Chapter 4

# INFORMATION QUALITY MANAGEMENT CHALLENGES FOR HIGH-THROUGHPUT DATA

Cornelia Hedeler and Paolo Missier
*School of Computer Science, The University of Manchester*

**Abstract** In post-genomic biology, high-throughput analysis techniques allow a large number of genes and gene products to be studied simultaneously. These techniques are embedded in experimental pipelines that produce high volumes of data at various stages. Ultimately, the biological interpretation derived from the data analysis yields publishable results. Their quality, however, is routinely affected by the number and complexity of biological and technical variations within the experiments, both of which are difficult to control.

In this chapter we present an analysis of some of these issues, conducted through a survey of quality control techniques within the specific fields of transcriptomics and proteomics. Our analysis suggests that, despite their differences, a common structure and a common set of problems for the two classes of experiments can be found, and we propose a framework for their classification. We argue that the scientists' ability to make informed decisions regarding the quality of published data relies on the availability of meta-information describing the experiment variables, as well as on the standardization of its content and structure. Information management expertise can play a major role in the effort to model, collect and exploit the necessary meta-information.

**Keywords:** Information quality, quality control, transcriptomics, proteomics.

## 1. MOTIVATION

With several genomes of model organisms now being fully sequenced and with the advent of high-throughput experimental techniques, research in biology is shifting away from the study of individual genes, and towards understanding complex systems as a whole, an area of study called *systems biology* (Ideker et al., 2001). Instead of studying one gene or protein at a time, a large number of genes or proteins are monitored simultaneously. Different kinds of experimental data are integrated and analyzed to draw biological conclusions, state new hypotheses, and ultimately generate mathematical models of the biological systems.

A single high-throughput experiment may generate thousands of measurements, requiring the use of data-intensive analysis tools to draw biologically significant conclusions from the data. The data and its biological interpretation are then disseminated through public repositories and journal publications. Once published, it can be used within the scientific community to annotate gene and protein descriptions in public databases, and to provide input to so-called *in silico* experiments, i.e., "procedures that use computer-based information repositories and computational analysis tools to test a hypothesis, derive a summary, search for patterns, or demonstrate a known fact" (Greenwood et al., 2003).

In the recent past, research into the quality of information available in public biology databases has been focusing mainly on the issue of data reconciliation across multiple and heterogeneous data sources (Lacroix and Critchlow, 2004; Rahm, 2004). In this area, it has been possible to adapt techniques and algorithms for which a largely domain-independent theoretical framework exists, notably for record linkage (Winkler, 2004) and for data integration in the presence of inconsistencies and incompleteness (Naumann et al., 2004; Motro et al., 2004).

Data reconciliation techniques, however, largely fail to address the basic problem of establishing the *reliability* of experimental results submitted to a repository, regardless of their relationships with other public data. This is a fundamental and pervasive information quality problem[1]: using unproven or misleading experimental results for the purpose of database annotation, or as input to further experiments, may result in wrong scientific conclusions. As we will try to clarify in this chapter, techniques and, most importantly, appropriate meta-data for objective quality assessment are generally not available to scientists, who can be only intuitively aware of the impact of poor quality data on their own experiments. They are therefore faced with apparently simple questions: are the data and their biological implications credible? are the experimental results sound, reproducible, and can they be used with confidence?

This survey offers an insight into these questions, by providing an introductory guide for information management practitioners and researchers, into the complex domain of post-genomic data. Specifically, we focus on data from transcriptomics and proteomics, i.e., the large-scale study of gene[2] and protein expression, which represent two of the most important experimental areas of the post-genomic era.

We argue that answering the scientists' questions requires a thorough understanding of the processes that produce the data, and of the quality control measures taken at each step in the process. This is not a new idea: a generally accepted assumption in the information quality community (Ballou et al., 1998; Wang, 1998) has been to consider information as a product, created by a recognizable production process, with the implication that techniques for qual-

ity control used in manufacturing could be adapted for use with data. These ideas have been embedded into guidelines for process analysis that attempt to find metrics for measuring data quality (English, 1999; Redman, 1996).

While we subscribe to this general idea, we observe that an important distinction should be made between business data, to which these methodologies have been applied for the most part (with some exceptions: see, e.g., (Mueller et al., 2003) for an analysis of data quality problems in genome data), and experimental scientific data. Business data is often created in few predictable ways, i.e., with human input or input from other processes (this is the case, e.g., in banking, public sector, etc.), and it has a simple interpretation (addresses, accounting information). Therefore, traditional data quality problems such as stale data, or inconsistencies among copies, can often be traced to problems with the input channels and with data management processes within the systems, and software engineering techniques are usually applied to address them.

The correct interpretation of scientific data, on the other hand, requires a precise understanding of the broad variability of the experimental processes that produce it. With research data in particular, the processes are themselves experimental and tend to change rapidly over time to track technology advances. Furthermore, existing quality control techniques are very focused on the specific data and processes, and are difficult to generalize; hence the wealth of domain-specific literature offered in this survey.

This variability and complexity makes the analysis of quality properties for scientific data different and challenging. In this domain, traditional data quality issues such as completeness, consistency, and currency are typically observed at the end of the experiment, when the final data interpretation is made. However, as the literature cited in this chapter shows, there is a perception within the scientific community that quality problems must be addressed at all the stages of an experiment.

For these reasons, we focus on the data creation processes, rather than on the maintenance of the final data output. We concentrate on two classes of experiments, microarray data analysis for transcriptomics, and protein identification for proteomics. In these areas, the quality of the data at the dissemination stage is determined by factors such as the intrinsic variability of the experimental processes, both biological and technical, and by the choice of bioinformatics algorithms for data analysis; these are often based on statistical models and their performance is in turn affected by experimental variability, among other factors. A brief background on these technologies is provided in Section 2.

As a matter of method, we observe that these two classes can be described using the same basic sequence of steps, and that the corresponding quality problems also fall into a small number of categories. We use the resulting framework to structure a list of domain-specific problems, and to provide references for the techniques used to tackle them. This analysis is presented in Section 3.

Although the quality control techniques surveyed are rooted in the context of post-genomics, and this survey does not discuss specific techniques or solutions in depth, a few general points emerge from this analysis regarding technical approaches to quality management. Firstly, the importance of standards for accurately modelling and capturing *provenance meta-data* regarding the experiments, i.e., details of the experimental design and of its execution. Secondly, the standardization of their representation, in order to deal with heterogeneity among different laboratories that adopt different experimental practices. These points are discussed in Section 4.

A further potentially promising contribution offered by the information quality community is the study of information as a product, already mentioned (Ballou et al., 1998; Wang, 1998). However, to the best of our knowledge, no general theory of process control for these data domains has been developed.

## 2. THE EXPERIMENTAL CONTEXT

We describe the general steps of a rather generic biological experiment, starting from the experimental design, leading to a publication, and further, to the use of published literature for the functional annotation of genes and proteins in databases. An overview of this abstraction is shown in Figure 4.1.

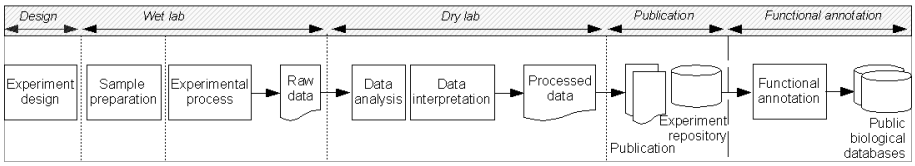


*Figure 4.1.* Sample high-throughput biological data processing pipeline

The experiment begins with the statement of a scientific hypothesis to be tested; along with constraints imposed by the laboratory equipment, this leads to the choice of an appropriate experimental design. The so-called *wet lab* portion is executed by the biologist, starting from the preparation of the sample, and usually leading to the generation of some form of raw data.

It is common to build elements of repetition into the experiment, to take into account both technical and biological variability, specifically: (i) *Technical repeats:* After preparation, the sample is divided into two or more portions and each portion is run through exactly the same technical steps, leading to separate measurements for each portion. This is done to account for the variability of the technical process; (ii) *Biological repeat:* Two or more samples are obtained from different individuals studied under exactly the same conditions. These samples are then prepared using the same protocol and run through the same technical process. These repeats allow for the estimation of biological variability between individuals.

The raw data generated in the lab is then analyzed in the so-called *dry lab*, a computing environment equipped with a suite of bioinformatics data analysis tools. The processed data is then interpreted in the light of biological knowledge, and scientific claims can be published. A growing number of scientific journals explicitly require that the experimental data be submitted to public data repositories at the same time (Editors, 2002).

The result of data analysis and interpretation is processed data, which can include not only the experimental data in analyzed form, but also additional information that has been used to place the data into context, such as the functional annotation of genes and proteins or pathways the proteins are involved in.

The repetition of this process for a large number of high-throughput experiments and over a period of time results in a body of literature about a particular gene or protein. This knowledge is used by *curators* as evidence to support the annotation of genes and proteins described in public databases, such as MIPS (Mewes et al., 2004) and Swiss-Prot (Apweiler et al., 2004). A protein annotation typically includes a description of its function, of the biological processes in which it participates, its location in the cell, and its interactions with other proteins.

Reaching conclusions regarding protein function requires the analysis of multiple pieces of evidence, the results of many experiments of different natures, and may involve a combination of manual and automated steps (Bairoch et al., 2004). In this chapter, we concentrate on two classes of experiments, microarray analysis of gene expression and protein identification; they share the general structure outlined above, and are relevant for their contribution to the knowledge used by the curation process.

We now briefly review some definitions regarding these experiments.

## 2.1 Transcriptomics

Transcriptome experiments use microarray technology to measure the level of transcription of a large number of genes (up to all genes in a genome) simultaneously, as an organism responds to the environment. They measure the quantity of mRNA produced in response to some environmental factor, for instance some treatment, at a certain point in time, by obtaining a snapshot of the gene activity at that time[3].

Here we only provide a brief introduction to the experimental steps involved and the data analysis. For recent reviews on this topic, see (Lockhart and Winzeler, 2000; Bowtell, 1999; Holloway et al., 2002). In addition, (Bolstad et al., 2004; Quackenbush, 2001; Leung and Cavalieri, 2003) provide reviews of methods to normalize and analyze transcriptome data.

An array is a matrix of spots, each populated during manufacturing with known DNA strands, corresponding to the genes of interest for the experiment.

When a sample consisting of mRNA molecules from the cells under investigation is deposited onto the array, these molecules bind, or *hybridize*, to the specific DNA templates from which they originated. Thus, by looking at the hybridized array, the quantity of each different mRNA molecule of interest that was involved in the transcription activity can be measured.

Among the many different array technologies that have become available within the last few years, we focus on the most common two, cDNA (Cheung et al., 1999) and oligonucleotide arrays (Lipshutz et al., 1999). The choice between the two is dictated by the available equipment, expertise, and by the type of experiment: an oligonucleotide array accepts one sample, and is suitable for measuring absolute expression values, whereas cDNA arrays accept two samples, labeled using two different fluorescent dyes, which may represent the state of the organism before and after treatment; they are used to measure ratios of expression levels between the two samples. To obtain ratios using oligonucleotide technology two arrays are necessary, and the ratios are computed from the separate measurements of each. This difference is significant, because the technical and biological variability of these experiments play a role in the interpretation of the results.

The measurements are obtained by scanning the arrays into a digital image, which represents the raw data from the wet lab portion of the experiment. In the dry lab, the image is analyzed to identify poor quality spots, which are excluded from further analysis, and to convert each remaining spot into an intensity value (the "raw readings" in Figure 4.2). These values are normalized to correct for background intensity, variability introduced in the experiment, and also to enable a comparison between repeats.

In the subsequent high-level data analysis, the normalized data is interpreted in the light of the hypothesis stated and the biological knowledge, to draw publishable conclusions. Typically, the goal of the analysis is to detect genes that are differentially expressed after stimulation, or to observe the evolution of expression levels in time, or the clustering of genes with similar expression patterns over a range of conditions and over time. Statistical and machine learning approaches are applied in this phase (Kaminski and Friedman, 2002; Dudoit et al., 2000; Quackenbush, 2001).

Each of these process steps involves choices that must be made (e.g., of technology, of experiment design, and of low-level and high-level data analysis algorithms and tools), which are inter-dependent and collectively affect the significance of the final result. We survey some of these factors in the next section.

## 2.2 Qualitative Proteomics

The term proteomics refers to large-scale analysis of proteins, its ultimate goal being to determine protein functions, and includes a number of areas of

investigation. Here we only consider the problem of identifying the proteins within a sample, a problem of *qualitative* proteomics; this involves determining the peptide masses and sequences of the proteins present in a sample, and matching those against theoretically derived peptides calculated from protein sequence databases. For in-depth reviews of the field, see (Aebersold and Mann, 2003; Pandey and Mann, 2000; Patterson and Aebersold, 2003).

This technology is suitable for experiments in which the protein contents before and after a certain treatment are compared, ultimately leading to conclusions regarding their function, the biological processes in which they are involved, and their interactions. The main steps of the experimental process are shown at the bottom part of Figure 4.2.

A sample containing a number of proteins (possibly of the order of thousands) undergoes a process of separation, commonly by two-dimensional electrophoresis (2DE), resulting in the separation of the proteins onto a gel based on two orthogonal parameters, their charge and their mass. The separated proteins spotted on the gel are then excised and degraded enzymatically to peptides. An alternative technique for peptide separation involves liquid chromatography (LC) (see (Hunter et al., 2002; de Hoog and Mann, 2004) for reviews). LC is used to purify and separate peptides in complex peptide mixtures and can be used without the extra step of protein separation on a gel before digestion (Pandey and Mann, 2000). Peptides are separated by their size, charge, and hydrophobicity.

To identify the proteins mass spectrometry (MS) is used to measure the mass-to-charge ratio of the ionized peptides. The spectrometer produces mass spectra, i.e., histograms of intensity vs. mass/charge ratio. For single-stage experiments, these are called peptide mass fingerprints (PMF). Additionally, a selection of these peptides can be further fragmented to perform tandem MS, or "MS/MS" experiments, which generate spectra for individual peptides. From these spectra the sequence tag of the peptide can be derived. Using sequence information of several peptides in addition to their masses is more specific for the protein identification than just the masses.

The key parameters for this technology are sensitivity, resolution, and the ability to generate information-rich mass spectra. The issue of resolution arises when one considers that every cell may express over 10,000 genes, and that the dynamic range of abundance in complex samples can be as high as $10^6$. Since 2DE technology can resolve no more than 1,000 proteins, clearly only the most abundant proteins can be identified, which creates a problem when interesting proteins are much less abundant (Pandey and Mann, 2000). Techniques have been developed to deal with these issues (Flory et al., 2002); in general, however, limitations in the technology translate into inaccuracies in the resulting spectra.

Finally, in the dry lab the mass spectra are compared with masses and sequences of peptides in databases. Here the experimenter is confronted with

more choices: a number of different algorithms, e.g., Mascot (Perkins et al., 1999), SEQUEST (Eng et al., 1994), exist to compute a score for the goodness of the match between the theoretical peptide sequences in the database and the experimental data. Also, these algorithms may be applied to different reference databases, and provide different indicators to assess the quality of the match. Examples of indicators are the hit ratio (the number of peptide masses matched, divided by the total number of peptide masses submitted to the search), and the sequence coverage (the percentage of the number of amino acids in the experimental sequence, to those in the theoretical sequence).

The quality of the scoring functions in particular is affected by experimental variability, and statistical and computational methods have been proposed to deal with the uncertainty of the identification process (see (Sadygov et al., 2004) for a review, as well as the references in Table 4.3).

## 3. A SURVEY OF QUALITY ISSUES

We begin our analysis by presenting a common framework, illustrated in Figure 4.2. At the top and the bottom are the main steps of the protein identification and of the microarray experiments, respectively. The figure shows their common structure in terms of the general *wet lab*, *dry lab*, and *dissemination* steps, and highlights the key quality concerns addressed by experimenters at each step. We use this high-level framework to provide structure for the analysis of domain-specific issues, and the current techniques and practices adopted to address them.

In Section 3.4, a separate discussion is devoted to the problem of annotating information after it has been submitted to proteomic databases; this has only been addressed in the past by relatively few and isolated experiments.

### 3.1 Variability and experimental design

Both transcriptome and proteome experiments consist of a number of steps, each of which can introduce factors of variability. However, it is not only the variability introduced in the experimental process (the so called technical variability) that can affect the quality of the results, but also biological variability. Systematic analyzes of variability in transcriptome studies (Bakay et al., 2002; Yang and Speed, 2002) and proteome studies (Molloy et al., 2003) have shown that biological variability may have a greater impact on the result.

#### Biological variability

This form of variability affects the results of both transcriptome and proteome experiments, it is of rather random nature and is hard to estimate. Examples include: (i) variability between individuals studied under the same experimental condition (Novak et al., 2002; Bakay et al., 2002) due to genetic differences

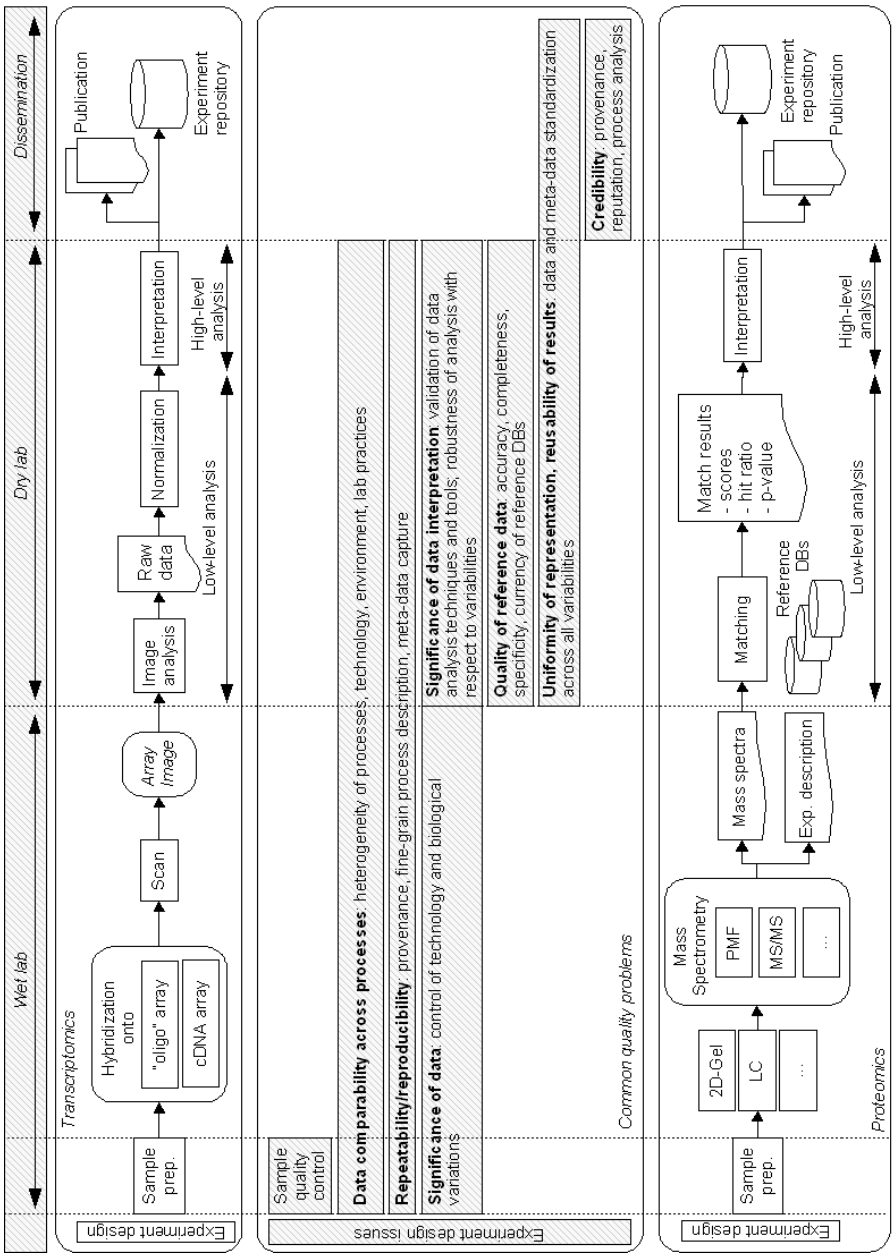


*Figure 4.2.* High-level biological experiment pipeline and common quality issues

(Molloy et al., 2003), minor infections resulting in inflammatory and immune responses of varying intensities (Yang and Speed, 2002), environmental stress, or different activity levels, but can also be due to tissue heterogeneity (varying distribution of distinct cell types in tissues); (ii) variability between individuals due to random differences in the experimental conditions, such as growth, culture, or housing conditions (Hatfield et al., 2003; Novak et al., 2002; Molloy et al., 2003); (iii) variability within individuals and within the same tissue due to tissue heterogeneity (Bakay et al., 2002; Leung and Cavalieri, 2003); (iv) variability within individuals in different tissues or cell types (Hatfield et al., 2003). In this case the differences are more distinct than within the same tissue.

These variabilities can obscure the variation induced by the stimulation of the organism (Novak et al., 2002), leading to results meaningless in the context of the stated hypothesis. Biological variability can be addressed in part at the early stages by proper experimental design, for example by a sufficient number of biological repeats (Bolstad et al., 2004; Novak et al., 2002; Yang and Speed, 2002) that can be used to average over the data and validate the conclusions over a range of biological samples. For further validation of the results, they should be confirmed using alternative experimental techniques on a number of biological samples (Novak et al., 2002).

**Technical variability**

This usually represents some kind of systematic error or bias introduced in the experimental process and once known can be corrected for in the data analysis, such as normalization. It can also be reduced by appropriate experimental design. Examples are mentioned in Table 4.1. To reduce technical variability, experimental protocols that result in reproducible, reliable results can be identified and then followed meticulously (Novak et al., 2002). Dye swap cDNA microarray experiments, in which the labeling dye of the samples is reversed in the repeat (Leung and Cavalieri, 2003; Kerr and Churchill, 2001), are used to account for dye-based bias. To estimate the influence of technical variability on results of both transcriptome and proteome experiments, technical repeats can be used (Bolstad et al., 2004; Novak et al., 2002; Leung and Cavalieri, 2003; Molloy et al., 2003). Formulae have been devised to determine the number of repeats and samples by taking into account the effects of pooling, technical replicates, and dye-swaps (Dobbin and Simon, 2005).

**Experimental design**

Experimental design not only includes the decision about the number of biological and technical replicates, it also includes all the decisions about sample preparation methods, experimental techniques and data analysis methods. All these decisions should be made to ensure that the data collected in the experiment will provide the information to support or reject the scientific hypothesis.

*Table 4.1.* Examples of technical variability introduced in transcriptomics and proteomics

| | *Wet lab* | | *Dry lab* |
|---|---|---|---|
| | **Sample preparation**<br>Variation in sample collection and preparation | **Experimental process**<br>Variation in experimental data collection processes | **Data analysis**<br>Variation in data processing and analysis |
| Transcriptomics | • RNA extraction and labeling (van Bakel and Holstege, 2004; Bakay et al., 2002; Bolstad et al., 2004; Hatfield et al., 2003; Novak et al., 2002). Variability in the sample preparation can result in change of the gene expression profile.<br>• Sample contamination (Leung and Cavalieri, 2003).<br>• Dye-based bias, i.e., one dye might be 'brighter' than the other dye (Kerr and Churchill, 2001; Leung and Cavalieri, 2003). | • Variation in hybridization process (van Bakel and Holstege, 2004; Bolstad et al., 2004; Hatfield et al., 2003; Novak et al., 2002; Yang and Speed, 2002). Variations introduced in the process can obscure changes caused by the stimulation of the organism, i.e., changes that the experiment actually seeks to determine. | • Different data processing approaches (van Bakel and Holstege, 2004; Hatfield et al., 2003). The wide range of available analysis approaches make it hard to assess the performance of each of them and to compare the results of experiments carried out in different labs. |
| Proteomics | • Variability in sample preparation and processing for LC/MS/MS can lead to differences in the number of low intensity peaks measured (Stewart et al., 2004). This can result in the identification of fewer peptides and proteins. | • Variability in tandem mass spectra collection (LC/MS/MS) (Venable and Yates, III, 2004; Stewart et al., 2004). Variability introduced here can lead to errors in search algorithms and ultimately to false positives in peptide identification.<br>• Quantitative variation between matched spots in two 2D-gels and fewer spots that can be matched in repeated gels (Molloy et al., 2003). | • Variability in tandem mass spectra processing (LC/MS/MS) (Venable and Yates, III, 2004; Stewart et al., 2004). Some of the search algorithms used to identify peptides might be more or less sensitive to the variability introduced during the collection of mass spectra (Venable and Yates, III, 2004), resulting in a different number of identified peptides in the same spectra using different algorithms. |

Badly designed experiments might not only not provide the answers to the questions stated, but might also leave potential bias in the data that might compromise the analysis and interpretation of the result (Yang and Speed, 2002). Reviews of experimental design of transcriptome and proteome experiments can be found in (Kerr and Churchill, 2001; Yang and Speed, 2002; Bolstad et al., 2004; Riter et al., 2005).

The number of variabilities that affect the outcome of an experiment make it hard to assess its quality. As we argue in the next section, an accurate record of the experimental design and of the environmental variables involved is a necessary, but hardly sufficient, condition to provide objective indicators that can be used to assess confidence in the experimental results.

## 3.2 Analysis of quality issues and techniques

Results of our survey analysis are presented in Tables 4.2 and 4.3 for transcriptomics and proteomics experiments, respectively. Each group of entries corresponds to one of the general quality concerns from Figure 4.2 (first column

in the table); for each group, specific problems are listed in the third column, and a summary of associated current practices and techniques including examples that illustrate the need to address the issues using those practices and techniques follows in the last column. An additional grouping of these issues by type of artifact produced during the process (second column) is provided where appropriate. For instance, "repeatibility and reproducibility" (second group) in Table 4.2 maps to two problems, of general adequacy of the process description for future reference, and of control of variability factors. For the latter, the issues are sufficiently distinct to suggest grouping them by artifact (hybridized array, raw data, interpretation of normalized data).

These tables, along with the selected references associated to the entries, are designed as a sort of "springboard" for investigators who are interested in a deeper understanding of the issues discussed in this chapter.

Quality issues that are not addressed in the process of the experiment may result in poor data quality in form of false positives or false negatives and may lead to incorrect conclusions. As these high-throughput experiments are frequently used not only to test hypotheses, but due to their scale also to generate new hypotheses, these new hypotheses might be wrong and follow-up experimental expenses and time to test these hypotheses may be wasted.

## 3.3 Specificity of techniques and generality of quality dimensions

Most of the techniques mentioned in the tables, on which we will not elaborate due to space constraints, are specific and difficult to generalize into a reusable "quality toolkit" for this data domain. While this may be frustrating to some quality practitioners in the information systems domain, we can still use some of the common terms for quality dimensions, provided that we give them a correct interpretation. A good example is the definition of "accuracy": in its generality, it is defined as the distance between a data value and the *real* value. When applied to a record or a field in a relational database table, this definition is specialized by introducing distance functions that measure the similarity between the value in the record and a *reference* value, for instance by computing the edit distance between strings. Further distinctions are made depending on the type of similarity that we seek to measure.

In the experimental sciences, the abstract definition for accuracy is identical (see for instance (van Bakel and Holstege, 2004)), however for a value that represents the numeric output of an experimental process, accuracy is interpreted in statistical terms, as a measure of systematic error, e.g., background noise in the experiment. Consequently, techniques for estimating accuracy, i.e., the equivalent of "distance functions", are grounded in the nature of the process, and are aimed at measuring and controlling noise. In (Fang et al., 2003), for example, a novel statistical model is proposed for the analysis of systematic errors

*Table 4.2.* Quality issues in transcriptomics experiments

| *Common quality issues* | *Artifact* | *Specific issues* | *Examples, techniques and references* |
|---|---|---|---|
| Quality of sample | Biological assay | RNA contamination control<br>biological variability | technical assessment of RNA quality (Imbeaud et al., 2005); low quality RNA may compromise results of data analyses |
| Process repeatability, results reproducibility | (General) | adequate process description | provenance, meta-data capture standards and techniques for fine-grain process description (Zhao et al., 2004; Greenwood et al., 2003) |
| | Raw data (image from hybridized array) | biological variability (Bakay et al., 2002)<br>technical variability: consistency of image quality control parameters | review: (Leung and Cavalieri, 2003; Hess et al., 2001)<br>experimental design (Kerr and Churchill, 2001; Dobbin and Simon, 2005) |
| | Normalized data | significance of interpretation given biological and technical variability | see "significance of data interpretation" |
| Data comparability | (General) | reproducibility across platforms, technologies, and laboratories | methods to accommodate variability across platforms and labs (Members of the Toxicogenomics Research Consortium, 2005; Larkin et al., 2005)<br>consistency of results across platforms (Wang et al., 2005; Petersen et al., 2005) |
| Significance of data | (General) | variability control | quantification of measurement errors (Huber et al., 2002) |
| | Raw data | image accuracy: interpretation of spots and their intensity levels, non-uniform hybridization | image analysis and quality control (Leung and Cavalieri, 2003; Hess et al., 2001)<br>bad spot detection, background identification, image noise modelling, manual inspection of spots; poor image quality may require costly manipulations and decrease the power of the analysis |
| | Normalized data | choice of normalization algorithms | review: (Leung and Cavalieri, 2003); choosing an inadequate normalization algorithm may lead to an incomplete removal of systematic errors and affect the power of the downstream analysis;<br>low-level data analysis (Bolstad et al., 2004); statistical error analysis, dye-bias control and reduction (Fang et al., 2003); algorithms to control signal-to-noise ratios (Seo et al., 2004) |
| Significance of data interpretation | Data interpretation | validity of data analysis techniques and tools; robustness of analysis with respect to variabilities | reviews on design and selection of clustering algorithms: (Kaminski and Friedman, 2002; Levenstien et al., 2003)<br>computational methods to take variabilities into account (Bakay et al., 2002; Hatfield et al., 2003)<br>algorithm performance analysis by cross-validation (Pepe et al., 2003)<br>identification of significant differences in gene expression: statistical analysis of replicated experiments (Dudoit et al., 2000); analysis of threshold choice to characterise disease vs. normality (Leung and Cavalieri, 2003; Pan et al., 2005); use of False Discovery Rate for generating gene expression scores (Pawitan et al., 2005; Reiner et al., 2003) |
| Quality of reference data | (General) | accuracy, completeness, specificity, currency of reference databases<br>functional annotations in reference databases | mostly based on practitioners' personal perception; systematic studies are needed (see Section 3.4 in main text) |
| Uniformity of representation, re-usability of results | Output data, publication | heterogeneity of presentation | data and meta-data standardization of content and presentation format (Leung and Cavalieri, 2003; Editors, 2002) (see also Section 4 in main text) |

*Table 4.3.* Quality issues in protein identification experiments

| *Common quality issues* | *Artifact* | *Specific issues* | *Examples, techniques and references* |
|---|---|---|---|
| Quality of sample | Biological assay | biological variability, contamination control | sample contamination with, e.g., human proteins from the experimenter may obscure the results of downstream analysis |
| Process repeatability, results reproducibility | (General) | adequate process description | data modelling for capturing experiment design and execution results (Fenyoe and Beavis, 2002)<br>the PEDRO data model (Taylor et al., 2003)<br>see also uniformity, below |
| | Raw data (mass spectra) | technical and biological variability | analysis of reproducibility (Stewart et al., 2004)<br>quantitative assessment of variability (Challapalli et al., 2004) |
| Data comparability | (General) | reproducibility across platforms, technologies, and laboratories | review: (Hancock et al., 2002) |
| Significance of data | (General) | variability control | review on statistical and computational issues across all phases of data analysis ((Listgarten and Emili, 2005))<br>review on analysis of sensitivity (Smith, 2002) |
| | Raw data (mass spectra) | sensitivity of spectra generation methods, dynamic range for relative protein abundance<br>technical and biological variability<br>limitations of technology for generating spectra | review on strategies to improve accuracy and sensitivity of PI, quantification of relative changes in protein abundance (Resing and Ahn, 2005)<br>studies on scoring models, database search algorithms, assessment of spectra quality prior to performing a search, analysis of variables that affect performance of DB search (Sadygov et al., 2004) (review), (Bern et al., 2004)<br>review on limitations of 2DE technology for low-abundance proteins (Flory et al., 2002) |
| Significance of data interpretation | Match results | significance and accuracy of match results, limitations of technology for accurate identification | definition (Colinge et al., 2003) and validation of scoring functions<br>review on limitations of technology: (Nesvizhskii and Aebersold, 2004)<br>statistical models (Nesvizhskii et al., 2003)<br>studies on matching algorithms (Sadygov et al., 2004) (review), (Zhang et al., 2002) |
| Quality of reference data | (General) | redundancy of reference DB (same protein appears under different names and accession numbers in databases)<br>accuracy, completeness, specificity, currency of reference databases | criteria for the selection of appropriate reference DB (Taylor et al., 2003): using a species-specific reference database will result in more real protein identifications than using a general reference database containing a large number of organisms; using the latter may result in a large number of false positives |
| Uniformity of representation, re-usability of results | Output data, publication | heterogeneity of presentation | need for representation standards (Ravichandran and Sriram, 2005)<br>the PEDRO proteomics data model (Taylor et al., 2003)<br>guidelines for publication (Carr et al., 2004)<br>standards for meta-data (Hancock et al., 2002) |

in microarray experiments. Here, errors that lead to low accuracy are detected and corrected by introducing different normalization techniques, whose effectiveness is compared experimentally; different statistical models are applied depending on the specific microarray experiment design used.

Information quality practitioners will probably be on more familiar ground when quality concepts like accuracy, currency, and timeliness are applied to reference databases used in the experiments, e.g., for protein-peptide matches, or to the last phase of our reference pipeline, when the differently expressed genes in a transcriptome experiment are functionally annotated. In this case, "accuracy" refers to the likelihood that a functional annotation is correct, i.e., that the description of the function of the gene or gene product corresponds to its *real* function[4]. As mentioned, annotations may be done either by human experts, based on publications evidence, or automatically by algorithms that try to infer function from structure and their similarity with that of other known gene products. In the first case, measuring accuracy amounts to supporting or disproving scientific claims made in published literature, while in the second, the predictive performance of an algorithm is measured.

In general, we observe a trade-off between the accuracy of curator-produced functional annotations, which have a low throughput, and the timeliness of the annotation, i.e., how soon the annotation becomes available after the gene product is submitted to a database. A notable example is provided by the Swiss-Prot and TrEMBL protein databases. While in the former, annotations are done by biologists, with great accuracy at the expense of timeliness, TrEMBL contains proteins that are automatically annotated, often with lower accuracy, but are made available sooner (Junker et al., 2000). This gives the scientist a choice, based on personal requirements. For well-curated database such as UniProt, claims of non-redundancy (but not of completeness) are also made (O'Donovan et al., 1999).

## 3.4 Beyond data generation: annotation and presentation

To conclude this section, we now elaborate further on the topic of functional annotations and their relationship to quality. The aim of annotation is, in general, to "bridge the gap between the sequence and the biology of the organism" (Stein, 2001). In this endeavour, three main layers of interpretation of the raw data are identified: nucleotide-level (where are the genes in a sequence?), protein-level (what is the function of a protein?), and process-level (what is the role of genes and proteins in the biological process? how do they interact?). The information provided by high-throughput transcriptomics and proteomics contributes to functional and process annotation. Thus, it participates in the cycle shown in Figure 4.3: publications are used by curators to produce functional annotations on protein database entries, which in turn may stimulate the

proposal of new experiments (automatic annotations use information from other databases, as well).

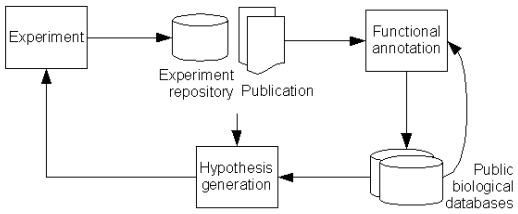


*Figure 4.3.* The annotation process

Although a bit simplistic, this view is sufficient to identify the critical issue with annotation: erroneous functional annotation based on biased results and conclusions due to unaccounted variabilities in experiments can propagate through public databases and further lead to wrong conclusions.

Most of the studies on the percolation effects of annotation errors have focused on automated annotations, in which protein function is determined computationally, based on sequence similarity to other proteins in the same domain (Karp et al., 2001; Gilks et al., 2002; Devos and Valencia, 2001; Wieser et al., 2004; Prlic et al., 2004). However, the issue of validating curated annotations that are based on published literature is more subtle. One approach is based on the observation that, when standard controlled vocabularies are used for annotation, the consistency of use of the terminology offered by these vocabularies in multiple independent annotations of the same data can be used as an indicator of annotation accuracy.

As an example, consider the Gene Ontology (GO), a well-known standard ontology for describing the function of eukaryotic genes (Consortium, 2000). GO is maintained by a consortium of three model organism databases, and consists of three parts: molecular function, biological process and cellular component (the sub-cellular structures where they are located). Up to the present, GO annotations have been used to annotate almost two million gene products in more than 30 databases. UniProt is the most prominent, accounting for almost 50% of the annotations.

The adoption of such a standard in biology has allowed researchers to investigate issues of annotation consistency. We mention two contributions here. The first (Lord et al., 2003) has studied measures of *semantic similarity* between SwissProt entries, based on their GO annotations. The authors hypothesize that valid conclusions about protein similarity can be drawn not only based on their sequence similarity (as would be done for instance by BLAST), but also from the semantic similarity of the annotations that describe the biological role of the proteins. The latter is described by metric functions defined on the GO

structure (GO is a directed acyclic graph). Based on statistical evidence, the authors conclude that the hypothesis is valid for various specific assumptions, e.g., that the data set is restricted to those proteins whose annotations are supported by published literature, as opposed to being inferred from some indirect data source.

The second contribution has studied the consistency of annotations among orthologues in different databases[5] (Dolan et al., 2005). Experiments on sets of mouse and human proteins resulted in a useful classification of annotation errors and mismatches, and in effective techniques for their detection.

These studies offer a partial, but quantitative validation of the main claim that standardization of terminology improves the confidence in the annotation process and facilitates the retrieval of information.

# 4. CURRENT APPROACHES TO QUALITY: META-DATA COLLECTION, STANDARDIZATION, VOCABULARIES

Partly to dominate the complexity of the domain and the broad variability of available techniques, the information management community has been adopting a general approach towards standardization based on (i) modelling, capturing and exploiting meta-data that describes the experimental processes in detail, known as *provenance*; and (ii) creating controlled vocabularies and ontologies used to describe the meta-data.

Information quality management may benefit greatly from this approach.

## 4.1 Modelling, collection and use of provenance meta-data

Throughout this chapter, we have mentioned a number of variability factors that affect the outcome of an experiment. The meta-information about these variables and their impact, i.e., the experimental design and details of experiment execution, is known as *provenance*. The importance of capturing provenance in a formal and machine-processable way has been recognized in the recent past, as a way to promote interoperability and uniformity across labs. The role of provenance in addressing quality issues, however, has not yet been properly formalized. Recent research efforts have been focusing on using provenance and other types of meta-data, to allow scientists to formally express quality preferences, i.e., to define decision procedures for selecting or discarding data based on underlying quality indicators (Missier et al., 2005).

Standards for capturing provenance are beginning to emerge, but much work is still to be done. Within the transcriptomic community, one initial response comes from the Microarray Gene Expression Data (MGED) society, which has proposed a standard set of guidelines called MIAME (Brazma et al., 2001), for Minimal Information About a Microarray Experiment, prescribing minimal

content for an acceptable database submission. Along with content standardization, the Microarray and Gene Expression (MAGE) group within MGED in collaboration with the Object Management Group (OMG) also defines MAGE-OM, an object model describing the conceptual structure of MIAME documents. The model has been mapped to MAGE-ML, an XML markup language for writing MAGE-OM documents, resulting in a complete standard for the preparation of MIAME-compliant database submissions.

Furthermore, MAGE prescribes that experiment descriptions be annotated using the MGED ontology, a controlled vocabulary for the gene expression domain. MGED is currently being redesigned, with the goal of encompassing a broader domain of functional genomics, and will hopefully include a structure and terminology for experimental variables, which is currently missing. Writing complete MAGE-ML documents is a lengthy process for non-trivial experiments. At present, adoption of the standard by the research community is driven mostly by the requirement that data submitted to major journals for publication be MIAME-compliant.

Similar efforts are under way in the proteomic field (Orchard et al., 2003), although accepted standards do not yet exist for data models and format (although some proposed data models like PEDRo are being increasingly adopted by the community (Garwood et al., 2004)). The Human Proteome Organisation (HUPO) provides updated information on its Proteomics Standards Initiative (PSI).

The challenge for these standardization efforts is the rapid development of functional genomics. This requires these standards to be specific enough to capture all the details of the experiments but at the same time to be generic and flexible enough to adapt and be extended to changes in existing or evolving of new experimental techniques. Furthermore, these standards need to cater for different communities within the large and diverse biological community. Examples of this diversity include the study of eukaryotes or prokaryotes, model organisms that have already been sequenced or non-model organisms with only limited amount of information available, inbred populations that can be studied in controlled environment or outbred populations that can only be studied in their natural environment[6].

To allow a systems biology approach to the analysis of data from different kinds of experiments, a further effort is undertaken by a number of standardization bodies to create a general standard for functional genomics (FuGE)[7]. This effort is based on the independent standards for transcriptomics and proteomics mentioned above and seeks to model the common aspects of functional genomics experiments.

One of the practical issues with provenance data is that, in the wet lab, the data capture activity represents additional workload for the experimenter, possibly assisted by the equipment software. The advantage in the dry lab is that

extensive information system support during experiment execution is available, in particular based on workflow technology, as proven in the myGrid project (Zhao et al., 2004; Greenwood et al., 2003). In this case, provenance can be captured by detailed journaling of the workflow execution.

## 4.2 Creating controlled vocabularies and ontologies

The second approach for a standardized representation of data and meta-data is the development of controlled vocabularies and ontologies. A large number of ontologies is being developed, including ontologies to represent aspects of functional genomics experiments, such as the MGED ontology for transcriptomics[8] or the PSI ontology for proteomics[9], both of which will form part of the Functional Genomics Ontology (FuGO, part of FuGE).

As for the development of standardized models for meta-data, the development of standardized controlled vocabulary faces similar challenges, such as the rapid development of the technologies that are described in the ontology or the knowledge presented in a controlled vocabulary. Furthermore, the representation of the ontologies varies, ranging from lists of terms to complex structures modeled using an ontology language, such as OWL[10].

# 5. CONCLUSIONS

We have presented a survey on quality issues that biologists face during the execution of transcriptomics and proteomics experiments, and observed that issues of poor quality in published data can be traced to the complexity of controlling the biological and technical variables within the experiment.

Our analysis suggests that, despite their differences, a common structure and a common set of quality issues for the two classes of experiments can be found; we have proposed a framework for the classification of these issues, and used it to survey current quality control techniques.

We argued that the scientists' ability to make informed decisions regarding the quality of published data relies on the availability of meta-information describing the experiment variables, as well as on standardization efforts on the content and structure of meta-data.

The area of information management can play a major role in this efforts, by providing suitable information management models for meta-data, and tools to exploit it. Although the literature offers many more results on these topics that can be presented here, we have offered a starting point for in-depth investigation of this field.

# ACKNOWLEDGEMENTS

We would like to thank Dr. Simon Hubbard and his group at the University of Manchester for help during the preparation of this manuscript and Dr. Suzanne

Embury and Prof. Norman Paton at the University of Manchester for valuable comments. This work is supported by the Wellcome Trust, the BBSRC and the EPSRC.

## NOTES

1 The term "information" is often used in contrast with "data", to underline the difference between the ability to establish formal correctness of a data item, and the ability to provide a correct interepretation for it. In this sense, assessing reliability is clearly a problem of correct interpretation, hence of *information quality*.

2 The term gene expression refers to the process of DNA transcription for protein production within a cell. For a general introduction to the topics of genomic and proteomics, see (Campbell and Heyer, 2003).

3 For a tutorial on microarrays, see http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html.

4 Assessing the *completeness* of an annotation is just as important; however, the intrinsic incompleteness of the biological interpretation of genes and gene products (King et al., 2003) makes this task even more challenging.

5 Orthologues are similar genes that occur in the genomes of different species.

6 See, e.g., http://envgen.nox.ac.uk/miame/miame_env.html for a proposal to extend the MIAME standard to take into account requirements of the environmental genomics community.

7 http://fuge.sourceforge.net/ and http://sourceforge.net/projects/fuge/

8 http://mged.sourceforge.net/ontologies/

9 http://psidev.sourceforge.net/ontology/

10 http://www.w3.org/TR/owl-features/

## REFERENCES

Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422:198–207.

Apweiler, R., Bairoch, A., Wu, C., et al. (2004). Uniprot: the universal protein knowledgebase. *Nucleic Acids Res*, 32:D115–D119.

Bairoch, A., Boeckmann, B., Ferro, S., et al. (2004). Swiss-prot: Juggling between evolution and stability. *Brief Bioinform*, 5:39–55.

Bakay, M., Chen, Y.-W., Borup, R., et al. (2002). Sources of variability and effect of experimental approach on expression profiling data interpretation. *BMC Bioinformatics*, 3:4.

Ballou, D., Wang, R., Pazer, H., et al. (1998). Modelling information manufacturing systems to determine information product quality. *Journal of Management Sciences*, 44.

Bern, M., Goldberg, D., et al. (2004). Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics*, 20:i49–i54.

Bolstad, B.M., Collin, F., et al. (2004). Experimental design and low-level analysis of microarray data. *Int Rev Neurobiol*, 60:25–58.

Bowtell, D.D.L. (1999). Options available - from start to finish - obtaining expression data by microarray. *Nat Genet*, 21:25–32.

Brazma, A., Hingamp, P., et al. (2001). Minimum information about a microarray experiment (miame)- towards standards for microarray data. *Nat Genet*, 29:365–371.

Campbell, A.M. and Heyer, L.J. (2003). *Discovering Genomics, Proteomics, and Bioinformatics*. Benjamin Cummings.

Carr, S., Aebersold, R., Baldwin, M., et al. (2004). The need for guidelines in publication of peptide and protein identification data. *Mol Cell Proteomics*, 3:531–533.

Challapalli, K.K. et al. (2004). High reproducibility of large-gel two-dimensional electrophoresis. *Electrophoresis*, 25:3040–3047.

Cheung, V.G., Morley, M., Aguilar, F., et al. (1999). Making and reading microarrays. *Nat Genet*, 21:15–19.

Colinge, J., Masselot, A., et al. (2003). A systematic analysis of ion trap tandem mass spectra in view of peptide scoring. In *Proc. Third International Workshop on Algorithms in Bioinformatics (WADI)*, Budapest.

Consortium, The Gene Ontology (2000). Gene ontology: tool for the unification of biology. *Nat Genet*, 25:25–29.

de Hoog, C.L. and Mann, M. (2004). Proteomics. *Annu Rev Genomics Hum Genet*, 5:267–293.

Devos, D. and Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends Genet*, 17:429–431.

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

Dolan, M.E., Ni, L., Camon, E., et al. (2005). A procedure for assessing go annotation consistency. *Bioinformatics*, 21:i136–i143.

Dudoit, S., Yand, Y.H., Callow, M.J., et al. (2000). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica*, 12:111–139.

Editors (2002). Microarray standards at last. *Nature*, 419:323.

Eng, J., McCormack, A.L., et al. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*, 5:976–989.

English, L.P. (1999). *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. John Wiley & Sons, 1 edition. ISBN: 0471253839.

Fang, Y., Brass, A., et al. (2003). A model-based analysis of microarray experimental error and normalisation. *Nucleic Acids Res*, 31:e96.

Fenyoe, D. and Beavis, R.C. (2002). Informatics and data management in proteomics. *Trends Biotechnol*, 20:S35–S38.

Flory, M.R., Griffin, T.J., et al. (2002). Advances in quantitative proteomics using stable isotope tags. *Trends Biotechnol*, 20:S23–S29.

Garwood, K., McLaughlin, T., Garwood, C., et al. (2004). Pedro: A database for storing, searching and disseminating experimental proteomics data. *BMC Genomics*, 5:68.

Gilks, W., Audit, B., Angelis, D. De, et al. (2002). Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, 18:1641–1649.

Greenwood, M., Goble, C., Stevens, R., et al. (2003). Provenance of e-science experiments - experience from bioinformatics. In *OST e-Science Second All Hands Meeting 2003 (AHM'03)*, Nottingham, UK.

Hancock, W.S., Wu, S.L., Stanley, R.R., et al. (2002). Publishing large proteome datasets: scientific policy meets emerging technologies. *Trends Biotechnol*, 20:S39–S44.

Hatfield, G.W., Hung, S., and Baldi, P. (2003). Differential analysis of dna microarray gene expression data. *Mol Microbiol*, 47:871–877.

Hess, K.R., Zhang, W., Baggerly, K.A., et al. (2001). Microarrays: handling the deluge of data and extracting reliable information. *Trends Biotechnol*, 19:463–468.

Holloway, A.J., van Laar, R.K., Tothill, R.W., et al. (2002). Options available - from start to fnish - for obtaining data from dna microarrays ii. *Nat Genet*, 32:481–489.

Huber, W., von Heydebreck, A., et al. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18:S96–S104.

Hunter, T.C., Andon, N.L., Koller, A., et al. (2002). The functional proteomics toolbox: methods and applications. *J Chromatogr B*, 782:165–181.

Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: Systems biology. *Annu Rev Genomics Hum Genet*, 2:343–372.

Imbeaud, S., Graudens, E., Boulanger, V., et al. (2005). Towards standardization of rna quality assessment using user-independent classifiers of microcapillary electrophoresis traces. *Nucleic Acids Res*, 33.

Junker, V., Contrino, S., Fleischmann, W., et al. (2000). The role swiss-prot and TrEMBL play in the genome research environment. *J Biotechnol*, 78:221–234.

Kaminski, N. and Friedman, N. (2002). Practical approaches to analyzing results of microarray experiments. *Am J Respir Cell Mol Biol*, 27:125–132.

Karp, P., Paley, S., and Zhu, J. (2001). Database verification studies of swiss-prot and genbank. *Bioinformatics*, 17:526–532.

Kerr, M.K. and Churchill, G.A. (2001). Experimental design for gene expression microarrays. *Biostatistics*, 2:183–201.

King, O.D., Foulger, R.E., Dwight, S.S., et al. (2003). Predicting gene function from patterns of annotation. *Genome Res*, 13:896–904.

Lacroix, Z. and Critchlow, T., editors (2004). *Bionformatics - Managing Scientific Data*. Elsevier.

Larkin, J.E., Frank, B.C., et al. (2005). Independence and reproducibility across microarray platforms. *Nat Methods*, 2:337–343.

Leung, Y.F. and Cavalieri, D. (2003). Fundamentals of cdna microarray data analysis. *Trends Genet*, 19:649–659.

Levenstien, M.A., Yand, Y., and Ott, J. (2003). Statistical significance for hierarchical clustering in genetic association and microarray expression studies. *BMC Bioinformatics*, 4:62.

Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R., et al. (1999). High density synthetic oligonucleotide arrays. *Nat Genet*, 21:20–24.

Listgarten, J. and Emili, A. (2005). Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics*, 4:419–434.

Lockhart, D.J. and Winzeler, E.A. (2000). Genomics, gene expression and dna arrays. *Nature*, 405:827–836.

Lord, P.W., Stevens, R.D., et al. (2003). Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19:1275–1283.

Members of the Toxicogenomics Research Consortium (2005). Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods*, 2:1–6.

Mewes, H.W., Amid, C., et al. (2004). Mips: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*, 32:D41–D44.

Missier, P., Embury, S., Greenwood, M., et al. (2005). An ontology-based approach to handling information quality in e-science. In *Proc. 4th e-Science All Hands Meeting*.

Molloy, M.P., Brzezinski, E.E., Hang, J., et al. (2003). Overcoming technical variation and biological variation in quantitative proteomics. *Proteomics*, 3:1912–1919.

Motro, A., Anokhin, P., and Acar, A.C. (2004). Utility-based resolution of data inconsistencies. In *Intl. Workshop on Information Quality in Information Systems 2004 (IQIS'04)*, Paris, France.

Mueller, H., Naumann, F., and Freytag, J.C. (2003). Data quality in genome databases. In *Proc. Eight International Conference on Information Quality (ICIQ03)*, Cambridge, MA. MIT.

Naumann, F., Freytag, J.C., et al. (2004). Completeness of integrated information sources. *Information Systems*, 29(7):583–615.

Nesvizhskii, A.I. and Aebersold, R. (2004). Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem ms. *Drug Discov Today*, 9:173–181.

Nesvizhskii, A.I., Keller, A., Koller, E., et al. (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, 75:4646–4658.

Novak, J.P., Sladek, R., and Hudson, T.J. (2002). Characterization of variability in large-scale gene expression data: implications for study design. *Genomics*, 79:104–113.

O'Donovan, C., Martin, M., Glemet, E., et al. (1999). Removing redundancy in swiss-prot and trembl. *Bioinformatics*, 15:258–259.

Orchard, S., Hermjakob, H., and Apweiler, R. (2003). The proteomics standards initiative. *Proteomics*, 3:1374–1376.

Pan, K.-H., Lih, C.-J., and Cohen, S.N. (2005). Effects of threshold choice on biological conclusions reached during analysis of gene expression by dna microarrays. *Proc Natl Acad Sci USA*, 102:8961–8965.

Pandey, A. and Mann, M. (2000). Proteomics to study genes and genomes. *Nature*, 405:837–846.

Patterson, S.D. and Aebersold, R.H. (2003). Proteomics: the first decade and beyond. *Nat Genet*, 33:311–323.

Pawitan, Y., Michiels, S., Koscielny, S., et al. (2005). False discovery rate, sensitivity and samples size for microarray studies. *Bioinformatics*, 21:3017–3024.

Pepe, M.S., Longton, G., et al. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics*, 59:133–142.

Perkins, D.N., Pappin, D.J.C., Creasy, D.M., et al. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551–3567.

Petersen, D., Chandamouli, G.V.R., Geoghegan, J., et al. (2005). Three microarray platforms: an analysis of their concordance in profiling gene expression. *BMC Genomics*, 6:63.

Prlic, A., Domingues, F., Lackner, P., et al. (2004). Wilma-automated annotation of protein sequences. *Bioinformatics*, 20:127–128.

Quackenbush, J. (2001). Computational analysis of microarray data. *Nat Rev Genet*, 2:418–427.

Rahm, E., editor (2004). *First International Workshop, DILS 2004*, volume 2994 of *Lecture Notes in Bioinformatics*.

Ravichandran, V. and Sriram, R.D. (2005). Toward data standards for proteomics. *Nat Biotechnol*, 23:373–376.

Redman, T.C. (1996). *Data quality for the information age*. Artech House.

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19:368–375.

Resing, K.A. and Ahn, N.G. (2005). Proteomics strategies for protein identification. *FEBS Lett*, 579:885–889.

Riter, L.S., Vitek, O., et al. (2005). Statistical design of experiments as a tool in mass spectrometry. *J Mass Spectrom*, 40:565–579.

Sadygov, R.G., Cociorva, D., and Yates, III, J.R. (2004). Large-scale database searching using tandem mass spectra: Looking up the answers in the back of the book. *Nat Methods*, 1:195–201.

Seo, J., Bakay, M., Chen, Y.-W., et al. (2004). Interactively optimizing signal-to-noise ratios in expression profiling: project-specific algorithm selection and detection p-value weighting in affymetrix microarrays. *Bioinformatics*, 20:2534–2544.

Smith, R.D. (2002). Trends in mass spectrometry instrumentation for proteomics. *Trends Biotechnol*, 20:S3–S7.

Stein, L.D. (2001). Genome annotation: From sequence to biology. *Nat Rev Genet*, 2:493–503.

Stewart, I.I., Zhao, L., Bihan, T. Le, et al. (2004). The reproducible acquisition of comparative liquid chromatography/tandem mass spectrometry data from complex biological samples. *Rapid Commun Mass Spectrom*, 18:1697–1710.

Taylor, C.F., Paton, N.W., Garwood, K.L., et al. (2003). A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat Biotechnol*, 21.

van Bakel, H. and Holstege, F.C.P. (2004). In control: systematic assessment of microarray performance. *EMBO reports*, 5:964–969.

Venable, J.D. and Yates, III, J.R. (2004). Impact of ion trap tandem mass spectra variability on the identification of peptides. *Anal Chem*, 76:2928–2937.

Wang, H., He, X., et al. (2005). A study of inter-lab and inter-platform agreement of dna microarray data. *BMC Genomics*, 6:71.

Wang, R. (1998). A product perspective on total data quality management. *Communications of the ACM*, 41(2).

Wieser, D., Kretschmann, E., and Apweiler, R. (2004). Filtering erroneous protein annotation. *Bioinformatics*, 20:i342–i347.

Winkler, W.E. (2004). Methods for evaluating and creating data quality. *Information Systems*, 29(7).

Yang, Y.H. and Speed, T. (2002). Design issues for cdna microarray experiments. *Nat Rev Genet*, 3:579–588.

Zhang, N., Aebersold, R., et al. (2002). Probid: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, 2:1406–1412.

Zhao, J., Wroe, C., et al. (2004). Using semantic web technologies for representing e-science provenance. In *Third International Semantic Web Conference (ISWC2004)*, Hiroshima, Japan.