Microarray data integration: frameworks and a list of underlying issues

Chintanu Kumar Sarmah*, Sandhya Samarasinghe

Chintanu Kumar Sarmah is a PhD student at Lincoln University, Canterbury, New Zealand. (phone: + 64-3-3218377; fax: +64-3-3253845; e-mail: sarmahc3@ Lincoln.ac.nz).

Sandhya Samarasinghe is an Associate Professor from Centre for Advanced Computational Solutions (C-fACS) at Lincoln University.

[* Corresponding author]

***Abstract* —** Microarray technology is expanding rapidly providing an extensive as well as promising source of data for better addressing complex questions involving biological processes. The ever increasing number and publicly available gene expression studies of human and other organisms provide strong motivation to carry out cross-study analyses. Besides, microarray technology provides several platforms to investigators that include arrays from commercial vendors like *Affymetrix*® (Santa Clara, CA, USA), *Agilent*® (Palo Alto, CA, USA), and other proprietorial arrays of various laboratories. Integration of multiple studies that are based on the same technological platform, or, combining data from different array platforms carries the potential towards higher accuracy, consistency and robust information mining. The integrated result often allows constructing a more complete and broader picture.

In this work, we highlight as well as exemplify two frameworks of microarray data integration approaches that are in practice. This follows a discussion on the important issues that may influence any microarray data integration attempt. The review, in general, intends to serve as a starting point for those interested in exploring this area of microarray study, while realizing the pertinent issues underneath.

***Keywords*—** Data integration, issues, microarrays, microarray technology, review.

I. Introduction

The methods to measure gene expression were revolutionized by Kary Banks Mullis's invention of the *in vitro* technique, *polymerase chain reaction* or PCR [1, 2] in 1985 that awarded him Nobel prize for Chemistry in 1993. While application of PCR-approaches could detect the expression of one gene within one reaction or to a maximum of a few genes in optimised state, high-throughput analysis of higher number of genes was time consuming requiring a lot of technical and personal power. In 1995, two seminal investigations, Schena et al. [3] and Smith et al. [4], led by Patric O. Brown of the *Howard Hughes Medical Institute*, ushered in the era of gene-expression microarray analysis and revolutionized the field of molecular biology. The technique of microarrays, which started off with simultaneous gene expression analysis of 45 genes within one experiment, now provides technological and conceptual advancement through its high throughput capability of simultaneously interrogating the RNA expression of the whole genomes.

In the relatively few years since its inception, microarray technology has improved dramatically, and becomes a widely used tool for studying global gene expression. Currently, it provides several platforms to researchers to carry out their investigations. In addition to arrays from commercial vendors like *Affymetrix*® (Santa Clara, CA), *Agilent*® (Palo Alto, CA), there are also proprietorial arrays used by various laboratories. Overall, microarrays have gained increasing use and acceptance to address a myriad of complex biological questions involving genetic and cellular processes. The increasing acceptance of microarrays is clearly demonstrated by the increasing number of citations appearing yearly in published literature. The researchers are particularly overwhelmed by the microarray-based methods as it confers the freedom to conduct large-scale gene expression profiling measurements giving rise to the resultant wealth of potentially valuable information within a very short span of time.

With the increase of the collection of microarray data especially in MIAME [5]-compliance public repositories such as *ArrayExpress*[i] [6], *GEO*[ii] [7], CIBEX[iii] [8], a growing number of investigators are looking at meaningful extraction of information by integration of various microarray experiments. As microarray study tends to explore specific areas of biological function, integration of data from multiple microarray experiments is considered to allow construction of a more complete as well as a broader biological picture. Integrated microarray data is potentially beneficial in several other ways including that it can compensate for the possible errors of individual experiments, amplify the sample-size, and can potentially lead to higher accuracy, consistency and robust information mining.

Integration of microarray studies can include integration of studies that are based on the same technological platform. Researchers around the world also combine data from different array platforms based on their needs. However, integration of microarray data remains to be a challenging problem as data from different investigations do not become readily comparable due to factors that can be attributed to biological and technical causes associated with the generation of these data [9, 10]. Nevertheless, with the accumulating amount of important microarray data generated from various microarray experiments, many investigators have taken up the challenging task of meaningful integration of microarray data as well as of overcoming the barriers of microarray platform-dependency in order to improve our understanding of biological processes, medical conditions, and diseases. In this paper, we highlight as well as exemplify two frameworks of microarray data integration. Besides, we have also listed the important issues that can influence any attempt of microarray data integration. Overall, this review is expected to serve as a starting point for those interested in exploring the area of microarray data integration. The article is also likely to be a 'one-stop shop' to the readers enabling them to realize and be aware of a range of possible underlying issues in the microarray data integration-pipeline.

II. Microarray data integration

Microarray technology has become an indispensable tool for monitoring genome wide expression levels of genes in a given organism. From the Patric Brown's lab, the technology has evolved representing both a technological and a conceptual advance of the field, and has expanded worldwide, where many laboratories are now making their own arrays, in addition to the availability of various commercial vendors. With the increasing number and availability of gene expression studies of various organisms, there has been a pressing need to develop approaches for integrating results across multiple studies.

In a cross-study analysis, the data, relevant results and statistics of several studies are combined. There are different practical advantages in such studies. Cross-study analysis has the potential to strengthen and extend the results gathered from the individual studies. This can turn an investigation towards higher accuracy and consistency, and thus, help in robust information mining. Moreover, output of such a study can provide a broader picture of gene-expression as the final 'integrated'-result emerges based on a set of individual studies. Cross-study analysis can also compensate for the possible data-errors in individual studies. The cost of such a study can be kept low by using the existing studies, as otherwise the setting up of each microarray investigation is not inexpensive. However, while attempting to carry out integration of microarray studies, there are much higher challenges and difficulties as genetic expressions of different studies are neither readily comparable nor can directly be combined.

There are several approaches to cross-study analysis, and they somewhat broadly fall into two categories – A. studies where integration occurs at the interpretative level, B. studies

[i] http://www.ebi.ac.uk/arrayexpress
[ii] http://www.ncbi.nlm.nih.gov/geo/
[iii] http://cibex.nig.ac.jp/index.jsp

where integration takes place with rescaling of the expression values.

A. Integration at the interpretative level

Meta-analysis is emerging as a standard way for the comparison of microarray studies at interpretative level. It involves comprehensive reanalysis of the primary data by merging data from multiple studies. Certain general reviews on meta-analysis include Hedges & Olkin [11], Cook et al. [12], Normand [13], Ghosh et al. [14] and Moreau et al. [15]. As broadly defined by Normand [13], meta-analysis is the quantitative review and synthesis of the results of related but independent studies. Despite having certain demerits of merged primary dataset as reviewed by Larsson et al. [16], the method is becoming useful in microarray studies with the expansion of the sheer volume of microarray data. The success of meta-analysis is dependent on the quality of the underlying data. When accuracy of one or more concerned microarray platforms is questionable, the outcome may become influenced. Nevertheless, browsing through the various studies, where the observation on accuracy, reliability and reproducibility of microarray platforms clearly ranges from relatively discouraging [17, 18] through cautiously optimistic [9, 19] to impressive [20, 21], the overall assessment of the usefulness of meta-analysis of similar microarray studies is cautious optimism. Moreover, the major sources that contribute to the discordance in this regard are mainly – random noise, biological and experimental variations in the samples being analysed, and the variation due to the technical methodology used in the platforms. It is possible to overcome the discordance to a greater extent with judicious and robust application of relevant statistical methods, standard reporting methods, as well as careful application of meta-analysis techniques.

The core objectives of meta-analysis are to increase efficiency in detecting an overall treatment effect, to estimate degree of benefit associated with a particular study, and to assess the amount of variability between studies etc. In the recent past, several statistical methods aiming at detecting differentially expressed genes among multiple conditions have been proposed in individual experiments [22, 23, 24, 25]. Pan [26] has published a comparative review on these statistical methods in replicated microarray experiments. However, most standard meta-analysis methods cannot be applied directly to microarray experiments as microarray technology is unique with its slew of issues, including its diverse experimental platforms, complicated data structures, presence of duplicate spots as well as often having a large number of genes tested for differential expression.

In 1925, a simple application of meta-analysis was implemented as Fisher's Inverse χ^2 test [27]. The method computes a combined statistic from the *P*-values obtained from the analysis of the individual datasets, $S = -2 \log(\Pi_i P_i)$. Here, S follows a Chi-square distribution with $2l$ degrees of freedom under the joint null-hypothesis. The approach does not require additional analysis, and is easy to use; however, it cannot estimate the average magnitude of differential expression in microarrays just by working with the p-values. The approach also remains highly dependent on the method used in the individual analysis.

Meta-analysis based on the t-statistic was reviewed by Normand [13] in the context of biostatistical applications. Choi et al. [28] adopted the classic biostatistical meta-analysis framework for microarray analysis, and implemented their methods as a Bioconductor [29]-package, *GeneMeta*[iv]. The approach of Choi et al. [28] was a model-based systematic integration of microarray datasets, where a hierarchical modeling approach to assess intra- and inter-study variation was used. The method estimated an overall effect size as the measure of differential expression for each gene through parameter estimation and model fitting. The effect size was a *t*-like statistic, which was the summary statistic for each gene from each individual dataset, and was defined to be a standardized mean difference between cancer and normal samples in a microarray data set. Integration of data using this meta-analysis method promoted the discovery of small but consistent expression changes and increased the sensitivity and reliability of analysis. Later, Hong and Breitling [30] found that this t-based meta-analysis method greatly improved over the individual analysis, however it suffered from potentially large amount of false positives when P-values served as threshold.

Based on the traditional effect size model [28], Hu et al. [31] proposed a model for implementing an efficient methodology for identifying genes that are differentially expressed between lung *adenocarcinoma* samples and normal samples by modeling the effect size and integrating information from two Affymetrix oligonucleotide studies. In this study, they presented a measure to quantify Affymetrix gene chip data quality for each gene in each study where the quality index measured the performance of each probeset in detecting its intended target. They extended the traditional effect size model by using the quality index as a weight for combining information from different Affymetrix chip types, and incorporating this weight into a random-effects meta-analysis model.

Rhodes et al. [32] proposed a statistical model for performing meta-analysis in their four prostate cancer microarray datasets, two of which were cDNA (also known as, *spotted arrays*) data and the remainder Affymetrix microarray data. The model was based on the statistical confidence measure rather than the expression levels, while avoiding direct comparisons of data sets and related cross-platform normalization issues. Each gene in each study was treated as an independent hypothesis, and significance was assigned based on random permutations. Then a meta-analysis model was implemented to assess the similarity of the findings between studies based on multiple inference statistical test for each possible combination of studies. This ultimately identified statistically reliable sets of over- and under-expressed genes in prostate cancer. A cohort of genes were

[iv] www.bioconductor.org/packages/bioc/html/GeneMeta.html

found to be consistently and significantly dysregulated in prostate cancer. The approach of Rhodes et al. is highly conservative because of the choice of null hypothesis; and therefore, the approach may not be recommendable. The data used by Rhodes et al. [32] were later used by Choi et al. [28], and they demonstrated that their method could lead to the discovery of small but consistent expression changes with increased sensitivity and reliability.

A Bayesian mixture model transformation of microarray data was proposed by Parmigiani et al. [33]. The modeling framework was used for molecular classification, and it provided both a statistical definition of differential expression and a precise, experiment-independent, definition of a molecular profile. It also generated natural similarity measures for traditional clustering and gave probabilistic statements about the assignment of tumors to molecular profiles.

The rank product is a non-parametric statistic, and was first proposed to detect differentially expressed genes in a single dataset [22]. To integrate multiple microarray studies from different platforms and/or different laboratories, a rank product meta-analysis algorithm was implemented as a Bioconductor package, *RankProd* [34]. The algorithm computed pairwise fold change (FC) with replicates for each gene between treatment and control in both directions, respectively. Then, it transformed FC into rank among all genes under study, searched for genes that were consistently top ranked across replicates, and finally generated a single significance measurement for each gene in the combined study. In this approach, converting FC into ranks increased robustness against noise and heterogeneity across studies.

Grutzmann et al. [35] performed a meta-analysis of four independent studies that applied high-density arrays for expression profiling of pancreatic cancer. They used a consensus set of UniGene clusters measured in all four studies, and applied a random effect model described by Whitehead & Whitehead [36], whereby expected values of individual study effects were assumed to be normally distributed. With the random effect model, an unbiased estimator for the PDAC (*Pancreatic ductal adenocarcinoma*) effect across all studies was measured, and was used to measure joint differential expression of a gene across all studies.

With three publically available breast cancer datasets having information on lymph node status, Garrett-Mayer et al. [37] compared the strength of evidence of gene–phenotype associations as well as combined effects across studies. For this, the three studies were first analyzed for reliability, and then, the comparability of results with regards to the genes associated with lymph node status was assessed. Instead of actually combining the data across studies, they mainly performed a comparative analysis making inferences based on the genes consistently measured in all studies, and finally estimated combined inferential statistics. Their proposed methods were implemented in the R [38]-library, *MergeMaid*[v] [39]. The novel addition in this work was the use of a reliability measure, which was extended to be applied for more than two studies.

Meta-analyses methods are useful; however, as Eysenck [40] mentioned, they require careful selection of inclusion criteria for participating studies and sound statistical models to avoid misleading conclusions. To date, a broader comparison across various integration approaches is not available. However, Hong and Breitling [30] compared performance of three widely used methods - Fisher's inverse Chi-square approach, t-like statistic of Choi et al. [28] and rank product method [22, 34], and found that the non-parametric rank-product method outperformed in terms of sensitivity and specificity.

In general, the overall framework used in all the above studies, where data integration occurs at the interpretative level can be outlined as shown in Fig. 1.

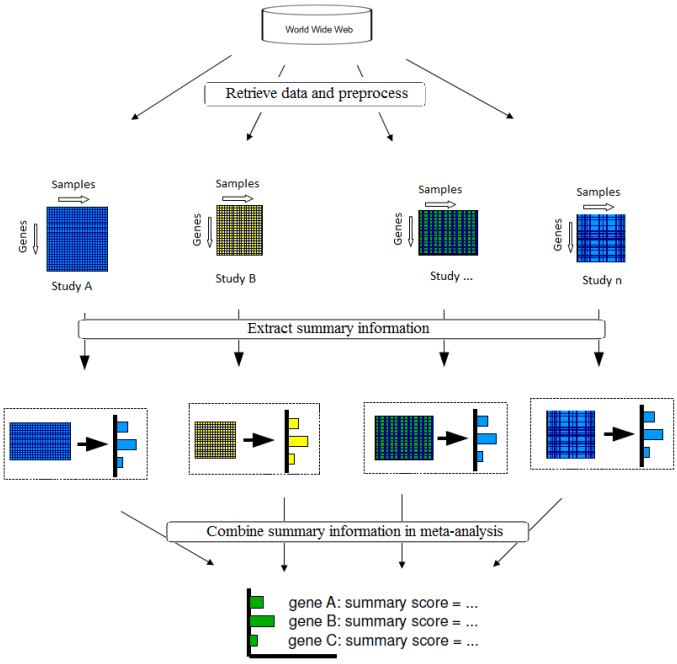

Fig. 1 Outline to the stages of microarray data integration at interpretative level

B. Integration with rescaling of the expression values

Contrary to the meta-analysis approaches, where the results of the individual studies are combined at an interpretative level, there are published researches where microarray expression data from various studies are integrated after transforming the expression values to numerically comparable measures. This is attained by deriving the genetic expression values from the individual platforms, and then, applying specific data transformation and normalization methods. The derived data from the individual studies are subsequently combined, which enlarges the sample size. Any further analysis, as required, is carried out on the new merged dataset. The cross-referencing of the genes between the platforms is usually achieved using UniGene database [41].

Ramaswamy et al. [42] reported rescaling of gene-

[v] http://astor.som.jhmi.edu/MergeMaid/

expression values of a common set of genes. The set of the common genes were from five microarray datasets generated by individual labs on different microarray platforms. The rescaled common genes were combined to produce a larger set of data. From the combined dataset, a gene expression signature was identified, which distinguished primary from metastatic tumors.

A standard normalization scheme can be used to combining cDNA and Affymetrix data. Hwang et al. [43] normalized the expression values of each gene across the samples for each platform so that the mean of each gene equals to zero and the standard deviation equals to unity, respectively. The normalized data were, then, combined to form a large dataset. Earlier, Cheadle et al. [44] proposed normalization and standardization of cDNA microarray intensity values within datasets using a *Z-score transformation* method. The method converted the raw intensity data from each experiment into $\log_{10}$, and then, Z-scores were calculated by the classical method, i.e., by subtracting the overall average gene intensity (within a single experiment) from the raw intensity data for each gene, and dividing that result by the standard deviation of all of the measured intensities. The application of this classical method in microarray normalization provided a way of standardizing data across a wide range of experiments, while allowing comparison of microarray data independent of the original hybridization intensities.

Based on the *distance weighted discrimination* (DWD) method of Marron & Todd [45], Benito et al. [46] integrated cDNA data with Agilent oligonucleotide data. DWD, which was basically an improvement method for *Support Vector Machines* in HDLSS (*High Dimension*, *Low Sample Size*) contexts, was used as an approach for removing systematic bias effects and then, merging the different data sets.

A gene-specific scaling factor was calculated in Bloom et al. [47], and was used to integrate microarray data from Affymetrix and cDNA platforms. Here, for each gene common to both platforms, expression levels for a reference RNA sample on the spotted arrays was averaged and compared to expression measured for the reference RNA sample on the appropriate Affymetrix *GeneChip* to calculate the scaling factor. This scaling factor was used to adjust the remaining data towards integrating the platforms.

Shen et al. [48] used a *two-stage Bayesian mixture modeling strategy* based method proposed by Parmigiani et al. [33]. This model was to integrate multiple independent studies addressing similar questions while considering different platforms – Affymetrix and *inkjet* oligonucleotides. The mixture modeling approach reportedly unified disparate gene expression data based on a probability scale of differential expression, the *poe*-scale [33], and derived an inter-study validated 90-gene 'meta-signature' that predicted relapse-free survival in breast cancer patients.

In addition to common data transformation and normalization procedures, Jiang et al. [49] added a distribution transformation (*disTran*) step in their study. The method transformed two microarray datasets belonging to two Affymetrix chip types so that the empirical distributions of two lung cancer datasets could become identical and be combined. The *disTran* method reportedly provided improved consistency in the expression patterns of the multiple datasets.

Two data integration methods, namely quantile discretization (QD) and median rank scores (MRS) were used in Warnat et al. [50] for direct integration of raw microarray data from six publicly available cancer microarray gene expression studies conducted by means of cDNA and oligonucleotide microarrays. In this study, comparable measures of gene expression from the independent data sets of the varied microarray platforms were numerically derived such that the different microarray data adhere to a common numerical range. These derived data were then integrated, and used to build SVM (support vector machine) classifiers for cancer classification. Similar to *disTran*, the quantile normalization technique, i.e., MRS, and QD of Warnat et al. [50] were used to transform the microarray data from diverse platforms so that their empirical distributions are identical. The approaches (*disTran*, MRS and QD) can significantly improve the comparability of cross-platform microarray data. These methods work well for classification tasks, but can suffer from information reduction, limiting their applicabilities other than classification.

Stafford & Brun [51] presented a calibration process for cross-laboratory and cross-platform microarray expression data. Using Agilent and Affymetrix expression platforms, they employed precision and sensitivity measurements along with biological interpretation for better selection of genes with respect to a particular outcome. Precision and sensitivity measurements were useful in finding the minimal detectable fold-change and raw performance values for a microarray platform. Gene Ontology and pathway analyses were considered in the study as a valuable way of examining and comparing the actual biological interpretation.

Xu et al. [52] used four independent breast cancer datasets, and identified a structured prognostic signature consisting of 112 genes organized into 80 pair-wise expression comparisons. They extended a previously proposed method [53], validated on a prostate cancer study, to predict distant metastases in breast cancer. The method of integration was based on the ranks of the expression values within each sample. Since the ranks of the features were invariant to all types of within-array preprocessing, there was no need to prepare the data for integration, in particular there was no need for data normalization.

XPN [54] is another method that deals with the problem of cross-study normalization: how to combine microarray datasets in order to produce a single, unified dataset to which standard statistical procedures can be applied. The method was based on a block linear model, and used three existing breast cancer datasets from Agilent oligonucleotide platform and Affymetrix GeneChip. The model assumed that the samples of each available study fell roughly into one of the statistically homogenous sample groups identified, and that each group was defined by an associated gene profile that was constant within each of the estimated gene groups. The proposed method applied sample standardization and gene median

centering before combining the data from the studies. To identify blocks (or, clusters) in the data, *k-means clustering* was applied independently to genes and samples of the combined data. Each gene expression value subsequently became a scaled and shifted block mean plus noise. XPN was reportedly preserved biological information according to ER (error rate) prediction error rates while removing systematic differences between platforms.

NLT or Normalized Linear Transform [55] is a method in which the samples of two microarray platform were linearly mapped such that the numerical range of the expression values of each gene became identical. The mapped data were, then, combined and normalized across samples to zero mean and unity standard deviation. Apparently, the approach avoids information reduction as it preserves the relative ranking order of the expression values for each gene.

The methods highlighted above pose important examples of integration of microarray datasets with rescaling of the gene expression values. Each of the approaches is unique; however, their overall stages follow a general framework as outlined in Fig. 2. A user should also be aware that all the methods do not support small sample size. In an attempt to extend a study of Sarmah et al. [56] where the authors used 7-Affymetrix and 7-cDNA samples, only DWD method [46], out of DWD, *poe* [33, 48] and XPN [54], could successfully integrate the results. It is also overall found during the current study that an elaborate comparative analysis on the performance of the individual microarray data integration approaches is still awaited.

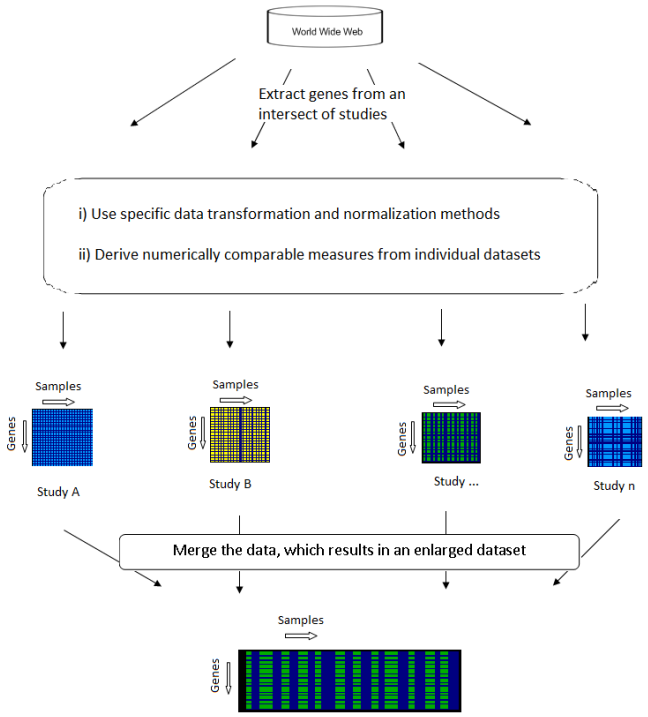

Fig. 2 Outline to the stages of microarray data integration with re-scaling of expression values

III. ISSUES WITH INTEGRATION OF MICROARRAY DATA

The microarray integration studies are valid only if the basic underlying issues are considered as well as addressed with proper attention during the conduct of these studies. We are discussing here a delineation of issues that are important to be reminded of while achieving throughput from such integration efforts.

A. Diversified platforms and noise

This is of fundamental knowledge that there is a range of available microarray platforms; besides, various laboratories also make their own arrays to carry out investigations. Stears et al. [57] provides a list of commercial microarray vendors. With the increasing number and accessibility of gene expression studies of various organisms, each available platform of this technology serves as a genomic readout along with unique characteristics that offer advantages and/or disadvantages in a given context. Further, due to the nature of the technology, noise is an inescapable phenomenon, which can infiltrate at any stage during the process. Draghici [58] compiles a list of major sources of noise. Basic understanding about the platforms, the associated noise and their effects can prove to be useful during any type of microarray investigation including data integration.

B. Preliminary data treatment

There is always some cleaning of the raw, individual datasets before proceeding to the core of the analysis. The cleaning usually extracts or enhances meaningful data characteristics. The goal of cleaning the data, with or without normalization, should be to retain the most accurate data rather than finding the most significant result. The attempt to find the latter can come any time after the end of this process.

C. The excluded genes – are they important ?

As commonly observed in the direct integration approaches where merging of microarray data is done with the rescaling of the gene expression measures, the approaches consider a consensus set of genes from the concerned studies. That way, some genes are excluded from the analysis pipeline right after the preliminary stage. Selecting and reassessing the excluded genes at some stage during the process of study may be important to ascertain that potentially significant genes have not been left out from the analysis.

D. Hunt for the best gene-selection method !

There is no gold-standard of methods which is completely fool-proof, and can efficiently select only the important genes from microarray studies. Therefore, a certain probability of error in such selection always remains as the strict methods would tend to leave out certain significant genes while others would leave room for the unimportant genes to appear in the experiment diluting the successive analysis stages. An example of it is to keep a high (strict) or a relaxed threshold

while using gene selection methods such as *fold change* (a method where an arbitrary ratio relating the expression level of a gene under control and experimental conditions is considered significant) or *unusual ratios* (method of selecting genes certain standard deviations away from the mean log-ratio). A high threshold would find a small number of genes as important, whereas a relaxed threshold would select a high number of genes. Thus, it is important to take extra care in the selection of genes so that an optimum number of important genes can be retrieved.

E. Microarray data repositories

Various microarray data repositories have been established worldwide to share microarray data. With every passing year, there is increase of repositories. However, the reliability of the data quality in each of these repositories is not secured – some repositories are undoubtedly of high quality, but it is doubtful whether the same applies to all available repositories. This makes it important that an investigator verifies her/his data quality beforehand so that once the data are integrated, the combined output gives accurate as well as meaningful results. Further, open source software packages, such as *Microarray Retriever* or *MaRe* [59], *WebArrayDB* [60], and A-MADMAN [61], have also started coming along that facilitate searching and data retrieval from public microarray repositories.

F. Considering appropriate data

Many times, the data in public repositories tend to act as convenience samples for the investigators. This makes the data to be combined inappropriate for the question(s) being asked. For example, while studying causes of a disease, if the data come from people already affected by that disease, it may not make sense because the resultant study may only reveal effects of the disease. Again, while studying a particular disease, it may not be worthwhile to use data from a different disease type or from healthy cases, unless there are reasons behind doing so. Thus, it is critical that selection of data must be carefully considered, and the selected data are appropriate for the question(s) being asked.

G. Data analysis

This is essential to bear in mind that publically available datasets of any two microarray investigations could have passed through different levels of data transformations including background correction, probeset summarization and normalization. All these lead to several combinations of possibilities, and high concordance in the outputs of these transformations is rarely found. In the MAQC project [62], effects of summarization and normalization strategies have been evaluated. Further, Gagarin et al. [63] reports 30% concordance in the differentially expressed gene lists while applying two different summarization methods in the same dataset.

H. Gene nomenclature

Any two microarray studies can be impossible to compare as their results may be reported in different gene nomenclatures. Results of microarray investigations can be reported in many ways, such as using *Genbank* [64], *Entrez Gene* [65], *EMBL* [66], *Unigene* [67], *RefSeq* [68], *OMIM* [69] and Affymetrix gene identifiers. Without using translation tools, like *DAVID* [70], *GoMiner* [71], *RESOURCERER* [72], *L2L* [73] and *LOLA* [74], it is very difficult to compare microarray studies.

I. Probe contents

The probe contents of different microarray platforms are varied. The probes for cDNA platform are often obtained from cDNA libraries. The cDNA libraries are common, but there are concerns about annotation, clone identity, and probe performance [75]. Commercially available oligonucleotide libraries are gaining acceptance as the annotation and identity of oligonucleotides in these libraries are reliable, and the hybridization characteristics of oligonucleotides are generally good [76].

J. Use of different splice-variants as probes:

Various tools, as mentioned earlier, help in translation between different nomenclatures. However, different microarray platforms use different splice forms of the transcripts. Ideally, we must know all the relevant splice-variants of the transcripts along with quantification of the sensitivity and specificity of the probes for nomenclature translation between different platforms. However, in practical, it is so far not yet completely achievable. MAQC project [62] shows that it could only cross-reference 12,091 transcripts between all of the major platforms, although some array platforms interrogate over 54,000 transcripts.

K. Time of sample collection

It is important to consider the time of collection of the tissue-sample. For studying a disease or treatment, the time when the tissue sample has been collected can alter the results significantly.

L. Biological questions

In normal circumstances, the common metaphor, ‘compare apple to apple’ holds relevance in microarray data integration in the sense that it is only reasonable to integrate microarray experiments that aim to address the same or similar biological questions.

M. Variations in samples

Variation in samples used may result in the variation in genes identified by different studies. Many times, similar investigations use different cell lines, tissues or specimens. On the other hand, even slight variation in the experimental conditions may influence investigations on identical cell lines.

All these can be influential in any microarray data integration attempt.

N. Organismal variation

There are gene expression variations from individual to individual within any population. Pritchard et al. [77] provides an example of such population variance where despite identical experimental parameters, the same organs from isogenic mice show distinct and detectable variability in expression. This form of organismal variation can only be examined by multiple repeats at the organismal level (i.e., multiple mice for each biological group). Thus, RNA from different tumors, each subjected to a microarray analysis, provides more information than replicate arrays run on a single batch of RNA extracted from either a single tumor or pooled tumors.

O. Time-series and steady state data

The steady state and time-series experiments are different. In the former, a snapshot of the expression of genes in different samples is measured while a temporal process is measured in time-series expression experiments. Moreover, while steady state data from a sample population are assumed to be independent identically distributed, time-series data exhibit a strong autocorrelation between successive points. Any attempt to combine these two types of microarray experiments must, therefore, have reason(s) to do so.

P. Biological factors

While focusing on microarray data integration, it is necessary to bear in mind that the results of microarray experiments cannot be trusted entirely. This is a technology which, in most cases, works at the mRNA level, and therefore, remains distanced from many underlying mechanisms. For example, in most cases, the microarrays measure the amount of mRNA specific to a particular gene as it is based on the premise that the expression level of the gene is directly proportional to its amount of mRNA. But it is not always true that the amount of mRNA accurately reflects the amount of protein. And, even if it is assumed that it does, a protein may require post-translational modification(s) to become active and perform its role in a cell.

Q. Software tools

There are several software tools available for microarray data analysis. Information on various software tools is available from sources like *SMD*[vi], Mark Fontenot's *Microarray Software List*[vii] at *Southern Methodist University*, Texas and Dresen et al. [78]. A few of the available tools are most widely used and comprehensive open source systems, such as the statistical analysis tools written in R [38] through the Bioconductor project [29], the TM4 software system [79] available from *The Institute for Genomic Research* (TIGR; Rockville, MD, USA), and *BioArray Software Environment* [80] developed at Lund University, Sweden (http://base.thep.lu.se). An useful review on microarray software tools is by Dudoit et al. [81]. Further, the open source web-applications, such as *WebArrayDB* [60], A-MADMAN [61], *microarray retriever* [59], are also available for retrieval and analysis of microarray data. Overall, there are several options available for a user to investigate microarray data. The downside of it, however, is that different software packages or tools may, at times, generate varying results for essentially the same analysis.

[vi] SMD or, *Stanford Micoarray Database* (http://tinyurl.com/24skjy2)
[vii] http://tinyurl.com/3xk83jn

IV. Conclusion

Microarray technology has strongly emerged due to the fact that it can provide a rapid snapshot of gene expression pattern of a tissue. It also helps in our understanding of global networks of bio-molecular interactions. Scientific areas including diagnosis, drug development, functional genomics, and comparative genomics are stimulated with the development of this high throughput technique resulting in avalanche of data from innumerable number of experiments.

With the emergence of microarray technology from the shadows of being ‘cautionary tale’ [82], the steps towards the growth in the area of microarray data integration have been initiated. The analysis carried out maintaining highest standards, use of sound study-design, application of robust statistical application as well as adoption of reporting standards can help in the maturity of this area, and to march towards unlocking the hidden treasures of biological knowledge.

References

1. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 1985 Dec 20;**230**(4732):1350-4.
2. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* 1986;**51 Pt 1**:263-73.
3. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;**270**(5235):467–70.
4. Smith V, Botstein D, Brown PO. Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. *Proc Natl Acad Sci USA* 1995;**92**(14):6479-83.
5. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. . *Nature Genetics* 2001;**29**(4):365–71.
6. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, et al. ArrayExpress--a public

repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 2003 Jan 1;**31**(1):68-71.

7. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002 Jan 1;**30**(1):207-10.

8. Ikeo K, Ishi-i J, Tamura T, Gojobori T, Tateno Y. CIBEX: center for information biology gene expression database. *Comptes rendus Biologies* 2003 Oct-Nov;**326**(10-11):1079-82.

9. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, et al. Multiple-laboratory comparison of microarray platforms. *Nat Methods* 2005 May;**2**(5):345-50.

10. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* 2002 Mar;**18**(3):405-12.

11. Hedges LV, Olkin I. *Statistical methods for meta-analysis*. New York: Academic Press; 1985.

12. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam consultation onmeta-analysis. *Journal of Clinical Epidemiology* 1995;**48**(1):167–71.

13. Normand SL. Tutorial in biostatistics-meta-analysis: formulating, evaluating, combining, and reporting. *Stat Med* 1999;**18**:321–59.

14. Ghosh D, Barette TR, Rhodes D, Chinnaiyan AM. Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Functional and Integrative Genomics* 2003 Dec;**3**(4):180-8.

15. Moreau Y, Aerts S, De Moor B, De Strooper B, Dabrowski M. Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends in Genetics* 2003 Oct;**19**(10):570-7.

16. Larsson O, Wennmalm K, Sandberg R. Comparative microarray analysis. *Omics* 2006;**10**(3):381–97.

17. Tan PK, Downey TJ, Spitznagel EL, Jr., Xu P, Fu D, Dimitrov DS, et al. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* 2003 Oct 1;**31**(19):5676-84.

18. Severgnini M, Bicciato S, Mangano E, Scarlatti F, Mezzelani A, Mattioli M, et al. Strategies for comparing gene expression profiles from different microarray platforms: application to a case-control experiment. *Anal Biochem* 2006 Jun 1;**353**(1):43-56.

19. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J. Independence and reproducibility across microarray platforms. *Nat Methods* 2005 May;**2**(5):337-44.

20. Canales RD, Luo Y, Willey JC, Austermiller B, Barbacioru CC, Boysen C, et al. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol* 2006 Sep;**24**(9):1115-22.

21. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* 2006;**24**(9):1151-61.

22. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters* 2004;**573**(1-3):83-92.

23. Efron B, Tibshirani R, Storey JD, Tusher V. Emperical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 2001;**96**(456):1151-60.

24. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001;**98**(9):5116–21.

25. Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 2004;**5**(2):155-76.

26. Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 2002;**12**:546–54.

27. Fisher RA. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd; 1925.

28. Choi JK, Yu U, Kim S, Yoo OJ. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* 2003;**Vol. 19** (Suppl. 1):84-90.

29. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;**5**(10):R80.

30. Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics* 2008;**24**(3):374-82.

31. Hu P, Greenwood CM, Beyene J. Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *BMC Bioinformatics* 2005;**6**(1):128-38.

32. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM. Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research* 2002 Aug;**62**(15):4427-33.

33. Parmigiani G, Garrett ES, Anbazhagan R, Gabrielson E. A statistical framework for expression-based molecular classification in cancer. *J R Statist Soc B* 2002;**64**(4):717-36.

34. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* 2006 Nov 15;**22**(22):2825-7.

35. Grutzmann R, Boriss H, Ammerpohl O, Luttges J, Kalthoff H, Schackert HK, et al. Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene* 2005;**24**(32):5079–88.

36. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine* 1991;**10**(11):1665-77.

37. Garrett-Mayer E, Parmigiani G, Zhong X, Cope L, Gabrielson E. Cross-study validation and combined analysis of gene expression microarray data. *Biostatistics* 2008;**9**(2):333-54.

38. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J Comput Graph Statist* 1996;**5**:299–314.

39. Cope L, Zhong X, Garrett E, Parmigiani G. MergeMaid: R tools for merging and cross-study validation of

gene expression data. *Stat Appl Genet Mol Biol* 2004;**3**:Article29.

40. Eysenck HJ. Problems with meta-analysis. In: Chalmers I, Altman DG, editors. *Systematic Reviews*. London: BMJ Publishing Group; 1995. p. 64-74.

41. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2000 Jan 1;**28**(1):10-4.

42. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. *Nat Genet* 2003 Jan;**33**(1):49-54.

43. Hwang KB, Kong SW, Greenberg SA, Park PJ. Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics* 2004 Oct 25;**5**:159.

44. Cheadle C, Vawter MP, Freed WJ, Becker KG. Analysis of microarray data using Z score transformation. *J Mol Diagn* 2003 May;**5**(2):73-81.

45. Marron JS, Todd MJ. Distance Weighted Discrimination. NY: Cornell University; 2002.

46. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, et al. Adjustment of systematic microarray data biases. *Bioinformatics* 2004 Jan 1;**20**(1):105-14.

47. Bloom G, Yang IV, Boulware D, Kwong KY, Coppola D, Eschrich S, et al. Multi-platform, multi-site, microarray-based human tumor classification. *American Journal of Pathology* 2004;**164**(1):9-16.

48. Shen R, Ghosh D, Chinnaiyan AM. Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics* 2004;**5**(1):94.

49. Jiang H, Deng Y, Chen H-S, Tao L, Sha Q, Chen J, et al. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics* 2004;**5**:81.

50. Warnat P, Eils R, Brors B. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics* 2005;**6**:265.

51. Stafford P, Brun M. Three methods for optimization of cross-laboratory and cross-platform microarray expression data. *Nucleic Acids Res* 2007;**35**(10):e72.

52. Xu L, Tan AC, Winslow RL, Geman D. Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics* 2008;**9**:125.

53. Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol* 2004;**3**:Article19.

54. Shabalin AA, Tjelmeland H, Fan C, Perou CM, Nobel AB. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* 2008 May 1;**24**(9):1154-60.

55. Xiong H, Zhang Y, Chen XW, Yu J. Cross-platform microarray data integration using the normalised linear transform. *Int J Data Min Bioinform* 2010;**4**(2):142-57.

56. Sarmah CK, Samarasinghe S, Kulasiri D, Catchpoole D. A simple Affymetrix ratio-transformation method yields comparable expression level quantifications with cDNA data. International Conference on Bioinformatics and Bioengineering; 2010; Cape Town, South Africa: World Academy of Science, Engineering and Technology, p. 78-83.

57. Stears RL, Martinsky T, Schena M. Trends in microarray analysis. *Nat Med* 2003 Jan;**9**(1):140-5.

58. Draghici S. Microarrays. *Data analysis tools for DNA microarrays*. Boca Raton, USA: Chapman & Hall/CRC; 2005. p. 15-32.

59. Ivliev AE, t Hoen PA, Villerius MP, den Dunnen JT, Brandt BW. Microarray retriever: a web-based tool for searching and large scale retrieval of public microarray data. *Nucleic Acids Res* 2008 Jul 1;**36**(Web Server issue):W327-31.

60. Xia XQ, McClelland M, Porwollik S, Song W, Cong X, Wang Y. WebArrayDB: cross-platform microarray data analysis and public data repository. *Bioinformatics* 2009 Sep 15;**25**(18):2425-9.

61. Bisognin A, Coppe A, Ferrari F, Risso D, Romualdi C, Bicciato S, et al. A-MADMAN: annotation-based microarray data meta-analysis tool. *BMC Bioinformatics* 2009;**10**:201.

62. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 2006 Sep;**24**(9):1151-61.

63. Gagarin D, Yang Z, Butler J, Wimmer M, Du B, Cahan P, et al. Genomic profiling of acquired resistance to apoptosis in cells derived from human atherosclerotic lesions: potential role of STATs, cyclinD1, BAD, and Bcl-XL. *J Mol Cell Cardiol* 2005 Sep;**39**(3):453-65.

64. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BFF, Rapp BA, et al. GenBank. *Nucleic Acids Research* 1999;**27**(1):12-7.

65. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* 2006;**00**(Database issue):D1–D6.

66. Stoesser G, Tuli MA, Lopez R, Sterk P. The EMBL Nucleotide Sequence Database. *Nucleic Acids Research* 1999;**27**(1):18-24.

67. Pontius JU, Wagner L, Schuler GD. UniGene: a unified view of the transcriptome. In: McEntyre J, Ostell J, editors. *The NCBI Handbook*. Bethesda (MD): National Library of Medicine (US); 2003.

68. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* 2006;**00**(Database issue):D1–D5.

69. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase ofhumangenes and genetic disorders. *Nucleic Acids Research* 2005;**33**(Database issue):D514–D7.

70. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 2003;**4**(5):P3.

71. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*

2003;**4**(4):R28.

72. Tsai J, Sultana R, Lee Y, Pertea G, Karamycheva S, Antonescu V, et al. RESOURCERER: a database for annotating and linking microarray resources within and across species. *Genome Biol* 2001;**2**(11):SOFTWARE0002.

73. Newman JC, Weiner AM. L2L: a simple tool for discovering the hidden significance in microarray expression data. *Genome Biol* 2005;**6**(9):R81.

74. Cahan P, Ahmad AM, Burke H, Fu S, Lai Y, Florea L, et al. List of lists-annotated (LOLA): a database for annotation and comparison of published microarray gene lists. *Gene* 2005 Oct 24;**360**(1):78-82.

75. Halgren RG, Fielden MR, Fong CJ, Zacharewski TR. Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones. *Nucleic Acids Res* 2001 Jan 15;**29**(2):582-8.

76. Woo Y, Affourtit J, Daigle S, Viale A, Johnson K, Naggert J, et al. A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. *J Biomol Tech* 2004 Dec;**15**(4):276-84.

77. Pritchard CC, Hsu L, Delrow J, Nelson PS. Project normal: defining normal variance in mouse gene expression. *Proc Natl Acad Sci USA* 2001 Nov 6;**98**(23):13266-71.

78. Dresen IM, Husing J, Kruse E, Boes T, Jockel KH. Software packages for quantitative microarray-based gene expression analysis. *Curr Pharm Biotechnol* 2003 Dec;**4**(6):417-37.

79. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, et al. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 2003 Feb;**34**(2):374-8.

80. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol* 2002 Jul 15;**3**(8):SOFTWARE0003.

81. Dudoit S, Gentleman RC, Quackenbush J. Open source software for the analysis of microarray data. *Biotechniques* 2003 Mar;**Suppl**:45-51.

82. Sherlock G. Of fish and chips. *Nature Methods* 2005;**2**(5):329-30.