GUIDELINES

Expression profiling — best practices for data generation and interpretation in clinical trials

*The Tumor Analysis Best Practices Working Group**

Microarrays are routinely used to assess mRNA transcript levels on a genome-wide scale. As use and acceptance increases, there is intensified focus on appropriate methods of data generation and interpretation, with important questions being asked about the best data analysis methods. The development of such 'best practices' is needed, as microarrays — in particular, Affymetrix oligonucleotide arrays — are becoming increasingly important in human clinical trials, both for differential diagnosis and monitoring of pharmacological efficacy. Here, representatives from high-volume microarray core centres consider the current status of 'best practices', focusing on the broadly used Affymetrix oligonucleotide arrays.

Microarrays represent a major technological advance in molecular biology. The introduction of any such advance is typically followed by a period of optimization and standardization. The latter is a crucial part of any maturing technology, as it allows an approach in which advances are made in parallel by individual researchers and companies who contribute new knowledge based on the existing standard. Any such standards must be constantly reassessed; stale or stagnant standards can inhibit the development of the technology.

Microarray-based mRNA-expression profiling can be considered to be the first mature genome-wide analysis technology, reflected in an increased interest in using microarrays as an endpoint in clinical trials. However, regulations of clinical trials require the development of clear standards for use and interpretation of microarray data (commonly referred to as quality control and standard operating procedures (QC/SOPs) and/or 'best practices'). Guidelines for reporting and annotation of microarray data from the Microarray Gene Expression Data (MGED) Society (see online links box) — using MIAME (Minimum Information About A Microarray Experiment) standards (BOX 1) and the MAGE-ML mark-up language[1,2] — represent an important step towards this goal. The efforts of this multinational academic–industry partnership has made it possible to develop databases that can house the many types of microarray data (see below) within the same data structure, enabling some data queries between experiments and experimental platforms. The ArrayExpress microarray database[3] (see online links box) is the first major publicly accessible database that adheres to this universal data-presentation platform, and some prominent journals (such as *Nature, Cell, EMBO Journal* and *The Lancet*) now demand that published microarray data conform to the MIAME standards. In addition, microarray manufacturers, such as Affymetrix, have implemented MIAME-compliant data output in their new software releases.

The MGED Society has effectively developed data-reporting guidelines, but it has not addressed issues of data generation and interpretation. The latter are more intimately coupled to the specific experimental platform. Of the three commonly used types of microarrays (spotted cDNA, spotted oligonucleotide and Affymetrix arrays), each has distinct methodologies associated with them; accordingly, the issues of data interpretation are also different (BOX 2). These differences make it difficult or impossible to develop cross-platform guidelines for data generation and interpretation. Best practices for spotted cDNA arrays are especially problematic because the manufacture of the arrays varies considerably from place to place. In addition, all spotted arrays use co-hybridization of a test RNA sample labelled with one colour FLUOROPHORE with a control RNA labelled with a different colour to which the test is compared on the same spot. The output is in the form of a ratio of hybridization signals that is comparable to other experiments only if the same control RNA is always used. Therefore, the development of standards in spotted arrays would require all laboratories to use the same control RNA solution before data could be easily compared.

Manufactured oligonucleotide arrays (both mechanically spotted and synthesized *in situ*) have the advantages of being centrally produced under controlled conditions. Affymetrix PHOTOLITHOGRAPHY-produced arrays have been available for nearly 10 years, whereas mechanically spotted oligonucleotide arrays have only very recently begun to appear in the marketplace. For example, Agilent Technologies (see online links box) recently released 17,000 60-mer oligonucleotides printed five times each on glass slides (85,000 FEATURES). Spotted oligonucleotide arrays typically have a single spot per gene (single probe measurement), whereas Affymetrix arrays provide multiple measurements — a series of independent or semi-independent oligonucleotides query each RNA in solution (the probe set) (BOX 2). Affymetrix probe sets are constructed from a series of perfect-match and paired-mismatch oligonucleotides, allowing some assessment of non-specific binding and performance of the probes. Overall, the Affymetrix probe sets provide a variety of measurements that allow robust measures of gene expression. The use of multiple perfect-match and mismatch probes for each gene enables the development of different methods of interpreting the hybridization patterns across the probe set and calculating a single 'expression level' or 'signal' that reflect the gene's relative expression level. A number of probe-set interpretation algorithms for Affymetrix arrays are available (see below for discussion).

> "[distinct methodologies] make it difficult or impossible to develop cross-platform guidelines for data generation and interpretation."

The increasing use of Affymetrix microarrays, and the emergence of this technology as an endpoint in clinical trials, has led to requests to develop, in both the pharmaceutical and academic research communities, best practices in data generation and analysis. Given the many differences between spotted cDNA, spotted oligonucleotide and Affymetrix arrays, the best practices need to be developed separately for each experimental platform; this is in contrast to data reporting that can be standardized across all platforms (BOXES 1 and 2). The Tumor Analysis Best

Box 1 | The MIAME guidelines for data reporting

The Microarray Gene Expression Data Society (MGED) is an international discussion group of microarray experts, with the primary goal of developing methods for data sharing between experimental platforms. The main output of this group has been the Minimum Information About A Microarray Experiment (MIAME) guidelines for microarray data annotation and reporting. The guidelines have been adopted by a number of scientific journals and have recently been endorsed for use by the US Food and Drug Administration and the US Department of Agriculture for pharmacogenomics projects.

The MIAME guidelines include descriptions of experimental design (number of replicates, nature of biological variables), samples used, extract preparation and labelling, hybridization procedures and parameters, and measurement data and specifications. These guidelines have been most important for the spotted cDNA and oligonucleotide experimental platforms (see BOX 2) in which the flexibility in microarray design and utilization also leads to considerable variation in array data generation and reporting between different laboratories. The guidelines do not attempt to dictate how experiments should be done, but rather provide adequate information associated with any published or publicly available experiment so that the experiment can be reproduced.

Box 2 | Microarray experimental platforms

There are three different types of microarray in common use: spotted cDNAs, spotted oligonucleotides and Affymetrix arrays.

Spotted cDNA arrays

Spotted cDNA arrays typically use sets of plasmids of specific cDNAs in gridded liquid aliquots. The inserts of each clone are typically amplified by PCR, and a few picolitres are physically spotted onto glass slides by liquid-handling robots. Robotic spotters can spot 100,000 spots per slide, and duplicate sets of clones are often spotted. The advantages of spotted cDNA arrays are that the content of each microarray is determined by the researcher, with complete flexibility in number and type of cDNA clones spotted. Also, the cost per array is relatively low, as the clone sets are a PCR-renewable resource and the glass slides are themselves inexpensive. The amount of the RNA that corresponds to each spot is determined relative to a second control RNA solution that is hybridized to the same spot, and a ratio is obtained.

Disadvantages of spotted cDNA arrays include the variable amount of DNA spotted in each spot, the 10–20% 'drop out' rate of failed PCR reactions or failed spots and mis-identification of clones (that is, the spot is not what you think it is). Also, there is no control over the actual sequence of the clone. As many gene-coding sequences contain regions of sequence that are shared with other genes, there are questions of specificity of the hybridization to the relatively large cDNA inserts. Spotted cDNA arrays were embraced by most academic centres, owing to their flexibility and relatively low cost.

Spotted oligonucleotide arrays

These arrays are also built by liquid handling on glass slides; however, the input solution is a synthetic oligonucleotide (often 60–70-mers). The resulting spotted material is typically of known concentration, of known sequence and is single stranded (all advantages relative to spotted cDNAs). Most of the process can be automated, leading to less sample mix-up and less drop-out of samples.

Disadvantages of spotted oligonucleotides include the relatively high cost of synthesizing large numbers of large oligonucleotides and the non-renewable nature of the resource. Spotted oligonucleotide arrays are becoming increasingly available.

Affymetrix GeneChips

These microarrays are factory designed and synthesized. Design is done using software to choose a series of 11 25-mer probes from the 3′ end of each transcript or predicted transcript in the genome; each of the 11 probes is then paired with a similar mismatch probe that is designed to contain a mutation in the centre. The latter serves as a form of control for hybridization specificity. Synthesis of arrays is done using light-activated chemistry and photolithography methods, and feature size can be reduced to approximately 8 μm^2, with about 1 million probes in a 1.2 cm^2 glass area. Probe-set algorithms interpret the signals from each 22-oligonucleotide probe set, and derive a single value (signal) from the patterns of hybridization to the 22 individual probes. This signal is then normalized to the entire microarray, or to the probe sets across an entire project.

For a more general discussion of normalization and analysis methods of different microarray platforms, the reader is referred to the excellent web information resource of the MGED group (see The MGED Data Transformation and Normalization Working Group in online links box).

Practices Working Group (see BOX 3) was convened to discuss and develop best practices for Affymetrix microarrays, including QC and SOPs for both data generation and data analyses. The first meeting was held in Santa Clara in March 2003, followed by a series of conference calls that focused on discussions of data generation and analysis standards for the Affymetrix oligonucleotide arrays. The Working Group deliberately focused on a platform that has widespread usage and is most likely to be used in clinical trials owing to the previously standardized manufacturing process. Here, we discuss recommendations for experimental design, probe-set analysis algorithms, signal/noise assessments and biostatistical methods.

Experimental design

Appropriate experimental design is a key aspect of all science, and microarray studies are no exception. The relatively high cost of some commercial microarray platforms is a frequently cited reason for suboptimal experimental design, especially with regards to the number of replicates. Data interpretation is inevitably compromised when replicates are decreased.

Replication in cross-sectional studies. The appropriate number of microarray replicates for any particular condition or time point depends on the source of biological variability in the study samples. Inter-individual variability is very large in outbred (genetically heterogeneous) humans, but is very small within inbred mouse strains. For example, expression profiles derived from muscle from different mice are not more variable than from muscles isolated from one mouse[4]. Defining the confounding variables that contribute to experimental variability, such as intra-subject, inter-subject, inter-group and technical variation (microarray protocol), is needed to design and statistically power a study, and to determine the number of replicates that are needed. In general, inbred mice require testing only three or four mice per group. We and others have found that five or six out-bred rats per group provide statistically robust results[5,6]. By contrast, human samples require considerably more individuals per group. Key variables in human samples include tissue heterogeneity, stage of disease and inter-individual variation, all of which have been found to be major confounding variables[7].

Replication in longitudinal studies. It has long been recognized that, in human clinical trials, LONGITUDINAL DESIGNS provide considerably greater power at lower numbers of replicates.

They best control for inter-individual variation because each subject serves as their own control. Serial blood sampling from single subjects is the least invasive[8] (see below for further discussion), and, for example, cancer patients are often longitudinally sampled[9]. Serial biopsies of other tissues are more invasive: however, a number of serial human muscle biopsy studies of healthy volunteers after different types of exercise training have begun to appear in the literature[10,11].

Expression profiling of blood samples (longitudinal or CROSS-SECTIONAL DESIGN) is the protocol that is most likely to be used in human clinical trials. One of the Working Group's goals was to establish SOPs for blood sample collection and RNA isolation in clinical trials. A specific follow-up report of these recommendations will be published elsewhere. Such a protocol must be easily adaptable to multiple trial sites, with relatively little need for resident expertise to carry out the isolation protocol. So far, standard methods for isolating peripheral-blood mononuclear cells have shown the most reproducibility, although others are being tested (see Affymetrix Technical Note in online links box). Cells isolated soon after collection can be flash frozen for storage and subsequent RNA isolation or an RNA stabilizing compound can be added if the samples need to be transported.

Tissue/cell heterogeneity. Tissue heterogeneity is a major confounding variable in most microarray experiments. In inbred mice, tissue heterogeneity is typically normalized by using whole organs. This is rarely possible in human experiments, and particularly not in clinical trials; the limited amount of human tissue that is available exacerbates heterogeneity. The mixed cell populations of peripheral blood can be thought of as a tissue heterogeneity problem similar to that encountered in all solid tissue and tumour biopsies. Indeed, a recent study showed that variation as a result of tissue variability in human muscle biopsies often exceeded inter-individual variability[12]. One potential solution to the tissue heterogeneity problem lies in bioinformatic methods. If computer software can be trained to recognize the expression profiles of each individual cell type within a mixed tissue sample, then it should be possible to subtract them from each other and to renormalize to obtain a set of cell-specific expression profiles derived from a single mixed profile. This will be most easily done on tumour biopsies, in which the main cells of interest are tumour versus contaminating normal tissue. Although there are no published examples so far, such methods are maturing rapidly.

Box 3 | **The Tumor Analysis Best Practices Working Group***

The Tumor Analysis Best Practices Working Group is a group of investigators who study the best practices of tumour analysis in humans taking part in clinical trials. The following authors are members of the Group:

- **Eric P. Hoffman is at the Research Center for Genetic Medicine, Children's National Medical Center, Washington DC 20010, USA. email: ehoffman@cnmcresearch.org**
- **Tarif Awad, John Palma, Teresa Webster, Earl Hubbell and Janet A. Warrington are at Affymetrix, Santa Clara, California 95051, USA. emails: tarif_awad@affymetrix.com; john_palma@affymetrix.com; teresa_webster@affymetrix.com; earl_hubbell@affymetrix.com; janet_warrington@affymetrix.com**
- **Avrum Spira is at The Pulmonary Center, Boston University Medical Center and the Bioinformatics Program, Boston University, Boston, Massachusetts 02118, USA. e-mail: aspira@lung.bumc.bu.edu**
- **George Wright is at the Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institute of Health, Bethesda, Maryland 20892, USA. e-mail: wrightge@mail.nih.gov**
- **Jonathan Buckley and Tim Triche are at the Children's Hospital, University of California, Los Angeles, California 90089, USA. e-mail: buckley@hsc.usc.edu; triche@hsc.usc.edu**
- **Ron Davis, Robert Tibshirani and Wenzhong Xiao are at Stanford University, Palo Alto, California 94303, USA. e-mails: dbowe@stanford.edu; tibs@stat.stanford.edu; wzxiao@pmgm2.stanford.edu**
- **Wendell Jones is at Expression Analysis Inc., Durham, North Carolina 27713, USA. e-mail: wjones@expressionanalysis.com**
- **Ron Tompkins is at Harvard University, Boston, Massachusetts 02115, USA. e-mail: rtompkins@partners.org**
- **Mike West is at the Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina 27708, USA. e-mail: mw@stat.duke.edu**

An experimental alternative to mitigate confounding tissue heterogeneity is to isolate pure cell populations for expression profiling. Many such methods are well developed in the research laboratory, including FLUORESCENCE-ACTIVATED CELL SORTING (FACS)[13], NEGATIVE CELL ISOLATIONS from blood (for example, Stem Cell Technologies RosetteSep)[14] and LASER CAPTURE MICRODISSECTION[15]. To research scientists, the profiles that are derived from isolated cell types are a more intuitive approach to define biologically relevant pathways. However, it should be noted that uses of array-based analysis of gene expression approved by the US Food and Drug Administration (FDA) will probably focus on reproducibility and robustness (as well as on predictive accuracy), rather than on biological interpretation or justification. The high-tech methods used to isolate specific cell types from clinical samples are unlikely to make their way into clinical trials unless tissues are procured in a highly centralized way.

"If computer software can be trained to recognize the expression profiles of each individual cell type within a mixed tissue sample, then it should be possible to subtract them from each other…"

Procedural variation. Beyond the usual issues of sampling and accrual, gene-expression data will be subject to many additional sources of error. For example, the surgical removal and processing of tumour tissue can vary considerably from site to site. Laboratory QC procedures in tissue handling, RNA extraction and processing, and variations in protocols for data management and processing will need to be addressed in any clinical trial design. In particular, prolonged tissue ISCHAEMIA prior to processing of surgically RESECTED tissue can significantly alter gene expression[16]. All tissue samples should be flash frozen within minutes of surgery and stored at –80°C or below. Samples should also be kept in small, airtight containers and kept from drying out during frozen storage by placing fragments of ice in with the sample.

Technical variability

The standard laboratory protocol for generating RNA profiles using Affymetrix microarrays involves a series of steps (FIG. 1).

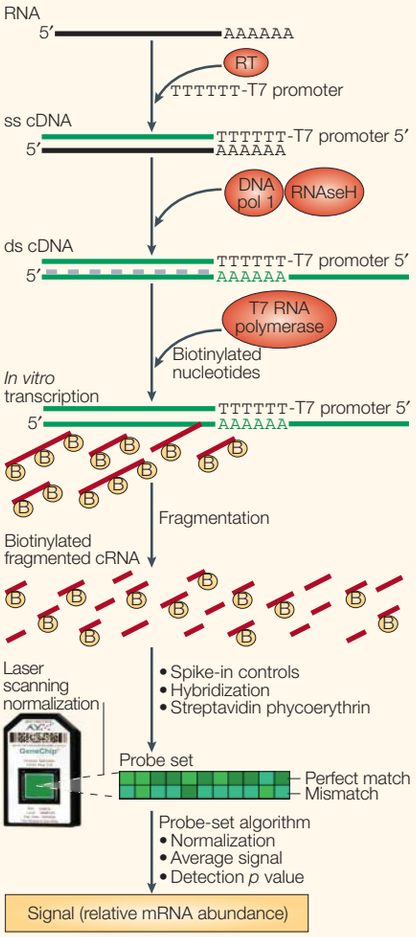


Figure 1 | **Sample processing and microarray interpretation of Affymetrix GeneChips.** Flash-frozen tissue (~50 mg) is homogenized to isolate total RNA. Single-stranded (ss) cDNA and then double-stranded (ds) cDNA is made from ~5 μg of total RNA. Double-stranded cDNA contains a T7 RNA-polymerase promoter adjacent to the 3′ polyA tail of each transcript. It is transcribed *in vitro* to generate more than 400 biotinylated cRNA molecules for each ds cDNA molecule. The biotinylated cRNA is fragmented and hybridized to the microarrays. Each transcript is queried by one or more probe sets of 11 perfect-match and 11 paired-mismatch oligonucleotides (the latter contain a centrally located point mutation as a form of hybridization specificity control). Currently available Affymetrix microarrays have ~54,000 probe sets on each 1.28 cm^2 glass microarray (~1.2 million 25-mer oligonucleotides on the HG-U133Plus 2.0 array). The biotinylated cRNA fragments hybridize to the appropriate oligonucleotide features. A laser scanner determines the amount of bound biotinylated cRNA indirectly through the streptavidin-conjugated phycoerythrin fluorescence at each feature within a probe set. The component probe pairs are interpreted and averaged to arrive at a single signal that reflects the relative abundance of the original mRNA. Probe sets are interpreted by any one of a number of probe-set algorithms, each providing a signal that reflects the relative hybridization intensity across the probe set. RT, reverse transcriptase.

RNA isolation. RNA quality and quantity is crucial to the success and reproducibility of the expression profiles. RNA quantity and quality is generally checked by complementary methods: UV 260/280 ratio >1.8, agarose gel electrophoresis or an Agilent Bioanalyzer to visualize clear 18S and 28S ribosomal RNA bands. Total RNA (5–10 μg) is input into the cDNA/cRNA reaction, with an expected corrected yield of biotinylated cRNA of between 4- and 10-fold greater than the total RNA input (so 5 μg of total RNA must yield at least 20 μg of biotinylated cRNA, or the sample is discarded). The biotinylated cRNA should be 500–3,000 base pairs (bp) in size. After fragmentation, the cRNA should be 50–200 bp. The Working Group recommends that samples that do not meet these criteria should be discarded.

If RNA amount is limiting — as is the case, for example, with laser capture microscopy samples, flow-sorted cell samples or small tissue samples — a two-round amplification protocol can be used. For example, 200 ng of total RNA is processed for *in vitro* transcription (IVT), with the same goal of 4–10-fold amplification (>800 ng of cRNA output). One hundred nanograms of this cRNA is then reverse transcribed into cDNA using random primers, after which a second IVT is done. The second round IVT should result in a 400-fold amplification.

Microarray controls. Hybridization controls include visualization of the image so that any abnormalities in hybridization patterns can be detected. ProbeProfiler from Corimbia Inc. is a program with extended capabilities for detecting defects in microarray manufacture. Affymetrix MAS 5.0 software adjusts the microarray-scanned image to a common target intensity by using a scaling factor. In addition, a general index of chip background and noise is represented by the percentage of 'present calls' (probe sets for which the hybridization to the perfect-match probes is significantly higher than mismatch hybridization). The Working Group believes that both the scaling factor and the percentage of present calls are important QC criteria. Considering MAS 5.0 chip analyses, the scaling factors to normalize chips within a project should lie within two standard deviations of the mean, with present calls being greater than 25% (BOX 4). The percentage of present calls is often lower when B or C arrays that contain higher proportions of more poorly characterized transcript units (expressed sequence tags or computer-predicted open reading frames) are used. The percentage of present calls across a set of samples should be consistent, within a range of 10%. Some software packages allow the identification of statistical 'outlier' microarrays in a group of microarrays in a given project, which additionally enables the experimenter to flag and exclude specific microarrays that are not acceptable for an analysis. In addition to these criteria, acceptable hybridizations must have adequately intact input RNA as shown by 3′ to 5′ ratios of hybridization within probe sets. A typical control is the glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) gene, which should have 3′ to 5′ ratios of less than 3 (BOX 4).

The QC criteria provided above are based on MAS 5.0 probe-set algorithms and data analyses. The measures of present calls and scaling factors are useful and serve as initial summary measures of the performance of a particular microarray. However, more focused statistical methods, coupled with routine visual inspection of images, hold promise for the continuing improvement of data quality and screening abilities.

Large-scale analyses of microarray data across laboratories have not yet been reported. However, the Working Group feels that adherence to the above QC criteria, using standard RNA isolation and processing methods, should yield data that are consistent between laboratories and intrinsically comparable. The same set of criteria can also be used as best practices for data generation in the design and conduct of clinical trials.

Standard clinical laboratory practice is to develop programmes for submission of known samples to different laboratories and assessment of comparability of results. Such programmes are under development within larger collaborative efforts, such as the National Heart, Lung and Blood Institute (NHLBI) Programs in Genomic Applications (see the HOPGENE Program for Genomic Applications in online links box) and the National Institute of General Medical Sciences (NIGMS) Glue Grant (see online links box).

Data analysis and interpretation

Signal generation versus statistical analyses. Two relatively distinct steps underlie all data analyses of Affymetrix oligonucleotide microarrays: the development of a normalized 'signal' for each transcript on each microarray and the subsequent statistical analysis of differences in signals between different arrays. The first step involves probe-set algorithms that use all, or part, of the component signals within a probe set and then derive a single signal that is representative of the relative abundance of each mRNA queried in each array. The second step is the

application of bioinformatic and statistical methods to identify interesting subsets of the assembled data of all arrays within a project. There is considerable debate about the best methods for both of these steps (see below for a discussion). Although the two steps are separable, it is clear that they have a marked influence on each other. It is in this realm that the bioinformatics of microarrays becomes avant-garde, and with the groundbreaking nature of research comes considerable debate as to what is appropriate in any specific situation.

Before discussing the different methods for probe-set analysis and data interpretation, it is important to point out that much of the debate in the field of bioinformatics about microarray interpretation revolves around signal/noise ratios. A common assumption is that signal/noise ratios across a microarray are homogeneous, or at least similar in magnitude. This might be true for general background hybridization, but not for the performance of probe sets. In any particular microarray, there are probe sets that give very strong and clear hybridization patterns and those that perform poorly. Many of the best performing probe sets (those with a highly significant probe-set detection *p* value) reflect highly expressed transcripts with no closely related sequences that might cross-hybridize. Low-level transcripts, or transcripts that belong to gene families with highly homologous sequences derived from distinct genes, often have corresponding probe sets that do not perform as well and might have a significant, if not overwhelming, noise component. The signal from such probe sets is difficult to interpret, and data interpretation can be limited to only the best performing probe sets, although arguably the most interesting data comes from the genes that are expressed at low levels but that still show significant differences between samples.

Determining adequate sensitivity of the signals and signal/noise responses relative to the absolute quantity of mRNA in clinical samples is crucial as microarrays become a component of clinical trials and diagnostic models. Affymetrix arrays provide a concentration of each mRNA queried relative to the genome-wide mRNA profile of the sample; it is assumed that the global mRNA content of a tissue as a whole does not change significantly, making relative mRNA quantification an accurate reflection of the response of the individual gene. This method differs from absolute quantification of specific mRNAs (such as S1 NUCLEASE PROTECTION and REAL-TIME PCR), or the isolated transcript ratio determined by co-hybridization of two samples to spotted cDNA or oligonucleotide arrays (BOX 2). Affymetrix arrays achieve considerable sensitivity through the inherent redundancy of the probe set; however, the Working Group acknowledged that some genes, such as some cytokines that are functional at very low expression levels, are probably below the limit of detection.

The Working Group agreed that each project will have its own signal/noise optimum, and analysis methods that prove best for one project might prove unsuitable for another. Ideally, a signal/noise ratio should be optimized for each project or trial using different probe-set algorithms and data-filtering methods, and some systematic efforts towards this end are beginning to appear in the literature[17].

> "… adherence to the above QC criteria … should yield data that are consistent between laboratories and intrinsically comparable."

After a signal is derived for each probe set, data is interpreted using statistical and visualization methods. All statistical methods run into two generic problems when faced with microarray data that are inter-related. The first is the curse of dimensionality — each gene is potentially related to every other gene, so all permutations of all available data must be considered, leading to an exponentially increasing number of possible associations in multidimensional space. The problem arises when associations (samples) become lost as the dimensionality increases — associations lose their local value and become generically global in statistical terms. Statistical models attempt to circumvent this curse by requesting larger and larger sample sizes, but fulfilling the requests becomes functionally impossible for the experimentalist. There is no easy answer to these problems and they remain a challenge for future bioinformatics research that uses microarrays[18].

Derivation of signal: probe-set algorithms and normalizations. One of the key advantages of the Affymetrix platform is the multiple measurements that are intrinsic to the probe set — most probes include 11 perfect-match and 11 paired-mismatch 25-bp oligonucleotides per gene (FIG. 1). Previous versions of GeneChip arrays used probe-set design methods that led to considerable overlap between probes, so that hybridizations to each feature/probe were not independent measurements; this led to considerable uncontrolled weighting of the contribution of any particular region of sequence to the resulting signal. All recent chips use a much more refined probe-set design with less overlap and considerably better performance of the probe set. Improvements in array and probe-set designs have been accompanied by an evolution in primary analysis algorithms and the supporting software provided by Affymetrix for data analysis and interpretation[19]. Affymetrix default algorithms are based on well-documented statistical methods, namely the robust TUKEY'S BI-WEIGHT ESTIMATOR and WILCOXON'S SIGNED RANK, to calculate the final probe-set signals and associated *p* values, respectively[19,20].

Box 4 | Quality Control metrics for Affymetrix microarrays

RNA quality
Optical density 260/280 of 1.8–2.1 | Agilent Bio-Analyzer | Gel electrophoresis

cDNA/cRNA efficiency
>4-fold amplification from total RNA | 500–3,000 bp prior to fragmentation | 50–200 bp after fragmentation

Chip hybridization
Image inspection for defects | Scaling factors within two standard deviations within a project | MAS 5.0 present calls >25% for the A-SERIES ARRAYS, including the HG-U133Plus 2.0 array | Percentage present calls for the B- AND C-SERIES ARRAYS are typically lower | 3′/5′ *GAPDH* ratios <3

Project normalization
The detection of statistical outliers for chips, probe sets or individual probe pairs requires normalization and analysis across an entire project. This is afforded by the dCHIP and ProbeProfiler, and other software packages. Data-analysis packages that rely on intra-chip normalization and scaling typically do not enable detection of statistical outliers.

Chip outliers
Probe-set outliers | Probe-pair outliers | Range in present calls <10%

Table 1 | **Comparisons of probe-set analysis algorithms**

Algorithm	Penalty for mismatch signal	Normalization method	Outlier detection and correction	Sensitivity*	Specificity‡
Affymetrix MAS 5.0	High	Individual chips	Little	Good	Excellent
dCHIP difference model	High	Cross-project	Moderate	Good	Excellent
dCHIP	None	Cross-project	Moderate	Excellent	Good
RMA	None	Cross-project	Moderate	Excellent	Good
ProbeProfiler	Moderate	Extensive	Extensive	Good	Good

*Sensitivity is based primarily on ROC (receiver operating characteristic) curves of spike-in mRNA data based on published reports (see http://www.bioconductor.org)[21,23].
‡Specificity measurements are based both on expectations from mismatch weights and published observations in experimental data sets[17,18].

Affymetrix has announced plans to continue to improve the software components of the GeneChip platform. The upcoming release of the GeneChip Operating System (GCOS) is expected to incorporate refinements in the user interface, data management and analysis algorithms. Software tools aside, the most significant development on the analysis front is arguably the decision by Affymetrix to release previously proprietary chip-design details, such as probe sequences, chip-design parameters and file APIs (applications programming interfaces). The goal is to encourage scientists to develop innovative analysis tools that can potentially derive more biological value from GeneChip expression data. The challenge of providing a constantly growing and evolving body of information associated with arrays has been solved in part with a web-based tool. The company's NETAFFX web site (see online links box) serves as the public portal for detailed information on chip design and has become a valuable resource for biological follow-up of GeneChip expression results. Third-party software developers can find additional support, including information on file APIs, through the Affymetrix Developers' Network (see online links box).

Encouraged in part by the openness of the platform and spawned by an increase in knowledge and experience in array data analysis, scientists are developing a number of alternative algorithms for probe-set analysis, with the goal to derive the best signal that is representative of the mRNA level for each gene. As each signal is relative to other signals in the experiment (both between arrays for the same gene and relative to all other genes on the array), the process of normalization is intimately tied to derivation of signal. The more commonly used alternative probe-set analysis algorithms include dCHIP[20], RMA[21] and ProbeProfiler (TABLE 1).

It is outside the scope of this article to discuss the nature of the different probe-set interpretation and normalization algorithms in depth, and the reader is referred elsewhere[22]. The algorithms differ in a number of important ways (TABLE 1). First, the PENALTY WEIGHT that is assigned to the mismatch probe varies — MAS 5.0 assigns a relatively heavy penalty for cross-hybridization to the mismatch probe, RMA assigns no weight and dCHIP gives the choice of providing weight or no weight. Second, the ability to discard outlier signals varies from package to package, with dCHIP and ProbeProfiler having refined methods to detect outliers at each level of analysis (probe, probe set and microarray). These packages are able to replace deviant probes with expected data based on the remainder of the probe set, and/or flag abnormal probe sets and arrays for possible exclusion from further data analysis. Third, the method of normalization varies from within a single array (MAS 5.0) to a project-based normalization (dCHIP, RMA and ProbeProfiler). Finally, MAS 5.0 provides a detection *p* value, in which a number is assigned to the confidence of the signal in question. This can be used to weight different probe-set signals in subsequent data interpretations.

"…robust feature selection for the purpose of diagnosis and molecular markers in clinical trials requires robust statistical methods…"

The output of all packages is a normalized signal (with or without an associated detection *p* value) for each probe set on each array. These signals are then fed into data interpretation packages for statistical analyses and data visualizations.

Different probe-set interpretation algorithms lead to different results. Members of the Working Group often encounter ~50% concordance in general data output in their own work between comparisons of two different algorithms. However, it is crucial to note that the large majority of discordant data lies in regions of relatively poor signal/noise ratios, and concordance deteriorates in experiments with high levels of confounding noise. In general, the programmes that put less weight on the mismatch show better sensitivity (linearity) when signals are noisy (TABLE 1). However, this increased sensitivity can come at a cost of substantial contaminating noise[17].

The Working Group recommends using at least two probe-set algorithms for comparison and prioritization of gene selection (for example, MAS 5.0 and the dCHIP difference model).

Data interpretation. Most published microarray papers could be considered data-poor in terms of replicates and systematic statistical analyses, but data-rich both in terms of amount of high-quality data generated and significant research findings. Below, we point out the most appropriate current bioinformatics methods and additional methods that require further development so that data can be more fully mined for information content.

A second general backdrop to the following discussion is that data visualization is one of the most powerful data interpretation tools, yet it rarely obeys statistical principles. The resolution of the human eye, coupled with the abstract computational power of the human brain, lies behind the popularity of hierarchical clustering and other non-statistical principles and visualization methods. However, the eye and brain are poorly suited to spontaneously deriving statistical support.

There are two general types of experimental design that lend themselves to different types of statistical and visual analysis: the cross-sectional study and the TIME-SERIES STUDY. The cross-sectional study typically has gene or pattern selection as the goal: the identification of one or more genes or patterns of expression that are diagnostic of the condition or state

under study. This 'gene selection' might be for truly diagnostic purposes (for example, differential diagnosis of leukaemia), or might be intended to identify relevant biochemical pathways. In both cases, the gene or pattern selection must be robust, usually implying a statistically principled approach, with subsequent validation by predictive computer modelling (internal cross-validation) or, preferably, prospective validation on new data.

Feature selection can be the main limiting factor in evaluation of the predictive performance of an analysis method when there are many predictors to select from. This was a 'mantra' for some of the senior statisticians involved in predictive modelling with gene-expression array data for several years, but only now do the non-statistical users and developers of predictive models from non-statistical perspectives begin to appreciate these issues. Proper validation of any model or algorithm that relies on explicit feature selection — such as choosing a subset of 70 genes from 20,000 — that underlies the resulting prediction simply must ensure that the analysis is tested by internal cross-validation that includes feature re-selection as part of the validation[23,24]. The Working Group acknowledged that prospective validation of any findings using new data is the acid test of predictive performance. The focus on feature or gene selection is vitally important when microarrays are used for differential diagnosis and has been best studied in cancer biopsy/tissue studies.

An increasing proportion of microarray studies focus on delineation of biochemical pathways that are modulated in response to some stimulus. In practice, these studies typically use feature selection to identify potential pathways that are involved in the response of the cells or tissues. Validation is then done on the identified biochemical pathways of interest, using mRNA (real-time PCR) or protein studies, often proving cause and effect in experimental models.

The Working Group notes that robust feature selection for the purpose of diagnosis and molecular markers in clinical trials requires robust statistical methods, as outlined below, and the burden of proof lies with statistical validation. For microarray experiments designed to delineate biochemical pathways, feature selection is used for generating a hypothesis and the burden of proof of the hypothesis lies with laboratory-based research, often at the protein level.

For feature selection, the Working Group recommends that users experiment with various statistical methods (such as standard parametric tests, nonparametric methods, false discovery rate and related methods[25], global or local shrinkage of raw signal intensities and Stanford's 'nearest shrunken centroids'[26]). Developments related to SURVIVAL DATA ANALYSIS are receiving increased attention because clinical trials will raise the need to move that way. As a corollary, analysis methods that focus on signatures of groups of genes (such as averages of clusters, Duke's metagenes[27–29] and Stanford's eigengenes[30]) seem worth stressing in predictive contexts.

Glossary

A-, B- AND C-SERIES ARRAYS
A series of human, rat and mouse Affymetrix arrays released in 2003, in which the A array contained the best-characterized genes, and B and C arrays contained less well-defined expressed sequence tags. In 2004, all probe sets have been condensed so that there is only one microarray per species that covers the entire genome.

CROSS-SECTIONAL DESIGN
The use of different subjects in an experimental and control group or groups. The statistical analysis compares the median and variation within each group relative to the other groups.

FEATURE
Typically one element (spot) on a microarray. In spotted cDNA or oligonucleotide arrays, features correspond to genes or transcripts; in Affymetrix arrays, there are typically 22 elements per probe set and often multiple probe sets per gene, so a feature might refer to a single oligonucleotide, a probe pair or a probe set, or a gene with multiple probe sets. In bioinformatics it is most often synonymous with a gene.

FLUORESCENCE-ACTIVATED CELL SORTING
(FACS). A method whereby dissociated and individual living cells are sorted, in a liquid stream, according to the intensity of fluorescence that they emit as they pass through a laser beam.

FLUOROPHORE
A small molecule, or a part of a larger molecule, that can be excited by light to emit fluorescence.

ISCHAEMIA
The loss of blood supply, and hence oxygenation, to a tissue or cells.

LASER CAPTURE MICRODISSECTION
A technique in which individual cells, or regions of tissue, are excised from a histological preparation, using specially equipped microscopes, and isolated for further study.

LONGITUDINAL DESIGN
The use of multiple samples from the same subject. With this design, each subject serves as their own control, eliminating confounding inter-individual variations at baseline; paired *t*-tests are used to interpret the data.

NEGATIVE CELL ISOLATION
The use of antibodies or other reagents to remove all unwanted cells from a mixed population of cells. In this method, the desired cells are not exposed to bound antibodies, thereby avoiding potential activation or other molecular alteration in the desired cells.

PENALTY WEIGHT
In Affymetrix arrays, hybridization to the 'mismatch' probe of a probe pair might or might not be considered as a form of measurement of noise or background, and can be factored into the signal seen with the paired 'perfect match' as a penalty weight.

PHOTOLITHOGRAPHY
The process of using light to either etch or activate regions of a surface (substrate). This method is used in microelectronics to create integrated circuits and processors.

REAL-TIME PCR
The quantification of the amount of PCR product during each cycle of a PCR reaction. The product concentration, as a function of cycle number, provides a good estimation of the relative quantity of the mRNA being tested.

RESECTION
Surgical removal of tissue, most commonly used for removing tumorous masses from surrounding tissue.

S1 NUCLEASE PROTECTION
An experimental method for determining mRNA transcript concentration in a tissue or cell RNA sample. It involves using labelled DNA probes that bind the RNA, with overhanging non-hybridized tails of the probe then being digested by the S1 nuclease. This creates a smaller labelled DNA probe that is indicative of the abundance of the mRNA being tested.

SURVIVAL DATA ANALYSIS
A battery of statistical methods applied to data when mortality is often the only, or best, measured outcome.

TIME-SERIES STUDY
The use of a series of samples taken at defined time points after a defined stimulus. In mice and rats, the samples at different time points are usually from different animals. In humans, time-series studies are necessarily longitudinal to avoid additional confounding noise.

TUKEY'S BI-WEIGHT ESTIMATOR
Many statistical tests require underlying definitions that are assumed to be valid (for example, tumour versus non-tumour), and require data that show a normal distribution. Microarray data, and the clinical information underlying the definition of samples, is often less exact, with genes or samples often performing as statistical outliers. Tukey's bi-weight estimator is one of the M-class of statistical models that is less sensitive to outliers and performs more gracefully when underlying assumptions are inexact.

WILCOXON'S SIGNED RANK
A statistical test that investigates the population median of paired differences. It is well suited for microarray work as it treats each gene as an independent variable and does not require normal distributions of the data.

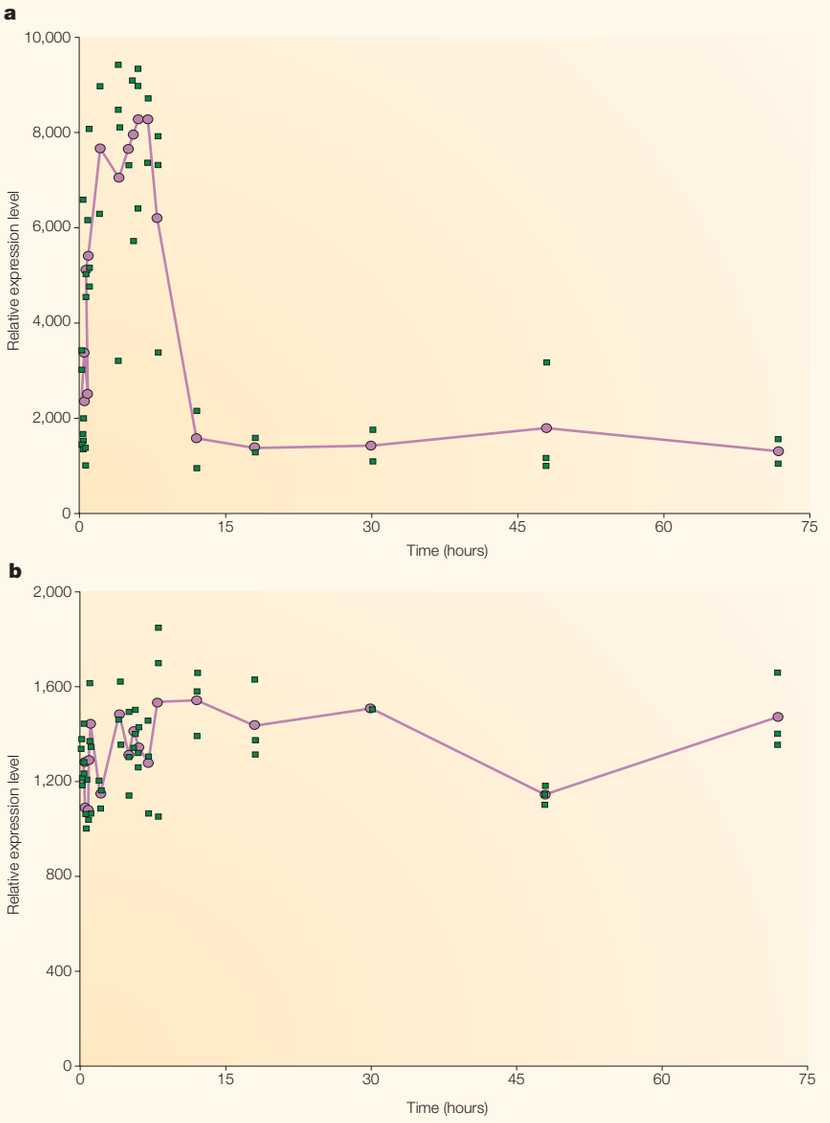


Figure 2 | **Dense time-series data with adequate replicates can provide robust visual interpretation of data.** Dynamic, single-gene queries can provide visually compelling results and avoid many issues that complicate statistical analyses of cross-sectional microarray studies. Dynamic queries of the *got1* transcript probe set from a time-series study is shown. The *x*-axis represents time (in hours) after rats were given a bolus of 3-methylprednisolone. The *y*-axis represents relative expression level. Individual rats were studied at each time point (green data points), with both liver (panel **a**) and muscle (panel **b**) tissue taken from the same animals. Averages of the replicates for each data point are shown (in magenta); the graph line is drawn between the averages. At baseline (time 0), the *got1* transcript has a normalized signal of about 2,400 in liver and 1,200 in muscle. The gene is clearly responsive to 3-methylprednisolone in liver (panel **a**), where expression rapidly increases within the first hour, plateaus between 1–7 hr, then quickly falls back to baseline by 12 hr. The replicates appear relatively consistent. By contrast, the same gene in muscle does not seem to respond to the drug; the variability in replicates is larger than any temporally-relevant change. Data from http://microarray.cnmcresearch.org/singlegenemain.asp and REFS 6,32.

Whatever the specific statistical model that is applied for prediction, using aggregate gene expression has important consequences: measures of aggregation of expression over a group of genes with related profiles can reduce dimension (thereby mitigating the curse of dimensionality). This can reduce multiplicities and, to some degree, ease the problems of gene selection, multiple testing and co-linearity, while improving signal estimation by averaging correlated noise components.

Data visualizations, time series and candidate genes. The above discussions of biostatistics all assume that the analysis is targeted towards a cross-sectional study, in which the primary goal is diagnostic gene discovery (gene or feature selection). In other words, a series of microarrays with a very large number of transcripts defines the very small minority of genes that are correlated and therefore predictive of the biological variable of interest. There are alternatives to this standard experimental design that use entirely different types of analysis, and the statistical issues are also quite different, as explained below.

The time-series study, if done with enough time points, can provide an effective antidote to the curse of dimensionality — the action of any gene during a time-series study should make biological sense, such that each signal is relatively easily discernible from noise. Visual query of a large time-series data set for single gene responses to the controlled variable either might meet expectations and is therefore valid, or might not meet expectations and is discarded as uninteresting. As an example, we show a time-series study in which rats are given a bolus of methylprednisolone, after which their liver and muscle are studied as a function of time (FIG. 2). In this case, the same gene (*got1*) is queried using a web-based dynamic visualization tool, first in liver (FIG. 2a) and then in muscle (FIG. 2b). The data in the top panel are visually compelling; *got1* in liver responds quickly and strongly to a bolus of 3-methylprednisolone, with relatively consistent replicates (each data point comes from a different animal) and a time course that is visually assuring so that complex statistical tests of the transcriptional response as a function of time are not needed. On the other hand, the same gene in muscle does not seem to respond to the drug[6,31] (FIG. 2b). Through such gene queries, the variability in replicates and the appropriateness of the action of the gene as a function of time can quickly be assessed. Another advantage of time-series data is that such profiles act as biomarkers that are amenable to analysis and interpretation using pharmacodynamic models that predict the underlying mechanisms of control of gene expression[32].

The Working Group agreed that data-rich, time-series experimental designs provide some latitude in reporting significant findings and that the query of individual genes within large data sets can circumvent complex issues of multidimensionality of data.

Future areas of development

The data-rich and highly dimensional nature of microarray data serves as a model for future dissection and understanding of biological

systems in general, including proteomics and integration of mRNA profiling and proteomics. The Working Group discussed data analysis needs within the microarray community and agreed that, along with the incorporation of QC, SOPs and optimized or customized signal/noise analyses in initial project signal generation, the back-end statistics needs to reach a commonly accepted method of dealing with the curse of dimensionality before microarrays can be reliably used in clinical trials. Statisticians need to focus more on representation of prediction results in terms of probabilities and associated measures of uncertainty, and reach a consensus on what is acceptable. In the meantime, it is likely that specific marker or diagnostic genes will be extracted from pilot profiling studies, and then only this small subset of genes will be used as a clinical trial endpoint. This data limitation approach removes much of the curse of dimensionality, but is liable to ignore the large majority of data, thereby decreasing the potential power of the study and bringing into question the use of microarrays in clinical trials.

A move towards the standardization of reporting of prediction accuracy would be desirable when assessing predictive accuracy through within-sample cross-validation. The Working Group suggests that one or more validation techniques be used when reporting predictive genes: leave-one-out and 10% cross-validation summaries, or true validation data sets. Communicating uncertainty about predictive performance is also key and will help evaluate results based on varying sample sizes. The Working Group suggests that until this information is routinely presented in published papers, it will be difficult to reach an acceptable consensus for use in clinical trials.

Conclusions

There are four key areas of optimization and standardization that are largely independent: study design, technical variability (QC/SOP of data generation), analysis method variation (signal/noise optimization using probeset algorithms and normalizations) and back-end statistical analyses. Statistics of clinical trial design is crucial: gene-expression data does not mitigate the need for sound and relevant design and analysis, nor does it challenge what we know about design. The field is quickly maturing from the small-chip-number hit-and-run type projects to those with a more robust study design. However, study design depends ultimately on appropriate powering of a study, which is greatly affected by both the chip-analysis algorithms that are used and the biostatistical data analysis.

Development of back-end statistical methods for data representation/summary and for high-level analysis remains an active area of research for both academic and commercial users, and is likely to remain so in the near future. We are some way from defining standards of summary signal intensities alone and even further from considerations of standardization of analytical methods for inference and prediction in clinical contexts. In regulated clinical studies, such standards will be enforced partly by the US FDA as submissions of medical test/device protocols emerge and increase in number. Even then, however, many approaches to data analysis and modelling will be used and developed, which is, of course, to be supported. It is very difficult to influence the research community, especially when the variety of problems that are encountered promotes the need for refined and new approaches.

*Members of The Tumor Analysis Best Practices Working Group are listed in Box 3.
Correspondence to Eric P. Hoffman at the Center for Genetic Medicine, Children's National Medical Center, 111 Michigan Avenue NW, Washington DC 20010, USA.
e-mail: ehoffman@cnmcresearch.org

doi:10.1038/nrg1297

1. Brazma, A. *et al*. Minimum information about a microarray experiment (MIAME) — toward standards for microarray data. *Nature Genet.* **29**, 365–371 (2001).
2. Spellman, P. T. *et al*. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3**, RESEARCH0046.1-0046.9 (2002).
3. Brazma, A. *et al*. ArrayExpress — a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**, 68–71 (2003).
4. Zhao, P., Iezzi, S., Sartorelli, V., Dressman, D. & Hoffman, E. P. Slug is downstream of myoD: identification of novel pathway members via temporal expression profiling. *J. Biol. Chem.* **277**, 20091–20101 (2002).
5. Di Giovanni, S. *et al*. Gene profiling in spinal cord injury shows role of cell cycle in neuronal death. *Ann. Neurol.* **53**, 454–468 (2003).
6. Jin, J. Y., Almon, R. R., DuBois, D. C. & Jusko, W. J. Modeling of corticosteroid pharmacogenomics in rat liver using gene microarrays. *J. Pharmacol. Exp. Ther.* **307**, 93–109 (2003).
7. Bakay, M. *et al*. Sources of variability and effect of experimental approach on expression profiling data interpretation. *BMC Bioinformat.* **3**, 4–15 (2002).
8. DePrimo, S. E. *et al*. Expression profiling of blood samples from an SU5416 Phase III metastatic colorectal cancer clinical trial: a novel strategy for biomarker identification. *BMC Cancer* **3**, 3 (2003).
9. de Vos, S. *et al*. Gene expression profile of serial samples of transformed B-cell lymphomas. *Lab. Invest.* **83**, 271–285 (2003).
10. Hittel, D. S., Kraus, W. E. & Hoffman, E. P. Skeletal muscle dictates the fibrinolytic state after exercise training in overweight men with characteristics of metabolic syndrome. *J. Physiol.* **548**, 401–410 (2003).
11. Zambon, A. C. *et al*. Time- and exercise-dependent gene regulation in human skeletal muscle. *Genome Biol.* **4**, R61 (2003).
12. Bakay, M. *et al*. Sources of variability and effect of experimental approach on expression profiling data interpretation. *BMC Bioinformat.* **3**, 4–15 (2002).
13. Cardozo, A. K. *et al*. Gene microarray study corroborates proteomic findings in rodent islet cells. *J. Proteome Res.* **2**, 553–555 (2003).
14. Chun, T. W. *et al*. Gene expression and viral prodution in latently infected, resting CD4+T cells in viremic versus aviremic HIV-infected individuals. *Proc. Natl Acad. Sci. USA* **100**, 1908–1913 (2003).
15. Kamme, F. *et al*. Single-cell microarray analysis in hippocampus CA1: demonstration and validation of cellular heterogeneity. *J. Neurosci.* **23**, 3607–3615 (2003).
16. Huang, J. *et al*. Effects of ischemia on gene expression. *J. Surg. Res.* **99**, 222–227 (2001).
17. Seo, J. *et al*. Interactive color mosaic and dendrogram displays for signal/noise optimization in microarray data analysis. *IEEE ICME* **3**, 461–462 (2003).
18. Somorjai, R. L., Dolenko, B. & Baumgartner, R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics* **19**, 1484–1491 (2003).
19. Mei, R. *et al*. Probe selection for high-density oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **100**, 11237–11242 (2003).
20. Li, C. & Hung Wong, W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* **2**, RESEARCH0032 (2001).
21. Irizarry, R. A. *et al*. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).
22. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
23. Ambroise, C. & McLachlan, G. J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA* **99**, 6562–6566 (2002).
24. West, M. *et al*. Predicting the clinical status of human breast cancer utilizing gene expression profiles. *Proc. Natl Acad. Sci. USA* **98**, 11462–11467 (2001).
25. Tusher, V., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5124 (2001).
26. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA* **99**, 6567–6572 (2002).
27. Huang, E. *et al*. Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nature Genet.* **34**, 226–230 (2003).
28. Black, E. P. *et al*. Distinct gene expression phenotypes of cells lacking Rb and Rb family members. *Cancer Res.* **63**, 3716–3723 (2003).
29. Huang, E. *et al*. Gene expression predictors of breast cancer outcomes. *Lancet* **361**, 1590–1596 (2003).
30. Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA* **97**, 10101–10106 (2000).
31. Chen, J. *et al*. The PEPR GeneChip data warehouse and implementation of a dynamic time series query tool (SGQT) with graphical interface. *Nucleic Acids Res.* **32**, D578–D581 (2004).
32. Almon *et al*. *In vivo* multitissue corticosteroid microarray time series available online at Public Expression Profile Resource (PEPR). *Pharmacogenomics* **4**, 791–799 (2003).

Acknowledgements

The authors thank their respective funding agencies, particularly the larger collaborative funding initiatives that make systematic and large-scale studies of the bioinformatics and biostatistics of genome-wide data sets possible from the Department of Defense and the Doris Duke Charitable Foundation CSDA. The authors also thank S. Hilmer, A. DeBiase and G. Miyada for their critique of the manuscript.

Competing interests statement

Some of the authors declare competing financial interests: see Web version for details.

Online links

DATABASES

The following terms in this article are linked online to:
LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink
GAPDH | *got1*

FURTHER INFORMATION

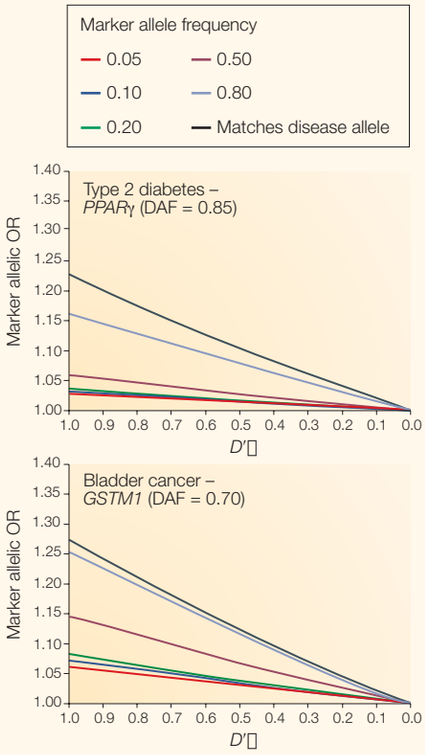
Affymetrix Developers' Network:
http://www.affymetrix.com/support/developer/index.affx
Affymetrix Technical Note:
http://www.affymetrix.com/support/technical/technotes/blood_technote.pdf
Agilent Technologies: http://www.chem.agilent.com
ArrayExpress microarray database:
http://www.ebi.ac.uk/arrayexpress
The Children's National Medical Center Microarray Center:
http://microarray.cnmcresearch.org/singlegenemain.asp
HOPGENE Program for Genomic Applications:
www.hopkins-genomics.org
MGED Data Transformation and Normalization Working Group: http://www.dnachip.org/mged/normalization.html
MGED Society: http://www.mged.org
NETAFFX web site:
http://www.affymetrix.com/analysis/index.affx
NIGMS Glue Grant:
http://www.gluegrant.org/whatsagluegrant.htm
Access to this interactive links box is free online.

THE COMPLEX INTERPLAY AMONG FACTORS THAT INFLUENCE ALLELIC ASSOCIATION

Krina T. Zondervan and Lon R. Cardon

Nature Reviews Genetics **5**, 89–100 (2004); doi:10.1038/nrg1270

In Figure 1, the axes of the two top panels were incorrectly labelled. The corrected version is shown below. This correction has been made to the online enhanced text and PDF version of this review.



ONLINE CORRESPONDENCE

Nature Reviews Genetics publishes items of correspondence online. Such contributions are published at the discretion of the Editors and are subject to peer review. Correspondence should be a scholarly comment on a specific Review or Perspective article that has been published in the journal. To view correspondence, please go to our home page at http://www.nature.com/reviews/genetics and select the link to New correspondence, or, alternatively, go to the archived correspondence at http://www.nature.com/nrg/correspondence.

The following correspondence has recently been published:

Biased gene repair needs unbiased review
By Kyonggeun Yoon, Vitali Alexeev and Olga Igoucheva

Reply:
By Eric B. Kmiec, Li Liu and Hetal Parekh-Olmedo
This correspondence relates to the article:

THE DEVELOPMENT AND REGULATION OF GENE REPAIR
Li Liu, Hetal Parekh-Olmedo and Eric B. Kmiec
Nature Reviews Genetics **4**, 679–689 (2003)

Arguments should be accurately represented
By Michael J. Behe

Reply:
By Scott F. Gilbert
This correspondence relates to the article:

OPENING DARWIN'S BLACK BOX: TEACHING EVOLUTION THROUGH DEVELOPMENTAL GENETICS
Scott F. Gilbert
Nature Reviews Genetics **4**, 735–741 (2003)