

Pathway and Ontology Analysis: Emerging Approaches Connecting Transcriptome Data and Clinical Endpoints

L. Yue* and W.C. Reisdorf

Bioinformatics, GlaxoSmithKline, USA

Abstract: The increasing use of gene expression profiling offers great promise in clinical research into disease biology and its treatment. Along with the ability to measure changing expression levels in thousands of genes at once, comes the challenge of analyzing and interpreting the vast sets of data generated. Analysis tools are evolving rapidly to meet such challenges. The next step is to interpret observed changes in terms of the biological properties or relationships underlying them. One powerful approach is to make associations between the genes that are under investigation and well-known biochemical or signaling pathways, and further to assess the significance of such associations. Similarly, genes can be mapped to standardized biological categories *via* an ontology resource. We discuss these approaches and several web-based resources and tools designed to facilitate such analyses. This information can be used to facilitate understanding and to help design more focused experiments for validating the relevance and importance of these biological pathways and processes in human disease and therapeutics.

Keywords: Pathway, ontology, microarray, transcriptome, data analysis, data sources, analysis tools, standard.

INTRODUCTION

Traditionally, measurement and determination of the effects of drugs on the human body, the status of a particular disease, or the classification of disease by subtypes rely on the measurement and determination of a handful of parameters. The genomic technology of microarray and subsequent analysis enables simultaneous observations for thousands of genes that reveal the status of our body. On one hand, this technology holds great promise as a tool for gaining a comprehensive understanding of drug effects and the mechanism of action and potential toxicity of pharmacophores through characterization of gene expression patterns on a whole genome scale. On the other hand, dissection and understanding of the output of hundreds of observations for thousands of parameters provide unprecedented challenges for clinical scientists. The complexity of data analysis and the task of extracting human comprehensible information that is relevant to the endpoints measured in a clinical study are daunting and often confusing.

There are many excellent reviews available that discuss the utility of microarrays and the fundamentals of data analysis step by step [1-5]. For major steps of microarray experiments and the major currently adopted methods for data analysis, readers are encouraged to refer to the excellent reviews [4-11]. In this review, we will briefly explain crucial steps in microarray data analysis and focus on the

emerging approaches that address the "post-analytical challenges" of microarray experiments, i.e., approaches that further the understanding of the observed changes in terms of pathway and biological ontology categories.

MICROARRAY ANALYSIS: STUDY DESIGN AND THE CHOICE OF ANALYTICAL METHODS

"Well begun is half done". Having a good study design is crucial for carrying out a meaningful microarray experiment in clinical studies. Clinical studies have many unique challenges that make good study design especially crucial. Examples include limitation of the sample size in the study, or the existence of many confounding factors such as disease status, previous medications taken that can not be changed or controlled due to practical limitations (see review in this issue of CMM by Hsu *et al.*). Such confounding factors will have differential effects on different genes and different samples collected. If the study design does not include statistical methods that allow the identification and, if possible, elimination of the effects of these confounding factors, it will be impossible to dissect out their effects from the factors that you wish to investigate. The study design will have great impact on how the data will be analyzed later. Thus, collaboration among statisticians, bioinformaticians and clinical scientists to gain a good understanding of the confounding factors that may impact your study and to derive a study design that maximizes the power of statistical analysis is a crucial first step. Other steps that will impact the quality of analysis output and methods of analysis to be chosen in a

*Address correspondence to this author at the Bioinformatics, GlaxoSmithKline Pharmaceuticals, 1250 Collegeville Road, Collegeville, PA 19426, USA; Tel: (610)-917-4947; E-mail: Lin.Yue-1@gsk.com

microarray experiment include: 1) experimental procedure (maintaining consistency for RNA extraction, fluorescent labeling of probes, microarray fabrication, and hybridization of probes); 2) image acquisition and analysis; 3) data pre-processing and normalization. Among these steps, special attention should be paid to the question of how data is normalized, since different data sets in a microarray experiment may have very different distributions of expression levels. These topics have been reviewed extensively [5-10].

After microarray data has been pre-processed and normalized, two classes of methods have been widely adopted in microarray analysis: supervised or unsupervised approaches. Supervised approaches are used to identify genes that fit a pre-determined pattern. In contrast, un-supervised approaches are used to uncover components and/or patterns internal to a data set without prior input or knowledge of an expected signal.

Supervised methods are generally used for finding genes with patterns that best match a designated query pattern [6,13-14]. For example, one ideal pattern might be a group of genes that is highly expressed in samples after a particular drug treatment, but expressed at a low level in samples without the drug treatment (differential gene expression). In the past, a large number of microarray experiments typically had a handful of samples, with each sample measured under a few conditions. In these situations, a fixed threshold cut-off method (e.g., two fold increased or decreased) has been used to find differentially expressed genes. However, the method is statistically insufficient, due to many systematic and biological variations that influence each individual sample. Statistical methods such as t-test and its variants, ANOVA, and many others have been used to evaluate the significance of the change. Still, setting a meaningful statistically significant cut-off value is difficult, due to the challenges in balancing false positives (Type I error) and false negatives (Type II error), and in performing multiple hypothesis-testing for tens of thousands of genes.

For large data sets and multiple conditions, comparing microarray data sets one pair at a time could miss trends that exist between various samples that were measured at different conditions. In these situations, one may wish to use nearest neighbors analysis, or support vector machine (SVM) methods. Nearest neighbors analysis allows the measurement of all genes in comparison to an idealized gene pattern and subsequently ranking of their similarities. It results in gene lists that might individually split two sets of microarray data, but may not find the smallest set of genes (e.g. the top two genes) that are most similar to the idealized pattern. Support vector machine (SVM) analysis, on the other hand, can address the problem of finding combinations of genes that can best split set of samples with the idealized pattern. By using mathematical operation-

derived functions, the SVM method expands the number of features available to be used for distinguishing samples. It is frequently used for finding genes that accurately predict a characteristic of a sample. Unfortunately, the biological significance of the features determined by mathematical operation-derived functions is often non-intuitive. Many more supervised methods are in use than can be enumerated here.

Un-supervised methods are often used to carry out exploratory data analysis to find internal structure or relationships in a data set, without the requirement of any prior knowledge to bias the process. Most commonly used methods include K-means clustering, hierarchical clustering, self-organizing maps (SOMs), principle component analysis (PCA) and related methods. For more details there are some excellent reviews readers can consult on various unsupervised methods and their applications in microarray data analysis [12-17,34]. We would like to emphasize the fact that different unsupervised methods will fit better for different situations, and no unsupervised data analysis methods will suit all situations. That is because different analysis methods or even different parameters of the same analysis may have very different sensitivities to the variations in the data set, thus revealing different aspects of the data. For example, different distance or similarity measures can assign the same genes to different clusters when one uses hierarchical clustering [1]. One commonly used similarity measure, Euclidean distance, cannot uncover the similarity between genes that are negatively associated [18, 19]. Thus, it is a good practice to apply several analysis methods and use different parameters during the data analysis. Neither statistical significance nor the biological knowledge alone is sufficient, but the two in combination are able to further guide the interpretation of results and the validation of new hypotheses.

Data analysis using supervised or unsupervised approaches usually result in the generation of list(s) of genes that are differentially expressed or related to each other in certain ways. Determining what the results actually mean is the rate-limiting step in many microarray experiments and is an extremely challenging process. First, detailed descriptive information concerning each gene in a particular list needed to be retrieved, yet many genes might not have any informative description available. Second, long list(s) of genes, no matter how accurate or informative their descriptions might be, are not very comprehensible to the human mind. Third, genes never act alone in a biological system. As a result, finding the links between the list(s) of the genes within a biological network context is required to reach a higher level of understanding of the system under investigation. To this end, much progress has been made in the past few years in the development of methods for associating data with biological

context, and we will focus our discussion in this domain in the following sections.

MICROARRAY ANALYSIS: PATHWAYS & BIOLOGICAL PROCESSES

First, we would like to review the available data sources that gather information for biological relationships. We then will review the available tools that can be applied to unravel existing or new biological relationships, or pathways. We will also review factors that will influence how effectively these resources and tools can be used to interpret or sort results in terms of pathway and biological ontology categories.

Pathway Data Sources

There are differences of opinion on what is meant by a "pathway". For our purposes we think of a pathway as a collection of genes, proteins, and metabolites involved in a particular biological process. In addition, these components are connected by an orderly series of events, and these connections are for the most part directed. This definition is meant to exclude entities such as those exclusively dedicated to protein-protein interaction networks, which are often much more complicated and can contain directed and undirected connections [20]. We will also not consider any time-varying aspect of pathways, such as reaction rates, metabolic fluxes or other aspects of pathway modeling [21].

We focus our attention on two types of biological pathways: metabolic and signaling. A third type of biological pathway, which we will not explicitly consider is transcriptional regulation, i.e. DNA-protein complexes [22, 23]. Although private or semi-public pathway data resources do exist, we will focus our discussion on the publicly accessible ones. Metabolic pathways are familiar to students of biochemistry. Some examples of biochemical pathways are glycolysis or fatty acid metabolism. These pathways are most often displayed as showing the transformation from one metabolite to another, with the connections between metabolites being the enzymes which catalyze each step of the process. Several groups have assembled large collections of metabolic pathways, although none of these collections are complete. The most basic features of metabolism are very well covered, but any organism-specific differences from the canonical view are often only to be found in the literature.

The best-known Internet-accessible collections of metabolic pathways are found in KEGG [24] and MetaCyc [25]. KEGG has very broad coverage, particularly of microbial organisms, with data covering approximately 150 prokaryotes. KEGG also contains pathway data for 13 eukaryotes, including humans, and the model organisms that are widely used for biomedical research (mouse, rat, fly, and worm).

However, the pathways are all originally based on *E. coli* and the same pathway diagrams are used for every organism. Overall there are approximately 140 metabolic pathways available in KEGG, although any specific organism will only contain a subset of the known metabolic pathways. For mammalian systems, there are slightly more than 90 pathways containing about 2000 genes per organism.

The other metabolic pathway data source, MetaCyc, is part of the BioCyc collection of pathway and genome databases developed by Peter Karp and colleagues at SRI. MetaCyc was curated from the biological literature and contains over 400 pathways and over 1000 enzymes from more than 100 organisms. For any organism whose genome has been sequenced, a software tool called PathoLogic predicts which MetaCyc pathways may be present. The pathway inference process adopted by PathoLogic is deliberately set to include pathways which are not complete, and is based on the genome sequence data for each organism. An advanced feature of the BioCyc family of databases concerns what to do about pathway "holes". Holes arise when there is no obvious candidate gene to catalyze a step in a pathway that is believed to be functionally intact in the organism. For example, sometimes the pathway is actually present and functional, but the enzyme catalyzing a specific step is so divergent from others of that type that it cannot be recognized based on simple sequence comparison. More detailed sequence/structure analysis may support the idea that the divergent enzyme can fill the "hole". However, the presence or absence of a predicted pathway should be verified experimentally, since it is possible that the pathway is nonfunctional in the organism, due to gene loss or mutation. The SRI team has developed bioinformatics methods to suggest candidate genes to fill pathway holes [26]. These predictions can be prioritized for experimental verification. The most recent addition to BioCyc is the human-centric database called HumanCyc. One major difference from KEGG is that the pathway figures in HumanCyc are not static diagrams, but are instead generated automatically, so they can be updated and expanded as needed.

Also commonly seen in the biomedical literature are signaling pathways (also known as signal transduction pathways), examples of which are Apoptosis or Cell-Cycle Regulation. Some well-known Internet-accessible sources of signaling pathways are BioCarta (<http://www.biocarta.com>), STKE (Signal Transduction Knowledge Environment, <http://stke.sciencemag.org>) and AfCS (Alliance for Cell Signaling-Nature Signaling Gateway, <http://signaling-gateway.org>). Each of these resources offers a variety of pathway diagrams, with links to additional information for the genes involved. BioCarta contains more pathways than the other two sources, with over 300 listed on their website at time of the writing. BioCarta has expanded its coverage to also include metabolism and protein complexes in

Table 1. Tools used to analyze pathway and biological relationship data.

Tool	Pathway/ Ontology	Statistical Methods	Visualization	Refs	URL
ArrayXPath	Pathway	F, M	Y	[38]	http://www.snubi.org/software/ArrayXPath/
Pathway Miner	Pathway	F	Y	[39]	http://www.biorag.org/pathway.html
Knowledge Editor	Both	None	Y	[53]	http://gscope.gsc.riken.go.jp/
EASE	Pathway	O	N	[47]	http://www.DAVID.niaid.nih.gov
GeneMerge	Both	H	N	[52]	http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html
MAPPFinder	Both	O	Y	[54]	http://www.GenMAPP.org
DAVID	Both	None	Y	[50]	http://www.DAVID.niaid.nih.gov
GFINDER	Both	H, F, C	N	[37]	http://www.medinfopoli.polimi.it/GFINDER/
OntoTools	Both	H, F, C, O	N	[51]	http://vortex.cs.wayne.edu/Projects.html
GOSurfer/ChipInfo	Ontology	C, M	Y	[48,49]	http://www.gosurfer.org
GOFish	Ontology	None	Y	[40]	http://llama.med.harvard.edu/Software.html
GOGet/GoView	Ontology	None	Y	[41]	http://db.math.macalester.edu/goproject
GOTree Machine	Ontology	H	Y	[46]	http://genereg.ornl.gov/gotm/
FatiGO	Ontology	F, M	N	[42]	http://fatigo.bioinfo.cnio.es
FuncAssociate	Ontology	F, M	N	[43]	http://llama.med.harvard.edu/Software.html
GOAL	Ontology	T, O	N	[44]	http://microarrays.unife.it
GOMIner	Ontology	F	N	[45]	http://discover.nci.nih.gov/gominer

H: hypergeometric; F Fisher exact test; T T-test; C Chi square; M Multiple testing correction; O Other

addition to signaling pathways. Each of the BioCarta pathways is based on an expert's summary of literature data and also provides a summary description. One limitation of this source is that only human and mouse pathways are available, covering about 1100 genes from each organism. Users interested in expanding the pathways to include rat data can do so by way of the Homologene resource [27].

STKE contains some 34 idealized or canonical pathways along with 19 more detailed organism-specific versions. AfCS is a consortium of labs dedicated to unraveling the biology of B lymphocytes and cardiac myocytes in the mouse, as a model for similar processes in humans. There are 10 pathway maps in AfCS at the time of this writing. Another source called GenMAPP has pathway collections that are complementary to KEGG and BioCarta, but perhaps less comprehensive at present [28]. Although currently no single data source exists that contains a full and up-to-date collection of metabolic, signaling and gene regulation, BioRag (Bio Resources for Array Genes, <http://www.biorag.org>) has a promising start to establish such a data source. It collects pathways from BioCarta, KEGG and GenMAPP, with plans to add more as they become available (STKE, AfCS etc.). Another effort to collect metabolic pathways, along with signaling pathways and other biological processes is the

Reactome project [29,30]. This collaboration between the Cold Spring Harbor Laboratory, the European Bioinformatics Institute, and the Gene Ontology Consortium is developing a curated resource for core pathways and reactions in human biology.

Complementary to pathways, biological ontologies are a more formally structured approach to organizing knowledge. Ontologies are formed from a hierarchical set of properties with increasing levels of detail, in combination with a set of controlled vocabulary terms to minimize the confusing overlap or redundancy of biological terms. One well-known example is the Medical Subject Heading (MeSH, see <http://www.nlm.nih.gov/pubs/factsheets/mesh.html>) terms that are assigned by curators at the National Library of Medicine to each paper indexed in the MEDLINE database. Another widely-used and accepted ontology is the Gene Ontology (GO), developed by the Gene Ontology™ (GO) Consortium. The goal of the GO Consortium is to "produce a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing" (see <http://www.godatabase.org>). GO intends to build a common annotation to describe genes and their products in any organism and assign biological properties to each gene in an organism [31]. There are three major branches of GO: Biological Process

(BP), Molecular Function (MF), and Cellular Component (CC). Much effort has gone into structuring these divisions so that most of basic biology is covered, and genes from any organism can be assigned to GO categories with high confidence. At the time of writing, the GO database covers a total of 17,733 terms and a total of 149,784 gene products from over 100 organisms. The Molecular Function category is the largest, containing 7401 terms and covering 101,079 genes. The Biological Process category contains 8884 terms and 99,849 genes, while the Cellular Component category contains 1448 terms and 80,819 genes overall. The information contained in GO is continuously growing. However, current GO is less well-suited for exploring connections between genes and human diseases directly, since its focus is on basic biology shared by all organisms. Another well-known biological ontology has been developed by the MIPS group [32]. MIPS began as a source for data on yeast biology, and now provides an integrated source for experimental, literature and computationally-predicted protein properties for a variety of complete genomes as well.

Analysis Tools

A number of groups have developed software to assist in analyzing gene expression data in a pathways context. We present a sampling of the most recent or interesting approaches, with apologies to the many authors whose tools we could not include due to space limitations and the rapid proliferation of analysis software. We leave discussion of some tools to a previously published software review [33]. One of the goals, perhaps the most important one, in analyzing transcriptome data in a pathway context is to select the key pathways or biological processes that have been changed due to the experimental manipulations carried out in the study. Three important factors will affect the outcome of the selection. First, the data sources that contain pathways or biological relationships used in a particular analysis. Second, the methods used in the tools to rank the pathway's importance to facilitate the selection of key pathways or biological processes. Third, "A picture is worth a thousand words". Good visualization will go a long way in helping scientists to comprehend the results of the analysis and to make the selection. Thus, as summarized in the Table 1, we divide the tools into three categories: those which apply primarily to pathways, or to GO, or both (second column: Pathway or Ontology). Within each category, the tools differ in which statistical methods (if any) are employed (third column: Statistical Methods), and whether or not the output is mapped back onto a visual figure, such as a pathway diagram (fourth column: Visualization).

Mathematical details about the statistical methods listed in the table are discussed in more detail by others [16,35,36]. Unfortunately there is no one

method that is best for every data set, and each method has its pluses and minuses. We will offer a few general guidelines, but we encourage researchers to consult with their statistics experts for more detailed information than can be provided here. Whatever microarray analysis methods are used, one is generally left at the end with a list of genes that are important or significant by some criteria. In mapping the genes of this list onto pathways, one question to be answered is how many genes in the pathways have been changed and whether or not a pathway is statistically over-represented in the gene list, as compared to what would be expected by chance alone. This will depend on such things as the number of genes in the pathway, the number of genes on the array (or in the genome), the number of genes in the input list, the number of hits found from your list to the pathway, and whether or not all genes in the pathway are present on the particular array in use. Methods such as chi-square, t-test, Fisher's exact test, binomial or hypergeometric test each take some or all of these variables into account, but in slightly different ways. Fisher's exact test is reported to be best when very few gene hits are expected [35] whereas the chi-square test cannot be used for small sample sizes [37]. The hypergeometric test is widely used, but is fairly computationally intensive, particularly for current DNA chips which contain tens of thousands of genes. However, as the size of the chip increases, the hypergeometric distribution tends toward a binomial distribution, which can be a useful simplification. Few of the methods used by the tools covered here take into account the statistical confidence or P-value associated with each gene in the list, or its rank order in the list. A further complication when analyzing such large data sets is that one is effectively doing thousands of comparisons of genes and pathways, and one should also include a correction factor for multiple hypothesis testing. Some of the tools listed in the table incorporate such a correction, but a general discussion of the methods and their relevant merits is beyond the scope of this review (and the expertise of the authors).

Of the tools primarily focused on pathways, ArrayXPath [38] appears to be the most comprehensive. The tool accepts a range of gene identifiers (Swissprot, GenBank, UniGene, LocusLink, etc.) and includes KEGG, BioCarta and GenMAPP as its data source. ArrayXPath uses Fisher's exact test along with the False Discovery Rate method for multiple hypothesis testing correction, and produces a very nice graphical display of results. Pathway Miner [39] can map the genes from up to four different experiments onto KEGG, BioCarta or GenMAPP pathways, and it can create a connection-graph on the fly using the Neato program from GraphViz (<http://www.research.att.com/sw/tools/graphviz/>). Pathway Miner uses Fisher's exact test, but without correction for multiple hypothesis testing.

There are many research groups developing tools for mapping genes to GO. Some of these tools are primarily visual in nature: GOFish [40], GOGet/GOView [41]. Other tools use some statistical measures to estimate the likelihood that a set of genes is enriched in one or more GO categories. Examples of such tools are: FatiGO [42], FuncAssociate [43]. GOAL [44], GOMiner [45], GOTree Machine [46], EASE [47] and ChipInfo/GOSurfer [48,49]. GOSurfer is one of the most comprehensive tools in this class. The developers address a number of statistical methods to deal with the complexity of the GO structure, such as the hierarchical nature of GO and the fact that a gene can be mapped to several locations simultaneously, depending upon the processes it participates in. The same group has also developed a microarray analysis package named dCHIP. EASE (Expression Analysis Systematic Explorer) uses one-sided Fisher exact test or a variant thereof, referred to as the "EASE score" which the authors claim favors more robust categorization. Although initially developed as a GO analyzer, EASE also links to the DAVID tool (described below) and can be used for visualizing genes in terms of their KEGG metabolic pathways.

Tools which cover both pathways and GO include DAVID (Database for Annotation, Visualization, and Integrated Discovery) [50], OntoTools [51],

GeneMerge [52], Knowledge Editor [53], GFINDER [37] and MAPPFinder [54]. MAPPFinder is a software tool associated with the data source GenMAPP [28], which uses z-scores as a statistical measure. MAPPFinder also contains a pathway editor which allows users to customize existing pathway figures or generate new pathway maps using GenMAPP/MAPPFinder software. The OntoTools software suite queries KEGG pathways and GO, along with several other data sources, with plans to add BioCarta in the near future. Included in OntoTools is the Onto-Translate tool which facilitates large scale translation from one type of sequence identifier to another (GenBank, UniGene, or Affymetrix IDs). Users can choose one of several statistical methods to determine how significant the matches are from input data lists to each pathway or GO term. The results are returned in tabular form, but there is not yet any way to map the results back onto pathway figures.

Since the results of microarray analysis are often very complex sets of table and lists, it is extremely helpful to have a way of displaying them visually. This facilitates recognition of trends or patterns that may not be readily apparent otherwise. One approach taken by a number of the tools is to map the analysis results back onto pathway figures (or onto the GO hierarchy). This enables the scientist to immediately see his or her data in the familiar context

Visualization of expression data

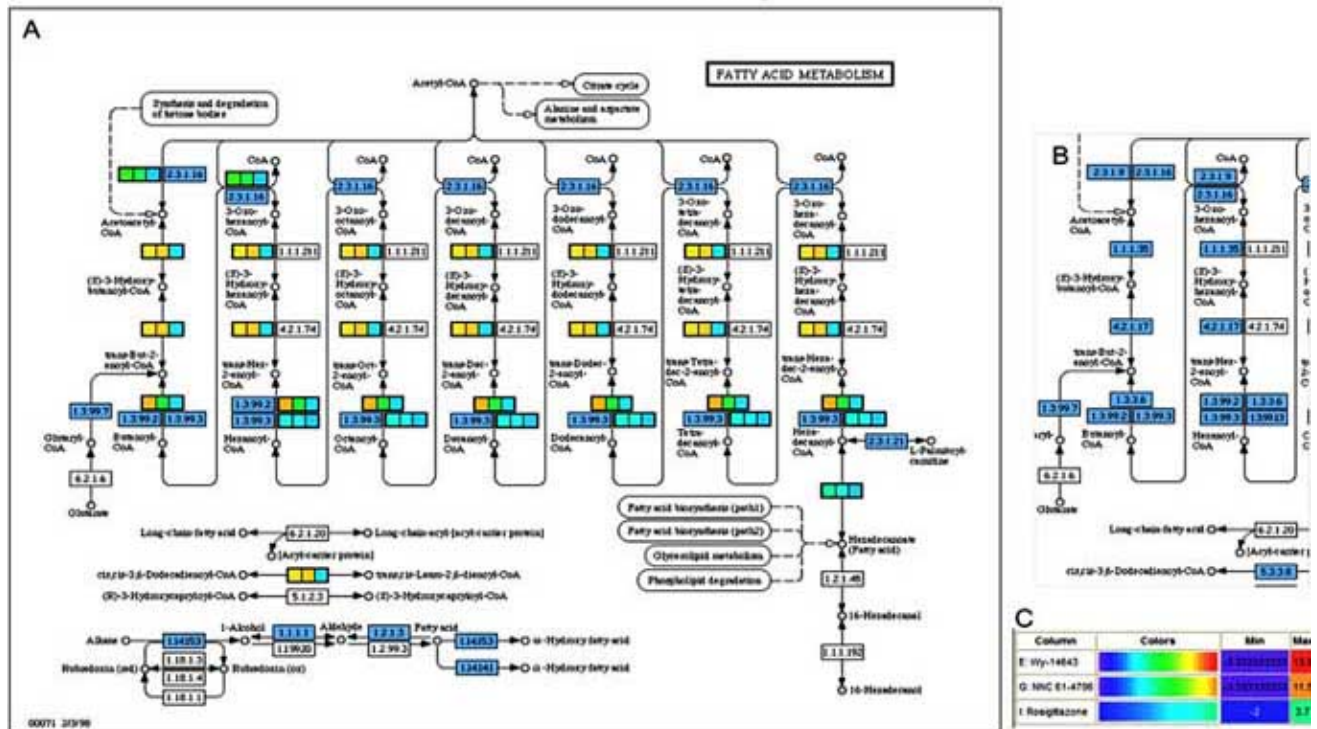


Figure 1. In the part A of the figure, enzymes that were affected by PPAR-alpha agonists were colored over. Right click the mouse, the name of the enzymes can be shown (part of that change is shown in the part B), and the enzyme EC number is linked to more information resources about the enzyme (KEGG EC record, associated gene information etc). Each colored enzyme has three colored boxes (three channels), and each channel represents different PPAR-alpha agonists and the color itself indicates the magnitude of the gene expression change, with the scale illustrated by the part C of the figure.

of a well-known pathway. Color can be used to distinguish the direction of the gene change, and one can use a continuous intensity scale to signify the magnitude of gene expression changes across various conditions, or a discrete scale to qualitatively represent the magnitude of change, such as low, medium or high. For data sets comparing trends over time, such as drug treatment or dosing, it is helpful if one can display several channels of data on the same pathway figure at once. Each channel can be used to represent different parameters of the data, such as the drug treatment, or the dose of drug given. We show an example of such an output in Figure 1, using a tool developed by our colleagues at GlaxoSmithKline. The pathway, Fatty Acid Metabolism, shown in the Figure 1 represents one of pathways affected by PPAR-alpha ligand-mediated physiological changes reported by Frederiksen *et al.* [55]. We mapped genes that were reported in Table 2 of their paper (regulated 2-fold or more by three of the PPAR-alpha agonists: Wy-14643, NNC 61-4706 and Rosiglitazone) to KEGG pathways. Fatty Acid Metabolism pathway is one of the top-ranked pathways that these genes mapped to. As one can easily see through this visualization, five different enzymes on this pathway were affected by the agonists. Agonists Wy-14643 and NNC 61-4706 affect the enzyme 1.1.1.35 and 4.2.1.17 to the similar degree, whereas Rosiglitazone affects these members of the pathway differently. On the other hand, the effect of the three agonists on enzyme 1.3.3.6 appears to be different. Such visualization can help further the understanding of the mechanism of drugs under investigation.

Recently Mootha *et al.* [56] presented an analytical technique designed to test a priori defined gene set(s) (genes in a pathway, for example) for association with disease phenotypes in a transcriptome profiling experiment. The technique, Gene Set Enrichment Analysis (GSEA), incorporates the information about the relationships among the genes in a gene set (e.g. in a pathway) in the statistical significance evaluation of gene expression data. The technique is to test for enrichment of pathway members among the most differentially expressed genes (taking the order of the genes in a list into account). GSEA computes a maximum enrichment score for a gene set and compares the gene set to a null distribution in which genes are randomly distributed. This "reverse pathway" analysis approach successfully identified modest but coordinated (i.e., gene sets, pathways) changes in gene expression that were missed when commonly used methods in microarray analysis such as SAM (Significance Analysis of Microarray) were employed. The approach is very interesting and complementary to the approaches discussed above. More work is needed to determine if the approach can be successfully applied to other diverse data sets. It may be a very useful approach to be applied for *in silico* validation of analysis results (see discussion below).

FACTORS THAT INFLUENCE THE EFFECTIVENESS OF PATHWAY/ONTOLOGY ANALYSIS

The pathway, ontology data sources and analysis tools discussed above establish a basis for finding links between lists of genes in their associated biological network context. Several factors will affect how effectively these resources and tools can be used to reach a higher level of understanding of the unifying biological themes underlying one's data.

First, different data sources and tools use different identifiers to track and analyze gene lists. In most tools mentioned above, the statistical significance of pathway ranking depends on parameters such as the number of genes in the pathway, the number of genes in the input list, and the number of hits found from the gene list to the pathway. If the conversion from one type of identifiers to another is not a one-to-one relationship, the resulting pathway ranking will be affected, and in some cases, the tools will not work. Thus, it is essential to have a solid understanding regarding the identifiers used in different data sources and tools: what the different identifiers stand for, and how to convert each type of identifier to another. Although currently there is no single tool and data source that is available to convert all different identifiers for genes and probes used in the transcriptome analysis from one type to another, Ensembl from Ensembl (<http://www.ensembl.org>), and the SOURCE database from Stanford University (<http://genome-www5.stanford.edu/cgi-bin/SMD/source/sourceBatchSearch>) can interconvert most types of identifiers (Affymetrix, clone identifiers, accession numbers, UniGene name, LocusLinkID etc.). We would also like to point out that dividing the genes into subsets can often be useful. For example, one can make a group of all genes that were differentially expressed, but one could also analyze separately the set of genes that were induced, and those which were repressed. Similarly for time-course data, analyzing the data at each time-point separately might give some biological insights not apparent when examining the trends over the entire experiment. In this respect, one should be creative, as there is no single "correct" choice that always works best.

Second, connections between pathways are important for the comprehensive understanding of relationships underlying transcriptomic data, but this perspective is often lost when looking at maps of individual pathways. An alternative is to create wall-chart diagrams linking metabolic or signaling pathways together. This is sometimes useful, but in many cases the results are still unsatisfactory. It is impossible to draw lines connecting all metabolites to every pathway in which they occur, and still retain a visually comprehensible figure. It is also not generally possible to make a planar figure in which all entities that are biologically connected are also near one another on the page. Such problems are the focus

of active research in computer science and bioinformatics. In our experience, metabolic pathways are relatively non-overlapping (a gene rarely appears in more than a small number of pathways). This is in contrast to signaling pathways which appear to be much more highly overlapping, having some proteins (e.g. second messengers, transcription factors) that are members of dozens of pathways.

Third, pathways that consistently appear in microarray data analysis are affected by diverse biological conditions (e.g. over a large sample of normal human tissues) or manipulations and whose genes show co-expression are indicative of processes that are regulated at the level of mRNA expression. Caution should be exercised when these pathways are identified as significantly affected by the biological phenomena under investigation (e.g., effects of a drug treatment). The observed change may have little or nothing to do with the biological phenomenon being investigated. Yang and colleagues [57] have reported a number of KEGG pathways in which the genes have highly correlated tissue expression patterns regardless of tissue type or origin. We have noticed similar results (unpublished observations). For example, we have confirmed that gene expression in the propanoate and butanoate metabolism pathways as well as the valine/leucine/isoleucine degradation pathway, are highly correlated over many tissues, whether normal or tumor-derived. We have also confirmed similar phenomena for protein complexes such as the ribosome and the proteasome as reported by Yang *et al.* [57]. In addition, particular pathways or biological processes may have different degrees of variation that are inherent to the pathways or biological process themselves, or due to regulation of pathways by metabolite concentrations, post translational modifications, or other factors such as the methods of experimental manipulation at various time points or in specific tissues. Separating the contribution of these factors from the effects for a particular pathway depends on future systematic examination of the relationship between these factors and pathways and biological processes.

Finally, one should also bear in mind that some pathway genes may not show up in any analysis of microarray data, simply because they are regulated at a level other than mRNA transcription. For example, many signaling proteins have their activity regulated by phosphorylation or other post-translational modifications. Such changes, while biologically highly-relevant, will not be apparent in transcriptome studies. Therefore it would be quite unusual to find every gene in a pathway was differentially expressed in any experiment. Additionally, some pathway genes may be absent from the chip or array, although this problem is decreasing as newer gene chips with nearly all known genes are developed.

MICROARRAY ANALYSIS: VALIDATION OF THE MICROARRAY EXPERIMENTS

Due to the challenges in balancing type I and type II error in the data analysis of the large sets of data in a microarray experiment and the challenges in performing multiple hypothesis-testing for tens of thousands of genes, many genes identified will be false positives and many biologically meaningful changes will not be found to be statistically significant. Thus, it is important to evaluate and validate the results from microarray experiments. When evaluating transcriptome data, it is important to determine if the results generated are accurate using independent approaches. Both laboratory-based and *in-silico* based approaches can be used to confirm the validity of results. Ribonuclease protection assay, in situ hybridization using tissue samples, Northern blot, semi-quantitative reverse transcription PCR (RT-PCR), and real-time RT-PCR are common laboratory-based methods that can provide independent confirmation of the transcription expression data. Among these methods, real-time RT-PCR has the promise to become the most dominant one in the future.

Since the number of genes identified in most transcriptome experiments will include lists with hundreds of unique entries, it is becoming increasingly important to derive ways to carry out validation experiments in a high throughput fashion. Taqman technology, with a detection system that is almost synonymous with real-time PCR, allows the determination of expression levels for hundreds of genes simultaneously. Once configured for optimized amplification conditions, it can be carried out in a format (such as 384-microwell plates) that can be automated. It is becoming the standard PCR platform in various laboratories [58]. Although the results from array-based technology and Taqman may differ, in general, the two techniques give qualitatively similar results [59]. Confirmation of the expression change for the phenomena under investigation by an independent approach such as Taqman will provide more confidence in selecting the pathway/biological processes that fundamentally describe the biological phenomena being investigated. *In silico*, computational analysis can be carried out for similar data sets from public or private data sources [61,62]. The task for comparative work may be extremely challenging, due to lack of common standards in carrying out experiments and in data analysis (see discussion below). Information from the literature can be retrieved and extracted, and further analyzed by text-mining technologies. Information can then be compared to the array results. Agreement between the array results and those documented in the literature and derived from other data validate the unique and novel discoveries made in a study.

In addition, we would like to suggest that the GSEA procedure/strategy reported by Mootha *et al.* [discussed above] has the potential to become a

general procedure for *in silico* validation of pathways selected. Pathways identified by the "forward" approach can be evaluated by GSEA. If the statistical significance of pathways identified by the "forward" approach is confirmed by GSEA, this would suggest that the pathway/biological processes likely describe the fundamental biological phenomena being investigated. On the other hand, certain pathways identified by the forward approach may have a low ranking, mainly due to a low number of statistically significant hits to the pathway. These pathways may have a high GSEA ranking, because of the true coordinated changes for each of the pathway members dictated by the underlying biology and the power of combining measurement across multiple members of the pathway by GSEA. Data sets from such pathways can then be further examined by technique such as Taqman to test various hypotheses concerning the member gene in a pathway (for example, individual variations of gene expression in response to drug treatment).

CONCLUDING REMARKS

Generation, analysis and interpretation of microarray data are challenging undertakings. As discussed, there are many statistical methods that can be applied to this task, depending on the study design and type of data collected. Many tools are available for analysis of transcriptome data in a pathways or ontology context, they are rapidly improving and no doubt many new tools are also under development at this time. In order to take full advantage of existing knowledge and to build upon it in a rational way, the time is probably right for more thought and effort towards effective standards. Not standards in the sense of having only one "approved" way to do things, but standards as guidelines to assist researchers in using and comparing different tools and methods. More importantly, standards will facilitate comparative analysis of the observations and conclusions derived from different transcriptome data sets, different array technology, and different analysis methods. Such comparative analysis will allow researchers to systematically examine the relationship between variations observed in the expression of genes and pathways in which the product of those genes play a role, and allow them to reach a higher level of confidence and understanding of the biological phenomena investigated and conclusions reached. Standards will also facilitate the efforts of obtaining functional annotation, in a high throughput mode, for those genes we know little about at the present time, i.e., annotating hundreds or thousands genes at a time, instead of one at a time in the traditional way. Standards are needed for all the steps of a transcriptome experiment. The measure of the quality of the extracted RNA, the quality of the fluorescent labeling and the quality of probes/cDNAs, microarray fabrication, hybridization of probes, and the process of image acquisition and

analysis, data pre-processing and normalization, etc. all need standards to allow meaningful comparison between different data sets generated. The Microarray Gene Expression Data (MGED) Society has made great strides in this area, including development of the MIAME standard. This offers guidelines for the minimal amount of information necessary to describe a microarray experiment, and has been adopted by many journals as a requirement for publication (see <http://www.mged.org> for more information). We will focus our discussion on the need for standards in the domain of pathway and ontology analysis.

In regards to future pathway and ontology tool development, one example of setting standards concerns the fact that the types of sequence identifiers accepted by each tool discussed above can be quite varied. Users frequently spend considerable time and effort to understand the relationship between different identifiers and convert them from one type to another. A standard is very desirable to require all tool developers to adopt one or few identifiers in the gene/probe identifiers used in their tools. Meanwhile, we encourage tool developers to use or work with groups that already provide such translation services (such as EnSEMBL and SOURCE) so that gene identifiers in their tools can be used interchangeably. Another useful step toward standardization is to agree on an accepted format for data input to these analysis tools. Ideally, it should be possible to take a publicly available data set and run it through a new software tool with minimal reformatting. This would greatly facilitate comparison of new and existing analysis tools. In regards to the data sources that gather information for biological relationships, we applaud the vast amount of effort which has gone into developing biological ontologies, particularly by the GO Consortium. They have provided a valuable resource and also a model for using biological expertise to develop a broadly applicable ontology. We look forward to a similar process in the pathway domain. The groups collaborating on the Reactome knowledgebase appear to be making progress in this direction and other efforts exist (see for example <http://www.biopax.org>). To be successful, any effort in setting standards in this area will have to involve close collaboration between biologists and bioinformaticians, championed by leaders in their respective fields.

Despite the advantage of being able to measure mRNA expression of a major proportion of the genes in the genome simultaneously, this massive set of data is only providing one aspect of the biology being investigated. Not every biological process is regulated at the level of mRNA transcription. Proteins are the key players in most processes. Post-translational modification (phosphorylation, acetylation, etc.) of proteins is also critical for regulating many processes and pathways (see the paper by Bilello in this issue). We would like to point out that the tools we have discussed are generally

applicable to the analysis of proteomic data to better understand the biological relationships among the proteins identified.

Current microarray experiments and analysis focus on mRNAs, that is, the protein coding part of the transcriptome. Recent progress in genome sequencing, large, systematic efforts in isolating cDNAs from wide range of tissue, developmental stages, and comparative analysis of genomes suggest that the mRNAs only consist of small part of the transcriptome [66-69]. Furthermore, recent research has demonstrated that many non-coding RNA molecules themselves play essential, diverse roles in the regulation of biological systems, ranging from catalyzing enzymatic reactions, controlling translation/degradation of other genes, and modifying the structure of chromatin so as to alter gene expression and even the genome itself [63-68]. In the coming years, we will see more work that examines the new part of transcriptome, or the "arrays" of non-coding RNA systematically. As more such experimental data becomes available, and new computational tools that can identify and analyze this data come along, we will have a more complete picture for the whole transcriptome of the cells, tissue, organisms, and human diseases [69-74]. RNAs themselves will not only be used primarily as a tool for target validation, but may become an important class of molecules for developing therapeutics as well [75].

ACKNOWLEDGEMENTS

We would like to thank James Brown and Pankaj Agarwal for their encouragement and support. We appreciate the comments and suggestions provided by Philippe Sanseau, James Brown, Mark Hurle, Kay Tatsuoka, Dilip Rajagopalan, Benjamin Hsu and James Butler. Special thanks also go to David Benton, James Butler, Scott Harker and Flip Fuma of the "BNII" (Biological Networks Integration Information) team at GlaxoSmithKline who developed the visualization tool used for generating the Figure 1.

REFERENCES

- [1] Rhodes, D.R., Chinnaiyan, A.M. (2002) *J. Invest. Surg.*, **15**, 275-9.
- [2] Wu, T.D. (2001) *J. Pathol.*, **195**, 53-65.
- [3] Dudda-Subramanya, R., Lucchese, G., Kanduc, D., Sinha, A.A. (2003) *J. Exp Ther Oncol.*, **3**, 297-304.
- [4] Quackenbush, J. (2001) *Nat. Rev. Genet.*, **2**, 418-427.
- [5] Leung, Y.F. and Cavalleri, D. (2003) *Trends in Genetics*, **19**, 649-659.
- [6] Butte, A. (2002) *Nat. Rev. Drug Disc.*, **1**, 951-960.
- [7] Slonim, D.K. (2002) *Nat. Genet. Suppl.*, **32**, 502-508.
- [8] Hess, K.R., Zhang, W., Baggerly, K.A., Stivers, D.N. and Coombes, K.R. (2001) *Trends Biotechnol.*, **19**, 463-468.
- [9] Nadon, R. and Shoemaker, J. (2002) *Trends Genet.*, **18**, 265-271.
- [10] Quackenbush, J. (2002) *Nat. Genet. Suppl.*, **32**, 496-501.
- [11] Moreau, Y., Aerts, S., De Moor, B., De Strooper, B. and Dabrowski, M. (2003) *Trends Genet.*, **19**, 570-577.
- [12] Narayanan, A., Keedwell, E. C., Olsson B. (2002) *Appl. Bioinformatics*, **1**, 191-222.
- [13] Brazma, A. and Vilo, J. (2001) *Microbes Infect.*, **3**, 823-829.
- [14] Hatfield, G.W., Hung, S.-P. and Baldi, P. (2003) *Mol. Microbiol.*, **47**, 871-877.
- [15] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 14863-14868.
- [16] Lee, J.M. and Sonnhammer, E.L.L. (2003) *Genome Res.*, **13**, 875-882.
- [17] Krajewski, P. and Bocianowski, J. (2002) *J. Appl. Genet.*, **43**, 269-278.
- [18] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 2907-2912.
- [19] Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L., Somogyi, R. (1998) *Proc. Natl. Acad. Sci. USA*, **6**, 334-9.
- [20] Hoffmann, R. and Valencia, A. (2003) *Trends Genet.*, **19**, 681-683.
- [21] Eungdamrong, N.J. and Iyengar, R. (2004) *Biol. Cell*, **96**, 355-362.
- [22] Li, H. and Wang, W. (2003) *Curr. Opin. Genet. Dev.*, **13**, 611-616.
- [23] Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M. and Teichmann, S.A. (2004) *Curr. Opin. Struct. Biol.*, **14**, 283-291.
- [24] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) *Nucleic Acids Res.*, **32**, D277-D280.
- [25] Krieger, C.J., Zhang, P., Mueller, L.A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S.Y. and Karp, P.D. (2004) *Nucleic Acids Res.*, **32**, D438-D442.
- [26] Green, M.L. and Karp, P.D. (2004) *BMC Bioinformatics*, **5**, 76.
- [27] Wheeler, D.L., Church, D.M., Edgar, R., Federhen, S., Helmberg, W., Madden, T.L., Pomtius, J.U., Schuler, G.D., Schrimm, L.M., Sequeira, E., Suzek, T.O., Tatusove, T.A. and Wagner, L. (2004) *Nucleic Acids Res.*, **32**, D35-D40.
- [28] Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C. and Conklin, B.R. (2002) *Nat. Genet.*, **31**, 19-20.
- [29] Joshi-Tope, G., Vastrick, I., Gopinath, G.R., Matthews, L., Schmidt, E., Gillespie, M., D'Eustachio, P., Jassal, B., Lewis, S., Wu, G., Birney, E. and Stein, L. (2003) *Cold Spring Harbor Symp. Quant. Biol.*, **68**, 237-243.
- [30] Robertson, M. (2004) *Drug Discov. Today*, **9**, 684-685.
- [31] Gene Ontology Consortium. (2004) *Nucl. Acids Res.*, **32**, D258-D261.
- [32] Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Güldener, U., Mannhaupt, G., Münsterkötter, M., Pagel, P., Strack, N., Stümpflen, V., Warfsmann, J. and Ruepp, A. (2004) *Nucleic Acids Res.*, **32**, D41-D44.
- [33] Guffanti, A., Reid, J.F., Alcalay, M. and Simon, G. (2002) *Trends Genet.*, **18**, 589-592.
- [34] Dudoit, S., Shaffer, J.P. and Boldrick, J.C. (2003) *Statist. Sci.*, **18**, 71-103.
- [35] Man, M.Z., Wang, X. and Wang, Y. (2000) *Bioinformatics*, **16**, 953-959.
- [36] Storey, J.D. and Tibshirani, R. (2003) *Proc. Natl. Acad. Sci. USA*, **100**, 9440-9445.
- [37] Masseroli, M., Martucci, D. and Pinciroli, F. (2004) *Nucleic Acids Res.*, **32**, W293-W300.
- [38] Chung, H.-J., Kim, M., Park, C.H. and Kim, J.H. (2004) *Nucleic Acids Res.*, **32**, W460-W464.
- [39] Pandey R., Guru, R. and Mount, D.W. (2004) *Bioinformatics*, **20**, 2156-2158.
- [40] Berriz, G.F., White, J.V., King, O.D. and Roth F.P. (2003) *Bioinformatics*, **19**, 788-789.
- [41] Shoop, E., Casaes, P., Onsongo, G., Lesnett, L., Petursdottir, E.O., Donkor, E.K.Y., Tkach, D. and Cosimi, M. (2004) *Bioinformatics*, [Epub ahead of print doi:10.1093/bioinformatics/bt h425].
- [42] Al-Shahrour, F., Díaz-Uriarte, R. and Dopazo, J. (2004) *Bioinformatics*, **20**, 578-580.
- [43] Berriz, G.F., King, O.D., Bryant, B., Sander, C. and Roth, F.P. (2003) *Bioinformatics*, **19**, 2502-2504.
- [44] Volinia, S., Evangelisti, R., Francioso, F., Arcelli, D., Carella, M. and Gasparini, P. (2004) *Nucleic Acids Res.*, **32**, W492-W499.
- [45] Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C.,

- Lababidi, S., Bussey, K.J., Riss, J., Barrett, J.C. and Weinstein, J.N. (2003) *Genome Biol.*, **4**, R28.
- [46] Zhang, B., Schmoyer, D., Kirov, S. and Snoddy, J. (2004) *BMC Bioinformatics*, **5**, 16.
- [47] Hosack, D.A., Dennis Jr. G., Sherman, B.T., Lane, H.C. and Lempicki, R.A. (2003) *Genome Biol.*, **4**, P4.
- [48] Zhong, S., Tian, L., Li, C., Storch, K.-F. and Wong, W.H. (2004) *Proc. IEEE Comp. Systems Biol. Conf.*
- [49] Zhong, S., Li, C. and Wong, W.H. (2003) *Nucleic Acids Res.*, **31**, 3483-3486.
- [50] Dennis Jr. G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. and Lempicki, R.A. (2003) *Genome Biol.*, **4**, R60.
- [51] Khatri, P., Bhavsar, P., Bawa, G. and Draghici, S. (2004) *Nucleic Acids Res.*, **32**, W449-W456.
- [52] Castillo-Davis, C.I. and Hartl, D.L. (2002) *Bioinformatics*, **19**, 891-892.
- [53] Toyoda, T. and Konagaya, A. (2003) *Bioinformatics*, **19**, 433-434.
- [54] Doniger, S.W., Salomonis, N., Dahlquist, K.D., Vranizan, K., Lawlor, S.C. and Cinklin, B.R. (2003) *Genome Biol.*, **4**, R7.
- [55] Frederiksen, K.S., Wulff, E.M., Sauerberg, P., Morgensen, J.P., Jeppesen, L. and Fleckner, J. (2004) *J. Lipid Res.*, **45**, 592-600.
- [56] Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D. and Groop L.C. (2003) *Nat. Genet.*, **34**, 267-273.
- [57] Yang, H.H., Hu, Y., Buetow, K.H. and Lee, M.P. (2004) *Genomics*, **84**, 211-217.
- [58] Walker, N. J. (2002), *Science*, **296**, 557-559.
- [59] Rajeevan, M.S., Vernon, S.D., Taysavang, N., Unger, E.R. (2001) *J. Mol. Diagn.*, **3**, 26-31.
- [60] Rajeevan, M.S., Ranamukhaarachchi, D.G., Vernon, S.D., Unger, E.R. (2001) *Methods*, **25**, 443-51.
- [61] Chuaqui, R.F., Bonner, R.F., Best, C.J.M., Gillespie, J.W., Flaig, M.J., Hewitt, S.M., Phillips, J.L., Krizman, D.B., Tangrea, M.A., Ahram, M., Linehan, W.M., Knezevic, V. and Emmert-Buck, M.R. (2002) *Nat. Genet. Suppl.*, **32**, 509-514.
- [62] Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D., Chinnaiyan, A.M. (2002) *Cancer Res.*, **62**, 4427-4433.
- [63] Couzin, J. (2002) *Science*, **298**, 2296-2297.
- [64] Kennedy, D. (2002) *Science*, **298**, 2283.
- [65] Zamore, P. (2002) *Science*, **296**, 1265-1269.
- [66] Mattick, J. and Gagen, M. J. (2001) *Mol. Biol. Evol.*, **18**, 1611-1630.
- [67] Mattick, J. (2001) *EMBO Rep.*, **2**, 986-991.
- [68] Morey, C. and Avner, P. (2004) *FEBS Lett.*, **567**, 27-34.
- [69] Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S. and Hayashizaki, Y. (2003) *Genome Res.*, **13**, 1324-1334.
- [70] Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. and Gingeras, T. R. (2002) *Science*, **296**, 916-919.
- [71] Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G., Wu, L.F., Altschuler, S.J., Edwards, S., King, J., Tsang, J.S., Schimmack, G., Schelter, J.M., Koch, J., Ziman, M., Marton, M.J., Li, B., Cundiff, P., Ward, T., Castle, J., Krolewski, M., Meyer, M.R., Mao, M., Burchard, J., Kidd, M.J., Dai, H., Phillips, J.W., Linsley, P.S., Stoughton, R., Scherer, S. and Boguski, M.S. (2001) *Nature*, **409**, 922-927.
- [72] Eddy, S.R. (2001) *Nat. Rev. Genet.*, **2**, 919-29.
- [73] Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., Haussler, D. (2004) *Science*, **28**, 304(5675), 1321-5.
- [74] Mattick, J. S. (2004) *Nat. Rev. Genet.*, **5**, 316-23.
- [75] Zaman, G. J., Michiels, P.J., van Boeckel, C.A. (2003) *Drug Discov. Today*, **8**, 297-306.