ANALYSIS

nature genetics

Repeatability of published microarray gene expression analyses

John P A Ioannidis[1–3], David B Allison[4], Catherine A Ball[5], Issa Coulibaly[4], Xiangqin Cui[4], Aedín C Culhane[6,7], Mario Falchi[8,9], Cesare Furlanello[10], Laurence Game[11], Giuseppe Jurman[10], Jon Mangion[11], Tapan Mehta[4], Michael Nitzberg[5], Grier P Page[4,12], Enrico Petretto[11,13] & Vera van Noort[14]

Given the complexity of microarray-based gene expression studies, guidelines encourage transparent design and public data availability. Several journals require public data deposition and several public databases exist. However, not all data are publicly available, and even when available, it is unknown whether the published results are reproducible by independent scientists. Here we evaluated the replication of data analyses in 18 articles on microarray-based gene expression profiling published in *Nature Genetics* in 2005–2006. One table or figure from each article was independently evaluated by two teams of analysts. We reproduced two analyses in principle and six partially or with some discrepancies; ten could not be reproduced. The main reason for failure to reproduce was data unavailability, and discrepancies were mostly due to incomplete data annotation or specification of data processing and analysis. Repeatability of published microarray studies is apparently limited. More strict publication rules enforcing public data availability and explicit description of data processing and analysis should be considered.

Microarray-based research is a prolific scientific field[1] where extensive data are generated and published. The field has been sensitized to the need for transparent design and public data deposition[2–5] and public databases have been designed for this purpose[6–8]. Issues surrounding the ability to reproduce published results with publicly available data have drawn attention in microarray-related research[9–11] and beyond. The reproducibility of scientific results has been a concern of the scientific community for decades and in every scientific discipline. In biomedical research, the Uniform Guidelines of the International Committee of Medical Journal Editors state that authors should "identify the methods, apparatus and procedures in sufficient detail to allow other workers to reproduce the results"[12]. Making primary data publicly available has many challenges but also many benefits[13]. Public data availability allows other investigators to confirm the results of the original authors, exactly replicate these results in other studies and try alternative analyses to see whether results are robust and to learn new things. Journals such as *Nature Genetics* require public data deposition as a prerequisite for publication for microarray-based research. Yet, the extent to which data are indeed made fully and accurately publicly available and permit confirmation of originally reported findings in many areas, including gene expression microarray research, is unknown.

In this project, we aimed to evaluate the repeatability of published microarrays studies. We focused specifically on the ability to repeat the published analyses and get the same results. This is one important component in the wider family of replication and reproducibility issues. We evaluated 18 articles published in *Nature Genetics* in 2005 or 2006 that presented data from comparative analyses of microarrays experiments that had not been previously published elsewhere. Detailed eligibility criteria and search strategies are presented in the Methods section. Of 20 initially selected articles[14–33], 2 were excluded[21,26] when they were found to use previously published data. The 18 evaluated articles[14–20,22–25,27–33] and the selected tables or figures we attempted to reproduce are shown in **Table 1**. They cover a wide variety of methods and applications, as expected from a multidisciplinary genetics journal. Of the 18 articles, 16 declare in either the primary article or its supplements that the gene expression profiling experimental data

[1]*Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece.* [2]*Biomedical Research Institute, Foundation for Research and Technology–Hellas, Ioannina 45110, Greece.* [3]*Center for Genetic Epidemiology and Modeling, Tufts Medical Center and Department of Medicine, Tufts University School of Medicine, Boston, Massachusetts 02111, USA.* [4]*Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA.* [5]*Department of Biochemistry, Stanford University School of Medicine, Stanford, California 94305, USA.* [6]*Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA.* [7]*Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA.* [8]*Genomic Medicine, Faculty of Medicine, Imperial College London, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK.* [9]*Department of Twin Research & Genetic Epidemiology, St. Thomas' Campus, King's College London, Lambeth Palace Road, London SE1 7EH, UK.* [10]*Fondazione Bruno Kessler, via Sommarive 18, 38100 Povo-Trento, Italy.* [11]*Medical Research Council Clinical Sciences Centre Microarray Centre, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK.* [12]*Statistics and Epidemiology Unit, RTI International, Atlanta, Georgia 30341, USA.* [13]*Department of Epidemiology, Public Health and Primary Care, Faculty of Medicine, Imperial College, Praed Street, London W2 1PG, UK.* [14]*European Molecular Biology Laboratory Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany. Correspondence should be addressed to J.P.A.I. (jioannid@cc.uoi.gr).*
Received 15 September 2008; accepted 4 November 2008; published online 28 January 2009; doi:10.1038/ng.295

Table 1 Selected articles and analyses to reproduce, mention of data availability and ability to download and correspond data to the published paper

Reference	Analysis selected	Mention of public availability of datasets	Completeness of coverage, ability to download and correspondence of datasets to published analyses
14	Table 1	Yes (GEO: GSE5800)	OK
15	Fig. 2a,b	Yes (GEO: GSE11324)	OK
16	Table 2	Yes (GEO: GSE5335)	OK
17	Fig. 3a	Yes (GEO: GSE5089)	OK
18	Fig. 1b	Yes (GEO: GSE3406)	OK
19	Fig. 5	Yes (GEO: GSE4189)	OK
20	Fig. 1	Yes (ArrayExpress: A-TIGR-9; E-TIGR-24 to E-TIGR-80; E-TIGR-111 to E-TIGR-116; E-TIGR-126; E-TIGR-127; Public database: TREX http://pga.tigr.org/)	Unable to find with certainty the data for Figure 1. Links are not available in the TREX website for the whole experiment. From ArrayExpress entries, we could not find the correct arrays for the experiment to reproduce
22	Fig. 1	Yes (GEO: GSE3031)	OK
23	Fig. 4b	No	Data not available
24	Fig. 2b,c	Yes (GEO: GSE3047)	OK
25	Fig. 2	Yes (http://splicing.rockefeller.edu/rawdata.php)	The project website provides 'chart' and 'raw' data, although the 'raw' data are actually processed data that require quite tedious editing. The website provides data for 370 records (corresponding to 290 genes), whereas the microarray contained 40,443 probe sets derived from 7,715 genes. Data are not shown for individual probe sets
27	Fig. 1	Yes (GEO: GSE4123)	Unable to correspond datasets to the published analyses
28	Fig. 1c	Yes (ArrayExpress: E-TABM-6)	OK
29	Fig. 1b,c	Yes (ArrayExpress: E-AFMX-9; E-TABM-17)	OK but had problems downloading E-AFMX-9. Contacted ArrayExpress curators who verified that updated data were in E-TABM-17 and had been reannotated. Annotation had been cleaned up but did not exactly match the annotation used in the paper
30	Fig. 3	No	Data not available
31	Fig. 1a	Yes (ArrayExpress: A-UMCU-3; E-UMCU-11; P-UMCU-11; P-UMCU-18 to P-UMCU-26)	OK
32	Table 1	The given website link is mistyped and thus nonfunctional; we eventually found the correct address through a web search	The URL lists multiple datasets, some in summary data form (GSEA), rather than individual-level reporter-specific data, and it was not easy to tell which set corresponded to each analysis. The data for the Table 1 analysis are not all provided in individual-level reporter-specific form to allow replication of the analysis
33	Fig. 2	Yes	No individual-level reporter-specific data are available to try to replicate the analysis

described are publicly available, and 13 of them had published GEO or ArrayExpress accession numbers.

Of the 16 articles with data in a public repository, one had deposited only summary analyzed information rather than complete unprocessed data that would be necessary to reproduce the analysis; one had submitted mostly summary-level data; and one had data provided for only 370 genes, although the microarray contained 40,443 probe sets derived from 7,715 genes. Additionally, for two articles, it was not possible to ascertain which dataset corresponded to which analysis, and for another article, ArrayExpress curators had created a new dataset with updated annotation that no longer matched exactly data annotation in the published article (**Table 1**). This left 10 of 16 articles with no problems in data availability, downloading and correspondence of deposited data annotation and published analyses, enabling attempts at reproduction of the data analyses described in these articles

Most of the 15 articles that had deposited individual-level reporter-specific datasets provided primarily processed information, and there were often ambiguities in the data processing and analysis path (**Table 2**). Scanned images were generally not available, with the exception of one article for which images had been deposited in Stanford Microarray Database (SMD). However, images are not required by current MIAME guidelines to be routinely deposited. Raw, unprocessed data were available for eight articles. The processing methods were described with very heterogeneous quality and quantity of information. Although four articles were judged to have sufficient detail at face value, in other cases there was limited or ambiguous information (see **Table 2** for examples). As a result, even though processed data were typically available, often it was unclear how exactly they had been derived from the raw data. For 2 of the 15 articles, it was not possible to identify the specific array platform with detailed descriptions of each array element, as specified by MIAME guidelines. Besides the array platform description, although typically there were no major MIAME violations, the completeness of detail in MIAME coverage varied considerably.

Information about the scanner used was unavailable in four articles, and in most articles where information was given, the exact model was not specified. Information on the software used and version was even more limited: only six studies mentioned the software and only one gave the version as well. Image processing parameters were often presented with limited information or none at all, a typical statement being that the manufacturer's procedures were followed (**Table 2**).

In the overall effort to reproduce the results of the 18 specified tables and figures (**Table 3** and **Fig. 1**), the selected analyses could be said to be 'reproduced in principle' (that is, the differences in results obtained by our teams of analysts and those reported in the original paper were felt to be minor) in only two cases; in another six cases, the analyses were reproduced with some discrepancies, and moreover, in one of these six, only partial reproduction of some of the data and panels was possible. Finally, in ten articles, the selected analyses could

not be reproduced. For 16 of the 18 articles, the two evaluating teams of analysts reached independently the same categorization for the overall repeatability. In the other two articles, they had identified the same problems or difficulties but had originally used a different categorization for their severity (see **Table 3** and Methods for details).

Inability to reproduce the analyses (**Fig. 1**) was mostly due to unavailability of data (no data at all, $n = 2$; no individual-level reporter-specific data, $n = 1$; data on a limited set of genes only, $n = 1$), inability to determine which data corresponded to which analyses ($n = 1$) or both ($n = 1$). In one other article[24], the major stumbling block was the unavailability of the GenRate algorithm and software in the public domain; in another article[31] that provided both raw and processed data, the documentation of preprocessing was insufficient to allow reproduction of the results and we met with problems of gene annotation in the processed data. Finally, in two articles[14,20], the raw cel files were not available and reproduction efforts were either considered impossible or gave very different results than the published ones when analysis was attempted, apparently because crucial analytical choices made were unknown. Details for the example of one article where results could not even be approximated by either team of assigned analysts are shown in **Supplementary Table 1** online. For contrast, **Supplementary Figure 1** online shows a figure from an article where it was possible to reproduce the results with some discrepancies.

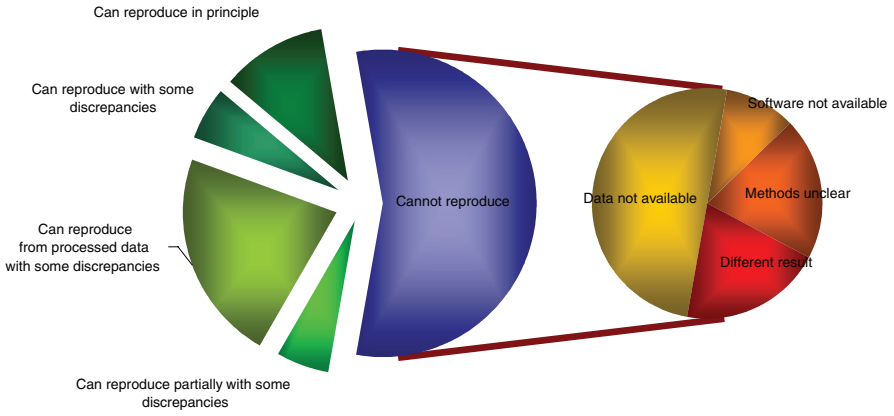

Figure 1 Summary of the efforts to replicate the published analyses.

Even in the eight articles where some replication of analysis was achieved, our results were not perfect matches with the published results. In two of these articles, the differences were felt to be minor and we judged that the results could be well reproduced in principle, even though minor details still suggested that neither team of analysts obtained exactly the same results as the published paper (**Supplementary Fig. 2** online). In the other six articles, there were more considerable discrepancies, resulting from insufficient documented detail on the processing and analytical options adopted by the authors. For example, in one article[17], in our attempt to use the authors' criteria, we found 120 eligible transcripts instead of the 162 published in the paper. Of those, we found 22 instead of the 33 published in the paper with an adjusted P value <0.01 and twofold enrichment. Details on all the discrepancies are shown in **Table 3**.

Our findings demonstrate that although microarrays are being used to produce consistent data with different platforms and/or at different laboratories[4,34,35], the task of repeating published microarray analyses requires much greater detail than that provided by descriptions of the platform used[36–38] or publicly available data. The complexity of experimental design, quantification, normalization, statistical analysis and computation issues involve many possible steps and different decisions. When data analysis steps are very complex and work intensive, it may be difficult or even impossible for even experienced teams of outsiders to reproduce published studies. However, our results also suggest that many, if not most, microarray analyses could potentially be largely reproduced if the data are available and adequately annotated and if the analytic steps and parameters are sufficiently described.

The lack of repeatability we observed should not be taken as evidence that the published analyses are wrong. Confirming or disproving these published analyses and evaluating their correctness was not our intention. In fact, we considered the published results as the gold standard—our experiment asked simply whether experienced independent analysts could reach the same results using the data and information that were publicly available. Moreover, we focused on one circumscribed analysis in each evaluated article. Each analysis was selected according to explicit rules that would maximize objectivity, reproducibility and transparency for our selection process. The analysis chosen was not necessarily the most important one presented in each paper, and judging which analysis was the most important would have been highly subjective.

We should also acknowledge that we focused on a single journal and a period of two years (2005–2006). *Nature Genetics* has implemented a strict policy requiring public data availability and compliance with MIAME guidance. It is not obvious that the quality of data availability and methods reporting would be better in journals with less strict requirements and policies—the opposite might be more plausible, if anything. Moreover, as the analyzed papers are recent, it seems unlikely that the situation has changed radically in the last two years.

The lack of sufficient data in the public domain reduces the options for efficient integration of information from many studies, and the option to use these data creatively to address additional research questions that may arise. Repeatability is only one part of a longer chain of other reproducibility issues. In some cases, independent replication of experiments happens and the conclusions from different studies can actually be compared to one another. However, when repetition of analysis steps is impossible, comparing the results of related studies and understanding any potential discrepancies is challenging. For the most part, few large-scale experiments such as those using microarrays are directly replicated, owing to expense and sometimes to unavailability of rare biological samples. Because it is not possible to rely on direct reproduction of most experiments, one should at least be able to understand and re-execute the data analysis steps described in the publication to satisfy healthy scientific scepticism.

Articles with more transparent availability or data and analyses may be able to achieve a higher impact in the literature, as more researchers may be able to use them. In an exploratory evaluation, we compared the number of citations catalogued by ISI as of the end of August 2008 for articles where some reproduction of at least part of the results was feasible (in principle or with some discrepancies) versus those where we could not reproduce the selected analyses. We found that the former group of articles had received more citations, after adjustment for the time of publication (median 29.8 per year (range 7.5–86.6) versus 12.4 per year (range 5.7–29.4), $P = 0.038$). The citation data should be considered with great caution, as there

Table 2 Availability of data and other pertinent information

Ref.	Raw data	Processing methods described	Processed data	Platform description	Scanner used	Software/ version	Image processing parameters
14	No cel files given (SOFT text file)	Normalized and modeled using the perfect match–mismatch difference model with negative values truncated to a default value of 10	Processed to some extent (SOFT text file)	Deposited in GEO	Affymetrix 3000	Not stated	Alludes to "manufacturer's recommended protocols"
15	Given	RMA (robust multichip averaging) without specifying software. A customized CDF was used to process the data but the version of this CDF was not mentioned	Yes, two RMA normalized datasets processed, one with a custom CDF and the other with a default Bioconductor Affy CDF are available from the authors website or GEO, respectively	Deposited in GEO	Hewlett-Packard GeneArray	Not stated	Not stated in sufficient detail
16	Not given	Log transformation followed by per-chip normalization (division by median value in each chip)	Yes	Deposited in GEO	Affymetrix	Not stated	Alludes to "standard Affymetrix," no further details
17	Not given	No details given: "After removal of low-quality, redundant, (partially) overlapping or ambiguous probes"	Yes	Deposited in GEO	Not stated	Not stated	Not stated in sufficient detail
18	Not given	Background intensity was subtracted using a Bayesian correction and the ratios of the two dyes were $\log_2$ transformed. $\log_2$ ratios were then corrected for intensity-dependent and spatial biases by subtracting a Lowess curve followed by a median filter	Yes	Deposited in GEO	ScanArray 4000	QuantArray 3	Information provided
19	Given	RMA normalization, no further details, but reference provided in supplement	Yes	Deposited in GEO	Not stated	Not stated	"Standard Affymetrix procedures," no further details
20	Not given	Described in detail	Some	Deposited in ArrayExpress	Axon Scan	Not stated	Not stated in sufficient detail
22	Given	Log base 2 (PM_635/PM_532) ratios were bi-weight mean centered	Yes	Deposited in GEO	Provided by NimbleGen Systems	Not stated	Not stated in sufficient detail
24	Not given	Described in detail	Yes	Deposited in GEO	Not stated	Not stated	Not stated in sufficient detail
25	Not given	Described in detail	Yes	"Custom Affymetrix," no exact description	Agilent Gene Array Scanner	Not stated	Not stated in sufficient detail
27	Given	Described in detail	No	Deposited in SMD and GEO	Axon	GenePix	Not stated in sufficient detail
28	Given	Described in detail	Yes	Deposited in ArrayExpress	ScanArray 5000	GLEAMS	Adaptive morphological detection method
29	Given	gcRMA R package, no version given	Yes	Deposited in ArrayExpress	Affymetrix	GeneChip	Refers to Affymetrix manual
31	Given	Refers to P-UMCU-11 but information is insufficient; article also mentions "background corrected and normalized" but without details	Yes	Deposited in ArrayExpress	Agilent G2565AA	Agilent Feature Extraction 7.5	Imagene 4.0 used, P-UMCU-25 does not provide full details
32	Given	Scaling to the median expression experiment	Yes	Commercial Affymetrix platforms	Affymetrix	GeneChip MAS4	Not stated in sufficient detail

Information is shown only for the 15 articles where at least individual-level data were available (2 articles had no data and another one had only summary data).

are many other factors that influence citations, but other authors have also reported on increased citation of publications with online-accessible material[39]. Although both MIAME guidelines and journal policies address public data availability, considerable room still remains for different interpretations of exactly which data should be available. The current public deposition databases should have no problem in allowing deposition of both raw and processed data and all the related platforms. Unambiguous connection of data to presented experiments and results is also important. Otherwise, even though some data may be publicly available, they may be either entirely unusable or their usability may be suboptimal.

A common experience of the analysts involved in this project was the difficulty reproducing the published analyses, even when data were available. There were often steps in the analysis process for which parameters or procedures were inadequately described and we had to make educated decisions about what to do. In cases where more than one step had more than one possible option, the combinatorial number of options that could be pursued became too burdensome to fully evaluate. Perhaps the appropriate combination of these options would have led to the exact published results, but this was typically not clearly recognizable from the published information. This lack of specific information led to considerable ambiguity; we needed to invest considerable effort to try to

Table 3 Summary of efforts at reproducing the published results

Ref.	Task	Reproduction summary	Comments on the reproduction exercise
14	**Table 1**	Cannot reproduce	Starting from SOFT text file available from GEO, it was not possible to reproduce the reported changes in gene expression as a multiple of WT control. For details, see **Supplementary Table 1**
15	**Figure 2a**,**b**	Can reproduce partially with some discrepancies	Although a heatmap of *t*-statistic values as in **Figure 2a** could be generated, the number of significant genes obtained was not consistent with the published report of 275 significant genes. The article does not specify what significance threshold (*P*-value cutoff) was used. The percentage of genes that are upregulated versus downregulated at 0 h versus 3 h at $P < 0.001$ was 50.1% and 49.9%, respectively, when we analyzed the processed data on the authors' website, and 51.5% and 48.5%, respectively, when we analyzed the processed data from GEO. These findings are close but not exactly equal to the published findings of 51.2% and 48.8%
16	**Table 2**	Can reproduce from processed data with some discrepancies	Although we found the same genes showing at least 25% expression difference between 'Aggressive' lines and 'Neutral' lines as the published paper, the expression fold change values were inconsistent. We checked the expression mean values we obtained for the four lines against the values displayed on the graphs of the **Supplementary Figure 2**. There seems to be no discrepancy of the mean expression values of the lines. We also observed some inconsistencies regarding the expression values included in the study. The authors reported that they did not use all A-flagged signals. However, over the 12 total samples, genes *CG31475* and *CG13252* had 10 and 12 A-flagged expression values, respectively, yet their expression values are included in **Table 2**. Because of this, one team of analysts gave an original categorization of the article as "cannot reproduce" and the other as "can reproduce partially with some discrepancies," and consensus was reached to categorize it as "can reproduce with some discrepancies"
17	**Figure 3a**	Can reproduce from processed data with some discrepancies	The trends are the same, and the reanalysis would lead to the same conclusions. There are some differences in the data included in the graphs. For Lam, 95% of the genes overlap the author selection, described in the **Supplementary Table 1**. For LamDeltaCAAX, 120 transcripts instead of the published 162 were selected using the author criteria, and 22 instead of 33 showed adjusted *P* value <0.01 and twofold enrichment; no overlapping comparison was possible
18	**Figure 1**	Can reproduce from processed data with some discrepancies	Plot of the reanalyzed data shows that the figures look more or less the same with the extremes of the figures removed. The *r* values obtained differ by up to 10% compared with the published *r* values (for example, 0.499 versus 0.554 for **Fig.1b**, item 3)
19	**Figure 5**	Can reproduce from processed data with some discrepancies	Reconstruction from the raw GenePix data was incomplete for lack of sufficient information. There was no clear unique identifier that could be used to cross-reference the data from the raw file with the data from the processed file (only ~2,000 genes out of ~16,000 match between the raw and processed file gene names). The reanalysis using the raw data was therefore not possible. The reconstruction from preprocessed data was, however, successful for plots. The results of the last panel seemed to have differences from the ones reported in the paper: the maximum frequency of Oct4 and Nanog found in the reanalysis was 10%, whereas it was reported in the original paper as ~17%. One team of analysts categorized this as "can reproduce in principle," whereas the other categorized it as "can reproduce with some discrepancies," and consensus was reached for the latter categorization (**Supplementary Fig. 1**)
20	**Figure 1**	Cannot reproduce	One team of analysts abandoned effort to reproduce, owing to inability to correspond data. The other team of analysts attempted reproduction, but no transcript was significantly associated with the infarct trait using the method and threshold used by the authors. The latter is not completely clear, as the authors state in the **Supplementary Methods** that "a gene was considered an excellent candidate for involvement in infarct size if >98% of the linear regressions exhibited an estimated *P* value <0.05." However, in the table, they report several genes with percentage occurrence of a *P* value <0.05 smaller than 98%. We did not observe any association after relaxing the threshold to 80% of the regressions showing estimated *P* value <0.05. It was not explicitly declared whether the expression data had been normalized and with which method, and it was not clear whether the phenotypic data had been transformed and/or outliers removed. Indeed, the total number of rats with cardiac data in PhysGen was slightly different than the number used in the paper (62 FHH instead of 22 and 105 SS instead of 70)
22	**Figure 1**	Can reproduce in principle	Almost perfect reconstruction was possible from raw data, but representative data samples, columns from the data files and condition on valid genes were unspecified and had to be found experimentally. There were a few nonconsequential differences in the scatter plots. For instance, one team of analysts found that in **Figure 1a**,**b** there were no values in the area below $x = 100$, $y = 100$, although some points lie in that area in the original figures; the other team of analysts found values in this area and almost perfectly identical data points overall (very minor displacement, if at all) (**Supplementary Fig. 2**)
23	**Figure 4b**	Cannot reproduce	Data were not publicly available
24	**Figure 2b**,**c**	Cannot reproduce	The authors provided the parameters used for the GenRate analysis. However, a more explicit definition of one parameter (sensitivity) should have been provided (995, "By varying θ, we obtained exon and CoReg FDERs varying from 0.13% to 32% and from 0.2% to 37%, respectively"). Analysis could not be reproduced from raw data, as results presented in **Figure 2b** (distributions of maximum probe signal intensities detected by GenRate) and **Figure 2c** (accuracy versus recall of GenRate) rely on application of GenRate algorithm to microarray data, and the GenRate algorithm/software does not appear to be publicly available
25	**Figure 2**	Cannot reproduce	Raw data appear to be partial or missing, so no direct reproduction seems possible. No direct information was given in text for this figure, although methods for previous figures were detailed. Some processed data may be available from the website. Data were not available for all genes. In addition, the data for each of the 370 genes presented on the website were stored in separate tables
27	**Figure 1**	Cannot reproduce	Figuring out which channel of which array was from which sample was a stumbling block for attempting a reproduction of the analyses
28	**Figure 1c**	Can reproduce in principle	Figures to be reproduced are heatmaps. The general theme of the heatmaps generated from the data available is identical to what is presented in the paper
29	**Figure 1b**,**c**	Can reproduce with some discrepancies (**Fig 1b** from raw data; **Fig 1c** from processed data)	The results we obtained for **Figure 1b** are different than those listed in the published paper. This may be because of either lack of details for the gcRMA normalization or ambiguous definition of tissue clusters. **Figure 1c** cannot be reproduced from the raw data based on information available in the paper, as the specific algorithm used for processing is not fully described. Processed gene expression data and marker gene list are nevertheless provided on the authors' website. Using these data, it was possible to reproduce (in part) **Figure 1c**. Some of the clusters of tissue-specific coexpressed genes are similar to what is depicted in the original **Figure 1c**. However, it was not possible to reproduce exactly the cluster pattern or the dendogram as ordered in the original figure owing to incomplete description of the gene filtering and clustering algorithm used
30	**Figure 3**	Cannot reproduce	The authors used an array comparative genomic hybridization approach to analyze palindrome formation at the genome scale in human cancer. No expression assay was done. The data are not available in detail to allow reanalysis. Even if available, it would not be very clear based on the methods how to do the *t*-test between cytogenic bands and how to handle the alluded *q*-values
31	**Figure 1a**	Cannot reproduce	Reproduction from raw data requires several decisions in pre-processing that are not adequately documented (background subtraction, normalization, application of error model). Gene filtering from spot morphology is not described. Reproduction from preprocessed data is also impossible for problems with gene annotations
32	**Table 1**	Cannot reproduce	Although some data are seemingly available, the GSEA data are not presented in sufficient detail to allow reproducing the analyses
33	**Figure 2**	Cannot reproduce	No individual-level data are available, only some summary-level information

reproduce a single focused analysis and usually obtained unsuccessful or at best approximate results. For some articles, we estimate that a team of analysts may have spent over a week's work trying to reproduce a single table or figure. Even though we kept no detailed time sheets, on average, each team of analysts spent approximately half a week per table or figure. Clearly, carrying out this process is not feasible for the average peer reviewer, who is reviewing a whole paper, not just a single table or figure.

Given that the lack of MIAME-compliant data is a key problem for several of the evaluated articles, the situation may be improved if MIAME compliance is scored explicitly for each paper to be published, as has been proposed by ArrayExpress as a service to reviewers and editors[40]. The other key problem, insufficient description of data analysis procedures, is not easy to dissect with quick, automated prepublication checks. In-depth probing of even a fraction of the analyses can take much time. Published articles should ideally have supplements that provide codes or protocols for the sequence of analytical choices and implementation of each of the major analyses presented in the paper. Statistical and bioinformatics research methods should be reproducible[41]. We should caution that even a detailed code would not make a novel and complicated analysis trivial to reproduce, nor would it totally eliminate the possibility for bias (for example, selective reporting of only one analysis among several undertaken[42]). It is not necessary for the code trail to be given in each minute detail, but important decision nodes should be described. The methods sections of the papers whose analyses we could reproduce (completely or in principle) provide examples[22,28] of successful descriptions of the analysis codes. These papers demonstrate that sufficiently describing a data analysis is feasible without unrealistic extra effort for the data submitter.

METHODS

Eligible articles. We selected articles published in *Nature Genetics* between January 2005 and December 2006 that had used profiling with microarrays in order to examine differential gene expression of two or more groups or samples, one group under two or more conditions, or two or more groups of genes in the same samples. We selected papers where the gene expression profiling was a substantial enough part of the whole investigation to be mentioned in the abstract and to have the respective results presented in at least one table or figure or part thereof. We selected articles regardless of the material used, number of samples, chip used, modeling approach (supervised or unsupervised) and species concerned, and regardless of whether other techniques were used and other types of data were also generated or not. Both cDNA and oligonucleotide probes were eligible. We excluded papers without primary data (editorials, reviews, comments), papers based on data integrated from previous studies and short papers without an abstract (letters to the editor and brief communications). We excluded tiling arrays, DNA methylation arrays, copy number arrays and SNP arrays, but included miRNA array–based expression studies, splicing expression arrays and exon expression arrays.

Articles were first identified in a PubMed search using the query strategy Nature Genetics [SO] AND (microarray* OR gene expression profiling) AND (2005 [DP] OR 2006 [DP]). The strategy did not aim to have perfect (100%) sensitivity in identifying all articles using microarrays technology for comparative gene expression analyses but rather to yield an objectively selected sample that would be enriched in eligible articles. The identified items were screened further for eligibility. Of the 56 items retrieved electronically, 20 articles were considered potentially eligible for the project.

Evaluation teams and selection of main analysis to reproduce. Four teams of analysts evaluated the 20 potentially eligible articles. The four teams were from University of Alabama at Birmingham (UAB), Stanford/Dana-Farber (SD), London (L) and Ioannina/Trento (IT). Each team was comprised of three to six scientists who worked together to evaluate each article. Each of the eligible articles was randomly allocated to two of the four teams, with randomization conducted by computer pseudorandom numbers.

Each team independently examined its allocated articles for data availability, completeness of available information, obvious errors in the presented analyses and ability to reproduce one specific main analysis from each article.

For each eligible article, the result to reproduce was set to be the first eligible analysis (comparison of groups for differential gene expression by microarrays, as above) reported in the results section by a table or figure of the main article (not supplements). By doing this, we ensured that we would focus on an analysis that was a main theme of the paper. When several tables or figures were dedicated to gene expression profiling results, we focused on the first mentioned results. If this first mention pertained to only some of many figure panels, we focused on the alluded figure panels.

The teams assigned to common papers agreed first on which table or figure and panels were to be evaluated but thereafter worked entirely independently to locate the pertinent data and conduct the analysis. There was intentionally no communication with the authors of the articles to request additional data or clarifications because we wished to rely only on the information that an investigator would have if they were unable to contact a primary author and receive her assistance. One might argue that a paper is only an 'advertisement', and that interested researchers can communicate with the primary authors to obtain the full data, algorithms and methods. However, this may not always work, and one cannot assume that the authors keep extended private records on their published studies for lengthy periods of time, nor can one assume that the authors will remain employed in a position where they can devote time to readers of previous publications. For the value of the publication to persist as long as possible, it is preferable that the full information in the published paper and related links and accession databases is publicly available. We also wanted to avoid any chance of interference by the primary investigators in the independent appraisals by the analysts. On closer scrutiny of the 20 articles, 2 of them were found to analyze only data that had already been previously published and thus were not eligible per our criteria. Therefore, we present the results of the replication of analyses for the other 18 eligible articles.

Evaluated items. For each eligible study, the allocated teams of analysts independently considered the following items. First, they determined whether there was any mention in the primary article or its supplements about public availability of datasets and protocols or methods for any of the gene expression profiling experiments described, and whether a GEO or ArrayExpress accession number was provided.

Second, they determined whether publicly available datasets covered all or only some of the gene expression profiling experiments described and whether they could be downloaded with sufficient information to determine which array corresponded to which experimental condition. If not, they determined what exactly was not covered and/or not clearly identified (in the case of datasets that do not correspond clearly to specific experimental conditions described in the paper).

Third, for each of the publicly available datasets, each team evaluated (i) whether scanned images were publicly available; (ii) whether raw, unprocessed data were available; (iii) whether the processing methods seemed described in sufficient detail (regardless of whether they were considered correct or wrong); (iv) whether processed data were available; (v) whether the array platform was described with detailed descriptions of each array element, as specified by MIAME; (vi) whether any other MIAME incompliance was noted (per MIAME items 1–6); and (vii) whether anything else was missing that is considered essential (that is, information regarding which scanner was used to scan images, which software and version was used and which image processing parameters were used).

Fourth, for the selected main analysis to be reproduced, each team determined whether the presented results and methods in the paper were given in sufficient detail for reproduction to be even considered or whether the results or methods were unclear, the methods were wrong or more than one of the previous problems existed. Each team determined whether the specific results of the main analysis are reproduced (i) entirely in principle (the same results were reproduced for all presented results), (ii) partially (some results were the same, others were not possible to reproduce at all) (iii) with some discrepancies, (iv) partially and with some discrepancies or (v) not at all.

Each team of analysts first tried to reproduce the selected main analysis starting from the raw, unprocessed data and the described processing methods. If this failed, the perceived reason(s) were recorded. A separate repeatability evaluation

starting from the presented processed data (after normalization, standardization and/or other processing) as supplied in each article's relevant databases was carried out to see whether the final results were reproducible.

Flow of information and consensus. The four teams of analysts independently sent their evaluations of the allotted papers to one of the authors (V.v.N.) who did not participate in the data extraction and repeatability efforts. This author collated all evaluations and examined whether there were any discrepancies in the overall final assessments. In two articles where discrepant overall assessments were observed between the teams evaluating the same article, the differing teams convened electronically and/or by teleconference to discuss their discrepancies. Consensus was reached on both articles after discussion. In both articles, the two teams of analysts found the same problems or difficulties in the reproduction effort but had originally ascribed different significance or categorization to them. For one article, one team categorized the same problems as "can reproduce partially with some discrepancies," the other as "cannot reproduce"; for another article, one team categorized the same problems as "can reproduce with some discrepancies," the other as "can reproduce partially in principle." For both cases, the consensus was to categorize as "can reproduce with some discrepancies." In one other case, one team had not noted that data were publicly available, but even with the publicly available data, there was not sufficient information to reproduce the published analysis. Finally, the lead author (J.P.A.I.) examined all the detailed data extractions and replication of analysis reports and generated the common final tables summarizing the findings. The final summary data were then perused again by all four teams to ensure that there were no errors in the entry and summarization of information.

URLs for public databases. J. Craig Venter Institute, http://pga.tigr.org/; Nova-dependent regulation of brain-specific splicing, http://splicing.rockefeller.edu/rawdata.php.

Accessions for evaluated databases. NCBI GEO:GSE5800; GSE11324; GSE5335; GSE5089; GSE3406; GSE4189; GSE3031; GSE3047; GSE4123. ArrayExpress: A-TIGR-9; E-TIGR-24 to E-TIGR-80; E-TIGR-111 to E-TIGR-116; E-TIGR-126; E-TIGR-127; E-TABM-6; E-AFMX-9; E-TABM-17; A-UMCU-3; E-UMCU-11; P-UMCU-11; P-UMCU-18 to P-UMCU-26.

Note: Supplementary information is available on the Nature Genetics website.

AUTHOR CONTRIBUTIONS

The protocol was designed with discussion among all authors. All authors except V.v.N. participated in evaluations of the eligible articles and their analyses. V.v.N. collected all the evaluations and examined if there were discrepancies among teams. J.P.A.I. wrote the manuscript, which was critically revised by all other coauthors. After the first author, the author order is alphabetical.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Schena, M. *Microarray analysis.* (John Wiley & Sons, Hoboken, New Jersey, 2003).
2. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
3. Allison, D.B., Cui, X., Page, G.P. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.* **7**, 55–65 (2006).
4. Shi, L. *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intra-platform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24**, 1151–1161 (2006).
5. Anonymous. Minimum compliance for a microarray experiment? *Nat. Genet.* **38**, 1089 (2006).
6. Ball, C.A. *et al.* Submission of microarray data to public repositories. *PLoS Biol.* **2**, e317 (2004).
7. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization assay repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
8. Brazma, A. *et al.* Array Express – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**, 68–71 (2003).
9. Larsson, O. & Sandberg, R. Lack of correct data format and comparability limits future integrative microarray research. *Nat. Biotechnol.* **24**, 1322–1323 (2006).
10. Dupuy, A. & Simon, R.M. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl. Cancer Inst.* **99**, 147–157 (2007).
11. Ioannidis, J.P., Polyzos, N.P. & Trikalinos, T.A. Selective discussion and transparency in microarray research findings for cancer outcomes. *Eur. J. Cancer* **43**, 1999–2010 (2007).
12. International Committee of Medical Journal Editors. *Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication.* <http://www.icmje.org/#prepare> (2008).
13. Ioannidis, J.P. Molecular evidence-based medicine: evolution and integration of information in the genomic era. *Eur. J. Clin. Invest.* **37**, 340–349 (2007).
14. Ingraham, C.R. *et al.* Abnormal skin, limb and craniofacial morphogenesis in mice deficient for interferon regulatory factor 6 (Irf6). *Nat. Genet.* **38**, 1335–1340 (2006).
15. Carroll, J.S. *et al.* Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.* **38**, 1289–1297 (2006).
16. Dierick, H.A. & Greenspan, R.J. Molecular analysis of flies selected for aggressive behavior. *Nat. Genet.* **38**, 1023–1031 (2006).
17. Pickersgill, H. *et al.* Characterization of the *Drosophila melanogaster* genome at the nuclear lamina. *Nat. Genet.* **38**, 1005–1014 (2006).
18. Tirosh, I., Weinberger, A., Carmi, M. & Barkai, N. A genetic signature of interspecies variations in gene expression. *Nat. Genet.* **38**, 830–834 (2006).
19. Loh, Y.H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38**, 431–440 (2006).
20. Malek, R.L. *et al.* Physiogenomic resources for rat models of heart, lung and blood disorders. *Nat. Genet.* **38**, 234–239 (2006).
21. Mehrabian, M. *et al.* Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nat. Genet.* **37**, 1224–1233 (2005).
22. Mito, Y., Henikoff, J.G. & Henikoff, S. Genome-scale profiling of histone H3.3 replacement patterns. *Nat. Genet.* **37**, 1090–1097 (2005).
23. Gupta, P.B. *et al.* The melanocyte differentiation program predisposes to metastasis after neoplastic transformation. *Nat. Genet.* **37**, 1047–1054 (2005).
24. Frey, B.J. *et al.* Genome-wide analysis of mouse transcripts using exon microarrays and factor graphs. *Nat. Genet.* **37**, 991–996 (2005).
25. Ule, J. *et al.* Nova regulates brain-specific splicing to shape the synapse. *Nat. Genet.* **37**, 844–852 (2005).
26. Schadt, E.E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37**, 710–717 (2005).
27. Denver, D.R. *et al.* The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans. Nat. Genet.* **37**, 544–548 (2005).
28. Van Driessche, N. *et al.* Epistasis analysis with global transcriptional phenotypes. *Nat. Genet.* **37**, 471–477 (2005).
29. Schmid, M. *et al.* A gene expression map of *Arabidopsis* thaliana development. *Nat. Genet.* **37**, 501–506 (2005).
30. Tanaka, H., Bergstrom, D.A., Yao, M.C. & Tapscott, S.J. Widespread and nonrandom distribution of DNA palindromes in cancer cells provides a structural platform for subsequent gene amplification. *Nat. Genet.* **37**, 320–327 (2005).
31. Roepman, P. *et al.* An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. *Nat. Genet.* **37**, 182–186 (2005).
32. Sweet-Cordero, A. *et al.* An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nat. Genet.* **37**, 48–55 (2005).
33. Oleksiak, M.F., Roach, J.L. & Crawford, D.L. Natural variation in cardiac metabolism and gene expression in Fundulus heteroclitus. *Nat. Genet.* **37**, 67–72 (2005).
34. Larkin, J.E. *et al.* Independence and reproducibility across microarray platforms. *Nat. Methods* **2**, 337–344 (2005).
35. Chen, J.J. *et al.* Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. *BMC Bioinformatics* **8**, 412 (2007).
36. Miron, M. & Nadon, R. Inferential literacy for experimental high-throughput biology. *Trends Genet.* **22**, 84–89 (2006).
37. Shields, R. MIAME, we have a problem. *Trends Genet.* **22**, 65–66 (2006).
38. Draghici, S., Khatri, P., Eklund, A.C. & Szallasi, Z. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* **22**, 101–109 (2006).
39. Piwowar, H.A., Day, R.S. & Fridsma, D.B. Sharing detailed research data is associated with increased citation rate. *PLoS ONE.* **2**, e308 (2007).
40. Brazma, A. & Parkinson, S. ArrayExpress service for reviewers/editors of DNA microarray papers. *Nat. Biotechnol.* **24**, 1321–1322 (2006).
41. Gentleman, R. Reproducible research: a bioinformatics case study. *Stat. Appl. Genet. Mol. Biol.* **4**, 2 (2005).
42. Ioannidis, J.P.A. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).