Studies in History and Philosophy of Biological and Biomedical Sciences 43 (2012) 37–51

Contents lists available at SciVerse ScienceDirect

Studies in History and Philosophy of Biological and Biomedical Sciences

journal homepage: www.elsevier.com/locate/shpsc

Too many numbers: Microarrays in clinical cancer research

Peter Keating[a], Alberto Cambrosio[b]

[a] Department of History, Université du Québec à Montréal, Case Postale 8888, Succursale Centre-ville, Montréal, Québec, Canada H3C 3P8
[b] Department of Social Studies of Medicine, McGill University, 3647 Peel Street, Montreal, Quebec, Canada H3A 1X1

> Technological progress in genome-wide measurements has changed the scientific discovery process to more data-driven rather than hypothesis-driven approaches.[1]
>
> The advent of global gene expression analysis changed molecular biology from the reductionist 'one gene, one post-doc' to a 'one experiment, one gigabit' model.[2]

1. Introduction

Microarrays allow researchers to simultaneously analyze the activity of thousands of genes and are thus a key technology of post-genomic biomedicine. They emerged around 1990, were first deployed in experimental research in 1995, but did not find their most important field of application, cancer clinical research, until the end of the century.[3] Microarrays and the deluge of data they produce provide an interesting case study of recent claims that data-driven research has replaced hypothesis-driven research. We find this claim wanting: while there is no denying that data-driven research—characterized by the development of 'high-throughput' molecular technologies that generate massive amounts of data stored in large databases—is now a reality, instead of replacing hypothesis-driven research it has added to it, engendering novel dynamics and complexities.

Microsoft researchers have announced with great fanfare the advent of data-intensive research, dubbed the 'fourth paradigm'.[4] The two epigraphs to this article express a similar sentiment, namely that post-genomic research in the biological and biomedical sciences inexorably leads to the marginalization of hypothesis-driven research to the advantage of data-driven research. Yet, we also hear dissonant voices. In a 2011 interview Janet D. Rowley, for instance, whose landmark findings on translocated chromosomes earned her the title of the 'matriarch of modern cancer genetics', claims that her career would not have been possible today: 'I was doing observationally driven research. That's the kiss of death if you're looking for funding today. We're so fixated now on hypothesis-driven research that if you do what I did, it would be called a "fishing expedition," a bad thing'.[5] The opposition between 'fishing expeditions' and 'hypothesis testing' is not new. The development of monoclonal antibodies and automated cell-sorting techniques in the 1980s and 1990s engendered a proliferation of novel entities, known as cell-surface markers, of unknown function: some researchers worried that this sort of 'stamp collecting' crowded out experiments that tested hypotheses about the function of a few, well-characterized markers.[6] On a grander scale, critics of the Human Genome Project expressed similar fears.[7] Outside of biomedicine, terms such as 'stamp collecting' have been used throughout the 20th century to ridicule natural history, which, according to the standard historical account, was in the process of being superseded by a triumphant experimentalism.[8] And yet, as Bruno Strasser has shown, the reassembling of biological and biomedical research at the turn of the 21st century is more the outcome of a hybridization of these two 'ways of knowing', rather than the replacement of one by the other.[9]

This hybridization process, 'especially as seen through the use of databases of experimental data, is now too widespread for historians to ignore'.[10] We ignore the extent to which historians have come to grips with this reality, but scientists certainly took notice: a simple PubMed search shows that the number of articles including

E-mail addresses: keating.peter@uqam.ca (P. Keating), alberto.cambrosio@mcgill.ca (A. Cambrosio)

[1] Pawitan, Michiels, Koscielny, Gusnanto, & Ploner (2005, p. 3017).
[2] Radich (2009, p. 165).
[3] Schena, Shalon, Davis, & Brown (1995), Perou et al. (1999), Alizadeh et al. (2000).
[4] Hey, Tansley, & Tolle, 2009.
[5] Dreifus (2011).
[6] Keating & Cambrosio (2003, pp. 186–187).
[7] Balmer (1996).
[8] Johnson (2007).
[9] Strasser (2008, 2011).
[10] Strasser (2011, p. 96).

1369-8486/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.shpsc.2011.10.004

in their title or abstract the term 'data-driven' has grown from 7 in 1990, to 41 in 2000, and 248 in 2010. Probably more significantly, the February 11, 2011 issue of *Science* was devoted to the problem of 'data deluge': a set of 10 articles explored the challenge of dealing with a quite unprecedented amount of data in domains such as genomics, where the increase in the sequencing output is overtaking computing and storage capacities.[11] While it is tempting to argue for the radical novelty of these events, we opt in this paper for a solution akin to Strasser's hybridization argument: we claim that in the domain of microarrays hybridization between statistical hypothesis testing and algorithm-driven data analysis is underway. In so doing, rather than speculating about epochal trends, we investigate the debate between bioinformaticians and biostatisticians as microarrays are transformed from an experimental technique to a clinical tool in the still largely programmatic field of personalized medicine. As a tool for the discovery of classes of genes defined by differential expression patterns, microarray data are not generally subject to the same strictures as clinical data. When, however, clinical material and potential downstream clinical uses enter the picture, so do biostatisticians. Firmly entrenched in clinical research after more than 40 years of fuelling statistical analyses for clinical trials, biostatisticians are quick to offer criticism and advice to their bioinformatics colleagues observing that the latter are too 'data-driven' and often lack 'inferential literacy', that is, the language and techniques of hypothesis-testing common to other fields of clinical research.[12] In other words, the extent of hybridization between data or hypothesis-driven analytical methods depends on the nature of the experiments to which those methods are applied.

During the early development of microarray technology, data analysis was conducted by an emerging group of practitioners, *bioinformaticians*, often trained in computer science and who specialize in developing algorithms for the storage, annotation, management and visualization of data generated by gene sequencing and gene expression profiling experiments. Historians of science have traced the development of bioinformatics back to the 1960s,[13] but the dramatic expansion of the domain—the creation of dedicated institutes and professional organization, the exponential growth in publications (Fig. 1) and the publication of numerous textbooks—is clearly a more recent event, linked to the emergence of gene cloning and sequencing technologies.[14] Indeed, one can argue that bioinformatics and post-genomic techniques such as microarrays and gene sequencing were co-produced. The algorithms implemented by bioinformaticians (such as hierarchical clustering) did not test hypotheses about the data: they created novel biomedical categories, such as subtypes of cancer, from the analysis of the results of high-throughput genomics. In the early years, microarray experiments were primarily concerned with discovery, and researchers placed little emphasis on problems of clinical validation and what that would mean for microarrays.

Soon, however, another group of practitioners, *biostatisticians*, entered the fray. Since the rise of clinical trials as a gold standard for medical evidence during the second half of the 20th century, biostatisticians have controlled the handling and meaning of clinical and epidemiological data, using and extending well-known statistical tools for that purpose, such as significance tests, *p*-values, odd-ratios and the like. As microarrays moved closer to the clinical domain in the early years of the new century, experts in clinical statistics became more closely involved in microarray experiments as part of multidisciplinary teams and offered both critiques of and techniques for the production of clinically useful and statistically valid microarray data. In particular, biostatisticians insisted that statistical methods for testing hypotheses should become a *sine qua non* for microarray experiments, and have since successfully enforced this point of view. Microarrays, however, and high-throughput technologies more generally, generate unprecedented amounts of data about thousands of genes, proteins and related confounding factors. This change of scale raised statistical challenges that could not be addressed by the simple application of time-honored statistical tools. Biostatisticians had to adapt them to the novel data landscape created by microarrays as well as to the new bio-pathological categories and processes they generated. In addition to displaying the impressive growth of publications using microarrays since the mid-1990s, Fig. 1 shows that during the first years of the new century the publication rate of statistical papers on microarrays paralleled the growth of the domain. It has since stabilized as a set of statistical methods adapted to microarray experiments have become entrenched in the domain.

The transition from a bioinformatics-centered to a biostatistically savvy practice is aptly illustrated by the following episode. When asked during an interview in 1999 if she planned to use microarrays in diagnostic and prognostic studies in clinical trials, a leading scientist for one of the major US cancer clinical trial networks answered 'No', adding: 'I would if we were able to send out a statistician with each microarray'.[15] This answer underscored a problem for both clinicians and researchers: microarrays generated a lot of data, but the interpretation of the data remained problematic. Nonetheless the use of microarrays in laboratory research exploded between 1999 and 2002 as registered not only by the aforementioned growth in the number of publications, but also by the fact that *Nature Genetics*, having produced a review of the technology in 1999, felt obliged to produce a second review a scant three years later.[16] By then, statisticians themselves had come to see microarray technology as an entirely new field of play, if not a specialty in its own right. Summarizing two NIH conferences held in 2004 to deal with a penury of biostatisticians, an article in the journal *Statistics in Medicine* observed: 'current biostatistics departments are competing for the same dozen top candidates, not just in the established methodology areas...but especially in the emerging fields such as microarray data and statistical genetics'.[17] In other words, within 5 years, microarrays had gone from a problem for clinical trialists to an 'emerging field' for statisticians.

At this point a terminological specification is needed. As used in this paper, the terms biostatistician and bioinformatician do not refer to a professional group or even to concrete individuals, but to a 'style of practice'[18]: a data-driven approach centered on processing and visualizing large amounts of data, and a hypothesis-driven approach focused on the testing of bio-clinical claims. Why, then, use those terms? One of the reasons is that our

[11] Kahn (2011).

[12] Miron & Nadon (2006).

[13] Hagen (2000).

[14] On the emergence of contemporary bioinformatics, see McMeekin & Harvey (2002), McMeekin, Harvey, & Gee (2004).

[15] Interview with a clinical researcher, New York, 30 May 1999.

[16] The chipping forecast (1999), The chipping forecast II (2001).

[17] DeMets et al. (2006, p. 3417).

[18] The notion of *style of practice*, as used by Fujimura & Chou (1994) and Keating & Cambrosio (2012), draws on a critical appraisal of the notion of *style of reasoning* developed by Hacking (1985) and (2002); see also Kusch (2010).

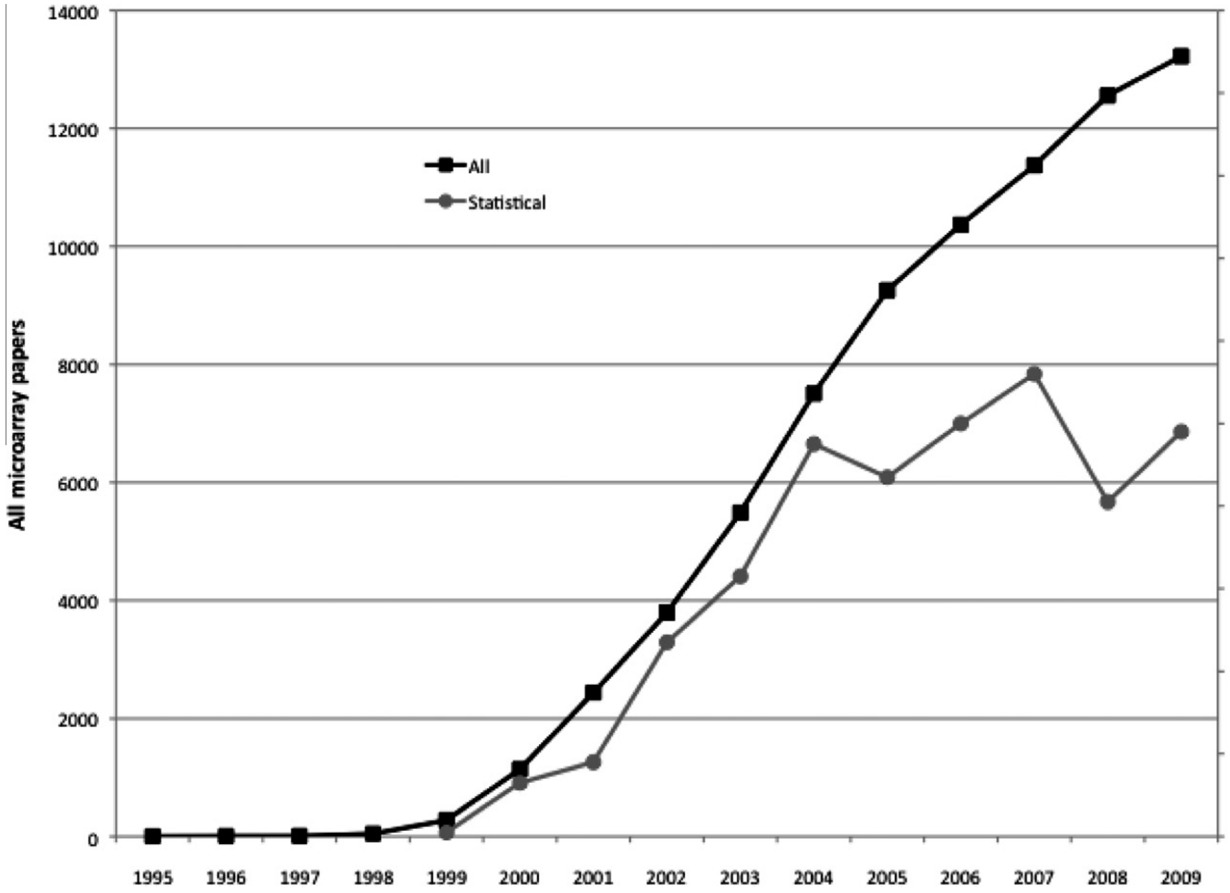

Fig. 1. Growth of microarray and microarray methodology literature listed in PubMed from 1995 to 2009. Updated version of a chart originally published in Mehta et al. (2004, p. 944).

informants do so, for instance when describing cases of outright conflict between bioinformaticians and biostatisticians.[19] But the term bioinformatician also covers a range of practitioners running from systems biologists and computer scientists to library scientists. Many of these practitioners refuse to identify themselves as bioinformaticians even when working as bona fide members of bioinformatics institutes or departments. In addition, as a result of a certain degree of cross-over and hybridization, bioinformaticians have now adopted many statistical routines and practices so that, in retrospect, clear lines between the types of practitioner are less easy to draw than between the two styles of practice: Bioinformaticians, moreover, are not passive bystanders watching biostatisticians (re)-gain control of the field: rather, they contribute to the development of tools that deal with some of the problems initially raised by biostatisticians. As for biostatisticians, they share a core of statistical knowledge and are held together in this story by a common interest in the application of statistical methods to the analysis of clinically relevant microarray data, but they also vary in terms of their professional and disciplinary provenance and standing. They are not restricted to university departments of biostatistics, as they also work in bioinformatics institutes and departments, cooperative oncology group statistical centers or for national cancer organizations like the NCI.

In what follows we will examine how statisticians adapted to the new world of microarrays, as well as some of the problems that created so much work for statisticians during the introduction of microarrays into clinical cancer research and practice.

2. Replication problems

One of the central problems of microarrays concerns the criteria for their reproducibility. Sociologists of science have long focused on reproducibility as a strategic domain for science studies:[20] as an ongoing concern for researchers, it provides a window onto the socio-technical arrangements that underlie experimental work. Microarray practitioners routinely instance three distinct kinds of problems: reproducing *experiments*, reproducing *data* and reproducing *results*. While the third is the focus of the present paper, we will begin by briefly describing all three. The first consists of those problems related to the replication of the microarray experiment itself within a single laboratory and with the same material. This sort of problem is common to all microarray experiments and statisticians further subdivide it into issues that arise from biological differences between different samples and those that arise from technical differences between individual experiments. In addition to biological differences between individual organisms, it is important to recall that within a single organism and within a single cell, every mRNA used in expression analysis is unique and has sequence and structural properties that will influence its behavior in a microarray experiment. Moreover, despite the best efforts of manufacturers, micro-

[19] The present article is part of a broader research project on cancer genomics that began in 2008 and is grounded in the systematic collection and analysis of published and unpublished documents, interviews (approximately 30 as of March 2011) with cancer clinicians, researchers, statisticians and the staff of biotech companies; and participant-observation at scientific meetings and within three clinical settings (in the US, Canada and France).

[20] The locus classicus is Collins (1992 [1985]).

array readings of the relative intensity of gene expression still admit of many variations due to fluctuations in dye intensity, spot formation and so on.[21] Even within a single laboratory these variations can be significant. According to one study at the turn of the century, measurements of the gene expression level of the same gene on two different places in a single microarray, for example, usually have a correlation of about 0.95. If researchers measure expression of that same gene on two different microarrays, the correlation falls somewhere in the range of 0.6 to 0.8. Correlation between samples taken from the same inbred mouse may fall to as low as 0.3.[22] Even when reproducibility is high, it does not by itself ensure accuracy: 'Unfortunately, a platform can have an excellent reproducibility without necessarily producing measurements that are accurate or consistent with other platforms.... Badly designed probes ... can easily provide highly reproducible and yet useless data'.[23]

The problems associated with the reproduction of a microarray experiment can be distinguished from those associated with the reproduction of microarray data.[24] Between 2000 and 2005, a number of studies compared platforms and laboratories and many found significant discrepancies.[25] Then, in 2005, three papers appeared in *Nature Methods* showing that given the proper precautions, measuring relative gene expression rather than absolute expression, and using the same statistical packages (other than those provided by manufacturers), comparisons between platforms were possible, if one also took into account the 'lab effect'.[26] In particular, one of the studies set out to control for both lab performance and platform, and showed that the effect of laboratory performance on inter-laboratory variability was larger and that results obtained in 'the best performing labs agree rather well'.[27] Editorialists in some journals such as *Nature Reviews Genetics*[28] breathed a sigh of relief. Others were not as sanguine. An editorial in *Trends in Genetics* commenting on the same three papers pointed to the repeated 'inconsistencies across platforms' and 'among laboratories that were using the same platform, and even using the same RNA samples',[29] thus advising caution. The principal authors of two of the three papers replied that the editorialist clearly did not appreciate the implications of their findings and asserted that 'to state that microarrays are not reliable seems to ignore the growing body of evidence that, on the whole, they are'.[30] In the reply to the reply, the editorialist invoked the in principle/in practice distinction: microarrays are reproducible in principle, but in practice they are not, because the rules of reproduction simply are not followed on a consistent basis. In other words, to say that microarrays are reliable is true only in 'the best laboratories using technical replicates.... But this is not the real world where results are often not reproducible from one laboratory to the next'.[31]

Finally, there are problems associated with the reproduction of the results of a given experiment—the subject of interest here. An attempt to reproduce data analyses of 18 articles published in *Nature Genetics* succeeded in only 2 cases.[32] The most common obstacles to reproduction lay in the lack of published or deposited data and/or software. While these secretarial omissions may be overcome through stricter publications standards, it is also the case that different software often gives different results (even software that embodies the same algorithm).[33] In any event, as we will see below, problems in the reproduction of results must first take into account the kind of study undertaken that, according to biostatistician Richard Simon, fall into three categories: *class discovery*, *class comparison* and *class prediction*. In terms of clinical cancer research these different types of microarray experiment would consist of the identification of new subtypes of tumors from within a single data set, the comparison of different classes of tumors, and the identification of classes of tumors with different prognoses or different reactions to therapy. Each kind of experiment generates unique problems. Before turning to those problems, however, we need a bit more history.

3. The first clinical cancer microarrays

As already mentioned, microarray technology was developed at the turn of the 1990s and the first experiment using microarrays was published in 1995. The clinical potential of the new technology appeared the following year when researchers at Stanford used the technique to compare normal and pathological tissue. In collaboration with researchers at the Laboratory for Cancer Genetics at the National Center for Human Genome Research (NIH), Patrick Brown and his associates published the results of a comparison between 870 different genes from a human melanoma cell line and normal cells. They found that 15 of the 870 genes had 'significantly diminished' activity and 63 of the 870 were 'significantly increased' in their actions.[34] Brown's work constituted proof in principle and set out the initial standards and techniques for the disposition of data in the field in part because his group gave fellow researchers access to their data, which were posted on the group's website, as well as access to biometric tools for their analysis and manipulation. In addition, their hierarchical clustering algorithm generated the well-known red/green 'heat maps' that have since become ubiquitous in the literature (Fig. 2). The Stanford computational expert Michael Eisen had originally trained in biophysics. Less concerned with statistics than 'data', Eisen commented in a 1999 interview: 'The Stanford "mantra" is quite simple: More data is good'.[35]

The molecular profiling of cancer received further stimulus two years later when the new director of the NCI, Richard Klausner, issued a 'Director's Challenge' that incited researchers to undertake the 'Molecular Classification of Tumors'. The first grants were awarded towards the end of 1999, the largest of which went to Patrick Brown's group at Stanford. In all, the program awarded 10 grants totaling $4 million for the development of tumor profiles. Program organizers explained at the time that the sheer amount of data that the program would produce was a cause for concern and that as a partial solution to the problem, 'investigators will work collaboratively, with the assistance of NCI staff, to identify ways to represent research data, so that other cancer researchers

[21] Knapen, Vergauwen, Laukens, & Blust (2009).

[22] Churchill (2002, p. 490).

[23] Draghici, Khatri, Eklund, & Szallasi (2006).

[24] See Lynch (2002) on the difficulties associated with the replication and adaptation of experiments to local laboratory conditions.

[25] Kothapalli, Yoder, Mane, & Loughran (2002, p. 22).

[26] Irizarry et al. (2005), Larkin, Frank, Gavras, Sultana, & Quackenbush (2005), Bammler et al. (2005).

[27] Irizarry et al. (2005, p. 345).

[28] Owens (2005).

[29] Shields (2006a).

[30] Quackenbush & Irizarry (2006, p. 472).

[31] Shields (2006b).

[32] Ioannidis et al. (2009).

[33] Interview with Robert Nadon, Montreal, 10 May 2010.

[34] DeRisi et al. (1996, p. 457).

[35] Marshall (1999, p. 447).

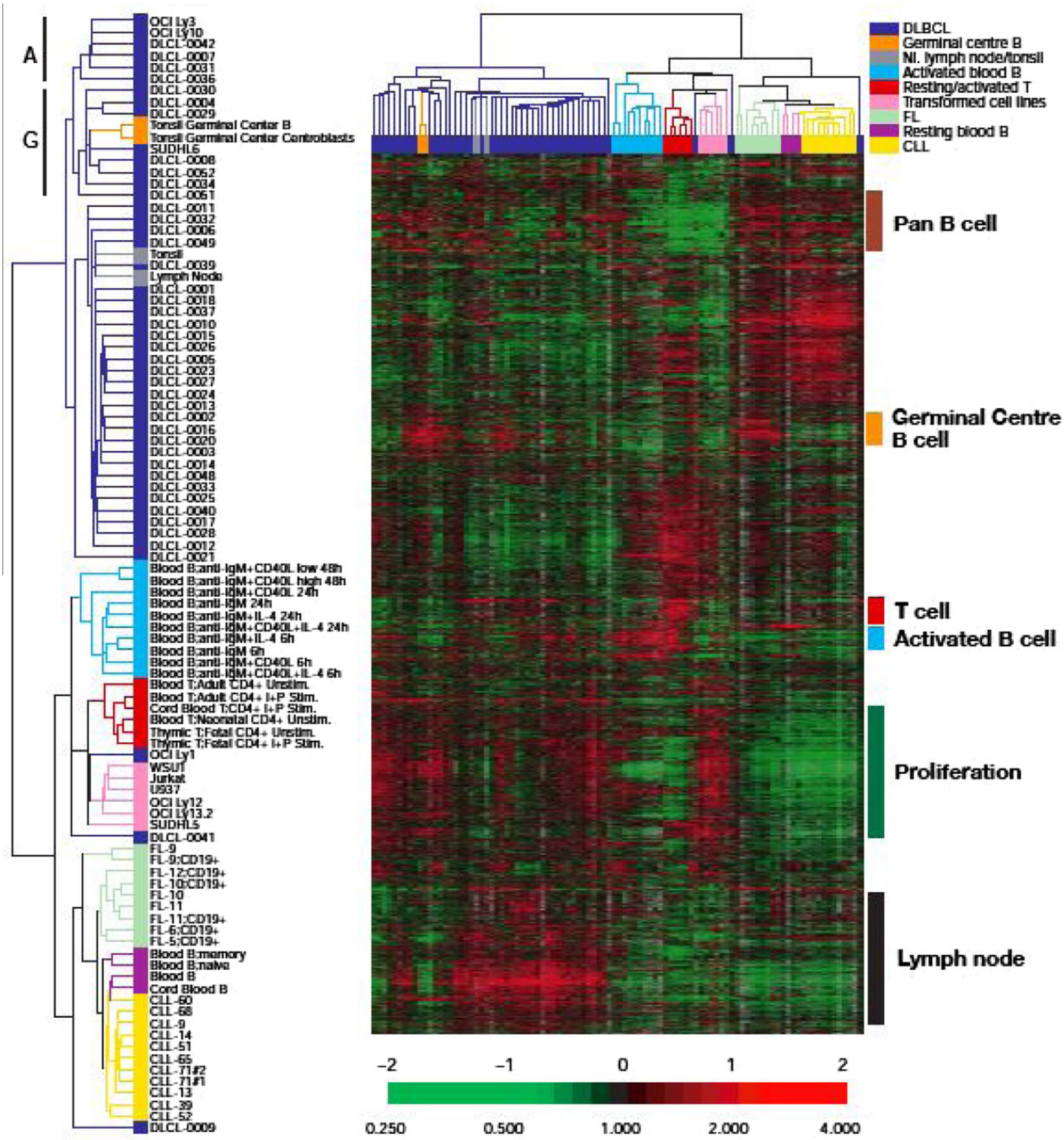

Fig. 2. Example of a microarray 'heat map'. The original legend reads in part: 'Hierarchical clustering of gene expression data. Depicted are the ~1.8 million measurements of gene expression from 128 microarray analyses of 96 samples of normal and malignant lymphocytes.... The results presented represent the ratio of hybridization of fluorescent cDNA probes prepared from each experimental mRNA samples to a reference mRNA sample. These ratios are a measure of relative gene expression in each experimental sample and were depicted according to the colour scale shown at the bottom.' For the original color version of this illustration see the online version of the present article. Reprinted by permission from Macmillan Publishers Ltd: NATURE, Alizadeh et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, 403, p. 505, copyright © 2000.

can interpret and analyze it. They will also develop strategies for publically releasing research data'.[36]

Klausner himself went on to become a member of the Lymphoma/Leukemia Molecular Profiling Project. The latter consisted of an international consortium of investigators led by Dr. Louis Staudt in the Molecular Biology of Lymphoid Malignancies Section of the NCI Center for Cancer Research. Including Brown's Stanford group, the consortium sought to pool resources in the molecular subtyping of blood cancers beginning with the lymphomas. The group had also developed a public–private partnership with Roche diagnostics to develop a custom microarray chip for lymphoma.[37] As commentators pointed out at the time, it was 'not by chance that the first applications to the diagnosis of cancer were made on leukemia and lymphomas', for samples are easily obtainable through

[36] NCI awards first grants ... (1999).

[37] http://dctd.cancer.gov/ProgramPages/brb/partnerships_international_leukemia.htm

venipuncture, no need to get the surgeons on board to do a biopsy.[38] Malignant cells can subsequently be easily purified using cells surface makers and the resultant extracts are relatively homogeneous insofar as they tend to be composed of single clones. This of course contrasts sharply with the vast majority of cancers that are solid tumors, i.e. heterogeneous masses that require a surgical biopsy for inspection.

The leukemia/lymphoma project also included the new head of the 'Molecular Statistics Section' and head of the Biometrics Research Branch of the NCI, Richard Simon. Originally called the 'Molecular Statistics and Bioinformatics Section', the latter had been created in 1999. Larry Hunter, a bioinformatician, had been the first head of the section. Recruited from the National Library of Medicine where he had developed a machine learning system, Hunter viewed his work at the NCI as a continuation of that line of research, telling the newsletter *BioInform* shortly after his appointment that he hoped to develop 'novel clustering tools for visualizing and analyzing gene expression data, and the use of text mining techniques in Medline to automatically develop organized knowledgebases [sic] for particular literature searches'.[39] None of this happened. By the end of 1999 Hunter had left the NCI and was replaced by Simon, a biostatistician, and 'bioinformatics' disappeared from the section title.

This is not to say that statistics superseded bioinformatics. Rather, Simon had already taken 'a bunch of courses' in genomics including a Harvard course in computational genomics.[40] In addition, he had been analyzing microarray data since the late 1990s, having collaborated with Jeffrey Trent at the National Human Genome Research Institute who, in 1996, was the only other biomedical researcher in the United States aside from Pat Brown with microarray technology.[41] The research program subsequently pursued by the molecular statistics section thus reflected a statistical approach to genomics. Simon saw this move into microarrays as part of a larger strategy for the biostatisticians at the NCI. Rather than restricting themselves to the clinical trials end of the research spectrum, the time had come to move further back the chain. Reflecting on this move in a 2003 interview, he observed: 'Now I feel that what statisticians need to do is move more up to the basic research level, particularly now with these technologies like microarrays that generate so much data. There's no hope that the biologists are going to make good use of that unless there's really mathematically trained people'.[42]

In the years that followed, Simon and his collaborators developed a number of statistical techniques and norms of application for those techniques that were sometimes published as editorials and quasi-guidelines. Simon's group also conducted extensive teaching within the NCI where Simon himself led a 'semi-monthly DNA microarray data analysis workshop for intramural investigators'.[43] The impact of their statistical take was further augmented by teaching beyond the walls of the NCI, the publication in 2003 of a textbook (that coincided with the publication of a beginner's guide to microarray analysis by three bioinformaticians that same year),[44] and the production of highly regarded computer software, BRB ArrayTools, for statistical analysis and data manipulation that is constantly updated.

Statisticians like Simon saw themselves as offering a different kind of expertise that would function as a corrective to the bioinformaticians' mistaken belief that the very nature of microarrays precluded the use of traditional statistical procedures such as the design of experiments, the framing of hypotheses and the calculation of significance levels or sample sizes. Indeed, as the biostatisticians Kerr and Churchill noted: 'With the rush to embrace microarray technology and its potential...a number of fundamental experimental principles have been neglected'.[45] In the course of enlightening practitioners at the beginning of the last decade, Simon outlined a number of bioinformatic 'myths' of microarray analysis then in circulation.[46] The first claimed that 'the only challenge is managing the mass of data'.[47] Second, was the idea that microarray data analysis consisted of looking for interesting patterns. Third myth: since microarray analysis was little more than a search for patterns, then cluster analysis was the most appropriate means. Finally, there was no problem that could not be solved by a prepackaged analysis tool. These myths, according to Simon, rested largely on the fact that bioinformaticians had lost sight of the biological questions behind microarray analysis. And yet most of the studies in which microarrays were used were ultimately motivated by some biological hypothesis, and that fact, in Simon's view, required collaboration with statisticians.[48]

Statisticians themselves were not exempt from Simon's criticism. Ignoring biology led many to develop methods that failed to use samples from multiple specimens to cancel out biological variability. More generally, according to Simon, some statisticians like their bioinformatician colleagues believed that microarray technology was merely a tool of description that did not test hypotheses concerning the mechanistic functions of specific genes and that as a consequence hypothesis generation and testing were not necessary. They also believed that the vast amount of data generated would by itself suffice to create 'important and unanticipated patterns in the data'.[49] Simon's formulation reflects the contrast between 'hypothesis testing' and 'fishing expeditions' that we discussed in the introduction to this article.

4. Managing the data deluge

Research groups using microarrays, like the Stanford and Staudt groups, often post their data and the software for manipulating the data on the web. By the turn of the century, the growth of microarray data had become a going concern as bioinformaticians predicted that: 'information being produced in this way is set to explode'.[50] The solution, they suggested, lay in the creation of public databases and standardized annotation that would allow researchers to retrieve and manipulate the gene expression data produced by others. The largest and most important data bank, known as the Gene Expression Omnibus (GEO) database, is run by the U.S. National Center for Biotechnology Information. Launched in 2000, as of 2006 it held over 50,000 individual submissions from over 1,000 laborato-

[38] Master & Lakhani (2000, p. 921).

[39] Burke (1999).

[40] Interview with Richard Simon, Rockeville (MD), 10 October 2003.

[41] Ibid.; Marshall (1999, p. 445).

[42] Interview with Richard Simon, Rockeville (MD), 10 October 2003.

[43] http://dctd.cancer.gov/ProgramPages/brb/partnerships_ccr.htm

[44] Simon et al. (2003), Causton, Quackenbush, & Brazma (2003).

[45] Kerr & Churchill (2001, p. 123).

[46] The 'myths' were presented in a talk is summarized in: NCI's Richard Simon...(2002).

[47] Ibid.

[48] Ibid.

[49] Simon, Radmacher, & Dobbin (2002, p. 21–22).

[50] Brazma, Robinson, Cameron, & Ashburner (2000, p. 699). For a similar argument, 10 years later, in connection to 'next-generation sequencing' see Fig. 2 in Stein (2010).

ries. A recognized public repository, many journals and grant committees often require a GEO accession number (indicating data deposit) before they will consider submissions.[51] The second major database is curated by the Microarray Informatics Group at the European Bioinformatics Institute and as of 2010 contained over 15,000 experiments.[52]

The data deposited in these banks pose a number of problems for would-be retrievers. First, the data generated by microarrays concern relative as opposed to absolute gene expression (see the caption of Fig. 2). In the common red/green microarrays, for instance, relative measurement refers to the ratio of gene expression (as measured by fluorescence) between experimental samples and controls: *up-regulation* of the former relative to the latter will be visualized as red and *down-regulation* as green. Intensity of color is proportional to the expression differential.[53] Measurements of relative values raise issues of quality control and assurance that, in general, require elaborate inter-laboratory arrangements spearheaded by regulatory initiatives (Keating & Cambrosio, 1998). Second, as we have seen, even with gene expression measurements made with the same technology and in the same laboratory, not all variables can be controlled and thus the data might not be comparable. In other words, it is possible to generate mountains of data (and numerous conclusions) that simply cannot be used by others. Moreover, the computer scripts used in analyzing the data are rarely deposited in the data banks. Often these scripts contain the algorithm used in the construction of the 'signature' (the specific configuration pattern of gene expression) that may be held secret and classified as proprietary.

When the first large data-banks opened at the beginning of the last decade, practitioners sought two kinds of standards in order to make comparisons between individual experiments: standards for technological platforms and standards for reporting practices. In theory, standard descriptions of experimental protocols should have sufficed to render reports with enough information to describe the experimental conditions under which the experiment took place. But since 'incidental and uncontrolled aspects of experimental conditions may also affect expression: for example the time of day, air humidity, or even the noise level of the laboratory', then it seemed 'realistic' to include the description of experiments 'largely in free text'.[54] In the meantime, in the absence of industry and laboratory standards, the only option seemed to be to get biologists 'to agree on as many sensible standards for microarray data and annotation as possible'.[55]

In the years that followed and as researchers awaited the 'data deluge' from microarrays,[56] the projected 'free text description' of experimental practices, which would have been a goldmine of information for future anthropologists, fell by the wayside and was superseded by the MIAME (Minimum Information About a Microarray Experiment) protocol.[57] Created by the Microarray Gene Expression Data Society (MGED), MIAME established the kinds of information that should be provided in the description of a microarray experiment. After lobbying the editors of scientific publications, MGED managed to get 'most scientific journals' to adopt MIAME requirements as a condition of publishing. Major databanks also decided to 'support' MIAME. In the latter case the support turned out to be somewhat problematic. MIAME only provided the information requirements; it did not supply the format or the software to encode the information. Although software was proposed shortly after the MIAME protocol, 'it did not gain popularity, largely due to its high complexity'.[58]

Standardizing the reporting of data did not solve all problems or limit the many forms of variability that pervade microarray data. Indeed, despite the multiplication of standards at all levels of the research process, a 2006 review concluded that it was still not 'possible to compare meaningfully results measured in different laboratories, at different times with different equipment'.[59]

5. Microarrays and statistics

Prior to 2000, researchers analyzing microarray data had been principally concerned with data display and clustering techniques.[60] The implication of statisticians in microarray experiments since the turn of the new century brought issues of statistical design and inference and hypothesis testing to the forefront. These techniques are often contrasted with application of algorithms for data exploration and display practiced by bioinformaticians. Indeed, notwithstanding Simon's remarks (Section 3), statisticians have not always been treated as valued members of microarray research teams and the contrast between biostatisticians and bioinformaticians has not always worked in their favor. According to a statistician, some biomedical researchers initially reasoned that:

> If the bioinformatician can take my data, run it through more sophisticated programs, produce longer gene lists and produce better looking graphics, what is the statistician contributing? The statistician will probably tell me it will take six weeks to complete the analysis and might start asking questions about whether I carefully designed my study with appropriate controls and safeguards against confounding effects. Invariably the statistician will tell me that the sample size was insufficient and insist that I pre-specify the questions I am trying to address with my study. It seems like statisticians just throw roadblocks in the way.[61]

Another respondent put it this way:

> Biologists say 'We want to do this', and the bioinformatician says 'I know how to do it and I can do it for you', whereas someone with biostatistical training will often say 'That's not what you want to do'. And I think that's the cultural clash.[62]

Part of the problem in the field, according to biostatisticians wary of published microarray studies, was that researchers often used the analytical techniques most familiar to them rather than choosing methods that were best suited to directly answer the study question. A good example is the overuse of clustering techniques for problems that required other statistical methods:

[51] Barrett & Edgar (2006).

[52] http://www.ebi.ac.uk/microarray-as/ae/

[53] We thank Lisa McShane for this formulation of 'relative gene expression'.

[54] Brazma et al. (2000, p. 700).

[55] Ibid.

[56] Hess, Zhang, Baggerly, Stivers, & Coombes (2001).

[57] Brazma et al. (2001). For a first analysis of MIAME and MGED, see Rogers & Cambrosio (2007).

[58] Brazma (2009, p. 421).

[59] Salit (2006, p. 75).

[60] Clustering techniques are a set of statistical methods used in many fields to group data points into strongly inter-related subgroups (called clusters).

[61] Interview with Lisa McShane, Rockville (MD), 16 February 2010.

[62] Interview with Robert Nadon, Montreal, 10 May 2010

> You could pick up almost any paper on gene expression microarrays, and you were likely to find cluster analyses, usually displayed in a heat map, even if the goal really wasn't to discover classes. If you have a predefined question such as how does the gene expression profile of a tumor change after it has been exposed to a drug, then the obvious thing to do is to evaluate which genes have changed and in what direction did they change. It doesn't really make much sense to cluster specimens in that situation because it was a designed experiment, but you will see people do it anyway.[63]

To understand the foundations of this criticism, let us return to the three types of microarray experiment invoked above: class discovery, class comparison and class prediction.

5.1. Class discovery

Class discovery experiments are a domain of choice for bioinformatics-based approaches. They commonly use the aforementioned 'cluster analysis' in a form that is termed 'unsupervised' because the data analysis involves only biological information that is collected in the course of the experiment and, in contrast to class comparison experiments (see below), does not use additional phenotype information linked to the samples such as that provided by pathology or clinical variables. One of the earliest and most spectacular (clinical) examples of this approach was the discovery of two subtypes of lymphomas by the Brown and Staudt collaboration in 2000.[64] In this experiment, the researchers analyzed cells of patients with B-cell lymphoma with a microarray known as the *Lymphochip*. The clustering algorithm sorted the patients into two groups based on the pattern of gene expression. The research team did not hypothesize what kind of subtypes or how many groups would emerge from this experiment. In a sense, the 'unsupervised' nature of the algorithm 'let the data do the talking'.

Many clustering techniques (over 40) have since been deployed in the field and considerable discussion exists as to which ones are the best.[65] Statisticians point out that in cluster analysis, 'the validity of specific solutions, algorithms, and procedures present significant challenges because there is no null hypothesis to test and no "right answer"'.[66] Validity lies therefore in reproducibility: do multiple samplings within the same population give the same profile? Statisticians from the University of Alabama evaluated four commonly used clustering techniques in 2005 on 37 different microarray datasets and concluded that sample sizes of less than 50 had poor reproducibility.[67] Given that most biological class discovery experiments proceed with far less samples, the question then becomes which clustering technique produces the most stable cluster. Although that may in many cases depend on the task at hand, the most popular technique remains the first and most visible: the hierarchical clustering technique.[68]

While assessment of cluster stability and validity are frequent concerns of bioinformaticians, statisticians have not been entirely absent. They have, in fact, devised a number of ways to assess the uncertainties associated with cluster formation, in particular cluster stability. These include introducing random noise into the data set, re-clustering the data and measuring the differences between the original clusters and the new clusters.[69] But, once again, bioinformatics-based approaches are paramount for class discovery experiments.

5.2. Class comparison

Compared to the previous category, this class of experiments has led to the controversial intersection of biostatistical with bioinformatics-based approaches. Class comparisons are supervised studies in the sense that outside information (e.g. 'normal' vs. 'diseased' tissue) is imposed upon the data. In these cases, statisticians argue, unsupervised clustering algorithms are not appropriate; rather, 'the state of the art…is comparing classes on a gene-by-gene basis using statistical tests and controlling at the margin'.[70] In other words, unlike discovery experiments, comparison experiments require calculations of statistical significance. This was not always done in the early days. Researchers generally contented themselves with reporting 'differential gene expression' between two samples rather than statistically significant differences in gene expression between samples.

As an example, let us return briefly to Brown's early melanoma study described above. As a sort of fishing experiment, the experiment did not rely on pre-set levels of significance or other standard statistical tests to determine whether or not the genetic differences between the two cell types (normal and melanoma) were significant. Rather, the selection of genes that differentiated the two types of tissue relied on several internal controls. For example, the team had included 'housekeeping' genes as controls, i.e. genes that should in principle have been equally expressed in both types of tissue. That turned out to be roughly the case and so they then decided to retain for further examination any genes whose fluorescence intensity differed by more than 3 standard deviations from the norm established by the housekeeping genes. They thus spoke of genes with 'significantly diminished' activity and genes whose activity was 'significantly increased'.[71] But this was not 'statistical' significance in the common use of the term as no test of significance had been conducted nor had a *p*-value been calculated. As there were thousands of genes, this would have required thousands of significance tests.

That in itself was not a problem since, these days, nobody calculates much of anything; modern software does the calculating. As we will see below, however, statistical significance for microarrays requires more than simply repeating standard tests thousands of times. Before turning to that question, let us briefly consider the software that does the work. In the field of microarrays, when researchers turn to problems of identifying genes that are significantly differentially expressed between classes of samples, two programs dominate the scene. The first is the Stanford group's SAM method (Significance Analysis of Microarrays). First described in 2001, the program has undergone a number of modifications in response to both critics and developments in the field.[72] The second is the Limma package that is part of a larger package known as Bioconductor. The latter uses the computer language 'R'. Described as a sort of Excel on steroids, R was derived from 'S', the com-

[63] Interview with Lisa McShane, Rockville (MD), 16 February 2010.

[64] Alizadeh et al. (2000).

[65] Thalamuthu, Mukhopadhyay, Zheng, & Tseng (2006), Andreopoulos, An, Wang, & Schroeder (2009).

[66] Garge, Page, Spargue, Gorman, & Allison (2005).

[67] Ibid.

[68] Until 2008, there was, moreover, no way to compare between clustering methods; see Liu, Lee, Casella, & Peter (2008).

[69] McShane et al. (2002).

[70] NCI's Richard Simon…(2002).

[71] DeRisi et al. (1996, p. 457).

[72] Goss Tusher, Tibshirani, & Chu (2001).

puter language and environment most commonly used by statisticians, developed at the Bell Laboratories.[73] Bioconductor has been available since 2001 as open-source software and has become 'a *de facto* standard for microarray data analysis'.[74] According to its curators, Bioconductor took off in 2004 following the publication of a description of the package in *Genome Biology*.[75] According to one of the creators of Bioconductor, a good deal of its success depends upon the fact that it uses R. Estimates of the number of R users in 2009 ranged from 250,000 to 2 million.[76]

Both SAM and Limma carry out a number of calculations. For the purposes of the present paper, two in particular stand out. First, they both allow researchers to test for differential gene expression using a statistical rather than a visual (i.e. clustering) approach. Both methods employ *shrinkage*, to address a problem common to most microarray experiments: while there are many genes, for each gene there are relatively few mRNA samples assayed. This means that even if you did thousands of *t*-tests, they would all be done on very small samples. Shrinkage corrects this deficiency by using information from the microarray as a whole and calculating the variance for all the genes, and by subsequently using this group variance to recalculate the test statistic. This adjustment decreases the false-negative rate or, in other words, increases the power of the test. While SAM and Limma contain the most popular shrinkage routines, there are many others and all such tests seem to work equally as well.[77] The important point in the present context, however, is that according to statisticians when it comes to class comparison statistical methods like shrinkage that allow researchers to produce *p*-values stand in sharp contrast to clustering methods that don't: 'The most popular problem in microarray data analysis is...: you have two groups and you want to find which genes are differentially expressed. If that is the question, clustering is absolutely not the tool for that. But the early papers would do clustering and try to see if the two samples would separate'.[78]

Packages like Limma and SAM also allow researchers to calculate *p*-values. But, here again, statisticians have intervened to adapt common statistical techniques to the exigencies of the data-dense microarray. Since microarrays typically compare tens of thousands of genes in the search for the dozens or even hundreds that are differentially expressed, from a statistical point of view one can expect a number of false positives, meaning that a microarray experiment that identifies, say, 500 genes as differentially expressed may be meaningless having, by chance, produced 500 'false discoveries'. To correct for these false discoveries statisticians have developed methods for calculating alternative forms of error rates such as *family-wise error* or *false discovery rate* (FDR). Briefly, the family-wise error rate is the probability that the identified list of significantly differentially expressed genes contains at least one false discovery. The aforementioned Limma offers a number of ways to calculate the family-wise error rate, but they come with their own downsides: a procedure known as the Bonferroni correction, for instance, 'drastically increases the false-negative rate'.[79] Now the whole point of a microarray discovery experiment is to generate a list of genes that is (initially at least) as inclusive as possible. So, statisticians reached back into their toolbox and withdrew the aforementioned concept known as FDR that is something like an expected proportion of false discoveries for the identified list of significantly differentially expressed genes. One of the advantages of the FDR over traditional tools such as p-values is that it offers an immediate understanding of the nature of the data, for in an era of data-intensive research 'it is not unusual to hear an investigator declare with desperation that he/she has several thousand "significant" genes, which make further steps in the experiment anything but clear'.[80]

Once again, as with cluster analysis, we have not even scratched the surface. The number of methods and permutations of those methods for determining whether or not a significant difference has been detected is constantly growing. Fun-sounding software such as BUM or SPLOSH, for example, can be used instead of SAM.[81] Our point here, however, is not to develop an exhaustive list of or even an overview of statistical methods, but, as noted above, to track the increasing implication of statisticians as microarray data moves towards the clinic. To use a slightly different formulation, the issue is no longer whether one should or should not perform statistical tests, but about the fine-tuning of statistical methods to confront the challenges raised by data-intensive domains. The statistical 'take-over' is even more evident in the case of class prediction.

5.3. Class prediction

This third category of experiments belongs to a domain in which biostatisticians have staked a strong claim, for attempts to create clinical class predictors (signatures) fall under the auspices of clinical trial methodology. The signatures or classifiers are constructed from a representative set of samples drawn from a known class (e.g., patients with stage I lung cancer). The idea is to select some clinical feature of the class, such as survival following surgery, and determine what configuration of genes is most closely associated with the presence and/or absence of that feature. Such studies, statisticians claim, are by analogy the same kind of study as prognostic studies in clinical medicine and they, in fact, suffer the same deficiencies as these studies. Indeed, statisticians have already characterized the prognostic study literature as 'unreliable' and 'inconsistent' due primarily to the fact that '[t]here is no written protocol, no established patient selection criteria, no clear and limited objectives, no critiqued analysis plan and no good practice standards'.[82] All these criticisms apply to microarray studies as do the standards of the field: 'For example, if the objective is to develop a predictive model for patients with well-staged stage I lung cancer treated with surgery alone, then the specimens included in the study should be from patients with well-staged stage I lung cancer who received surgery alone'.[83] While this may seem rather obvious, statisticians insist that the invasion of the biomarker field by genomic data has been in some cases a disaster:

> We see studies with terrible designs in which samples are selected haphazardly with no thought to the target population, or technical artifacts such as batch effects[84] confound major

[73] http://www.r-project.org/

[74] Salit (2006, pp. 74–75).

[75] Gentleman et al. (2004).

[76] Vance (2009a,b).

[77] Cui, Hwang, Qiu, Blades, & Churchill (2005).

[78] Interview with R. Irizarry, Baltimore, 31 March 2010.

[79] Nadon & Shoemaker (2002, p. 269).

[80] Pawitan et al. (2005, p. 3023).

[81] Pounds & Morris (2003), Pounds & Cheng (2004).

[82] Simon et al. (2002, p. 26).

[83] Ibid.

[84] The term 'batch effects' refers to how 'measurements are affected by laboratory conditions, reagent lots and personnel differences'; batch effects are 'widespread' and they can obviously lead to invalid conclusions; Leek et al. (2010, p. 733). See also Lewitter & Bell (2010).

> experimental factors of interest. Many researchers have no clue how to do sample size calculations. Sample size is often picked out of the air, reasoning that, 'As long we have lots of variables, we'll find something'. Traditional biomarker research, looking at only one or a few biomarkers at a time, has had its own long history of problems due to use of poor study designs, inappropriate use of statistical methods, over-analysis of data, over-fitting models, and selective reporting of the 'most significant' results. Researchers were already getting into plenty of statistical trouble handling just a few biomarkers, and now they get into deeper trouble even faster with high-dimensional genomic data.[85]

Once again, while there are many statistical methods available to develop such predictors, 'none of these methods was developed in the context of studies in which the number of candidate predictors is at least one order of magnitude larger than the number of cases'.[86] The earliest cancer studies thus developed their own biometric methods for developing predictors. Statisticians then compared these methods amongst themselves and with standard prediction methods, and found that most worked well enough.[87] Moreover, the simplest methods generally work better than the more complicated ones that deploy neural nets and other forms of artificial intelligence. The latter, in fact, appear to statisticians to be more trouble than they are worth: 'Artificial intelligence sells to journal reviewers and peers who cannot distinguish hype from substance when it comes to microarray data analysis'.[88] Once a signature or a class predictor has been constructed (i.e. once the algorithm that selects the genes for the signature has been finalized), the central problem of class prediction becomes the validation of the predictor or signature. In other words, does the classifier work on other samples and with what rate of success or error? There are many class predictor validation strategies. The most rigorous form of validation requires testing the performance of the predictor on an independent external data set. However, independent data sets are often not available and investigators rely instead on internal validation methods, such as data re-sampling or sample splitting that 're-use' data from the initial set of samples.

As with class discovery and class comparison, there are no standards for the creation of class predictors. Class predictors, however, are close to clinical application and so there has been an international effort to set standards and guidelines. Many class predictors are developed to predict prognosis, so guidelines applicable to conventional prognostic markers should be relevant. The REMARK recommendations[89] published in 2005/6, for example, set out the criteria for reporting tumor marker prognostic studies. According to one of the authors, the reporting elements are quite straightforward:

> The REMARK guidelines consist of 20 reporting elements which, quite frankly, should be obvious. The recommended reporting elements include information such as how patients were selected, inclusion/exclusion criteria, description of analysis methods, including method of marker cutpoint determination (if applicable), and so forth. With complete information about how a study was designed, conducted, and analyzed, one can better judge the usefulness of the data and understand the context in which the conclusions apply.[90]

The REMARK authors stress that the REMARK guidelines advise what to report, and thus make it easier to detect poorly designed, executed or analyzed studies. They do not dictate how a study should be designed or analyzed. This last comment refers us back to one of the requirements invariably stressed by biostatisticians, namely that they should be associated from the very outset to the design and planning of an experiment: 'analysis and design are inseparable…[to paraphrase one of the founders of modern statistics, R.A. Fisher] going to see a statistician after the data have been gathered is like doing a post-mortem: you can tell what the experiment died of'.[91] Indeed, according to the same respondent a defining characteristic of biostatisticians vis-à-vis bioinformaticians is that while the latter do a lot of analysis with data they get off databases, the former 'consider it absolutely fundamental to know how the data were generated and to be involved in those findings'.[92] Less a gratuitous attempt to become the arbiters of scientific method by investing epistemic authority in statistics rather than in experimental technique—an epistemic 'coup d'état', so to speak—the biostatisticians' attitude reflects the intersection of two processes: first, the evolution of their status within the clinical trial process, whereby statisticians have become full-fledged collaborators with, rather than consultants to, clinical researchers (Keating & Cambrosio, 2012); and, second, the advent of high-throughput technologies such as microarrays that mark a shift from a situation where experiments generated relatively few data in support of mechanistic hypotheses to one where an onslaught of numbers required probabilistic approaches.

5.4. Sample size

A difficult question for all three kinds of studies and long the bread and butter question for statisticians is how large a sample is needed for the production of valid findings. As far as class comparison and class prediction are concerned, '[t]here is generally no accepted theory for planning sample size for developing multivariate predictors with numerous candidate variables. Rules of thumb that are sometimes used, such as having 5–10 cases for each candidate variable, would suggest that tens of thousands of cases are needed for microarray studies. This is clearly not practical'.[93] In the absence of theory, it is possible to offer ballpark figures for sample size based on a simulation that takes into account a number of assumptions such as the expected degree of variability between samples. The University of Alabama at Birmingham Biostatistics Department has made things even easier for investigators wondering how large a sample is necessary and how to control for power. Researchers can simply consult the web-based resource *PowerAtlas*,[94] either by entering the microarray type and species under study to find similar studies whose sample size and power measurements they can adopt, or, alternatively, by using the software to estimate sample size and power using their own preliminary data.[95]

The sample size problem differs obviously according to the type of study conducted. For class discovery, the issue is further complicated by the fact that different cluster algorithms can be used.

[85] Interview with Lisa McShane, Rockville (MD), 16 February 2010.
[86] Simon et al. (2003, p. 96).
[87] Dudoit, Fridlyand, & Speed (2002).
[88] Simon (undated), slide 19.
[89] McShane et al. (2005).
[90] Interview with Lisa McShane, Rockville (MD), 16 February 2010.
[91] Interview with Robert Nadon, Montreal, 10 May 2010.
[92] Ibid.
[93] Simon et al. (2002, p. 32).
[94] Page et al. (2006).
[95] http://www.poweratlas.org/

Their effect on sample size remains nebulous according to biostatisticians:

> There is no all-purpose method for determination of sample size for class discovery investigations. Adequate sample size will depend on factors such as the degree of separation between the true clusters and the likely number and size of clusters. If you feel pretty certain that there will be only two or three clusters and that they are probably well separated, then a sample size of two or three-dozen samples may be sufficient. But if you expect large numbers of relatively small clusters with subtle differences between them, then much larger sample size will be required to detect all of the different clusters and tease them apart. This may all sound like sensible advice until you ask the question of what exactly do I mean by a 'cluster.' If I apply K-means clustering method, I'll get clusters defined in a certain way; if I apply hierarchical clustering methods, I'll get clusters defined in another way. So thinking about the fact that the notion of cluster is somewhat vague, how can you talk about sample size for detecting clusters if the things aren't well defined in the first place?[96]

Thus, class discovery using cluster analysis can be compared to a liminal domain where bioinformatic and statistical approaches intersect in open-ended ways, without the latter necessarily dictating its stringent rules and requirements to the former, as it has become the case in class comparison and prediction.

6. Controversy

> Sometimes the glamour of the technology or the sheer volume of omics data seem to make investigators forget basic scientific principles.[97]

Between 1999 and 2005, the number of microarray studies within the field of oncology grew substantially. This growth was accompanied by a rising concern of the extent to which the results of these individual studies could be reproduced by other researchers, the issue being less one of *technical* replication than of *biological* replication or, to put it otherwise, of *statistical* significance leading to *biological* significance.[98] The two-year period 2003–2005 saw the publication of a number of articles that pointed out the poor transition of microarrays from the lab to the clinic. A 2003 editorial in the *Lancet* opined: 'The initial hyperbole surrounding microarray technology has quietened as researchers grapple with the more mundane but critical issues of transition to clinical utility. Expect to see papers in the future in which novel clinical insights take precedence over technology'.[99] And yet, the problem remained. In 2005, a comment accompanying an article in the *Lancet* concluded that five of the seven largest microarray studies in cancer prognosis gave results 'that were no better than flipping a coin' whereas the other two 'barely beat horoscopes'.[100] According to the editorial: 'That scenario is scary: this noise is so data-rich that minimum, subtle and unconscious manipulation can generate spurious "significant" biological findings that withstand validations by the best scientists, in the best journals'.[101] In a domain where the issue of inherent random error that characterizes biological data is compounded by the 'high-dimensionality'[102] of the data generated by high-throughput technologies, the editorial highlighted two distinct problems: the possibility of producing spurious findings and the difficulty of showing them to be spurious.

In the case of findings involving the treatment of actual patients, the debate has moved out of the rarified atmosphere of highly technical debates between statisticians and bioinformaticians and, in one prominent case, the 'Duke genomics scandal', has entered the pages of the *New York Times*.[103] The scandal involved a series of single-institution clinical trials testing genomic signatures to predict clinical outcomes, and led to the suspension and, subsequently, resignation of one of the two principal investigators, the retraction of a high-profile article and the closure of three phase II clinical trials.[104] We will skip the somewhat arcane details of the scandal, and concentrate instead on the form of 'forensic bioinformatics' that sparked it off. Since 2005, Keith Baggerly from the MD Anderson Cancer Center in Houston has produced a number of articles and almost as many controversies in his efforts to reconstruct the methods that must have been employed in individual studies from the raw data that journals now require for deposit. His most notorious investigation concerned the results initially published in 2006 by Duke researchers. Analyzing the NCI's human-tumor cell lines used to screen drugs with microarrays, the Duke team claimed to have generated a signature for drug sensitivity that could be used to predict patient response to common chemotherapeutic substances.[105] The publication garnered *Discover* magazine's accolades as one of 'The top 6 genetics stories of 2006'.[106] But Baggerly's analysis of the raw data uncovered a number of clerical errors (mislabeling of cell lines, 'off-by-one' mislabeling of genes, reused test data, mixed up group labels, etc.) and more serious software problems: 'Their software does not maintain the independence of training and test sets, and the test data alter the model. Specifically, their software uses "metagenes": weighted combinations of individual genes'.[107]

While admitting to some of the clerical errors, the Duke researchers responded that the MD Anderson group had not followed their methods 'precisely' and that using the original signatures to predict patient response in additional datasets led to consistent results.[108] Thus undeterred, they moved forward into clinical trials. Equally undeterred, the MD Anderson forensic team, having already racked up 1500 hours of investigation, continued to probe the Duke data and assertions. In 2009, they reiterated their original claims, added new ones and charged that the clinical trials instituted by the Duke researchers constituted a danger to patients.[109] Then, the situation became more confused: Duke Univer-

[96] Interview with Lisa McShane, Rockville (MD), 16 February 2010.

[97] Lisa McShane quoted in Goldberg (2011, p. 8).

[98] Lewitter & Bell (2010).

[99] Winegarden (2003, p. 1428).

[100] Ioannidis (2005, p. 454).

[101] Ibid.

[102] The term 'dimension' refers to the number of coordinates needed to describe a point in a mathematical object (a line is one-dimensional, a surface two-dimensional, a volume three-dimensional). High-dimensionality refers to abstract spaces with a dimensionality substantially greater than three. On 'high dimensionality' in biology see Mehta, Tanik, & Allison (2004).

[103] Singer (2010).

[104] Goldberg (2010b).

[105] Potti et al. (2006); this paper was followed by a second one, Hsu et al. (2007), now retracted.

[106] Barone (2007).

[107] Coombes, Wang, & Baggerly (2007, p. 1276).

[108] Potti & Nevins (2007, p. 1278).

[109] Baggerly & Coombes (2009).

sity authorities suspended the clinical trials and created an anonymous committee of cancer center directors to investigate the charges.[110] The committee's confidential report led to the decision to reactivate the trials, but the MD Anderson team maintained that substantial errors remained that would impact the performance of the predictors. In the midst of the investigation in November of 2009, the Duke team posted additional data in support of claims published in the 2007 *Journal of Clinical Oncology* article that had been used to support clinical trials. Forensic examination showed that 'at least 43' of the 59 test samples 'were mislabeled. We say "at least" because the genes were so thoroughly mislabeled for the remaining 16 samples that we could not identify the sample of origin'.[111] Baggerly's persistence in the end paid off and led to the aforementioned debacle. One of the key lessons of this episode, according to the forensic team, is that:

> One characteristic of high-dimensional predictive 'signatures' is that we have little intuition about what 'makes sense'. We have to trust that the underlying analyses are correct, or at least checkable. . . . Before a signature we can't intuitively grasp is introduced, we contend that the data and code used to generate the signature should be assembled with sufficient clarity for an independent group to easily run the code and confirm the predictions.[112]

Taking as a starting point the Duke debacle, a group of over 30 bioinformatics and biostatistics specialists (notice the coalition front), wrote to the NCI Director 'expressing concern about the reproducibility of the [Duke] data', and subsequently gathered under the banner of 'Scientists for Reproducible Research' to lobby journal editors to enforce transparency in articles publishing genomics research.[113] The Institute of Medicine, the health arm of the National Academy of Sciences, has in the meantime been asked to examine the Duke episode, a somewhat unusual task for this high-profile institution, and has broadened its mandate to 'identify appropriate evaluation criteria for tests based on "omics" technologies . . . that are used as predictors of clinical outcomes' . . . [and to] identify criteria for the analytical validation, qualification, and utilization components of test evaluation'.[114]

Farther from the limelight, other controversies flare up. Building on a previous study[115] of all microarray publications on cancer prognosis that appeared between 1995 and 2003, biostatisticians at the Gustave-Roussy Institute, France's premier cancer research centre, selected a subset of the original sample for further study. Retaining publicly available data from the seven studies that measured survival-related outcomes on at least 60 patients, they undertook a re-analysis of the data. They randomly divided the original samples into a training set and a validation set, using the former to generate a signature composed of the genes most highly correlated with the prognosis in question and the latter to validate the signature. They repeated the procedure 500 times and were led to conclude that 'every training set of patients led to a different list of genes in the signature', and that 'five of the seven largest published studies did not classify patients better than chance'. [116] The smaller the number of patients, the more unstable the signature; hence the conclusion: 'Studies with larger samples are needed before gene expression profiling can be used in the clinic'.

The article generated considerable discussion in the letters section of *The Lancet*. Researchers at Italy's National Institute of Tumors pointed out that they, too, had reworked data and found a different signature, but they surmised that the stability problem was more than a question of sample size. Lack of common standards was also a powerful source of variability or in other words, 'the underlying issue in microarray studies is the lack of standard methods for design, data analysis, and performance assessment according to clinical aims'. This quintessentially socio-technical problem had a social solution: 'Achieving these goals requires cooperative efforts between multidisciplinary research networks'.[117] That was 2005. In the five years following its publication, the review became 'a point of reference for application of DNA microarrays in cancer prognosis'.[118] Work with microarrays has nonetheless continued unabated, and in 2007 the US Food and Drug Administration cleared the first microarray based commercial prognostic test for breast cancer.[119] In 2010, the issue was revisited by a consortium of Chinese and FDA researchers who charged that the original review of the microarray studies published in the *Lancet* suffered its own flaws, namely that the techniques used to analyze the data were only one of a number of possible methods, and that the limited sample sizes of the seven data sets studied were not large enough to generate the power needed to reach the conclusions set out in the 2005 publication. Using other methods and generating a larger data set through permutation, the authors concluded that 'on a conservative estimation', four of the seven data sets 'yielded classifiers performing better than chance'.[120] 2010 was also the publication year of a report by a large consortium (36 research teams) called MAQC-II that has set out to 'assess the capabilities and limitations of microarray data analysis methods . . . in identifying gene signatures representatives of a specific pathological condition'. The report argued that the different models tested were 'remarkably similar in predicting outcomes, irrespective of the approach used' but that the overall success of the classifier depended on the bio-pathological material examined: highly heterogeneous tumors such as breast cancer lead to worse predictions than liver toxicology studies. The quality and skills of data analysis teams also appeared to ensure more accurate predictions, a finding that led to the conclusion that 'MAQC-II was as much an exercise in sociology as in technology'.[121]

7. Conclusion

A number of observers have underscored the fact that despite what is now more than 15 years of use, microarrays have yet to penetrate clinical medicine to any significant degree. With regards to class prediction studies, 'the prognostic value of the gene signatures identified seems to have been oversold, maybe because of the enormous investments and because of the high expectations in a new technology'.[122] Early reviews of the domain complained that studies suffered significant deficiencies, such a lack of independent

[110] Goldberg (2009).
[111] Goldberg (2010a, p. 4).
[112] Baggerly & Coombes (2010).
[113] Tuma (2010).
[114] Goldberg (2010b).
[115] Ntzani & Ioannidis (2003).
[116] Michiels, Koscielny, & Hill (2005, p. 489 and 491).
[117] Biganzoli, Lama, Ambrogi, Antolini, & Boracchi (2005, p. 1683).
[118] Fan et al. (2010, p. 629).
[119] Kohli-Laven, Bourret, Keating, & Cambrosio (2011).
[120] Fan et al. (2010, p. 630).
[121] MAQC-II: analyze that! (2010).
[122] Michiels, Koscielny, & Hill (2007).

validation or even of any attempt to validate,[123] a criticism that resonated with the biostatisticians' crusade to bring (statistical) order to a domain where the deluge of data opened up avenues for the unbridled exertion of bioinformatic skills. But criticism continues. In a 2010 'Perspective' article, Serge Koscielny summarized the previous fifteen years of microarray research from a clinical point of view, concluding that '[g]ene microarrays have brought little progress to the clinical management of cancer since Schena et al.'s 1995 publication [Patrick Brown's team first microarray publication in *Science*]' and that '[t]he field urgently needs a breakthrough in the way we analyze [prognostic microarray] data, or we will end up with a collection of data sufficient to explain everything but unable to predict anything'.[124]

Koscielny's criticisms lie at one end of the spectrum of attitudes that have greeted microarrays since their inception and thus run from 'extreme skepticism' (biostatistician Richard Simon's term) to unrestrained enthusiasm. In the field of cancer, these forms of reception have been particularly acute. Enthusiasts initially accused cancer pathologists of living in the past, comparing them to 'primitive tribes' and arguing that they were 'to some extent still guided by fairly unsophisticated determinants, such as what organ the tumor arose from, how big it is, and what it looks like after it has been imbedded in wax and colored with different dyes'.[125] They predicted that microarrays would destroy everything in their path. An editorial accompanying a retrospective of microarray use in breast cancer diagnosis suggests that the situation is somewhat more complicated: 'molecular subtyping should not be used instead of morphology and immunohistochemistry but rather in addition to these classic approaches, to increase clinical relevance and robustness'.[126] Between the enthusiasts and the skeptics lie the pathologists themselves whose evaluation of the contributions of microarray technology to their practice combines elements of both points of view, arguing that while microarrays 'have led to a paradigm shift in the way breast cancer is perceived and how breast cancer research is carried out', 'their clinical applications are still limited'.[127] The impediments to wider acceptance cited by the pathologists are similar to those instanced above: poor reproducibility, poor overlap between different signatures and frequent lack of clinical validation.

This last remark leads us back to the twin themes of our paper, namely the shifting analytical infrastructure of the microarray domain—from bioinformatics-centered to biostatistics-centered—and its characterization as data-driven or hypothesis-driven. The manifold strictures entailed by clinical work have led to a situation where rather than a choice between the aforementioned alternatives we witness their hybridization. The handling and processing of the massive data generated by microarrays has made bioinformatics a must, but has not exempted the domain from becoming answerable to statistical requirements. The centrality of statistical analysis emerged diachronically, as the field moved into the clinical domain, and is re-specified synchronically depending on the kind of experiments one carries out. But statistical analysis must also engage the change of scale introduced by the large amount of data produced by microarrays. As shown by the controversial episodes discussed in the previous section, bioinformaticians and biostatisticians have in fact joined forces to produce a 'reproducible science', a task that as mentioned in relation to MAQC-II involves as much 'sociology' as technology. Reproducibility appears in this case to be less a problem of skills and tacit knowledge, as in early sociological accounts of this issue,[128] than a question of the proper assemblage of statistical tests and data-handling algorithms. But the story cannot be reduced to these human and technological components. As the vagaries of tumor biology and pathology re-emerge, the early enthusiasm about genomic technologies as a master key to unlock the underlying molecular mechanisms—the tumors' 'true nature'—has given way to a more guarded attitude: it is not uncommon for investigators to lament that 'I still haven't seen as many results as I had expected to a few years back'. No wonder, then, that many researchers continue to assert the need to articulate hypotheses about the biological significance of experimental results with the data-intensive findings produced by post-genomic technologies.

Acknowledgments

Research for this paper was made possible by grants from the Canadian Institutes of Health Research (MOP-93553), the Fonds québécois de la recherche sur la société et la culture (SE-124896) and the Social Sciences and Humanities Research Council of Canada (410-2008-1833). We would like to thank all the biostatisticians and bioinformaticians who kindly accepted to be interviewed and to comment on early drafts of the text.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature, 403*, 503–511.

Andreopoulos, B., An, A., Wang, X., & Schroeder, M. (2009). A roadmap of clustering algorithms: Finding a match for a biomedical application. *Briefings in Bioinformatics, 10*, 297–314.

Baggerly, K. A., & Coombes, K. R. (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics, 3*, 1309–1334.

Baggerly, K., & Coombes, K. (2010). Retraction based on data given to Duke last November, but apparently disregarded. *The Cancer Letter, 36*(39), 1. and 4–6.

Balmer, B. (1996). Managing mapping in the Human Genome Project. *Social Studies of Science, 26*, 531–573.

Bammler, T., Beyer, R. P., Bhattacharya, S., Boorman, G. A., Boyles, A., Bradford, B. U., et al.Members of the Toxicogenomics Research Consortium. (2005). Standardizing global gene expression analysis between laboratories and across platforms. *Nature Methods, 2*, 351–356.

Barone, J. (2007). New tests help chemotherapy hit the mark. *Discover*, January; at: http://discovermagazine.com/2007/jan/genetics/.

Barrett, T., & Edgar, R. (2006). Mining microarray data at NCBI's Gene Expression Omnibus (GEO). *Methods in Molecular Biology, 338*, 175–190.

Biganzoli, E., Lama, N., Ambrogi, F., Antolini, L., & Boracchi, P. (2005). Prediction of cancer outcome with microarrays. *Lancet, 365*, 1683.

Brazma, A. (2009). Minimum Information About a Microarray Experiment (MIAME). Success, failure, challenges. *The Scientific World Journal, 9*, 420–423.

Brazma, A., Robinson, A., Cameron, G., & Ashburner, M. (2000). One-stop shop for microarray data. *Nature, 403*, 600–700.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., et al. (2001). Minimum information about a microarray experiment. Towards standards for microarray data. *Nature Genetics, 29*, 365–371.

Burke, A. (1999). NCI creates new molecular statistics and bioinformatics section, appoints Hunter. *Bioinform*.

Causton, H. C., Quackenbush, J., & Brazma, A. (2003). *Microarray gene expression data analysis: A beginner's guide*. Maldon, MA: Blackwell.

Churchill, G. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics, 34*(Suppl. 4), 490–495.

Collins, H. M. (1992 [1985]). *Changing order: Replication and induction in scientific practice*. Chicago: University of Chicago Press.

Coombes, K. R., Wang, J., & Baggerly, K. A. (2007). Microarrays: Retracing steps. *Nature Medicine, 13*, 1276–1277.

Correa Geyer, F., & Reis-Filho, J. S. (2009). Microarray-based gene expression profiling as a clinical tool for breast cancer management: Are we there yet? *International Journal of Surgical Pathology, 17*, 285–302.

[123] Ntzani & Ioannidis (2003, p. 1441).

[124] Koscielny (2010, p. 3).

[125] He & Friend (2001, p. 658).

[126] de Ronde, Wessels, & Wesseling (2010, p. 307).

[127] Correa Geyer & Reis-Filho (2009, p. 302).

[128] Collins (1992 [1985]).

Cui, X., Hwang, J. T., Qiu, J., Blades, N. J., & Churchill, G. A. (2005). Improved statistical test for differential gene expression by shrinking variance component estimates. *Biostatistics, 6*, 59–75.

de Ronde, J., Wessels, L., & Wesseling, J. (2010). Molecular subtyping of breast cancer: Ready to use? *Lancet, 11*, 306–307.

DeMets, D. L., Stormo, G., Boehnke, M., Louis, T. A., Taylor, J., & Dixon, D. (2006). Training of the next generation of biostatisticians: A call to action in the US. *Statistics in Medicine, 25*, 3415–3429.

DeRisi, J., Penland, L., & Brown, P. O. (Group 1); Bittner, M. L., Meltzner, P. S., Ray, M., Chen, Y., Su, Y. A., & Trent, J. M. (Group 2) (1996). Use of cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics, 14*, 457–460.

Draghici, S., Khatri, P., Eklund, A. C., & Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. *TRENDS in Genetics, 22*, 101–109.

Dreifus, C. (2011). A conversation with the matriarch of modern cancer genetics. *New York Times*, February 8, D2; at: http://www.nytimes.com/2011/02/08/science/08conversation.html.

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for classification of tumors using DNA microarrays. *Journal of the American Statistical Association, 97*, 77–87.

Fan, X., Shi, L., Fang, H., Cheng, Y., Perkins, R., & Tong, W. (2010). DNA microarrays are predictive of cancer prognosis: A re-evaluation. *Clinical Cancer Research, 16*, 629–636.

Fujimura, J. H., & Chou, D. Y. (1994). Dissent in science. Styles of scientific practice and the controversy over the cause of AIDS. *Social Science & Medicine, 28*, 1017–1026.

Garge, N., Page, G. P., Spargue, A. P., Gorman, B. S., & Allison, D. B. (2005). Reproducible clusters from microarray research: Whither? *BMC Bioinformatics, 6*(Suppl. 2), 1–11.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology, 5*, R80.

Goldberg, P. (2009). Duke University suspends two clinical trials after journal paper questions assay results. *The Cancer Letter, 35*(37), 1. and 7–8.

Goldberg, P. (2010a). Duke in process to restart three trials using microarray analysis of tumors. *The Cancer Letter, 36*(3), 1. and 2–4.

Goldberg, P. (2010b). IOM review of Duke genomics trials to focus on validation, scientific criteria. *The Cancer Letter, 36*(38), 1–2.

Goldberg, P. (2011). IOM committee will probe Duke scandal together with other 'omics' case studies. *The Cancer Letter, 37*(1), 1–8.

Goss Tusher, V., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS, 98*, 5116–5121.

Hacking, I. (1985). Styles of scientific reasoning. In J. Rajchman & C. West (Eds.), *Post-analytic philosophy* (pp. 145–165). New York: Columbia University Press.

Hacking, I. (2002). Inaugural lecture: Chair of philosophy and history of scientific concepts at the Collège de France, 16 January 2001. *Economy and Society, 31*, 1–14.

Hagen, J. B. (2000). The origins of bioinformatics. *Nature Reviews Genetics, 1*, 231–236.

He, Y. D., & Friend, S. H. (2001). Microarrays: The 21st century divining rod. *Nature Medicine, 7*, 658–659.

Hess, K. R., Zhang, W., Baggerly, K. A., Stivers, D. N., & Coombes, K. R. (2001). Microarrays: Handling the deluge of data and extracting reliable information. *Trends in Biotechnology, 19*, 463–468.

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The fourth paradigm: Data-intensive scientific discovery*. Redmond: Microsoft Research.

Hsu, D. S., Balakumaran, B. S., Acharya, C. R., Vlahovic, V., Walters, K. S., Garman, K., et al. (2007). Pharmacogenomic strategies provide a rational approach to the treatment of cisplatin-resistant patients with advanced cancer. *Journal of Clinical Oncology, 25*(2007), 4350–4357.

Ioannidis, J. P. A. (2005). Microarrays and molecular research: Noise discovery? *Lancet, 365*, 454–455.

Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., et al. (2009). Repeatability of published microarray gene expression studies. *Nature Genetics, 41*, 149–155.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., et al. (2005). Multiple-laboratory comparison of microarray platforms. *Nature Methods, 2*, 345–349.

Johnson, K. (2007). Natural history as stamp collecting: A brief history. *Archives of Natural History, 34*, 244–258.

Kahn, S. D. (2011). On the future of genomic data. *Science, 331*, 728–729.

Keating, P., & Cambrosio, A. (1998). Interlaboratory life: Regulating flow cytometry. In J. P. Gaudillière & I. Löwy (Eds.), *The invisible industrialist: Manufacturers and the construction of scientific knowledge* (pp. 250–295). London: Macmillan. New York: St. Martin's Press.

Keating, P., & Cambrosio, A. (2003). *Biomedical platforms. Realigning the normal and the pathological in late-twentieth-century medicine*. Cambridge, MA: MIT Press.

Keating, P., & Cambrosio, A. (2012). *Cancer on trial: The rise of oncology as a new style of practice*. Chicago: The University of Chicago Press.

Kerr, M. K., & Churchill, G. A. (2001). Statistical design and the analysis of gene expression microarray data. *Genetic Research, Cambridge, 77*, 123–128.

Knapen, D., Vergauwen, L., Laukens, K., & Blust, R. (2009). Best practices for hybridization design in two-color microarray analysis. *Trends in Biotechnology, 27*, 406–414.

Kohli-Laven, N., Bourret, P., Keating, P., & Cambrosio, A. (2011). Cancer clinical trials in the era of genomic signatures: Biomedical innovation, clinical utility, and regulatory-scientific hybrids". *Social Studies of Science, 41*, 487–513.

Koscielny, S. (2010). Why most gene expression signatures of tumors have not been useful in the clinic. *Science Translational Medicine, 2*(14), 1–3.

Kothapalli, R., Yoder, S. J., Mane, S., & Loughran, T. P. Jr., (2002). Microarray results: How accurate are they? *BMC Bioinformatics, 3*, 22.

Kusch, M. (2010). Hacking's historical epistemology: A critique of styles of reasoning. *Studies in History and Philosophy of Science, 41*, 158–173.

Larkin, J. E., Frank, B. C., Gavras, H., Sultana, R., & Quackenbush, J. (2005). Independence and reproducibility across microarray platforms. *Nature Methods, 2*, 337–343.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics, 11*, 733–739.

Lewitter, F., & Bell, G. (2010). Maximizing an experiment. *Genome Technology* (June); http://www.genomeweb.com/maximizing-experiment.

Liu, X., Lee, S.-C., Casella, G., & Peter, G. F. (2008). Assessing agreement of clustering methods with gene expression microarray data. *Computational Statistics and Data Analysis, 52*, 5356–5366.

Lynch, M. (2002). Protocols, practices, and the reproduction of technique in molecular biology. *British Journal of Sociology, 53*, 203–220.

MAQC-II: analyze that! (2010). *Nature Biotechnology, 28*, 761.

Marshall, E. (1999). Do-it-yourself gene watching. *Science, 286*, 444–447.

Master, J. R. W., & Lakhani, S. R. (2000). How diagnosis with microarrays can help cancer patients. *Nature, 404*, 921.

McMeekin, A., & Harvey, M. (2002). The formation of bioinformatics knowledge markets: an 'economies of knowledge' approach. *Revue d'Économie Industrielle, 101*, 47–64.

McMeekin, A., Harvey, M., & Gee, S. (2004). Emergent bioinformatics and newly distributed innovation processes. In M. McKelvey, J. Laage-Hellman, & A. Rickne (Eds.), *The economic dynamics of modern biotechnology* (pp. 236–261). Oxford: Oxford University Press.

McShane, L. M., Altman, D. G., Sauerbrei, W., Taube, S. E., Gion, M., & Clark, G. M.Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. (2005). Reporting recommendations for tumor marker prognostic studies (REMARK). *JNCI, 97*, 1180–1184.

McShane, L. M., Radmacher, M. D., Freidlin, B., Yu, R., Li, M. C., & Simon, R. (2002). Methods for assessing the reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics, 18*, 1462–1469.

Mehta, T., Tanik, M., & Allison, D. (2004). Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nature Genetics, 36*, 943–947.

Michiels, S., Koscielny, S., & Hill, C. (2007). Interpretation of microarray data in cancer. *British Journal of Cancer, 96*, 1155–1158.

Michiels, S., Koscielny, S., & Hill, C. (2005). Prediction of cancer outcome with microarrays: A multiple random validation strategy. *Lancet, 365*, 488–492.

Miron, M., & Nadon, R. (2006). Inferential literacy for experimental high-throughput biology. *Trends in Genetics, 22*, 84–89.

Nadon, R., & Shoemaker, J. (2002). Statistical issues with microarrays: Processing and analysis. *Trends in Genetics, 18*, 265–271.

NCI awards first grants for Director's Challenge. (1999). Press release, NCI Office of Cancer Communications, 3 December; at: http://rex.nci.nih.gov/massmedia/pressreleases/molec_chal_grants.html.

NCI's Richard Simon: The state of the art in microarray informatics. (2002). *BioArray News*, 10 May; at: http://www.genomeweb.com/arrays/nci-s-richard-simon-state-art-microarray-informatics.

Ntzani, E. E., & Ioannidis, J. P. A. (2003). Predictive ability of DNA microarrays for cancer outcomes and correlates: An empirical assessment. *Lancet, 362*, 1439–1444.

Owens, J. (2005). Gene expression: Do microarrays match up? *Nature Reviews Genetics, 6*, 432–433.

Page, G. P., Edwards, J. W., Gadbury, G. L., Yelisetti, P., Trivedi, P., & Allison, D. B. (2006). The *PowerAtlas*: A power and sample size atlas for microarray experimental design and research. *BMC Bioinformatics, 7*, 84.

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., & Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics, 21*, 3017–3024.

Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., et al. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *PNAS, 96*, 9212–9217.

Potti, A., & Nevins, J. R. (2007). Reply. *Nature Medicine, 13*, 1277–1278.

Potti, A., Dressman, H. K., Bild, A., Riedel, R. F., Chan, G., Sayer, R., et al. (2006). Genomic signatures to guide the use of chemotherapeutics. *Nature Medicine, 12*, 1294–1300.

Pounds, S., & Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics, 20*, 1737–1745.

Pounds, S., & Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of *p*-values. *Bioinformatics, 19*, 1236–1242.

Quackenbush, J., & Irizarry, R. A. (2006). Response to Shields: MIAME, we have a problem. *Trends in Genetics, 22*, 471–472.

Radich, J. (2009). The promise and pitfalls of gene expression studies. *Best Practice and Research Clinical Haematology, 22*, 165–167.

Rogers, S., & Cambrosio, A. (2007). Making a new technology work: The standardization and regulation of microarrays. *Yale Journal of Biology and Medicine, 80*, 165–178.

Salit, M. (2006). Standards in gene expression microarray experiments. *Methods in Enzymology, 411*, 63–78.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science, 270*, 467–470.

Shields, R. (2006a). MIAME, we have a problem. *Trends in Genetics, 22*, 65–66.

Shields, R. (2006b). The emperor's new clothes revisited. *Trends in Genetics, 22*, 463.

Simon, R. (undated). Guidelines on Statistical Analysis and Reporting of DNA Microarray Studies of Clinical Outcome, PowerPoint presentation; at: http://linus.nci.nih.gov/techreport/Simon-Dos&Donts.ppt.

Simon, R. M., Korn, E. L., McShane, L. M., Radmacher, M. D., Wright, G. W., et al. (2003). *Design and analysis of DNA microarray investigations*. New York: Springer.

Simon, R. M., Radmacher, M. D., & Dobbin, K. (2002). Design of studies using DNA microarrays. *Genetic Epidemiology, 23*, 21–36.

Singer, N. (2010). Duke suspends researcher and halts cancer studies. *The New York Times*.

Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome Biology, 11*, 207.

Strasser, B. J. (2008). GenBank: Natural history in the 21st century? *Science, 322*, 537–538.

Strasser, B. J. (2011). The experimenter's museum: GenBank, natural history, and the moral economies of biomedicine, 1979–1982. *Isis, 102*, 60–96.

Thalamuthu, A., Mukhopadhyay, I., Zheng, X., & Tseng, G. C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics, 22*, 2405–2412.

The chipping forecast. (1999). *Nature Genetics, 21*(Suppl. 1).

The chipping forecast II. (2001). *Nature Genetics, 34*(Suppl. 2).

Tuma, R. S. (2010). Calls for new reporting standards, quality control in microarrays. *JNCI, 102*, 1380–1381.

Vance, A. (2009a). Data analysts captivated by R's power. *New York Times*.

Vance, A. (2009b). R you ready for R? *New York Times*.

Winegarden, N. (2003). Microarrays in cancer: Moving from hype to clinical reality. *Lancet, 362*, 1428.