DRUG DISCOVERY TODAY TARGETS

# '*The 39 steps*' in gene expression profiling: critical issues and proposed best practices for microarray experiments

Sandrine Imbeaud and Charles Auffray

**Gene expression microarrays have been used widely to address increasingly complex biological questions and to produce an unprecedented amount of data, but have yet to realize their full potential. The interpretation of microarray data remains a major challenge because of the complexity of the underlying biological networks. To gather meaningful expression data, it is crucial to develop standardized approaches for vigilant study design, controlled annotation of resources, careful quality control of experiments, robust statistics, and data registration and storage. This article reviews the steps needed in the design and execution of valid microarray experiments so that global gene expression data can play a major role in the pursuit of future biological discoveries that will impact drug development.**

Despite great advances in deciphering the human genome sequence, much of the natural world remains to be explored at the molecular level. There is a gap between the structure of the genome and biology, as a vast majority of identified genes have as yet unknown functions and roles in physiology or disease.

To understand gene function, it is helpful to know when, where and to what extent a gene is expressed, and under what circumstances its expression level is affected. Traditionally, analysis of the regulation and function of genes has been performed through step-by-step studies of individual genes and proteins. With the advent of large-scale biology triggered by the Human Genome Project, it has become clear that identification of global sets of gene transcripts (the transcriptome) and proteins (the proteome) is a necessary step, however insufficient towards understanding the inherent complexity of biological systems, which requires a multi-disciplinary systems approach [1].

Soon after the discovery of RNA in the 1960s, assessment of RNA complexity was performed by DNA–RNA hybridization re-association kinetic measurements. Array hybridization developed gradually during the next two decades as a semi-quantitative, semi-automated technology [2]. The conjunction of novel chemistries, robotics and computation tools brought array technology to a new stage of development and to its explosive growth during the past decade [3–5].

As the use of DNA microarrays expanded dramatically, they became an effective tool and a standard technique for gene expression profiling, allowing interrogation of thousands of genes simultaneously. Array technology found a wide range of applications, such as experimental annotation of the human genome [6], discovery of gene functions [7], analysis of complex diseases [8,9], biological-pathway dissection [10], tumor profiling [10,11], diagnostic and prognostic predictions for various types of cancers [10], drug-target validation [10,12], biomarker identification [10,13] and compound-toxicity studies [14].

However, although initially exciting, microarray research soon became highly frustrating, generating

**Sandrine Imbeaud***
**Charles Auffray**
Array s/IMAGE,
Genexpress,
Functional Genomics and
Systems Biology for Health,
LGN - UMR 7091 - CNRS and
Pierre and Marie Curie
University – Paris VI,
Villejuif,
France
*e-mail: imbeaud@vjf.cnrs.fr

1359-6446/05/$ – see front matter ©2005 Elsevier Ltd. All rights reserved. PII: S1359-6446(05)03565-8

BOX 1

### *'The 39 steps'* in gene expression profiling

**• Experimental design**

1. Objectives articulation
2. Resources allocation
3. Study design
4. Power and confidence
5. Pilot collection
6. Statistical design validation
7. Data collection

**• Gene collection**

8. Gene coverage
9. Probe selection
10. Resources annotation
11. Confidence metrics
12. Probe synthesis and purification
13. Quality controls and metrics

**• Sample collection**

14. Sample selection
15. Resources cataloguing
16. Sample preparation
17. Quality controls and metrics

**• Array preparation**

18. Instruments calibration
19. Slides printing and treatment
20. Quality controls (QC)

**• Target synthesis**

21. Spike RNA controls
22. Biochemical reactions
23. Quality controls and metrics

**• Hybridization**

24. Hybridization and mixing
25. Washing and drying
26. Quality controls (QC)

**• Data transformation**

27. Image acquisition
28. Image segmentation
29. Data subtraction and filtering
30. Data normalization
31. Data reduction (replicates)
32. QC metrics, descriptive statistics of the measures

**• Knowledge extraction**

33. Ratio- or intensity-statistics
34. Genes modules identification (clustering and classification)
35. Functional annotation
36. Networks definition
37. 'Omics' integration (inter-disciplinary validations)

**• Data storage**

38. Data warehouse (proprietary or public repositories)
39. Data integrity and standards maintenance

The following 39 steps constitute a guideline in gene expression profiling to collect reliable measures, improve knowledge extraction and facilitate information sharing. Five key sections are highlighted, which are reviewed in this paper, including vigilant experimental design (green box), controlled annotation of resources (red box), robust practical handling (blue box), knowledge extraction for identification of biologically-relevant information (yellow box) and data storage and exchange with microarray data standards and public repositories (grey box).

an uncontrollable flow of data that was extremely difficult to manage and, in the absence of standards and public data repositories, overwhelmed even the experts [15]. Thus, several questions were raised on its validity and accuracy, even on its usefulness. Here, we review the 39 key steps that we have found essential for a generic quality assurance process in gene expression profiling, based on sound principles of careful planning, experimental design, analysis and interpretation required to achieve the promise of microarrays and to turn mountains of data into credible and useful biological knowledge (Box 1).

## Experimental design (green box, Box 1)

Properly designed microarray studies must begin and end with the definition of well-characterized objectives and associated experimental requirements [16]. R.A. Fisher stated that 'To call in the statistician after the experiment is done may be no more than asking him to perform a *post-mortem* examination: he may be able to say what the experiment died of' [17]. An experiment that is well designed can be used to answer the questions that it is specifically designed to address (hypothesis testing) but also to mine for useful information that was not previously anticipated (hypothesis generating).

The challenge is to identify differential gene expression with a high degree of success and a low rate of false positives. Biological replicate measurements performed at the population level, so that each replication represents RNA from a different individual, improve accuracy of fold-change estimates, thus increasing the chance of observing biologically significant gene expression differences while decreasing error rate. The key, however, is to determine which biological resources are needed and how much replication is necessary and sufficient to address the objectives of a study.

The number of independent samples needed is clearly dependent on the particular biological model investigated, the magnitude of anticipated biological variability, the specific genes examined and the laboratory performing the experiments. With this multitude of variables, it is easy to generate a high percentage of false positives, leading to expensive and time-consuming validation. This is aggravated by the high cost of microarrays and often by the difficulty of obtaining biological or clinical samples in large enough numbers or quantities. Thus, to avoid data overfitting, it is essential to determine the required sample size using two additional measures of statistical reliability that include a chosen level of sensitivity (or power, $1\text{-}\beta$) and specificity (or confidence level, $1\text{-}\alpha$). Approaching the problem in this manner is equivalent to a classical power study used in clinical trials.

To date, very few studies have assessed power and sample size requirements in microarray experiments [18]. In those reports, a representative subgroup of individual samples is examined and results extrapolated to the whole population. If the number of samples is not a faithful representation of the population and its intrinsic variance (Box 2), the distribution of parameters will be biased toward those specific for the type of samples collected. Reasonable estimates of parameters needed for subsequent calculations are obtained through small-scale pilot experiments or derived from similar studies.

BOX 2

### Sources of biological variations in gene expression profiling

Biological variations are the largest source of non-systematic error. Most biological samples contain cell mixtures of different cell types, all of which might behave differently. In the case of minor cell subsets, cell mixtures were reported to convincingly mirror changes in gene expression profiles, whereas a greater than 75% pure sample was found to be indistinguishable from a 100% pure sample [66]. Even when genetically 'identical' cells cultured under 'identical' conditions are compared, substantial differences in gene expression levels are observed [67]. Thus, replicate arrays using the same RNA have limited value in reducing total variability. In an effort to reduce the effect of biological variations, some studies proposed isolation of small numbers of cells of a common type or laser-controlled micro-dissection (LCM) [66]; others have reported pooling samples across subjects [68]. Pooled designs were attractive because they have the potential to decrease cost due to the fact that a large number of individual samples can be evaluated using relatively few arrays. Recent strategies compromise between pooling everything and only considering individual biological samples on individual arrays [68]. However, reducing the number of arrays often results in decreasing accuracy.

BOX 3

### Sources of experimental variations in gene expression profiling

Experimental variations come from undesirable systematic error introduced during the many technical steps. One common problem is handling errors such as accidental exchange of different probes during array production, or incorrect association with biological information (for example: probe sequence, IDs, gene name) and quality controls (e.g. sequence verification, probe purity, spot shape). Significant changes in gene expression are also observed during sample collection [69]; all attempts should be made to process tissue both rapidly and uniformly. Then brightness, relative binding affinity and concentration of the dye labels often produce artifacts in gene expression measurements [70]. Universally applicable approaches are the use of exogenous control genes and experimental replications, which will allow monitoring that the microarrays are working properly [71]. Reverse fluor replicates will be necessary to interpret individual ratios from single dual channel array experiments, whereas conventional approaches with a single dye often result in over 20% of inaccurate conclusions. However, performing both forward and reverse labeling systematically for all the samples was reported to be unnecessary [72]. Hybridization can highly influence array hybridization and data reproducibility. Using current static hybridization methods, diffusion transport of RNA targets near probes is rate-limiting. Dynamic hybridization using mixing by chaotic advection or periodic flow convection significantly improves the molar and temporal efficiencies of hybridization [73].

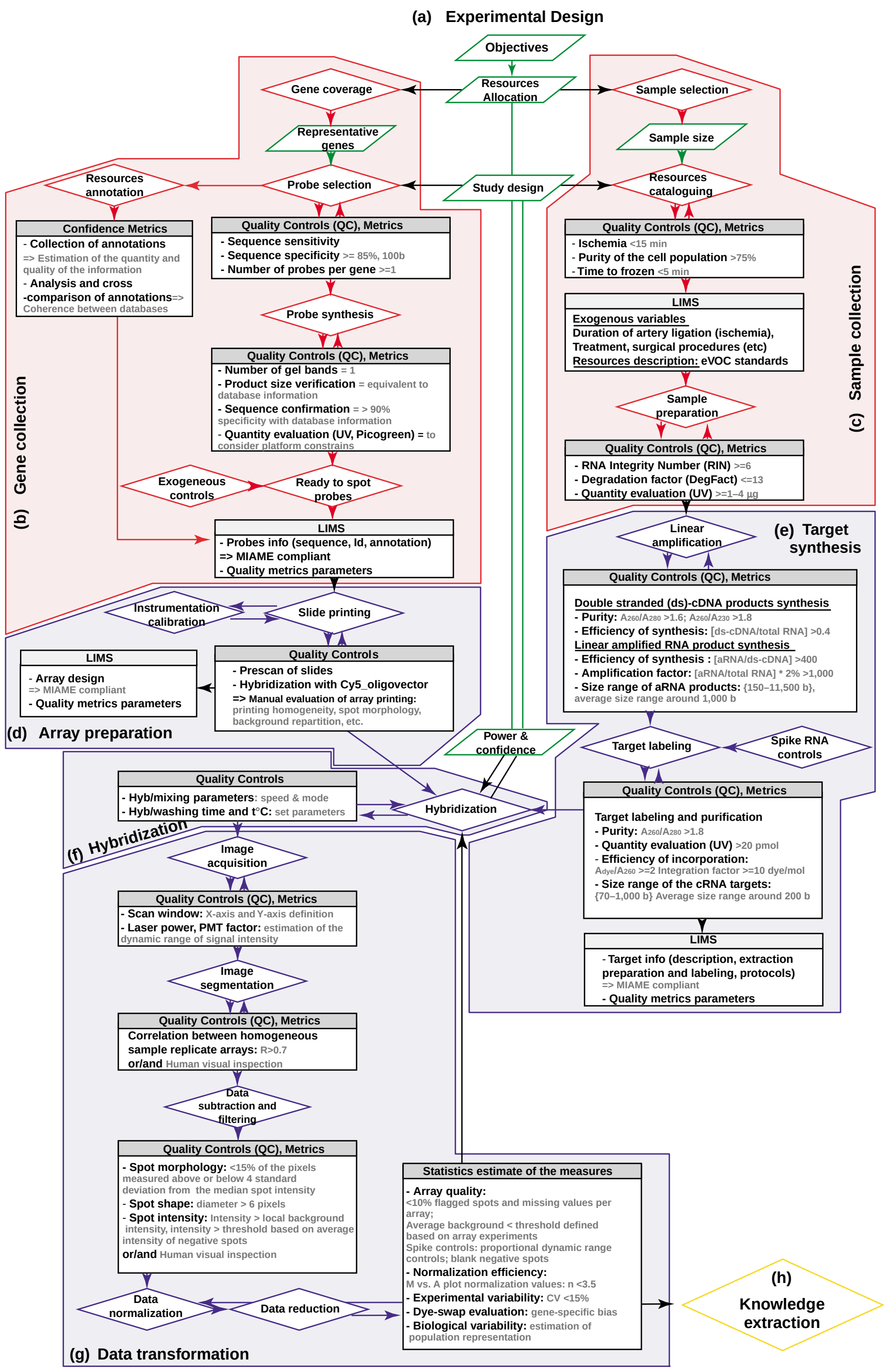
## **Annotation of resources** (red box, Box 1)

### *Gene collection*

It is vital that probe sequences selected for incorporation into an array, their length and location, are chosen with optimal sensitivity, that is, the ability of the probe to bind strongly to its target sequence, and specificity, that is, the inability of the probe to bind strongly to non-target sequences, and then that these are well characterized and that complete, accurate and up-to-date information is available.

Availability of a 'complete' sequence of the human genome has yielded neither a proven method for gene identification nor a definitive count of human genes [19]. Direct clustering of human expressed sequence tags (EST) has predicted as many as 120,000 genes [20], whereas sampling and sequence-similarity-based methods have predicted far lower numbers, ranging from 27,000 to 35,000, and hybrid approaches an intermediate number [21,22]. Thus, when novel genes are characterized and new sequence information is identified, first-generation microarrays become outdated and new ones need to be constructed, allowing genome-wide coverage of gene transcripts and functional pathways.

Despite the existence of rich public resources, no standardized method is available that allows the selection of relevant clusters and corresponding representative sequences, while avoiding selection of clones and/or sequences with mistaken orientation and identity, due to inaccurate annotation such as multiplicity in gene naming, redundancy of gene sequences or phage contamination. Another problem is that 5′ or 3′ sequences derived from the same cDNA clone can get distributed into different EST clusters; conversely, it is very difficult to distinguish between two genes that share a high degree of sequence similarity. Numerous software packages developed for probe selection, mostly relying on the same BLAST algorithm, were found to be non-intuitive, limited in flexibility and often inappropriate when considering cross-hybridizing non-target sequences (reviewed in [23]). If probe selection is not optimized, DNA microarray hybridization can generate false-positive data due to non-specific cross-hybridization to highly similar sequences [23], gene families [24], or alternatively spliced variants [25]. The time and effort required to invalidate erroneous expression results due to ambiguous array design and annotation (Box 3) would be better spent if lists of genes coupled with probe sequence information and comprehensive annotation, ranked according to expected accuracy and confidence, were available. Innovative strategies were recently presented that incorporate gene position-specific scoring, use composite probe design with multiple specific non-contiguous subsequences or maximize efficiency of the duplex formation [26,27]. Furthermore, automated systems have been proposed to standardize and improve array annotation [28,29], emphasizing importance of biological database synchronization and cross-referencing. The amount of available annotations and assessment of reliability and coherence between individual databases are also critical. Therefore, it is a high priority to develop a scoring system (Figure 1) considering both technology-related parameters and genomic knowledge to rank all the information collected for each probe from various databases.


(a) Experimental Design
Objectives
Resources Allocation
Gene coverage
Sample selection
Representative genes
Sample size
Resources annotation
Probe selection
Study design
Resources cataloguing
Confidence Metrics
- Collection of annotations
=> Estimation of the quantity and quality of the information
- Analysis and cross -comparison of annotations=> Coherence between databases
Quality Controls (QC), Metrics
- Sequence sensitivity
- Sequence specificity >= 85%, 100b
- Number of probes per gene >=1
Quality Controls (QC), Metrics
- Ischemia <15 min
- Purity of the cell population >75%
- Time to frozen <5 min
LIMS
Exogenous variables
Duration of artery ligation (ischemia), Treatment, surgical procedures (etc)
Resources description: eVOC standards
Probe synthesis
Quality Controls (QC), Metrics
- Number of gel bands = 1
- Product size verification = equivalent to database information
- Sequence confirmation = > 90% specificity with database information
- Quantity evaluation (UV, Picogreen) = to consider platform constrains
Sample preparation
Quality Controls (QC), Metrics
- RNA Integrity Number (RIN) >=6
- Degradation factor (DegFact) <=13
- Quantity evaluation (UV) >=1–4 µg
(c) Sample collection
(b) Gene collection
Exogeneous controls
Ready to spot probes
LIMS
- Probes info (sequence, Id, annotation)
=> MIAME compliant
- Quality metrics parameters
Linear amplification
(e) Target synthesis
Quality Controls (QC), Metrics
Double stranded (ds)-cDNA products synthesis
- Purity: A260/A280 >1.6; A260/A230 >1.8
- Efficiency of synthesis: [ds-cDNA/total RNA] >0.4
Linear amplified RNA product synthesis
- Efficiency of synthesis : [aRNA/ds-cDNA] >400
- Amplification factor: [aRNA/total RNA] * 2% >1,000
- Size range of aRNA products: {150–11,500 b}, average size range around 1,000 b
Instrumentation calibration
Slide printing
LIMS
- Array design
=> MIAME compliant
- Quality metrics parameters
Quality Controls
- Prescan of slides
- Hybridization with Cy5_oligovector
=> Manual evaluation of array printing: printing homogeneity, spot morphology, background repartition, etc.
(d) Array preparation
Power & confidence
Target labeling
Spike RNA controls
Quality Controls
- Hyb/mixing parameters: speed & mode
- Hyb/washing time and t°C: set parameters
Hybridization
Quality Controls (QC), Metrics
Target labeling and purification
- Purity: A260/A280 >1.8
- Quantity evaluation (UV) >20 pmol
- Efficiency of incorporation: Adye/A260 >=2 Integration factor >=10 dye/mol
- Size range of the cRNA targets: {70–1,000 b} Average size range around 200 b
(f) Hybridization
Image acquisition
Quality Controls (QC), Metrics
- Scan window: X-axis and Y-axis definition
- Laser power, PMT factor: estimation of the dynamic range of signal intensity
LIMS
- Target info (description, extraction preparation and labeling, protocols)
=> MIAME compliant
- Quality metrics parameters
Image segmentation
Quality Controls (QC), Metrics
Correlation between homogeneous sample replicate arrays: R>0.7
or/and Human visual inspection
Data subtraction and filtering
Quality Controls (QC), Metrics
- Spot morphology: <15% of the pixels measured above or below 4 standard deviation from the median spot intensity
- Spot shape: diameter > 6 pixels
- Spot intensity: Intensity > local background intensity, intensity > threshold based on average intensity of negative spots
or/and Human visual inspection
Statistics estimate of the measures
- Array quality:
<10% flagged spots and missing values per array;
Average background < threshold defined based on array experiments
Spike controls: proportional dynamic range controls; blank negative spots
- Normalization efficiency:
M vs. A plot normalization values: n <3.5
- Experimental variability: CV <15%
- Dye-swap evaluation: gene-specific bias
- Biological variability: estimation of population representation
(h) Knowledge extraction
Data normalization
Data reduction
(g) Data transformation

Drug Discovery Today

**FIGURE 1**

**Workflow of standard operating procedures for gene expression profiling.** Quality controls (QC), key steps and metrics that help building a decision tree in a workflow of Quality Assurance (QA) standard operating procedures.

### *Sample collection*

The quality of data obtained in a microarray experiment is highly dependent on an adequate choice of the biological sample collection in relation to the objectives pursued. Integrity of the information on the samples is essential to minimize artifactual detection.

Biological samples should be selected in close collaboration with clinicians and biologists and coupled to comprehensive documentation, protected by stringent ethical criteria [30], including standardized descriptions of pathological examination, micro-dissection, extensive clinical information, patient background and laboratory data. Controlled vocabularies, such as the eVOC ontologies, based on four orthogonal, mutually exclusive knowledge domains (anatomical system, cell type, developmental stage and pathology) can be highly valuable to standardize biological resource description [31] and are offered as choices for describing the minimal information about a microarray experiment (MIAME) BioSource properties [30]. Additionally, as many factors are known, and many more suspected to affect data, such as drug treatment, surgical procedures and anesthesia (Box 3); all exogenous variables must be documented.

Biological samples are easily degradable, and differences in handling cells or tissues can greatly influence microarray results. Although moderately degraded samples might still lead to reasonable expression profiles [32], vigilant assessment of sample quality must become a routine part of any gene expression workflow under quality assurance. Newly developed user-independent RNA quality metrics were found to be highly reliable to classify RNA samples in groups of integrity [33]. Comparative analysis using such quality metrics showed that the use of RNA samples of disparate integrity might correlate with ratio discrepancies in gene expression measurements, and therefore with false-positive and false-negative rates of differential gene expression, whereas samples of similar, even poor integrity are highly comparable [33]. Such quality metrics can be easily integrated in a sample tracking information (Figure 1), offering the opportunity to establish a standard exchange language and operating procedures for cataloguing each sample in an informative comparative group of integrity, thus improving the scheme of meaningful experiments.

## **Practical robustness** (blue box, Box 1)

### *Array preparation, target synthesis and hybridization*

Microarray analysis represents a classic ‘precision in, precision out’ technology. Many factors influence extrapolation from observed signal intensity levels to real gene expression levels. Although statistical power analysis is critical in experimental design, it is also important to consider the stability of the results obtained, that is, the tendency for the results to remain the same as the experimental replication level is changed. Distinguishing between genes that are truly differentially regulated and genes that are simply affected by measurement errors remains a challenge. Accuracy and reproducibility are essential and standardization with respect to the implementation of quality controls is a desirable prerequisite. It will provide the confidence to enable the full power of the experiments to be exploited, although no single method has proven free from ambiguity.

The problem of making microarray data more reproducible is currently tackled from various angles, including experimental logistics and procedures [34]. Variation can be laboratory- and biological system-dependent and can be divided into two major categories. Systematic errors (Box 3) arise reproducibly as a result of experimental procedures, whereas non-systematic errors (Box 2) originate from inherent biological variability. Data 'normalization' can help controlling some of this bias (reviewed in [35]), but is often imperfect. Quality controls can be introduced to detect 'outliers' or 'contaminants' that confound data interpretation. Each laboratory needs to analyze the source of variations in its data, and then assign weights to individual factors when defining quality measures (Figure 1).

### *Data transformation*

Ideally, a microarray system should measure gene expression levels in terms of natural biological units, such as mRNA copies per cell with an estimated error model. In practice, raw data are represented as an image generated by measuring signals in defined areas (pixels) across the entire hybridized area. Raw data are then transformed into a defined unit by determining the molar relationship between amplitude of a given array probe signal and its corresponding target, based on the amount of array features present and the DNA fragment sequence hybridized.

A key group of steps is the effectiveness of data transformation that correspond to: (1) scanning the array (image acquisition); (2) extracting raw intensities from images and identifying the foreground 'feature' (image segmentation); (3) correcting intensities for a variety of defects that occur during manufacturing and processing (data subtraction/filtering, e.g. perturbation of spot positions, irregular spot shapes, variable background); and (4) adjusting data for bias effects accounting for part of experimental and biological variations (data normalization).

Image acquisition and segmentation systems should handle printing imperfections (e.g. irregular spots, blank spots) and misplacement of array image patterns. Many current methods and software packages offer automatic image processing (reviewed in [36]). Though useful, such softwares can introduce errors; for instance, misalignments of spotted arrays can cause incorrect coordinates to be assigned to each feature, as annotations are tracked by grid coordinates. Subsequent visual inspections and manual corrections are still necessary. Interestingly, an unsupervised mathematical method establishes correlation of intensities between arrays based on coordinates location to determine if spot finding is done correctly [37]. Using conventional segmentation methods (e.g. fixed and adaptive shapes or histograms), the greatest challenge remains identifying whether or not a gene is detected at a given spot, while dealing with artifactual and blank spots (reviewed in [38]). A new model-based clustering of pixels method was recently presented, which was found to be highly robust [39].

Then, proper data subtraction and/or filtering and normalization associated with appropriate quality controls are essential (Figure 1). The problem of normalization has already generated an enormous amount of literature (reviewed in [35]) and is fast becoming a statistical discipline in itself. Comparison of different methods available (e.g. intensity-based and non-linear approaches) indicates that complex methods do not necessarily perform better than simpler ones [40], and that excluding weak spots greatly improves normalization [41]. Conversely, few standardized models were proposed that provide a quantitative approach for data filtering [42], although it would be advantageous to attach quality metrics to each ratio to capture expression variations (Figure 1). Such metrics must be designed and implemented at an early stage of the analysis because information is lost when ratios are extracted from the image and forwarded to higher data processing levels. Complete automation of those steps might remove the need for human intervention [43]. The key factor of success is that variations are understood, so that they can be modeled.

## **Knowledge extraction for identification of biologically-relevant information** (yellow box, Box 1)

Biological knowledge is derived from gene expression measurements, assuming that a set of gene products displaying coordinated expression levels is involved in a functional module, that is, a regulatory unit consisting of a set of genes that are co-regulated in a specific cellular context [44]. Extracting this information involves two distinct phases: first identifying the functional modules through 'ratio or intensity statistics' and 'genes modules identification' and then understanding their roles through 'functional annotation', 'networks definition' and 'omics integration'.

### *Ratio or intensity statistics*

The first step is now abundantly studied; approaches for individual and combined processing and analysis steps have recently been reviewed [45]. Considering A. Einstein stated that 'Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted', the challenge is to distinguish genuine expression changes from background 'noise'. The ratio-statistics approach is often used because it is simple and intuitive

[46]. A common difficulty is that there are thousands of genes but only a small number of samples. Mathematically, this problem is characterized by high data dimensionality, resulting in a very complex multiple hypothesis-testing problem (considering a 10K array, n=10,000 tests are performed simultaneously; thus, for a 1% threshold, we expect 0.01x10,000=100 false positive genes). Numerous correction procedures are available that focus on p-value adjustments (reviewed in [47]), however, they rarely identify the correct number of differentially expressed genes. Re-sampling tests with cross-validation (for example by permutation and perturbation tests) were found to greatly help in selecting and ranking genes for significance in differential expression. Moreover, it is now known that simply using fold changes is unreliable and inefficient, as a great deal of valuable information can be contained in the absolute expression levels, including the large number of those of low magnitude [48].

### *Genes modules identification*

The following and central step is the identification of groups of genes that exhibit similar expression patterns, as they are considered likely to be co-regulated [49]. The expectation is that the clusters identified will point to relevant sites of expression, functional and regulatory networks, and normal or pathological functions in organs and whole organisms (e.g. capturing portraits of the *in vivo* transcriptome and deducing mechanistic targets) or discover disease subtypes (e.g. cancer diagnosis and drug discovery). Different clustering methods have been developed that usually produce different solutions, but no objective guideline is available [50]. Two essential aspects of this problem are (1) to estimate the number of clusters and (2) to allocate biological samples to these clusters, and assess the confidence of cluster assignments for individual samples. Several strategies have addressed the first issue and cluster validity indices have been proposed; then numerous solutions have been presented to evaluate systematically the quality of clusters [51,52]. Both approaches should help to better define how to measure the similarity 'distance' that estimates the number of clusters, and where to cut the cluster dendrogram to allocate biological samples to clusters.

Among genes involved in the same regulatory processes or functions, some genes were also reported not to display similar expression patterns [53]. Those genes will be missed by traditional clustering techniques. To address this problem, a second-order expression analysis was recently suggested to extract essential features of regulatory networks from multiple microarray datasets, while capturing combinatorial co-regulations for genes that do not exhibit identical expression patterns (reviewed in [54]).

### *Functional annotation*

The next phase of microarray data analysis requires systematic retrieval of functional information to identify the general biological processes associated with lists of selected genes. Unveiling the biological relevance of gene expression profiles is initially performed by reading relevant preexisting biomedical literature. Availability of published literature describing genes and their functions in computer-interpretable form is a potentially rich source of information [55]. Given that there are currently >12 million research articles registered in PubMed, this is extremely challenging [48]. In the last few years, there has been a lot of interest in analytical literature- and text-mining tools that help finding the most relevant information. Frequently, information retrieval incorporates keyword-based approaches (reviewed in [55]). Such strategies often suffer from several drawbacks and do not yet allow taking into account user's context, as the underlying processes are static and not user-trainable. Synonymy, the many ways in which to refer to the same object and polysemy, the fact that a given word may have multiple meanings, can hamper finding records that are truly related to the gene analyzed (reduced call) and instead retrieve records that are not related to the gene (reduced precision). As a result, they sometimes produce misleading interpretations [56].

To overcome such problems, librarians and information scientists developed the notion of controlled vocabularies or ontologies (reviewed in [57]). Gene Ontology (GO) has been recently developed for annotating genes in various organisms [58]. It constitutes a standardized nomenclature for expressing and sharing biological concepts, and formalizing knowledge about biological processes, molecular functions and cellular components. Information is structured into highly interconnected hierarchical structures that reflect relationships between GO terms and associated genes [58]. Relationship is one to many (a GO term can have multiple parent terms), as gene products can be involved in several processes, or carry out multiple molecular functions in alternate cellular locations. Several dedicated tools have been developed that link microarray experimental data to GO terms in an automated way (see Gene Ontology website: www.geneontology.org/GO.tools.shtml).

This opens new perspectives in functional annotation to quantify the difference between terms associated with gene clusters by scoring them according to the functional coherence of corresponding gene groups. Recent studies reported that GO functional groups can be used as a framework for numerical clustering to create more meaningful clusters [59].

### *Networks definition and 'omics' integration*

The next step is to relate gene expression patterns with physiology, pathology or clinical outcome, through the functional networks and regulatory mechanisms in which the relevant genes are involved. Due to underlying network complexity and combinatorial expansion in the number of potential network structures, available methods often decompose networks into basic functional and structural units. For instance, biological pathways are networks of relationships between biological entities, which were established steadily through a century of biochemistry and molecular biology. With the advent of genome-scale biology, systematic cataloguing and storage of pathways has recently gained increasing importance. A wide variety of resources and tools exists to display pathway information, enabling presentation of microarray data in the framework of documented biological pathways [60].

As actual knowledge about biological networks is often incomplete, exact solutions will not be determined, but a solution space can be generated, containing all possible cellular functions that are consistent with the information assembled and the biological context. This constitutes a starting point for inferring new knowledge by integrating other 'omic' spaces, based on data from additional transcriptomic, genomic, proteomic, metabolomic, phenomic or other studies, and forms the basis for data-driven hypothesis generation and independent verification processes of biological significance [61–63].

## Data storage, benefits from microarray data standards (grey box, Box 1)

Microarray datasets are valuable only if they are annotated by sufficiently detailed but 'not too many' experimental descriptions. In many databases, a substantial number of these annotations are stored in different formats, levels of detail and locations [64]. It is often in a free-text format, which is not readily accessible for multivariate statistical methods, and few of them can 'talk to' each other. Thus it is virtually impossible to conduct data analysis across these databases.

Expression profiling data should be made fully available in a standardized way using community-defined vocabularies in a form that allows the basic conclusions to be evaluated independently. As for DNA sequences, it will guarantee the compatibility and comparability of data across different projects. With MIAME, the Microarray Gene Expression Data (MGED) Society (www.mged.org) has proposed a standard to describe the minimum information required to unambiguously interpret and verify microarray experiments. It was published in December 2001 [30] and was finally adopted in the summer 2002 by an increasing number of publishers (for example: Nature Publishing Group, Oxford University Press, Biomed Central). Over the past three years, such standards have been adopted generally and used widely in public resources and are increasingly implemented by companies whose products are related to microarrays. In addition, as an effort to make public availability of microarray data compulsory for publication, researchers are urged to deposit experimental data into a public repository such as the Gene Expression Omnibus (GEO) at NCBI, the Stanford Microarray database, ArrayExpress at EBI or the Center for Information Biology gene Expression Database (CIBEX) at DDBJ [65]. Accessibility to microarray raw data will be instrumental to approach the level of standardization necessary to realize the full potential of expression profiling data.

## Conclusion

Microarray technology represents an increasingly powerful but not fully mature research tool for gene expression profiling. Strategies for the management and interpretation of the vast amounts of data generated are under intensive investigation, using a plethora of statistical and computational methods. Careful study design and standardization are essential to enable collection of reliable measures of gene expression, and to allow meaningful comparisons across platforms and conditions. This is a prerequisite to realize the full potential of the technology to contribute to a systems approach towards understanding biological complexity and to the development of novel therapeutics to tackle human diseases.

## Acknowledgements

The authors would like to thank the members of the Genexpress team for helpful discussion. We are very grateful to Virginie Boulanger for manuscript editing.

## References

1 Auffray, C. *et al.* (2003) From functional genomics to systems biology: concepts and practices. *C. R. Biol.* 326, 879–892
2 Gros, F. (2003) From the messenger RNA saga to the transcriptome era. *C. R. Biol.* 326, 893–900
3 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470
4 Pietu, G. *et al.* (1999) The Genexpress IMAGE knowledge base of the human brain transcriptome: a prototype integrated resource for functional and computational genomics. *Genome Res.* 9, 195–209
5 Lockhart, D.J. and Winzeler, E.A. (2000) Genomics, gene expression and DNA arrays. *Nature* 405, 827–836
6 Schadt, E.E. *et al.* (2004) A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biol.* 5, R73
7 Joshi, T. *et al.* (2004) Genome-scale gene function prediction using multiple sources of high-throughput data in yeast Saccharomyces cerevisiae. *OMICS* 8, 322–333
8 Greenberg, S.A. (2001) DNA microarray gene expression analysis technology and its application to neurological disorders. *Neurology* 57, 755–761
9 Grant, G.M. *et al.* (2004) Microarrays in cancer research. *Anticancer Res.* 24, 441–448
10 Clarke, P.A. *et al.* (2004) Gene expression microarray technologies in the development of new therapeutic agents. *Eur. J. Cancer* 40, 2560–2591
11 Mocellin, S. *et al.* (2005) DNA array-based gene profiling: from surgical specimen to the molecular portrait of cancer. *Ann. Surg.* 241, 16–26
12 Sausville, E.A. and Holbeck, S.L. (2004) Transcription profiling of gene expression in drug discovery and development: the NCI experience. *Eur. J. Cancer* 40, 2544–2549
13 Hu, Y.F. *et al.* (2005) From traditional biomarkers to transcriptome analysis in drug development. *Curr. Mol. Med.* 5, 29–38
14 Searfoss, G.H. *et al.* (2005) The role of transcriptome analysis in pre-clinical toxicology. *Curr. Mol. Med.* 5, 53–64
15 Grant, G.R. *et al.* (2003) Maintaining data integrity in microarray data management. *Biotechnol. Bioeng.* 84, 795–800
16 Yang, Y.H. and Speed, T. (2002) Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* 3, 579–588
17 Fisher, R.A. (1938) In *Indian Statistical Congress*
18 Wei, C. *et al.* (2004) Sample size for detecting differentially expressed genes in microarray

experiments. *BMC Genomics* 5, 87

19 Anonymous (2004) A treasury of exceptions. *Nat. Genet.* 36, 1239

20 Liang, F. *et al.* (2000) Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* 25, 239–240

21 Roest Crollius, H. *et al.* (2000) Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat. Genet.* 25, 235–238

22 Imanishi, T. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* 2, e162

23 Wu, C. *et al.* (2005) Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Res.* 33, e84

24 Evertsz, E.M. *et al.* (2001) Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. *Biotechniques* 31, 1182, 1184, 1186 passim

25 Modrek, B. and Lee, C. (2002) A genomic view of alternative splicing. *Nat. Genet.* 30, 13–19

26 Chou, C.C. *et al.* (2004) Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Res.* 32, e99

27 Dondeti, V.R. *et al.* (2004) In silico gene selection strategy for custom microarray design. *Biotechniques* 37, 768-770, 772, 774-766

28 Ringner, M. *et al.* (2004) ACID: a database for microarray clone information. *Bioinformatics* 20, 2305–2306

29 Dai, H. *et al.* (2004) Dynamic integration of gene annotation and its application to microarray analysis. *J. Bioinform. Comput. Biol.* 1, 627–645

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

31 Kelso, J. *et al.* (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res.* 13, 1222–1230

32 Schoor, O. *et al.* (2003) Moderate degradation does not preclude microarray analysis of small amounts of RNA. *Biotechniques* 35, 1192–1196, 1198-1201

33 Imbeaud, S. *et al.* (2005) Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces. *Nucleic Acids Res.* 33, e56

34 Shi, L. *et al.* (2004) QA/QC: challenges and pitfalls facing the microarray community and regulatory agencies. *Expert Rev. Mol. Diagn.* 4, 761–777

35 Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat. Genet.* 32(Suppl.), 496–501

36 Yang, Y.H. *et al.* (2002 on) Comparison of methods for image analysis cDNA microarray data. *J. Comp. Graph. Stat.* 11, 108–136

37 Marzolf, B. and Johnson, M.H. (2004) Validation of microarray image analysis accuracy. *Biotechniques* 36, 304–308

38 Ahmed, A.A. *et al.* (2004) Microarray segmentation methods significantly influence data precision. *Nucleic Acids Res.* 32, e50

39 Li, Q. *et al.* Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics* (in press)

40 Park, T. *et al.* (2003) Evaluation of normalization methods for microarray data. *BMC Bioinformatics* 4, 33

41 Dozmorov, I. *et al.* (2004) Statistical monitoring of weak spots for improvement of normalization and ratio estimates in microarrays. *BMC Bioinformatics* 5, 53

42 Wang, X. *et al.* (2003) Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. *Bioinformatics* 19, 1341–1347

43 Chrimes, D. (2005) How can data quality and automation enhance confidence in microarray data? *Drug Discov. Today* 10, 675–677

44 Ihmels, J. *et al.* (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics* 20, 1993–2003

45 Anonymous (2002) The chipping forecast II. *Nat. Genet.* 32(Suppl. 2), 461–552

46 Chen, Y. *et al.* (2002) Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* 18, 1207–1215

47 Reiner, A. *et al.* (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19, 368–375

48 Auffray, C. *et al.* (2003) Self-organized living systems: conjunction of a stable organization with chaotic fluctuations in biological space-time. *Philos. Transact. A Math. Phys. Eng. Sci.* 361, 1125–1139

49 Allocco, D.J. *et al.* (2004) Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* 5, 18

50 Datta, S. (2003) Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* 19, 459–466

51 Dudoit, S. and Fridlyand, J. (2003) Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* 19, 1090–1099

52 Famili, A.F. *et al.* (2004) Evaluation and optimization of clustering in gene expression data analysis. *Bioinformatics* 20, 1535–1545

53 Zhou, X. *et al.* (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. U. S. A.* 99, 12783–12788

54 Imbeaud, S. and Auffray, C. (2005) Extracting functional and regulatory order from microarrays. *Mol. Syst. Biol.* 1, E1–E2

55 Krallinger, M. *et al.* (2005) Text-mining approaches in molecular biology and biomedicine. *Drug Discov. Today* 10, 439–445

56 Rebholz-Schuhmann, D. *et al.* (2005) Facts from text-is text mining ready to deliver? *PLoS Biol.* 3, e65

57 Bard, J.B. and Rhee, S.Y. (2004) Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.* 5, 213–222

58 Harris, M.A. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32 (Database issue), D258–261

59 Adryan, B. and Schuh, R. (2004) Gene-Ontology-based clustering of gene expression data. *Bioinformatics* 20, 2851–2852

60 Curtis, R.K. *et al.* Pathways to the analysis of microarray data. *Trends Biotechnol.* (in press)

61 Goesmann, A. *et al.* (2003) Building a BRIDGE for the integration of heterogeneous data from functional genomics into a platform for systems biology. *J. Biotechnol.* 106, 157–167

62 Toyoda, T. and Wada, A. (2004) Omic space: coordinate-based integration and analysis of genomic phenomic interactions. *Bioinformatics* 20, 1759–1765

63 Penny, M.A. and McHale, D. (2005) Pharmacogenomics and the Drug Discovery Pipeline: When Should it Be Implemented? *Am. J. Pharmacogenomics* 5, 53–62

64 Gardiner-Garden, M. and Littlejohn, T.G. (2001) A comparison of microarray databases. *Brief. Bioinform.* 2, 143–158

65 Ball, C. *et al.* (2004) Standards for microarray data: an open letter. *Environ. Health Perspect.* 112, A666–A667

66 Szaniszlo, P. *et al.* (2004) Getting the right cells to the array: Gene expression microarray analysis of cell mixtures and sorted cells. *Cytometry A* 59, 191–202

67 Blake, W.J. *et al.* (2003) Noise in eukaryotic gene expression. *Nature* 422, 633–637

68 Kendziorski, C. *et al.* (2005) On the utility of pooling biological samples in microarray experiments. *Proc. Natl. Acad. Sci. U. S. A.* 102, 4252–4257

69 Spruessel, A. *et al.* (2004) Tissue ischemia time affects gene and protein expression patterns within minutes following surgical tumor excision. *Biotechniques* 36, 1030–1037

70 Cox, W.G. *et al.* (2004) Possible sources of dye-related signal correlation bias in two-color DNA microarray assays. *Anal. Biochem.* 331, 243–254

71 Benes, V. and Muckenthaler, M. (2003) Standardization of protocols in cDNA microarray analysis. *Trends Biochem. Sci.* 28, 244–249

72 Dobbin, K. *et al.* (2003) Statistical design of reverse dye microarrays. *Bioinformatics* 19, 803–810

73 Adey, N.B. *et al.* (2002) Gains in sensitivity with a device that mixes microarray hybridization solution in a 25-microm-thick chamber. *Anal. Chem.* 74, 6413–6417