In press: IARC Scientific Publication on Mechanisms of Carcinogenesis, 2002

DESIGN AND ANALYSIS ISSUES IN STUDIES OF GENE-ENVIRONMENT INTERACTIONS

Vineis, Paolo (1), Schulte, Paul A. (2), Carreón, Tania (2),
Bailer, A. John (2,3), Medcalovic, Mario (4)

(1) University of Torino and ISI Foundation. Address for reprint request: Dipartimento di Scienze Biomediche e Oncologia Umana, via Santena 7 10126 Torino, Italy. E-mail paolo.vineis@unito.it

(2) National Institute for Occupational Safety and Health, Cincinnati, OH

(3) Miami University, Oxford, OH

(4) University of Cincinnati, Cincinnati, OH

Acknowledgements: this work has been partially funded with a grant of the Compagnia di San Paolo to the ISI Foundation and with a grant of the Associazione Italiana per le Ricerche sul Cancro.

Various study designs have been identified for the detection of gene-environment interaction. These include: case-only designs, case-control study designs using unrelated controls, case-control designs using related controls, case-control designs using relatives of cases and population- or hospital-based controls, twin-study designs, family-study designs, and combined segregation and linkage analyses.

To assess gene-environment interaction, data could be displayed in a 2 x 4 table, where exposure is classified as being either present or absent and the underlying susceptibility genotype is also classified as present or absent. Using unexposed subjects with no susceptibility genotype as the reference group, odds ratios can be computed for all other groups (Khoury *et al.*, 1993; Botto and Khoury, 2001). Power and sample size are critical in the statistical analysis of these models of interaction. A moderate number of cases and controls is required to detect an interaction when both the exposure frequency and the proportion of susceptible individuals are close to 50%. However, a large sample size will be required, although not necessarily attainable, when the exposure frequency is very low or very high, or when the proportion of risk-increasing alleles at the susceptibility locus is very rare or very common (Hwang *et al.*, 1994). Even larger numbers will be needed if more than one genetic and environmental factor are studied. Statistical models to assess interaction have been developed, but rarely applied, beyond the 2x2x2 design because sample sizes in the cells become small and hence, limit the power (Selvin, 1991). Assessment of interaction can also be affected by exposure and genotype misclassification. Modest error in exposure assessment may result in substantial increases in sample size requirements and this problem is compounded by even small errors in genotype assessment (Rothman *et al.*, 1993; Garcia-Closas *et al.*, 1999; Clayton and McKeigue, 2001).

Alternatively, frequency matching, for known environmental risk factors with a low prevalence in the population, has been proposed to gain power in the study of gene-environment interactions (Stommer and Brenner, 2000). In the mean time, matching is far less

Making sense from population studies incorporating genes and environment has a rich history, but one fraught with problems. The nature-nurture debate has given way to the fundamental realization that both genes and environment play a role in disease, especially cancer. Interaction has been defined in several ways, for example as "the coparticipation of two or more agents in the same causal mechanism leading to disease development" (Yang and Khoury, 1997). According to a more generally accepted definition, interaction occurs when the effect of multiple factors differs from the effect predicted on the basis of the separate effects of each of the factors. Interactions need to be explicitly considered in the design and analysis of epidemiologic studies.

If only the association between a genotype and disease is examined, the effect of biologic interaction cannot be appreciated (Khoury et al., 1993). Epidemiologists need to measure relevant risk factors, in addition to genotype(s). Also, it is likely that each exposure interacts with genes along a pathway (typically: DNA repair genes), so that multiple interactions are likely to be the norm. Evidence for a three-way interaction was showed by Taylor *et al.* (1998) for *NAT2* and *NAT1* genotypes and smoking as risk factors for bladder cancer. Depending on the type of model for gene-environment interaction, risk estimates can vary dramatically. At least five different types of gene-environment interaction have been described (Khoury and Wagener, 1995). These have been generally considered with the dichotomous conditions, a single susceptibility genotype (present or absent), and a single environmental factor (present or absent). For example, in one analysis (Ottman, 1996) the risk to an exposed person with a rare (1% prevalence) susceptibility genotype ranges from 8% (under a multiplicative model of interaction) to 95.6% (when we assume that the exposure has no effect in susceptibles, but the genotype raises risk in unexposed as well as exposed persons).

efficient than an unmatched design, and statistical adjustment can achieve unbiased estimates without such loss of efficiency.

Gene-environment interactions and the case-only design

Epidemiological studies usually involve the comparison between different subpopulations, for example subjects exposed and subjects non-exposed to an environmental agent, or cancer patients and healthy controls (for design issues in molecular epidemiology studies see Schulte and Perera (1993), Toniolo et al.(1997)). Given the recent emphasis on gene-environment interactions, that seem to be the norm and not the exception in the study of chronic diseases including cancer, a different study design has been proposed, that avoids the use of a comparison series. The "case-only" design is based on the study of patients only, affected by the disease at issue, and is exclusively valuable for the investigation of interactions, in particular gene-environment interactions (Yang et al., 1999; Khoury and Flanders, 1996; Umbach and Weinberg, 1997). The basic idea is that, in the presence of interaction, the association between a certain exposure and a genetic trait departs from a predefined model, and to test such hypothesis just the case series is sufficient. If the odds ratio for disease associated with exposure in the absence of the genetic trait of interest (ExOR) is 5, and the odds ratio for disease associated with the genetic trait in the absence of the exposure (GeOR) is 2, then under a multiplicative model we expect that the combination of the two will be 10 (the reasons for choosing a multiplicative model will be discussed below). The simple mathematics of the case-only design are reported, with an example, in Table 1.

The limitations of such approach are obvious, but its advantages are not trivial. The first limitation is related to the impossibility of studying the exposure and the genetic trait independently; in other words, the etiologic role of exposure must have been ascertained already. The case-only approach will allow the researcher to establish, under a plausible

example, we already knew that smoking is a well-established cause of bladder cancer, and the object of the case-only analysis was to study interaction with the NAT2 genotype. A second limitation is related to the selection of the cases. If, for any reason, the genetic trait has a frequency among the cases that does not reflect the expected frequency – irrespective of the etiological relationship at issue–, then any inference based on the case-only approach will be biased. This occurs, for example, if prevalent cases are recruited: since some genetic traits modify the life expectancy, we will find more subjects with a certain allele among those still alive. This is clearly shown by the apolipoprotein E e4 polymorphism: since in smokers the presence of such polymorphism considerably increases the risk of dying from cardiovascular disease, the recruitment of prevalent (i.e., surviving) patients with MI will be associated with an abnormally low prevalence of the e4 allele, particularly among non-smokers. Clearly, this second limitation can be overcome by recruiting incident cases. A third limitation is related to an association between the genetic trait and the exposure among the controls: for example, if the genetic trait induces the subjects to drink more alcohol (such as in the case of the ADH2 polymorphism), then any "interaction" found in the cases only is uninterpretable. If an assumption of independence between genetic and environmental factors is violated, the case-only design will perform poorly causing multiplicative interactions to be highly distorted (Albert *et al.*, 2001).

Let us consider now the advantages of this design. The case-only approach allows us to study interaction with at least the same statistical power as with the more conventional case-control design: in fact the power is usually greater just because of the lack of controls: their variability is in fact eliminated (Khoury and Flanders, 1996). Also confidence intervals tend to be narrower with this approach. The case-only design starts with the plausible assumption of independence between exposure and genotype in the population, and requires fewer case subjects than a case-control design. As a rule of thumb, to study interaction in a

only approach we need the same number of cases but controls are not needed. A second advantage is related to the well-known difficulties associated with the choice of controls in epidemiological studies. In a population-based study controls should be a random sample of the source population from which the cases originated, while in a hospital-based design controls are patients hospitalized for diseases other than the one under investigation. In the former design, one frequent drawback is the low response rate (often around 50% or lower) of the population controls, while in the latter the most common problem is the non-comparability of cases and controls (Wacholder *et al.*, 1992). The case series is usually less affected by low response rates, and its representativeness of all cases of the disease at issue is more easily attained. Therefore, a study design based on cases only has several advantages. It should be clear, however, that it allows us to explore just one specific scientific curiosity, whether some characteristic (usually a genetic trait) modifies the effect of another (usually an environmental exposure). In addition it is worth noting that in the absence of knowledge – from other sources – about the role played by exposure, it is impossible to conclude whether the effect of exposure is adverse in genetic subgroup 1 vs. subgroup 2 or it is protective in subgroup 2 vs. subgroup 1.

Now comes the key question, i.e., what is the expected value for the interactive term we model in the context of the case-only study. In the example in Table 1 we assumed that the null hypothesis of no interaction corresponds to a multiplication of the odds ratios for the environmental trait and the genetic trait separately. Other models for the lack of interaction have been proposed, for example an additive model. While according to the multiplicative model the null hypothesis of no interaction is OR(joint)=ExOR x GeOR, according to the additive model OR(joint)=ExOR+GeOR-1 (Rothman *et al.*, 1980). In general, it has been stressed that any statistical model can be used to generate the null hypothesis, depending on

the underlying biological hypotheses. For example, exposures that act on the same target can have less than additive effects if they compete for a receptor (an example could be PAH and dioxins in relation to the Ah receptor). Conversely, an agent that impairs DNA repair can interact more than multiplicatively with an agent that mutates a cell-cycle gene. However, the multiplicative model is a special case that deserves a central role. The fact that the odds of disease (when jointly assessing the environmental and genetic trait) equals the product of the exposure-only and genetic-trait-only odds ratios indicates the statistical independence, i.e., lack of effect modification. If the multiplicative model is not taken by default as the expression of the null hypothesis, alternate statistical treatment of the data would be required, including techniques for analysing special designs such as the case-only approach. In the case-control study, the absence of interaction is expressed by the same GeOR in exposed and unexposed populations, i.e., the same odds of disease related to genetic trait regardless of exposure status. In the case-only approach the same concept is expressed by an "interaction OR" (Table 1) of 1.0. However, there can be good biological reasons to express the null hypothesis under a non-multiplicative mathematical model. While the case-only study is used to assess departures from multiplicative effects, it does not allow the investigation of the independent effects of the exposure alone or the genotype alone.

Population stratification

In a pooled analysis of the studies on CYP1A1 polymorphisms and the risk of lung cancer (Vineis P et al, submitted) we have observed the following odds ratios for the CYP1A1*2 (MspI) polymorphism:

ALL ETHNICITIES	Heterozygotes	homozygotes
OR	0.88	0.88
95%CI	0.77-1.01	0.68-1.14

CAUCASIANS

OR	1.02	2.36
95%CI	0.84-1.24	1.16-4.81
ASIANS		
OR	1.06	1.14
95%CI	0.81-1.41	0.78-1.69

It is clear that, while mixing ethnicities conceals any relationship (odds ratios are lower than 1 and not statistically significant), when we consider Caucasians and Asians separately there is an association at least with the homozygous genotype in Caucasians. This phenomenon is an example of confounding, since Asians have a much higher frequency of the variant genotype, in comparison with Caucasians, and smoke less, thus fulfilling the criteria for confounding (in this case negative confounding, since the association disappears when mixing the populations).

Population stratification occurs when any factor –genetic or environmental– with a variable frequency between ethnic groups with different risks of disease, appears to be related to disease; this is noted even if there is no causal relationship. Wacholder *et al.* (2000) have shown that population stratification can cause a spurious allele-disease association, but is likely to occur rarely. This phenomenon is observed when allele frequency and disease rates not only differ substantially across ethnic groups, but also correlate strongly with each other and the true risk factor that is responsible for the disease rate difference. However, the true risk factor may be unknown, thereby investigators may fail to account for it. For epidemiologists it is nothing new, being just an issue of confounding. However, it is a particularly important type of confounding in the era of the Genome project and of extensive studies on genes and disease. As Altshuler *et al.* (1998) have stressed, if we do not stratify by ethnicity we risk to attribute to genes a causal responsibility that is in fact related to other characteristics associated with ethnicity, such as other genetic traits or environmental

prevalence of hypertension

polymorphisms in comparison with Caucasians. Therefore, in an unstratified study, in which "population admixture" occurs, one could find a spurious association between smoking-related diseases or hypertension and some polymorphisms; the association would disappear after stratification by ethnicity. However, the problem is not so simple. In fact, it is probably not sufficient to stratify by ethnicity, since genetic heterogeneity also occurs within a certain ethnic group.

There is currently a debate whether bias from population stratification (the mixture of individuals from heterogeneous genetic backgrounds) undermines the credibility of epidemiologic studies designed to estimate the association between genotype and risk of disease. However, Wacholder *et al.* (2000) found only a small bias from stratification in a well-designed case-control study of genetic factors that ignored ethnicity among non-Hispanic, U.S. Caucasians of European origin. In general, there are good reasons to argue that population admixture only rarely can distort the estimates. First, in the example above, it is very unlikely that the investigators ignore the simple Caucasian/Asian stratification. Second, the greater the degree of admixture within a population, and the smaller the difference in allele frequency or baseline disease risk, the less likely that population stratification leads to confounding (Wacholder et al. 2000). Third, when important confounding caused by population stratification does occur, it should be controllable by the usual design and analytical features employed by epidemiologists.

Finally, genetic studies are becoming more and more sophisticated, and the genetic background of populations can be investigated in several ways, for example by stratifying by microsatellite polymorphisms as markers of genetic heterogeneity within a population. Devlin et al (2001) have proposed complex statistical methods. For example, the method called genomic control exploits the fact that the population substructure generates

Description of the two dominant microarray technologies

DNA microarrays are glass slides on which a large number of DNA probes, each corresponding to a specific mRNA species, are placed at predefined positions. DNA probes are either synthesized in-situ (Lockhart *et al.*, 1996) or are pre-synthesized and then spotted on the slide (DeRisi *et al.*, 1997). RNA is extracted from the biologic sample, reverse transcribed into cDNA and labelled in the process. Such labelled cDNA representing the transcriptome of the biologic sample is hybridized on the microarray. The amount of the label cDNA that hybridizes to each probe on the microarray is proportional to the relative abundance of the corresponding cDNA. The expression of all genes is then quantified by intensity of the color used to label the RNA. The most common experimental protocol used with spotted microarrays consists of labeling two RNA extracts with different colors and co-hybridizing the samples to the same microarray. While this approach introduces some restrictions on the experimental design (Kerr and Churchill, 2001), the overall principles of the two major technologies are the same. Quantification of individual gene expressions proceeds by various normalization procedures whose role is to remove systematic biases and rescaling the measurements on different arrays to be directly comparable. The development of an appropriate normalization procedure is still an active research topic (Colantuoni *et al.*, 2002; Hill *et al.*, 2001;Yang *et al.*, 2000).

Applications of microarrays

Although the most common application of microarrays is in monitoring gene expression, other applications, corresponding to the analysis of genomic DNA, encompass the Loss of Heterozygosity (LOH) analysis (Pinkel *et al.*, 1998; Pollack *et al.*, 1999), and genotyping and re-sequencing applications (Iwasaki *et al.*, 2002; Warrington *et al.*, 2002). Gene expression data generated using microarrays is generally used to identify genes that are

of overdispersion generated by population substructure can be estimated and taken into account.

Observations related to the analysis of large datasets on gene expression/polymorphism

New high throughput technologies (such as gene and expression microarrays) will present a number of issues that amplify the challenges of combining genetic and environmental variables. First, will be the large amount of data. This will be a challenge in terms of validity, reduction, summarization, analysis, and interpretation. Many early studies have shown different sets of genes perturbed for similar exposures. This may be due to a lack of standardization of approach to guide analysis and interpretation. Studies where expression of genes is measured under different conditions need rules for determining what is an outlier (a potentially important deviation in expression). In the context of gene expression data, this is sometimes described as the challenge of detecting "signal", a true difference in expression, among the "noise" of natural variability in gene expression. There is a need to determine whether variability of experience for each gene can be measured in the same or different scale (Wittes and Friedman, 1999). Replication is necessary to determine whether tested genes have the same natural scales of measurement. Even if the scale of measurement is appropriate for different genes, the question arises whether an effective method has been used for detecting outliers (Wittes and Friedman, 1999). Other fundamental questions will need to be answered before interpretations of the deluge of data will begin to be possible. These include the ability to empirically define "housekeeping" genes, to identify reproducible artifacts, and to distinguish homoeostatic from pathologic perturbations.

differentially expressed under different experimental conditions, identify groups of genes with similar expression profiles across different experimental conditions (co-expressed genes) and classify the biologic sample based on the pattern of expression of all or a subset of genes on the microarray. Identified differentially expressed genes have been successfully used to hypothesize which pathways were involved in a particular biologic process. Same is the case with groups of co-expressed genes identified in a cluster analysis. Additionally, clusters of co-expressed genes can be used to hypothesize the functional relationship of a clustered gene and as a starting point of dissecting regulatory mechanisms underlying the co-expression. Expression based classification has been used to successfully classify tumor tissues as well as to characterize new tumor sub-types.

Detecting differentially expressed genes across different experimental conditions

Statistical methods for identifying differentially expressed genes have come a long way from initial heuristic attempts (Schena *et al.*, 1996), through the realization that rigorous statistical analysis of replicated data is needed (Claverie, 1999), to sophisticated statistical modeling of various modeling using frequentist and Bayesian approaches. Generic statistical methods of the analysis of variance (Kerr *et al.*, 2000) and mixed models (Wolfinger *et al.*, 2001) are complemented with specialized maximum likelihood approaches (Ideker *et al.*, 2000a), and Bayesian analysis flavored approaches (Efron *et al.*, 2001; Baldi and Long, 2001; Tusher *et al.*, 2001). While the consensus about the optimal method has still not been reached, the intense statistical research is a promising sign. One of the most daunting issues in the process of identifying differentially expressed genes is the severe problem of multiple comparisons. Presently, expression levels of up to more than 20,000 different genes can be assessed on a single microarray. Searching for genes whose expression change is statistically significant corresponds to testing 20,000 hypotheses simultaneously. The traditional significance level of alpha=.05, unadjusted for multiple comparisons, will yield on average

result in the adjusted significance [illegible]

unattainable in simple experiments with very few experimental replicates. The False Discovery Rate (FDR) adjustment (Benjamini and Hochberg, 1995) keeps the balance between the specificity and the sensitivity of microarray data analysis (Tusher *et al.*, 2001). In contrast with traditional adjustments that control the probability of a single false positive in the whole experiment, the FDR approach controls the proportion of false positives among the implicated genes. For example, if 20 genes are selected using FDR=.05, one of them will on average be a false positive regardless of the total number of genes. The traditional (e.g., Bonferroni) adjustment will limit the probability of a single false positive to .05, resulting in an overly conservative testing procedure.

Identifying clusters of co-expressed genes

Overview of clustering approaches

The high-dimensional nature of microarray data has prompted the widespread use of various multivariate analytical approaches aimed at identifying and modeling patterns of expression behavior. In this context, patterns of expression behavior are defined by groups of genes with similar expression levels at different experimental conditions. The major premise of interpreting results of such analyses is that the co-expression is the result of some underlying commonality between mechanisms of expression regulation of such genes. Results of the cluster analysis can be used either to infer common biologic function of co-expressed genes or to serve as a starting point for dissecting the common mechanism of co-regulation that drives the observed co-expression. Since the advent of the microarray technology virtually all traditional clustering approaches have been applied in this context and numerous brand new clustering approaches have been developed.

distances or similarities between the gene profiles. Various correlation coefficients are the most commonly used measures of similarity. Hierarchical agglomerative methods generally proceed by grouping genes based on such pairwise measures of similarity and one of the several ways of establishing distance between groups of genes using still pairwise distance measures (single-linkage corresponding to the minimum pairwise distance between genes in two different groups, complete-linkage corresponding to the maximum distance, k-nearest neighbors principle corresponding to the k^{th} largest distance and average-linkage corresponding to the average distance) (Everitt, 1993).

Partitioning approaches, on the other hand, work by iteratively re-assigning profiles in a pre-specified number of clusters with the goal of optimizing an overall measure of fit. Two most commonly used traditional approaches are k-means algorithm and self-organizing map method, first applied in this context by Tavazoie *et al.* (1999) and Tamayo *et al.* (1999), respectively. One of the problems with clustering methods that are based on pairwise distances of expression profiles is that, at least in the initial steps, only data from two profiles at a time is used. That is, the information about relationships between the two profiles and the rest of the profiles is not taken into account although these relationships can be very informative about the association between them. The major drawback of partitioning approaches is the need to specify the number of clusters.

In a model-based approach to clustering, the probability distribution of observed data is approximated by a statistical model. Parameters in such a model define clusters of similar observations. The cluster analysis is performed by estimating these parameters from the data. In a Gaussian mixture model approach (McLachlan and Basford, 1987), similar individual profiles are assumed to have been generated by the common underlying "pattern" represented by a multivariate Gaussian random variable (Yeung *et al.*, 2001; Ghosh and Chinnaiyan,

2002; McLachlan *et al.*, 2002). In the situation where the number of clusters is not known, this approach relies on one's ability to identify the correct number of mixture components. A mixture based method for clustering expression profiles that produces clusters by integrating over models with all possible number of clusters was developed by Medvedovic and Sivaganesan (2002). This method is based on the Bayesian non-parametric approach based on the Dirichlet process mixtures (Ferguson, 1973). The joint distribution of the data is modeled by a specific hierarchical Bayesian model and the posterior distribution of clusterings is generated using a Gibbs sampler.

Assessing statistical significance of observed patterns

A reliable assessment of reproducibility of observed expression patterns and gene clusters is one of the burning issues in cluster analysis. Since cluster analysis has generally been used as an exploratory analysis tool and not a hypothesis testing tool, establishing statistical significance of observed results has not been a priority. The problem is exacerbated by the heuristic nature of the majority of clustering procedures and the complexity of the clustering space. Even for the model-based approaches, designing the appropriate null distribution for establishing the statistical significance of specific features of created clusters is difficult in most situations. Two exceptions are the significance of the existence of the overall clustering structure and the significance of pairwise associations between individual profiles. In both situations, the null distributions can be constructed by simulating "randomized" data sets and comparing the observed distribution of the overall measure of the model fit, or pairwise similarities, to their distribution in "randomized" data. In the finite mixture model approach, the confidence in obtained patterns and the confidence in individual assignments to particular clusters are assessed by estimating the confidence in corresponding parameter estimates. Assuming that the number of mixture components is correctly specified, this approach offers reliable estimates of confidence in assigning

individual profiles to particular clusters. All conclusions are then made assuming that the correct number of clusters is known. Consequently, estimates of the model parameters do not take into account the uncertainty in choosing the correct number of clusters. The ability of Bayesian approaches to directly assess the probability of any particular feature of the constructed clusters seems to be an obvious advantage in these situations.

Modeling experimental variability

Precision of measurements of gene expressions generally varies across different genes. A portion of such variability is attributable to the association between the absolute expression levels and the variance of the corresponding estimate. Such variability can also be artificially introduced when the number of replicated measurements used to calculate gene expression vary for different genes. When identifying differentially expressed genes, this heterogeneous variability is taken into account by analyzing each gene separately (Wolfinger *et al.*, 2001) or by modeling directly gene specific variability due to different levels of expression (Newton *et al.*, 2001; Baggerly *et al.*, 2001). Unfortunately, such precise characterizations of experimental variability have not been common in the context of cluster analysis. Disregarding information about such variability can result in a biased clustering, and the Bayesian mixture model that can directly model gene-specific variability has been shown to perform better in such situations (Medvedovic and Sivaganesan, 2002).

Reducing the dimensionality of the data

Traditional statistical approaches to identify the coordinate system in which the most of the variability in the data is associated with only a few coordinates, such as principal components analysis, have been applied in the analysis of microarray data. The dimensionality reduction can be applied to either the number of genes or to the number of experiments. Genes are considered replicates and experimental conditions variables when the

problem. In this case the number of non-zero components that can be used as gene surrogates is limited by the number of microarrays. Yeung and Ruzzo (2001) demonstrated that such dimensionality reduction of the experimental condition space may or may not be beneficial when clustering genes. Principal components reduction of the number of genes seems to have been more beneficial when constructing efficient classifiers of tumor samples (West *et al.*, 2001). Other variance decomposition approaches, such as partial least squares (Nguyen and Rocke, 2002) and correspondence analysis (Fellenberg *et al.*, 2001) have been also applied in such analysis with an apparent success. In particular, the partial least squares method appears to have certain advantages over the principal components method when reducing the number of genes to be used in the classification of tumors, because the method chooses an alternative coordinate system that accentuates the relationship between projections on its and the grouping of microarrays. Another approach to identify the most informative projections of the data on a low dimensional (1 or 2 dimensions) space that is yet to be applied in this context is the method of projection pursuit (Friedman, 1987). In contrast to the principal components approach, which assumes that projection on the plane determined by few largest principal components separates data well because of the large spread, projection pursuit directly searches for the projection that best separates data.

Drawing conclusions about co-regulation based on the co-expression

Transcriptional regulation is one of crucial mechanisms used by a living system to regulate protein levels. It is estimated that 5-10% of the genes in eukaryotic genomes encode transcription factors (Brivanlou and Darnell, Jr., 2002) that are dedicated to the complex task of deciding where, when and which gene is to be expressed. Mechanisms applied by these factors range from the recruitment and the activation of the transcriptional pre-initiation

factors that interact with DNA regulatory modules and each other. However, the exact nature of the interactions between various components of the regulatory mechanism is still largely unknown. Identification of co-expressed genes by a cluster analysis of gene expression profiles has been often utilized as a first step in identifying factors regulating expression of different genes (Chu *et al.*, 1998; Tavazoie *et al.*, 1999). On the other hand, using information about presence of known regulatory elements can be applied to refine the cluster analysis of expression profiles and the simultaneous identification of known regulatory elements causing such co-regulation (Pilpel *et al.*, 2001).

An indirect indication of co-regulation of co-expressed genes is the tendency of co-expressed genes to participate in the same biologic pathway (Tavazoie *et al.*, 1999) as well as their tendency to code for proteins that interact with each other (Fuchs *et al.*, 2001). In both of these situations, the mechanism of co-regulation might not be at the level of common *cis*-regulatory elements, yet the need for co-regulation and the actual presence of them is obvious. Actually, it has been shown that the particular expression regulatory mechanism can sometimes be a better determinant of the protein function than even its three dimensional structure (Fessele *et al.*, 2002). All these suggest that analytical methods capable of integrating information about regulatory sequences, biologic pathways and protein-protein interactions, and expression data generated in microarrays experiments will be better able of creating biologically meaningful clusters of genes than clustering expression data alone.

Gene expression based classification

Classification of biologic samples based on their transcriptome profile has been shown to have a tremendous potential for clinical applications in the areas of tumor

classification and toxicity screens of potential drug compounds (Thomas *et al.*, 2001; Waring *et al.*, 2001b; Waring *et al.*, 2001a). Dudoit *et al.* (2002) performed a comprehensive comparison of the performance of traditional classification methods on microarray data based tumor classification. Among other things, their results support the need of basing classifiers on statistical theory. For example, they showed that the maximum likelihood-based classifier clearly outperforms a popular heuristic equivalent described by Golub *et al.* (1999). Due to the large number of variables (genes) used to classify a relatively small number of samples (tumors), most traditional statistical approaches require sub-setting of the genes. Genes are usually selected based on their association with the disease state. However, when designing a Support Vector Machine classifier, Ramaswamy *et al.* (2001) argued that the maximum precision is achieved when all genes are used.

Modeling genetic networks

The traditional approach to characterize roles of different cellular components has been to collect information on the single gene, single protein or single interaction at a time. However, some characteristics of behavior of the complex network of biochemical interactions defining the living system are unlikely to be recovered by such local approaches (Niehrs and Meinhardt, 2002; Ideker *et al.*, 2001). For example, the functional role of a gene whose expression is regulated by several transcription factors can not be fully understood without simultaneously monitoring for the presence and/or activation status of all of them. The ability of the modern high-throughput assays to generate in the same time measurement on a large number of molecules participating in such a network allows us to tackle the problem of reconstructing the whole networks. The complete strategy for such analysis consists of a mathematical model describing interactions of various components of the network, experimental approaches to perturb the network, biologic assays for quantitating the

effects of such perturbation and the inference procedures for estimating parameters of the assumed model. (Ideker *et al.*, 2000b).

Mathematical models of genetic networks

Mathematical models describing dynamics of biochemical networks include the deterministic ordinary differential equation (o.d.e.) based models of kinetics of coupled chemical reactions (Voit, 2002;Voit and Radivoyevitch, 2000), stochastic generalization of such models following the Gillespie's algorithm (Gillespie, 1977) for simulation approach to the chemical master equations (Kierzek, 2002; McAdams and Arkin, 1997; McAdams and Arkin, 1998; McAdams and Arkin, 1999; Gillespie, 1992), Boolean network models which reduce the information about the abundance of various interacting molecules to a binary variable representing on/off (0/1, present/absent) states (D'haeseleer *et al.*, 2000), and hierarchical Bayesian models (Friedman *et al.*, 2000). All of these mathematical models have certain advantages and disadvantages depending on the goal of the analysis, available data and the knowledge about interactions of various molecules in the network. The specification of an o.d.e. model requires detailed knowledge of the modeled interactions and is intended for examination of the overall dynamics of the system when individual relationships between components of the network are more or less known. In this context, the data is more commonly used for checking the predictions based on such models than for reconstructing the networks themselves. Similar conclusions can be made about various stochastic approaches to simulating behavior of such network. The stochastic approaches are likely to offer a more realistic result in biochemical networks involving molecules of very low abundance, as is the case in gene expression regulation (McAdams and Arkin, 1997;McAdams and Arkin, 1998). In contrast to these models, the Boolean network (Toh and Horimoto, 2002) model approach offers a relatively straightforward approach to reconstructing the topology of the network based on discretized data. However, it has been

Networks in particular, seem to be capable of capturing the rich topological structure, integrating components operating on various scales and the stochastic component of both underlying biologic processes and the noise inherent in the data. In this statistical approach, the behavior of the network is expressed as the joint probability distribution of measurements that can be made on elements of the network. The structure of the network is described in terms of the Directed Acyclic Graph (DAG) (Cowell *et al.*, 1999). Nodes in the network correspond to the elements of the network and directed edges specify the dependences between the components. DAG specifies the dependence structure of the network through the Markov assumption that the node is statistically independent of its non-descendants given its parents. Well-established inferential procedures allow for the data-driven reconstruction of the network topology and specific interactions along with the corresponding measures of confidence in the estimated structure and model parameters (Heckerman, 1998; Friedman *et al.*, 1999). The ability to incorporate various levels of prior knowledge through the informative priors about the structure and the local probability distributions (Heckerman *et al.*, 1995) allows Bayesian networks to potentially serve as the model of choice for encoding the current knowledge and the analysis of new data on the background of the current knowledge. Predictions about the future behavior of the network can take into account all sources of uncertainties: uncertainty about the estimated parameters of the networks, structure of the network and the stochastic nature of the biologic system modeled by the network. The application of Bayesian networks for constructing genetic networks based on microarray data has been demonstrated (Friedman *et al.*, 2000). The combined approach of clustering gene expression profiles first and then fitting a probabilistic network using average cluster profiles (Toh and Horimoto, 2002) as well as the use of probabilistic Boolean networks, which are

Integrating current knowledge and various types of data in the analysis of experimental data

Formal methods of integrating accumulated knowledge and information in the analysis have been developed. Zien *et al.* (2000) integrated the existing pathway information with microarray data and developed a method for scoring likelihood of whole pathway involvement in the process under investigation based on the expression levels of genes involved in the pathway. Holmes and Bruno (2000) developed the gene clustering procedure based on the statistical model that integrates microarray data and the genomic regulatory sequence data models. Barash and Friedman (2001) developed a similar approach using higher level information about known regulatory motifs in the genomic data. Ideker *et al.* (2001) demonstrated how to use joint proteomic and functional genomic data after perturbing a biologic system to reverse engineer the underlying network of molecular interactions. They used Maximum Likelihood approach to identify genes with significantly changed expression levels (Ideker *et al.*, 2000a) and Boolean Network model for modeling the network (Ideker *et al.*, 2000b). Benefits of integrating genomic, functional genomic and proteomic data have been demonstrated (Boulton *et al.*, 2002; Ge *et al.*, 2001; Matthews *et al.*, 2001; Vidal, 2001). In terms of human studies, tumors represent a naturally perturbed genomic system. Concurrent genomic and functional genomic investigations of tumors by the high-throughput microarray approaches can be used to dissect genetic networks involved in the process of tumor genesis. In this context, microarrays can be used to both characterize the genomic aberrations and the gene expression in different tumors. Several models mentioned before are capable of integrating such information into a single powerful analysis.

Conclusion

The combination of genetic and environmental variables, always somewhat problematic, has been made more so by new biotechnologies. Issues that need consideration pertain to the design, analysis, and interpretation of studies with genetic and environmental variables. Prior to those considerations should also be ones about the rationale for such studies. Do they represent the judicious use of scarce resources and the diversion of prevention efforts? The potential conflict is between a focus on mechanistic studies at the expense of research on preventive or control efforts. Clearly both activities have their values. With limited budgets, it may be more appropriate to use research funds to effect the greatest health benefit. However, powerful new technologies have great promise is properly explored and harnessed.

As new information about the underlying molecular and genetic determinants becomes available, we are faced with questions about their impact on disease definition, classification, and diagnosis. The high throughput technologies have the potential to define subcategories of disease, and this could contribute considerably to the understanding of disease etiology if the subcategories have distinct etiologies. Even more important is the potential for new definitions of disease. In the biological sciences, speculation exists that a new molecular-based theoretical biology will emerge and lead to the identification of the coordinated action from the products of multiple genes. Such coordinated functions can be viewed as a genetic circuit representing the cellular "wiring diagrams"–the collections of different gene products that together are required to execute a particular function (Palsson, 1997). Others have referred to it as a genome operating system that translates any alteration in the quantity and configuration of a cellular protein into a signal that causes up- or down-regulation of either individual proteins or batteries of proteins (Strohman, 1997). In addition to the molecular, biochemical, physiologic, and medical analyses of the elements and hierarchies that constitute a person and that are involved in disease, the need may exist to add

a systems science approach to disease definition (Palsson, 1997; Strohman, 1997). Seemingly diverse diseases such as cancer, atherosclerosis, and arthritis might have common cellular and molecular features (Satyanarayana and Schoskes, 1997). A better classification system might be one that links clinical observations with cellular and molecular biology data. In addition, artificial neural networks are being explored increasingly as alternatives to discriminant analyses in evaluating diagnostic parameters (e.g., presence of chest pain, serum creatine kinase concentration, etc.) and the best diagnosis available (e.g., whether myocardial infarction was present or not) (Schultz, 1996). This approach may also allow for the inclusion of multiple molecular markers although the importance of comprehensive and extensive training sets for such methods cannot be overstated. The impact of these developments on epidemiologic research is unknown, but they are likely to have an effect. Disease definitions may change and hence, so will diagnostic procedures and ultimately therapeutic approaches (Strohman, 1997; Gyorkos and Coupal, 1995; Crooke, 1996). What constitutes a case or a control for epidemiological studies might be based on probabilities or distributions rather than on the presence or absence of a particular disease endpoint (similar to analyses conducted in receiver-operating-characteristic studies).

smokers (e=0); genetic trait: slow acetylator (g=1) or rapid acetylators (g=0).

Conventional analysis

Gene	Exposure	Controls	Cases	OR
0	0	225	215	
0	1	412	601	1.53
1	0	214	233	
1	1	400	950	2.18

Interaction:

2.18/1.53=1.43 (95% Confidence Interval 1.04-1.96)

Case-only approach

	g=0	g=1
e=0	215	233
e=1	601	950

Odds Ratio (for interaction)=1.46 (95% CI 1.18-1.80)

only design for identifying gene-environment interactions. *Am. J. Epidemiol.* 154: 687-693.

Altshuler, D., Kruglyak, L. and Lander, E. (1998) Genetic polymorphisms and disease. *N. Engl. J. Med.* 338:1626

Baggerly, K.A., Coombes, K.R., Hess, K.R., Stivers, D.N., Abruzzo L.V. and Zhang, W. (2001) Identifying differentially expressed genes in cDNA microarray experiments. *J Comput Biol* (In Press).

Baldi, P. and Long, A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* 17:509-519.

Barash, Y. and Friedman, F. (2001) Context-specific Bayesian clustering for gene expression data. *Proc Fifth Annual Inter Conf on Computational Molecular Biology (RECOMB 2001).*

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 57:289-300.

Botto, L.D. and Khoury, M.J. (2001). Commentary: facing the challenge of gene-environment interaction: the two-by-four table and beyond. *Am. J. Epidemiol.* 153:1016-1020

Boulton, S.J., Gartner, A., Reboul, J., Vaglio, P., Dyson, N., Hill, D.E. and Vidal, M. (2002) Combined functional genomic maps of the *C. elegans* DNA damage response. *Science* 295:127-131.

Brivanlou, A.H. and Darnell, J.E., Jr. (2002) Signal transduction and the control of gene expression. *Science* 295:813-818.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. and Herskowitz, I. (1998) The transcriptional program of sporulation in budding yeast. *Science* 282:699-705.

Claverie, J.M. (1999) Computational methods for the identification of differential and coordinated gene expression. *Hum Mol Genet* 8:1821-1832.

Clayton, D., McKeigue, P.M. (2001) Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 358:1356-1360.

Colantuoni, C., Henry, G., Zeger, S. and Pevsner, J. (2002) Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. *Biotechniques* 32:1316-1320.

Cowell, R.G., Dawid, P.A., Lauritzen, S.L. and Spiegelhalter, D.J. (1999) *Probabilistic Networks and Expert Systems*, Springer, New York.

Crooke, S.T. (1996) New drugs and changing disease paradigms. *Nat Biotechnol.* 14, 238-241.

Devlin, B., Roeder, K., Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theor. Popul. Biol.* 60:155-166

DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680-686.

D'haeseleer, P., Liang, S. and Somogyi, R. (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16:707-726.

Dudoit, S., Fridlyand, J. and Speed, T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *JASA* 97:77.

Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *JASA* 96:1151-1160.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863-14868

Everitt, B.S. (1993) *Cluster Analysis*. Edward Arnold, London.

Fellenberg, K., Hauser, N.C., Brors, B., Neutzner, A., Hoheisel, J.D. and Vingron, M. (2001) Correspondence analysis applied to microarray data. *Proc. Natl. Acad. Sci. USA.* 98:10781-10786.

Ferguson, T.S. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Stat.* 1:209-230.

Fessele, S., Maier, H., Zischek, C., Nelson, P.J. and Werner, T. (2002) Regulatory context is a crucial part of gene function. *Trends Genet.* 18:60-63.

Friedman, J.H. (1987) Exploratory projection pursuit. *JASA* 82:249-266.

Friedman, N., Goldszmidt, M. and Wyner, A. (1999) Data analysis with Bayesian networks: a bootstrap approach. *Proc. UAI.* 196-205.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7:601-620.

Fuchs, T., Glusman, G., Horn-Saban, S., Lancet, D. and Pilpel, Y. (2001) The human olfactory subgenome: from sequence to structure and evolution. *Hum Genet* 108:1-13.

Garcia-Closas, M., Rothman, N., Lubin, J. (1999). Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. *Cancer Epidemiol. Biomarkers Prev.* 8(12):1043-50.

Ge, H., Liu, Z., Church, G.M. and Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* 29:482-486.

Ghosh, D. and Chinnaiyan, A.M. (2002) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics* 18:275-286.

Gillespie DT (1977) Exact Stochastic Simulation of Coupled Chemical Reactions. *J Phys Chem* 81, 2340-2361.

Gillespie DT (1992) A Rigorous Derivation of the Chemical Master Equation. *Physica A* 188, 404-425.

(2001) Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biol.* 2:RESEARCH0055.

Holmes, I. and Bruno, W. (2000) Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 8:202-210.

Hwang, S.-J., Beaty, T.H., Liang, K.-Y., Coresh, J. and Khoury, M.J. (1994) Minimum sample size estimation to detect gene-environment interaction in case-control designs. *Am. J. Epidemiol.* 140, 1029-1037.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R. and Hood, L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292:929-934.

Ideker, T., Thorsson, V., Siegel, A.F. and Hood, L.E. (2000a) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.* 7:805-817.

New York: Oxford University Press.

Khoury, M.J. and Flanders, W.D. (1996) Non-traditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Am. J. Epidemiol.* 144, 207-213.

Khoury, M.J. and Wagener, D.K. (1995) Epidemiological evaluation of the use of genetics to improve the predictive value of disease risk factors. *Am. J. Hum. Genet.* 56, 835-844.

Locker J (2001) *Transcription Factors*. Academic Press. San Diego.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14:1675-1680.

Mareus, P.M., Hayes, R.B., Vineis, P., Garcia-Closas, M., Caporaso, N.E., Autrup, H., Branch, R.A., Brockmöller, J., Ishizaki, T., Karakaya, A.E., Ladero, J.M., Mommsen, S.,

Okkels, H., Romkes, M., Roots, I. and Rothman, N. (2000) Cigarette smoking, *N*-acetyltransferase 2 acetylation status, and bladder cancer risk: a case-series meta-analysis of a gene-environment interaction. *Cancer Epidemiol. Biomark. Prev.* 9, 461-467.

Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., Garrels, J., Vincent, S. and Vidal, M. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res.* 11:2120-2126.

McAdams, H.H. and Arkin, A. (1997) Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 94:814-819.

McAdams, H.H. and Arkin, A. (1998) Simulation of prokaryotic genetic circuits. *Ann. Rev. Biophys. Biomol. Struct.* 27:199-224.

McAdams, H.H. and Arkin, A. (1999) It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet.* 15:65-69.

McLachlan, G.J., Bean, R.W. and Peel, D. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* 18:413-422.

McLachlan, J.G. and Basford, E.K. (1987) *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

Medvedovic, M. and Sivaganesan, S. (2002) Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics (Accepted For Publication)*.

Newton, M.A., Kendziorski, C.M., Richmond, C., Blattner, F.R. and Tsui, K.W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.* 8:37-52.

Nichts C and Meinhardt H (2002) Modular Feedback. *Nature* 417, 35-36.

Nguyen, D.V. and Rocke, D.M. (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 18:39-50.

Ottman, R. (1996) Gene-environment interaction: definitions and study designs. *Prev. Med.* 25, 764-770.

Palsson, B.O. (1997) What lies beyond bioinformatics? *Nat. Biotechnol.* 15, 3-4.

Pilpel, Y., Sudarsanam, P. and Church, G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 29:153-159.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, B.M., Gray, J.W. and Albertson, D.G. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* 20:207-211.

Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D. and Brown, P.O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* 23:41-46.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S. and Golub, T.R. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U.S.A.* 98, 15149-15154.

Rothman, K.J., Greenland, S. and Walker, A.M. (1980) Concepts of interaction. *Am. J. Epidemiol.* 112, 467-470.

Rothman, N., Stewart, W.E., Caporaso, N.E. and Hayes, R.B. (1993) Misclassification of genetic susceptibility biomarkers: implications for case-control studies and cross-population comparisons. *Cancer Epidemiol. Biomark. Prev.* 2, 299-303.

Satyanarayana, K. and Schaskes, D.A. (1997) A molecular injury-response model for the understanding of chronic diseases. *Molecular Medicine Today* 331-334.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O. and Davis, R.W. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. U.S.A.* 93:10614-10619.

Schulte, P.A. and Perera, F.P. (1993) *Molecular Epidemiology: Principles and Practices*. San Diego: Academic Press.

environment interactions by matching in case-control studies. *Genet. Epidemiol.* 18, 63-80.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96:2907-2912.

Taylor, J.A., Umbach, D.M., Stephens, E., Castranio, T., Paulson, D., Robertson, C., Mohler, J.L. and Bell, D.A. (1998) The role of *N*-acetylation polymorphisms in smoking associated bladder cancer: evidence of a gene-gene-exposure three way interaction. *Cancer Res.* 58, 3603-3610.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.* 22:281-285.

Thomas, R.S., Rank, D.R., Penn, S.G., Zastrow, G.M., Hayes, K.R., Pande, K., Glover, E., Silander, T., Craven, M.W., Reddy, J.K., Jovanovich, S.B. and Bradfield, C.A. (2001) Identification of toxicologically predictive gene sets using cDNA microarrays. *Mol Pharmacol.* 60:1189-1194.

Voit EO (2002) Metabolic Modeling: a Tool of Drug Discovery in the Post-Genomic Era. *Drug Discov Today* 7, 621-628.

Voit EO and Radivoyevitch T (2000) Biochemical Systems Analysis of Genome-Wide Expression Data. *Bioinformatics* 16, 1023-1037.

Wacholder, S., Rothman, N. and Caporaso, N. (2000) Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J. Natl. Cancer Inst.* 92, 1151-1158.

Wacholder, S., Silverman, D.T., McLaughlin, J.K. and Mandel, J.S. (1992) Selection of controls in case-control studies. II. Types of controls. *Am. J. Epidemiol.* 135, 1029-1041.

Waring, J.F., Ciurlionis, R., Jolly, R.A., Heindel, M. and Ulrich, R.G. (2001a) Microarray analysis of hepatotoxins in vitro reveals a correlation between gene expression profiles and mechanisms of toxicity. *Toxicol. Lett.* 120:359-368.

Waring, J.F., Jolly, R.A., Ciurlionis, R., Lum, P.Y., Praestgaard, J.T., Morfitt, D.C., Buratto, B., Roberts, C., Schadt, E. and Ulrich, R.G. (2001b) Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles. *Toxicol. Appl. Pharmacol.* 175:28-42.

Warrington, J.A., Shah, N.A., Chen, X., Janis, M., Liu, C., Kondapalli, S., Reyes, V., Savage, M.P., Zhang, Z., Watts, R., DeGuzman, M., Bemo, A., Snyder, J. and Baid, J. (2002) New developments in high-throughput resequencing and variation detection using high density microarrays. *Hum. Mutat.* 19:402-409.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Jr., Marks, J.R. and Nevins, J.R. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. USA* 98:11462-11467.

Wittes, J. and Friedman, H.P. (1999) Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data. *J. Natl. Cancer Inst.* 91, 400-401.

Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R.S. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *Submitted.*

Yang, Q. and Khoury, M.J. (1997) Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. *Epidemiol. Rev.* 19, 33-43.

Yang, Q., Khoury, M.J., Sun, F. and Flanders, W.D. (1999) Case-only design to measure gene-gene interaction. *Epidemiology* 10, 167-170.

Yang, Y., Dudoit, S., Luu, P. and Speed, T. (2000) Normalization for cDNA microarray data. *SPIE BIOS 2001, San Jose, California*

Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. and Ruzzo, W.L. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17:977-987.

Yeung, K.Y. and Ruzzo, W.L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics* 17:763-774.

Zien, A., Kuffner, R., Zimmer, R. and Lengauer, T. (2000) Analysis of gene expression data with pathway scores. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 8:407-417