1593–9.
8. Dunning MJ, Smith ML, Ritchie ME, *et al*. beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics* 2007;**23**:2183–4.
9. Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput* 2000;455–66, available at: http://www.bepress.com/sagmb/
10. Eisen MB, Spellman PT, Brown PO, *et al*. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998;**95**:14863–8.
11. Sanges R, Cordero F, Calogero RA. oneChannelGUI: a graphical interface to Bioconductor tools, designed for life scientists who are not familiar with R language. *Bioinformatics* 2007.
12. Barnes M, Freudenberg J, Thompson S, *et al*. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res* 2005;**33**:5914–23.
13. Irizarry RA, Wu Z, Jaffee HA. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* 2006;**22**:789–94.
14. Hubbell E, Liu WM, Mei R. Robust estimators for expression analysis. *Bioinformatics* 2002;**18**:1585–92.
15. Irizarry RA, Bolstad BM, Collin F, *et al*. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003;**31**:e15.
16. Wu Z, Irizarry RA. Preprocessing of oligonucleotide array data. *Nat Biotechnol* 2004;**22**:656–8; author reply 658.
17. Seo J, Hoffman EP. Probe set algorithms: is there a rational best bet? *BMC Bioinformatics* 2006;**7**:395.
18. Lim WK, Wang K, Lefebvre C, *et al*. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics* 2007;**23**:i282–8.
19. Hughes TR, Mao M, Jones AR, *et al*. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* 2001;**19**:342–7.
20. Dunning M. Quality control and low-level statistical analysis of Illumina BeadArrays. *Revstat* 2006;**4**:30.
21. Gentleman RC, Carey VJ, Bates DM, *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;**5**:R80.
22. Inoue M, Nishimura S, Hori G, *et al*. Improved parameter estimation for variance-stabilizing transformation of gene-expression microarray data. *J Bioinform Comput Biol* 2004;**2**:669–79.
23. Durbin BP, Rocke DM. Variance-stabilizing transformations for two-color microarrays. *Bioinformatics* 2004;**20**:660–7.
24. Rocke DM, Durbin B. Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics* 2003;**19**:966–72.
25. Seo J, Gordish-Dressman H, Hoffman EP. An interactive power analysis tool for microarray hypothesis testing and generation. *Bioinformatics* 2006;**22**:808–14.
26. Bolstad BM, Irizarry RA, Astrand M, *et al*. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003;**19**:185–93.
27. Yang MC, Ruan QG, Yang JJ, *et al*. A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. *Physiol Genomics* 2001;**7**:45–53.
28. Boes T, Neuhauser M. Normalization for Affymetrix GeneChips. *Methods Inf Med* 2005;**44**:414–7.
29. Workman C, Jensen LJ, Jarmer H, *et al*. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* 2002;**3**:research0048.
30. Bylesjo M, Eriksson D, Sjodin A, *et al*. Orthogonal projections to latent structures as a strategy for microarray data normalization. *BMC Bioinformatics* 2007;**8**:207.

31. Futschik A, Posch M. On the optimum number of hypotheses when the number of observations is limited. *Stat Sinica* 2005;**15**:841–55.
32. von Heydebreck A, Huber W, Gentleman R. Differential expression with the Bioconductor Project. Bioconductor Project Working Papers 2004, available at: http://www.bepress.com/bioconductor/
33. Pounds S, Cheng C. Statistical development and evaluation of microarray gene expression data filters. *J Comput Biol* 2005;**12**:482–95.
34. Calza S, Raffelsberger W, Ploner A, *et al*. Filtering genes to improve sensitivity in oligonucleotide microarray data analysis. *Nucleic Acids Res* 2007;**35**:e102.
35. Brown MP, Grundy WN, Lin D, *et al*. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* 2000;**97**:262–7.
36. Bellazzi R, Zupan B. Towards knowledge-based gene expression data mining. *J Biomed Inform* 2007.
37. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.
38. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. *Brief Bioinform* 2006;**7**:256–74.
39. McClintick JN, Edenberg HJ. Effects of filtering by present call on analysis of microarray experiments. *BMC Bioinformatics* 2006;**7**:49.
40. Bosotti R, Locatelli G, Healy S, *et al*. Cross platform microarray analysis for robust identification of differentially expressed genes. *BMC Bioinformatics* 2007;**8**(Suppl 1):S5.
41. Spugnini EP, Cardillo I, Verdina A, *et al*. Piroxicam and cisplatin in a mouse model of peritoneal mesothelioma. *Clin Cancer Res* 2006;**12**:6133–43.
42. Lo Iacono M, Di Costanzo A, Calogero RA, *et al*. The Hay Wells syndrome-derived TAp63alphaQ540L mutant has impaired transcriptional and cell growth regulatory activity. *Cell Cycle* 2006;**5**:78–87.
43. Olivero M, Ruggiero T, Saviozzi S, *et al*. Genes regulated by hepatocyte growth factor as targets to sensitize ovarian cancer cells to cisplatin. *Mol Cancer Ther* 2006;**5**:1126–35.
44. Spellman PT, Sherlock G, Zhang MQ, *et al*. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 1998;**9**:3273–97.
45. Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* 2006;**7**:359.
46. Hosack DA, Dennis G, Jr, Sherman BT, *et al*. Identifying biological themes within lists of genes with EASE. *Genome Biol* 2003;**4**:R70.
47. Choe SE, Boutros M, Michelson AM, *et al*. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol* 2005;**6**:R16.
48. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001;**98**:5116–21.
49. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 2001;**17**:509–19.
50. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004;**3**: Article3.
51. Breitling R, Armengaud P, Amtmann A, *et al*. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* 2004;**573**:83–92.
52. Delmar P, Robin S, Daudin JJ. VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics* 2005;**21**:502–8.
53. Xie Y, Pan W, Khodursky AB. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics* 2005;**21**:4280–8.
54. Guo X, Pan W. Using weighted permutation scores to detect differential gene expression with microarray data. *J Bioinform Comput Biol* 2005;**3**:989–1006.
55. Zhang S. A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance. *BMC Bioinformatics* 2007;**8**:230.
56. Breitling R, Herzyk P. Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *J Bioinform Comput Biol* 2005;**3**:1171–89.
57. Bar-Joseph Z. Analyzing time series gene expression data. *Bioinformatics* 2004;**20**:2493–503.
58. Nueda MJ, Conesa A, Westerhuis JA, *et al*. Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA. *Bioinformatics* 2007;**23**:1792–800.
59. Conesa A, Nueda MJ, Ferrer A, *et al*. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics* 2006;**22**:1096–102.
60. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 2003;**19**:368–75.
61. Draper N, Smith H. *Applied Regression Analysis*. New York: Wiley, 1998.
62. Angelini C, De Canditiis D, Mutarelli M, *et al*. A Bayesian approach to estimation and testing in time-course microarray experiments. *Stat Appl Genet Mol Biol* 2007;**6**:Article 24.
63. Tai Y, Speed TP. A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann Stat* 2006;**34**:2387–412.
64. Leek JT, Monsen E, Dabney AR, *et al*. EDGE: extraction and analysis of differential gene expression. *Bioinformatics* 2006;**22**:507–8.
65. Fischer EA, Friedman MA, Markey MK. Empirical comparison of tests for differential expression on time-series microarray experiments. *Genomics* 2007;**89**:460–70.
66. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* 2003;**4**:210.
67. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. *Genet Epidemiol* 2002;**23**:70–86.

68. Shi Y, Klustein M, Simon I, *et al*. Continuous hidden process model for time series expression experiments. *Bioinformatics* 2007;**23**:i459–67.
69. Kohler J, Munn K, Ruegg A, *et al*. Quality control for terms and definitions in ontologies and taxonomies. *BMC Bioinformatics* 2006;**7**:212.
70. Park YR, Park CH, Kim JH. GOChase: correcting errors from Gene Ontology-based annotations for gene products. *Bioinformatics* 2005;**21**:829–31.
71. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. *Bioinformatics* 2007;**23**:257–8.
72. Pavlidis P, Qin J, Arango V, *et al*. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res* 2004;**29**:1213–22.
73. Alexa A, Rahnenfuhrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 2006;**22**: 1600–7.
74. Schlicker A, Domingues FS, Rahnenfuhrer J, *et al*. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* 2006;**7**:302.
75. Lord PW, Stevens RD, Brass A, *et al*. Semantic similarity measures as tools for exploring the gene ontology. *Pac Symp Biocomput* 2003;601–12, available at: http://www.bepress.com/sagmb/
76. Lord PW, Stevens RD, Brass A, *et al*. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 2003; **19**:1275–83.
77. Wolting C, McGlade CJ, Tritchler D. Cluster analysis of protein array results via similarity of Gene Ontology annotation. *BMC Bioinformatics* 2006;**7**:338.
78. Grosu P, Townsend JP, Hartl DL, *et al*. Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks. *Genome Res* 2002;**12**:1121–6.
79. Pandey R, Guru RK, Mount DW. Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics* 2004;**20**: 2156–8.
80. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**:15545–50.
81. Tian L, Greenberg SA, Kong SW, *et al*. Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci USA* 2005;**102**:13544–49.
82. Markowetz F, Bloch J, Spang R. Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics* 2005;**21**:4026–32.
83. Sandve GK, Drablos F. A survey of motif discovery methods in an integrated framework. *Biol Direct* 2006;**1**:11.
84. Kellis M, Patterson N, Endrizzi M, *et al*. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 2003;**423**:241–54.
85. Kreiman G. Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucleic Acids Res* 2004;**32**:2889–900.
86. Cora D, Herrmann C, Dieterich C, *et al*. Ab initio identification of putative human transcription factor binding sites by comparative genomics. *BMC Bioinformatics* 2005;**6**:110.
87. Huber BR, Bulyk ML. Meta-analysis discovery of tissue-specific DNA sequence motifs from mammalian gene expression data. *BMC Bioinformatics* 2006;**7**:229.
88. Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics* 2000;**16**:16–23.
89. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* 1987;**193**: 723–50.
90. Bussemaker HJ, Foat BC, Ward LD. Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu Rev Biophys Biomol Struct* 2007;**36**: 329–47.
91. Foat BC, Morozov AV, Bussemaker HJ. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* 2006; **22**:e141–9.
92. Schneider TD, Stormo GD, Gold L, *et al*. Information content of binding sites on nucleotide sequences. *J Mol Biol* 1986;**188**:415–31.
93. GuhaThakurta D. Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res* 2006;**34**:3585–98.
94. Sandelin A, Alkema W, Engstrom P, *et al*. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 2004;**32**:D91–4.
95. Matys V, Fricke E, Geffers R, *et al*. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 2003;**31**:374–8.
96. Pavesi G, Mauri G, Pesole G. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* 2001;**17**(Suppl 1):S207–14.
97. Sinha S, Tompa M. YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* 2003;**31**:3586–8.
98. Hu J, Li B, Kihara D. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res* 2005;**33**: 4899–913.
99. Caselle M, Di Cunto F, Provero P. Correlating over-represented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes. *BMC Bioinformatics* 2002;**3**:7.
100. Stormo GD, Hartzell GW, 3rd. Identifying protein-binding sites from unaligned DNA fragments. *Proc Natl Acad Sci USA* 1989;**86**:1183–7.
101. Bailey TL, Gribskov M. Methods and statistics for combining motif match scores. *J Comput Biol* 1998;**5**: 211–21.
102. Quandt K, Frech K, Karas H, *et al*. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* 1995;**23**:4878–84.
103. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 1999;**15**:563–77.
104. Kel A, Konovalova T, Waleev T, *et al*. Composite module analyst: a fitness-based tool for identification of transcription factor binding site combinations. *Bioinformatics* 2006; **22**:1190–7.

105. Waleev T, Shtokalo D, Konovalova T, *et al*. Composite module analyst: identification of transcription factor binding site combinations using genetic algorithm. *Nucleic Acids Res* 2006;**34**:W541–5.

106. Kel A, Voss N, Jauregui R, *et al*. Beyond microarrays: finding key transcription factors controlling signal transduction pathways. *BMC Bioinformatics* 2006; **7**(Suppl 2):S13.

107. Werner T. Target gene identification from expression array data by promoter analysis. *Biomol Eng* 2001;**17**:87–94.

108. Werner T. Cluster analysis and promoter modelling as bioinformatics tools for the identification of target genes from expression array data. *Pharmacogenomics* 2001;**2**: 25–36.

109. Tamada Y, Kim S, Bannai H, *et al*. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics* 2003;**19**(Suppl 2):ii227–36.

110. Wingender E, Chen X, Fricke E, *et al*. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res* 2001;**29**:281–3.

111. Wingender E, Dietze P, Karas H, *et al*. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 1996;**24**:238–41.

112. Wingender E, Chen X, Hehl R, *et al*. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* 2000;**28**:316–9.

113. Cartharius K, Frech K, Grote K, *et al*. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* 2005;**21**:2933–42.

114. Tompa M, Li N, Bailey TL, *et al*. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005;**23**:137–44.

115. Barenco M, Tomescu D, Brewer D, *et al*. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biol* 2006;**7**:R25.