Combining Gene Expression and Molecular Marker Information for Mapping Complex Trait Genes: A Simulation Study

Miguel Pérez-Enciso,*[,1] Miguel A. Toro,[†] Michel Tenenhaus[‡] and Daniel Gianola[§]

*Station d'Amélioration Génétique des Animaux, INRA, BP 27, 31326 Castanet-Tolosan, France, [†]Departamento de Mejora Genética Animal, INIA, 28040 Madrid, Spain, [‡]HEC School of Management, 78352 Jouy-en-Josas, France and [§]Department of Animal Sciences, University of Wisconsin, Madison, Wisconsin 53706

Manuscript received January 6, 2003
Accepted for publication April 4, 2003

ABSTRACT

A method for mapping complex trait genes using cDNA microarray and molecular marker data jointly is presented and illustrated via simulation. We introduce a novel approach for simulating phenotypes and genotypes conditionally on real, publicly available, microarray data. The model assumes an underlying continuous latent variable (liability) related to some measured cDNA expression levels. Partial least-squares logistic regression is used to estimate the liability under several scenarios where the level of gene interaction, the gene effect, and the number of cDNA levels affecting liability are varied. The results suggest that: (1) the usefulness of microarray data for gene mapping increases when both the number of cDNA levels in the underlying liability and the QTL effect decrease and when genes are coexpressed; (2) the correlation between estimated and true liability is large, at least under our simulation settings; (3) it is unlikely that cDNA clones identified as significant with partial least squares (or with some other technique) are the true responsible cDNAs, especially as the number of clones in the liability increases; (4) the number of putatively significant cDNA levels increases critically if cDNAs are coexpressed in a cluster (however, the proportion of true causal cDNAs within the significant ones is similar to that in a no-coexpression scenario); and (5) data reduction is needed to smooth out the variability encountered in expression levels when these are analyzed individually.

A powerful tool for monitoring gene expression in parallel is cDNA microarray technology. At present, microarrays are being used for improving our knowledge about disease classification as well as for unraveling complex genetic regulation networks (Knudsen 2002). So far, massive expression data have been mostly utilized *per se*, without regard to marker information. However, combining both sources of information may yield a more accurate picture of genetic processes underlying complex traits than that currently obtained by using them separately. For example, expression data can perhaps be used to improve estimates of location of genes affecting complex traits or quantitative trait loci (QTL). Seemingly, this issue has not been addressed, although it has been suggested (Jansen and Nap 2001) that genomics and genetics should be merged into "genetical genomics." This field would involve expression profiling, marker genotyping, and the statistical tools that have been developed for QTL analysis.

There are several potentially useful alternatives to combining microarray and marker data. For instance, one may study the genetic basis of the individual expression levels themselves; Eaves *et al.* (2002) gives an illustration. In this setting, expression levels are regarded as phenotypes and analyzed one by one separately, *i.e.*, treated as any quantitative trait in a usual QTL analysis (Brem *et al.* 2002; Schadt *et al.* 2003). This approach encounters several difficulties, such as the problem of assigning correct significance levels when multiple statistical tests are conducted or the presence of skewed distribution of gene expression measurements. In addition, many genes are regulated and expressed in concerted action (Caron *et al.* 2001), so a gene-by-gene analysis may not be insightful enough. Further, a huge number of simultaneous QTL analyses would be hard to interpret biologically.

An arguably more powerful and appealing approach may consist of detecting some underlying pattern of expression that is correlated with the trait of interest. This implies that some sort of data reduction would be needed. Techniques for this purpose include, *e.g.*, principal components, canonical analysis, and partial least squares (PLS). Principal components, a widely used technique in multivariate analysis, has been already applied to expression data (Alter *et al.* 2000; Holter *et al.* 2000, 2001; West *et al.* 2001). PLS, on the other hand, may be viewed as a compromise between multivariate regression and principal component analysis (Tenenhaus 1998; Hastie *et al.* 2001). The objective here is to find some linear combination of the original expression measurements, or "supergene," that maximizes the correlation with some response variable of interest, such

[1]*Corresponding author:* SAGA-INRA, BP 27, 31326 Castanet-Tolosan, France. E-mail: mperez@toulouse.inra.fr

as the phenotype for a disease trait. In PLS, each new supergene is obtained such that it is orthogonal to all previously defined supergenes (Tenenhaus 1998; Nguyen and Rocke 2002). In PLS, all variables (gene expression levels and phenotypes) are used to arrive at the supergenes, whereas only the expression measurements are used in principal component regression. Enlightening comparisons of PLS, principal component regression, and ridge regression have been published (Frank and Friedman 1993).

If some pattern of expression correlated with the trait of interest can be identified, the microarray data could be used to refine our knowledge about the genetic basis of a complex trait (*e.g.*, a disease), instead of being viewed merely as an additional set of phenotypes to be analyzed as any other quantitative trait. For instance, expression data could be used to improve QTL mapping if the following two conditions were met: (1) some of the gene expression levels must be under (at least partial) genetic control of the QTL and (2) some of these heritable gene expression levels must be related to the disease. Otherwise, accommodating expression data in a statistical model would reduce the power of tests (due to an additional, unneeded, level of parameterization) and increase experimental costs. There is evidence that both conditions can be met, at least in some situations. For instance, p53 mutations lead to a differential gene expression in breast cancer-affected and -unaffected individuals (Sorlie *et al.* 2001). Likewise, the levels of heat-shock protein differ between congenic strains in rats, which suggests a genetic basis for the observed difference in expression (Dumas *et al.* 2000).

Large-scale experiments involving both microarray and marker genotyping are not foreseeable in the immediate future. Rather, we envisage trials where a relatively small number of individuals, say <100, have their gene expression levels monitored as well as genotyped for molecular markers; there may be additional individuals whose genotypes are known but are not microarrayed. Two of the most promising experimental approaches involve recombinant inbred lines and association studies, where controls and cases are carefully stratified to avoid confounding effects. Use of recombinant lines is possible only with laboratory species (Eaves *et al.* 2002), whereas case/control studies constitute one of the most typical research protocols in humans. Although we concentrate on case/control designs, the principles outlined in this work apply to other statistical methods and/or designs.

Our objective is to study the issue of whether or not cDNA microarray data can be used to refine genomic position estimates of genes that affect a complex trait, such as a disease. The impact of the gene expression information is quantified under a range of plausible genetic architectures, including presence or absence of gene expression clustering, different QTL effects and frequencies, and varying number of expression levels affecting disease susceptibility.

MATERIALS AND METHODS

Underlying genetic model: It is assumed that the probability that a disease affects an individual depends on the value of some latent, unobservable, variable (often referred to as liability). The relationship between the probability of disease and liability may not be linear. We express liability as some unknown linear combination of gene expression levels. Considering a single QTL affecting the disease, the allelic variants at the QTL are assumed to produce a shift in mean liability, thus affecting the risk of individuals carrying a given mutation. Note that the effect of the gene is mediated through the relevant expression levels; *i.e.*, its impact on the probability of an individual contracting the disease is indirect.

Simulation strategy: Current knowledge about possible statistical distribution(s) followed by gene expression levels measured with microarray technology is scant. Further, it has been noted that expression levels may be intercorrelated in a complex manner, which would require posing some multivariate distribution. Hence, a standard simulation of expression levels would be probably unrealistic, at least given present knowledge. To circumvent this problem here we propose, instead, to use available real data and simulate the underlying liability *conditionally* on observed expression levels contained in real data, thus reducing dramatically the arbitrariness in the simulation. Suppose the "true" liability of the ith individual is $h_i = \boldsymbol{\omega}'\mathbf{x}_i$, where $\boldsymbol{\omega}$ is a vector of unknown weights given to each of the gene expression levels, with the latter contained in vector $\mathbf{x}_i$. It is reasonable to suppose that most of the values in $\boldsymbol{\omega}$ would be zero because the majority of the genes will not affect the trait of interest. Assume now that probability of disease is related to liability via a logistic function, so that the chance of individual i being affected ($y_i = 1$) is given by $P(y_i = 1|h_i) = \exp(h_i)/[1 + \exp(h_i)]$. Hence, given h_i, the disease status for each individual can be simulated using a Bernoulli distribution with probability $P(y_i = 1|h_i)$. The logistic transformation was chosen because it is widely used for modeling and analyzing binary data (Hosmer and Lemeshow 2000).

Different plausible scenarios of gene interaction models were considered to generate weights $\boldsymbol{\omega}$. First, we allowed gene expression levels included in the liability to be independent or not. In the first case (referred to as "diffuse"), the cDNA clones having an effect on liability were selected independently and with equal probability within those whose expression level had been measured in the microarray. In the second case ("clustered") the first cDNA clone was chosen randomly, and the rest were selected with a probability that was proportional to the absolute value of the correlation of expression levels between the first cDNA and the other candidates. We generated weights using either a uniform (0, 1) distribution or an exponential distribution with mean = variance = 1. Signs (+/−) of the weights were selected at random in the diffuse case and had the same sign as the correlation in the clustered case. cDNAs that were not selected received a weight of zero. Thus, there was a total of four hypothetical scenarios for eliciting the true weight vector $\boldsymbol{\omega}$: diffuse/uniform (D/U), diffuse/exponential (D/E), clustered/uniform (C/U), and clustered/exponential (C/E). The four scenarios are briefly described in Table 1. Note that the variance of liabilities changes according to the scenario and the number of expression levels in the liability.

In all scenarios, the set of weights $\boldsymbol{\omega}$ was such that the frequency of affected individuals, $P(y = 1)$, in the whole population was bounded between 45 and 55%. This condition stems from assuming a case control study, where the population is sampled such that the number of affected individuals is roughly equal to the number of controls. Weights were determined by trial and error sampling; usually less than three

TABLE 1

Gene expression scenarios considered

Scenario	Description	$\bar{\sigma}_h^{2\,a}$
Diffuse/uniform (*D*/*U*)	Clones in *h* chosen at random,[b] weights sampled from uniform (0, 1)	1.8, 5.5, 8.5, 17.8
Diffuse/exponential (*D*/*E*)	Clones in *h* chosen at random, weights sampled from exponential $\mu = 1$	6.5, 19.3, 49.0, 92.9
Clustered/uniform (*C*/*U*)	Clones in *h* chosen proportional to correlation, weights sampled from uniform (0, 1)	—,[c] 8.9, 17.2, 39.9
Clustered/exponential (*C*/*E*)	Clones in *h* chosen proportional to correlation, weights sampled from exponential $\mu = 1$	—,[c] 26.8, 56.9, 168.7

[a] Variance of true liabilities averaged over replicates when 1, 5, 10, and 20 genes are included in the liability, respectively. Disease incidence was 50% in all scenarios.

[b] Underlying true liability.

[c] Same as without clustering.

draws of weights were required because a 50% incidence is simply ensured when the average of the weights is close to zero. Weights were scaled using the standard deviation of each cDNA level.

We further assumed a biallelic additive QTL, where a mutant allele shifts the mean of the underlying susceptibility. Individuals carrying this allele are more prone to contracting the disease, but this relationship is not perfect (incomplete penetrance). In the context of our model, this means that the mutation may affect several cDNA levels to a different extent, depending on the values of the elements of the vector $\boldsymbol{\omega}$. It was assumed that the distribution of the liabilities given the genotype (g) could be approximated by a normal distribution $f(h|g) = N(\mu_g, \sigma^2)$; the standardized QTL additive effect was defined as $a = (\mu_{g=AA} - \mu_{g=BB})/2\sigma$, with *A* and *B* denoting the two QTL alleles. Given h, the probability of an individual i having genotype k is, applying Bayes' theorem,

$$P(g_k|h_i) = P(g_k)f(h_i|g_k)/\sum_{j=1}^{3} P(g_j)f(h_i|g_j), \qquad (1)$$

where $P(g_k)$ is the frequency of genotype k, $k = 1, 2, 3$ for the biallelic QTL. Equation 1 allows us to assign a genotype probability to an ith individual, given its observed microarray data, the weights $\boldsymbol{\omega}$, the genotype frequencies $P(g)$, and the parameters of the normal distribution. However, one needs to specify μ_g and σ^2. The mean of the distribution follows directly from the desired standardized QTL effect, a. The variance of the liabilities is the variance of a mixture and can be written as $\mathrm{Var}(h) = E_g[\mathrm{Var}(h|g)] + \mathrm{Var}_g[E(h|g)]$. Given a standardized genotypic effect, $a = (\mu_{g=AA} - \mu_{g=BB})/2\sigma$, we solved for σ^2 using an iterative algorithm such that $\mathrm{Var}(h)$ was equal to the observed variance of the liabilities in our sample.

Once an individual's genotype was obtained, the rest of the haplotype was simulated. Ten biallelic markers [single-nucleotide polymorphisms (SNPs)] were generated every 0.5 cM, following a simple model for linkage disequilibrium decay. Briefly, a founder haplotype was chosen, sampling a combination of SNP alleles at random. This was assumed to be the original haplotype where the QTL mutation occurred. Then, for individuals that had one or two mutant QTL alleles, one or two haplotypes carrying the mutation were simulated. As generations proceed, the probability that at least one recombination occurs within the 0.5-cM region surrounding the QTL will increase and thus the homology with the founder haplotype will disappear gradually. The length of the nonrecombinant region starting from the QTL was sampled, knowing that the probability of no recombination between the QTL and a position at δ morgans is $1 - \exp(-\tau\delta)$, where τ is the number of generations since the mutation (McPeek and Strahs 1999). For those haplotypes not carrying the mutation, the SNP alleles were sampled assuming linkage equilibrium between markers. We used a frequency of 0.7 for the most common allele for all SNPs, and we set $\tau = 500$. The QTL was in position 0.

With respect to the expression data, a breast cancer data set (Sorlie *et al.* 2001) was used; at the time it was one of the largest data sets publicly available at the Stanford microarray public database (http://genome-www5.stanford.edu/microarray/SMD/). It consists of 85 samples and the expression levels of 456 cDNA clones, what the authors called the "intrinsic data set" (Perou *et al.* 2000). The data reported are the log2 ratios between the mean intensities of the test sample and of a control sample that consisted of a pool of tissues. The log transformations were used to make distributions more "normal," and the base 2 is convenient because it makes interpretation easier. Full details of the experimental and statistical protocols are available online at the web page cited above. Only the 71 cDNAs that did not have any missing record were eligible to enter into the true liability. Thus, a number n_g of cDNA levels was chosen at random out of the 71, and the weights $\boldsymbol{\omega}$ were adjusted as specified above for each of the n_g expression levels. The values of number of genes studied were $n_g = 1$, 5, 10, or 20. The QTL effects were $a = 0.5$, 1, and 1.5 SD units. The QTL genotype frequencies $P(g)$ were chosen to represent two extreme distributions, 0.25/0.50/0.25 and 0.5/0.0/0.5 for the *AA*/*AB*/*BB* genotypes, respectively. The latter frequencies correspond to a case/control study where the disease allele is recessive and at very low frequency; in this case, all affected individuals are homozygous and the frequency of heterozygous individuals in the normal population is negligible. Five hundred simulation replicates were carried out for each scenario. In each replicate a new set of n_g causal cDNAs was chosen, and new values for $\boldsymbol{\omega}$, QTL genotypes, phenotypes, and haplotypes were simulated, always conditionally on Sorlie *et al.*'s (2001) data.

Figure 1 summarizes the main steps in the simulation. First, a series of weights $\boldsymbol{\omega}$ are chosen according to any of the four scenarios described (Table 1), and the individual liabilities are obtained; the obtained liability population is a mixture, wherefrom the QTL genotypes are sampled using Equation 1 for each individual; the haplotypes are obtained; and phenotypes are sampled from binomial processes depending on individual h values.

Partial least-squares and analysis strategy: A first analysis

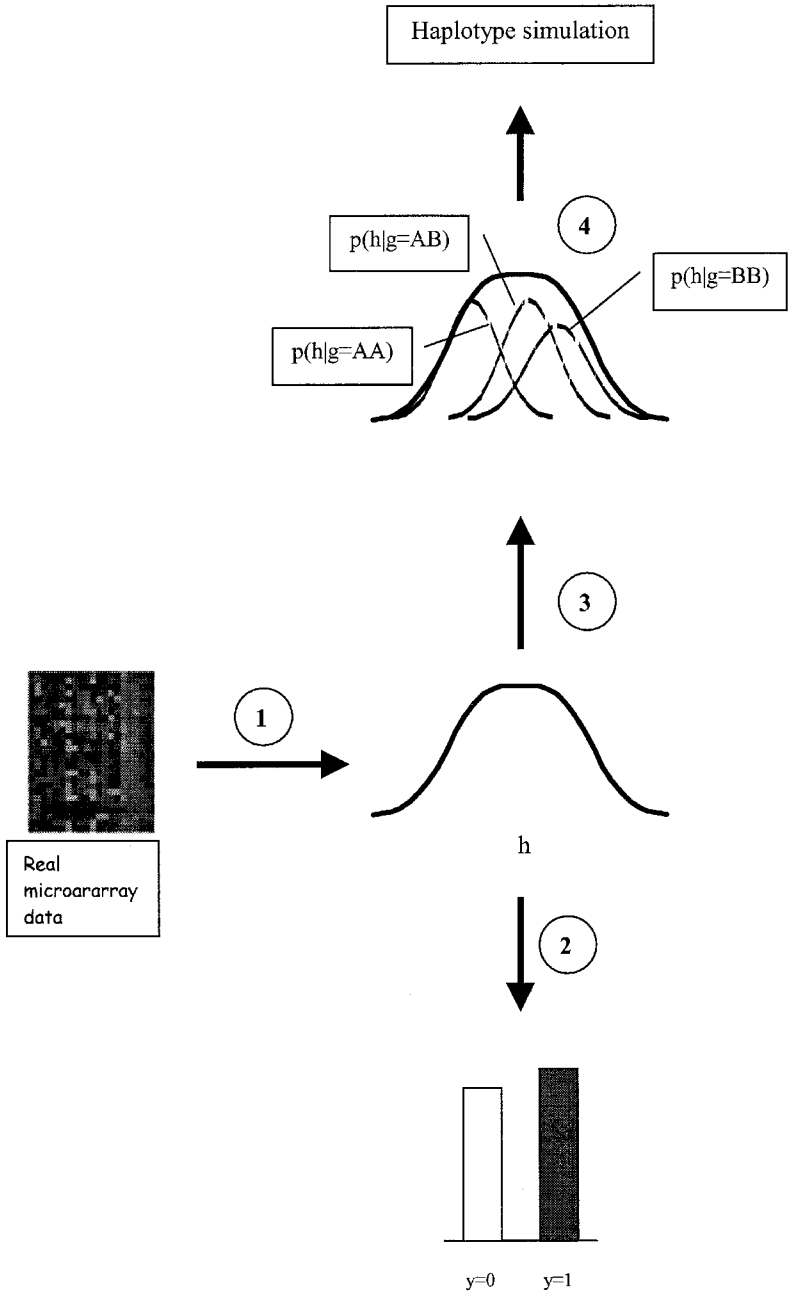

FIGURE 1.—Simulation scheme. (1) Choose weights assigned to cDNA clones for obtaining individual liabilities, after fixing a number of cDNA clones and the gene expression scenario; (2) simulate binary phenotypes from the liability using the logistic distribution; (3) simulate QTL genotypes given liability, QTL effect, and QTL genotype frequencies using Equation 1; and (4) simulate haplotypes from QTL genotype, marker allele frequencies, and number of generations since mutation.

was carried out at the "true" position of the QTL using either the phenotypic information only or the phenotypes and the microarray data. An analysis of variance (ANOVA) was used to test differences in phenotype (y) or estimated liability ($\hat{h}$) among *AA*, *AB*, and *BB* genotypes. The liability for individual i was estimated as $\hat{h}_i = \sum_{k=1}^{\nu} b_k t_{ik}$, with b and t obtained by logistic PLS regression as explained below, and where ν is the number of components fitted. Subsequently, an ANOVA was performed at each of the 10 SNPs using the marker genotype as classification factor. The difference between P values of the ANOVA F-tests using either the phenotype or the estimated liability provides an indication of the relative power of the different sources of information for locating a QTL. The PLS analysis was done with all 456 expression levels, rather than with only the 71 with no missing data.

Suppose that the matrix $\mathbf{X} = \{x_{ij}\}$ contains the expression levels x_{ij} of the jth gene (cDNA) for the ith individual, $j = 1, 2, \ldots, q$ and $i = 1, 2, \ldots, n$. The goal of PLS logistic regression is to find a linear combination of the expression levels for modeling

$$P(y_i = 1) = \exp\left[\sum_{k=1}^{\nu} (\mathbf{w}_k'\mathbf{x}_i)\, b_k\right] \Big/ \left\{1 + \exp\left[\sum_{k=1}^{\nu} (\mathbf{w}_k'\mathbf{x}_i)\, b_k\right]\right\}$$

$$= \exp(h_i)/[1 + \exp(h_i)],$$

where ν is the number of PLS components, $\mathbf{w}_k$ is a q-dimensional vector containing the weights given to each original variable in the kth component (defining a "supergene"), $\mathbf{x}_i$ is the vector containing the q expression levels for individual i, b_k is the regression coefficient of the underlying variable on the kth component variable, and h is the underlying liability. The elements of $\mathbf{w}$ and $\mathbf{b}$ can be obtained as follows (ESPOSITO-VINCI and TENENHAUS 2001):

1. For each variable $j = 1, 2, \ldots, q$ compute its significance in a logistic regression, each variable in turn using the model $P(y_i = 1) = \exp(b_0 + \beta_{1j}x_{ij})/[1 + \exp(b_0 + \beta_{1j}x_{ij})]$.
2. Select those variables that are significant; The first supergene is defined, for each ith individual, as $t_{1i} = \mathbf{w}_1'\mathbf{x}_i$, with $w_{1j} = \beta_{1j}/\sqrt{\Sigma_{j\in\Re^1}\beta_{1j}^2}$, where the sum of j is over the significant cDNAs. An extremely useful property of this approach is that it can deal with missing data in the regressors x_{ij}, a common phenomenon with microarray data. Suppose a subset of x_{ij} are actually measured in the ith individual, the weights are given by $\beta_{1j}/\sqrt{\Sigma_{j\in\Re^{1*}}\beta_{1j}^2}$, where $\Re^{1*}$ is the subset of significant variables present for that individual, and the superscript 1 indicates the significant subset in the first PLS component.
3. The regression coefficient b_1 is obtained from fitting $P(y_i = 1) = \exp(b_0 + b_1t_{1i})/[1 + \exp(b_0 + b_1t_{1i})]$.
4. The next PLS component is obtained by testing again each of the original q variables plus the previous surpergene $P(y = 1) = \exp(b_0 + b_1t_1 + \beta_{2j}x_j)/[1 + \exp(b_0 + b_1t_1 + \beta_{2j}x_j)]$, $j = 1, 2, \ldots, q$. Once the new set of significant variables is determined, the second supergene is obtained from $t_{2i} = \mathbf{w}_2'\mathbf{x}_i$, with $w_{2j} = \beta_{2j}/\sqrt{\Sigma_{j\in\Re^2}\beta_{2j}^2}$, applying identical considerations as before with missing observations.

This process is repeated until no new variable (expression level) is found to be significant. It should be noted that a given gene expression level may be significant in only one PLS component, whereas others may be significant in more than one component. Moreover, it is also possible (actually, this is often the case) that a variable is not significant in component k but significant in component $k + 1$. Thus, it is wise not to discard a set of variables fully from the first component. In this study, a variable was declared significant if its estimate divided by its standard deviation, which is approximately normally distributed, was >3.27, a two-tailed 0.1% significance level. Logistic coefficients can be estimated using a variety of algorithms and software. Here we employed the publicly available subroutines of A. Miller (http://users.bigpond.net.au/amiller/).

RESULTS AND DISCUSSION

One of the main issues arising when microarray data are analyzed is the "excess" of potential regressors relative to the much smaller number of individuals arrayed. Here, we have proposed to combine linearly a set of expression levels instead of studying each cDNA clone separately. Among the many available techniques in multivariate analysis, we have chosen the partial least-squares approach (Wold *et al.* 1983). This technique is quite popular in chemometrics but much less so in genetics. The main advantages of PLS lie in its simplicity (it can be implemented using standard statistical tools); its versatility (*e.g.*, generalized linear models can be fitted, as in the present study); and in the fact that the components are derived using both the regressors (**X**) and the dependent variable, the latter being disease status here. Tests of hypotheses are carried out using standard techniques. We used Wald's test to ascertain whether a given expression level was significant but other tests, such as using the deviance, can be applied as well (Hosmer and Lemeshow 2000). Bootstrapping techniques also appear in the PLS literature and are implemented in some commercial packages (Umetrics 2001). Nevertheless, the problem caused by multiple tests cannot be overemphasized. Here, we used a rather high significance level because the number of clones was relatively small but more stringent levels should be used in larger data sets. The false discovery rate can be a useful alternative to the usual Bonferroni corrections employed with multiple testing (Storey and Tibshirani 2003). Frank and Friedman (1993) have shown how PLS, principal component regression, and ridge regression can be interpreted in terms of applying a penalty on the usual least-squares estimates; *i.e.*, these are shrinkage estimators. It has been recently discussed (Gianola *et al.* 2003) how classical shrinkage estimators can be superseded by Bayesian counterparts for marker-assisted selection. We are not aware of the existence of any equivalent of PLS in the Bayesian context; this could be an interesting area of research either by itself or in how it relates to microarray analysis. Nonetheless, a drawback of most of the dimension-reducing techniques, PLS included, is that the results are usually difficult to interpret biologically.

A potentially important application of microarray experiments is the identification of genes whose expression is affected by a given disease, in the hope of finding

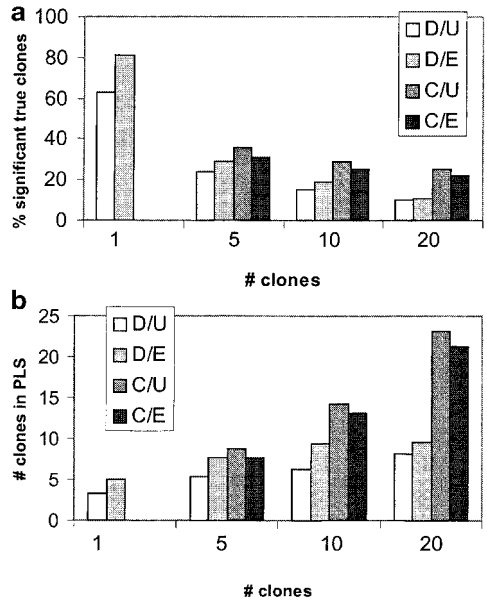

Figure 2.—(a) Percentage of cDNA clones that do influence true liability found to be significant (and included in any of the PLS components) as a function of the number of clones that affect underlying liability and according to the gene expression scenario. (b) Number of cDNA clones in the PLS component, as a function of the number of clones that affect underlying liability and according to the gene expression scenario. The results are the average over 500 simulation replicates and across QTL frequencies and effects, which were very similar. Inset: D/U, diffuse/uniform scenario; D/E, diffuse/exponential; C/U, cluster/uniform; C/E, cluster/exponential (Table 1). Note that a cluster scenario is not defined for a single clone.

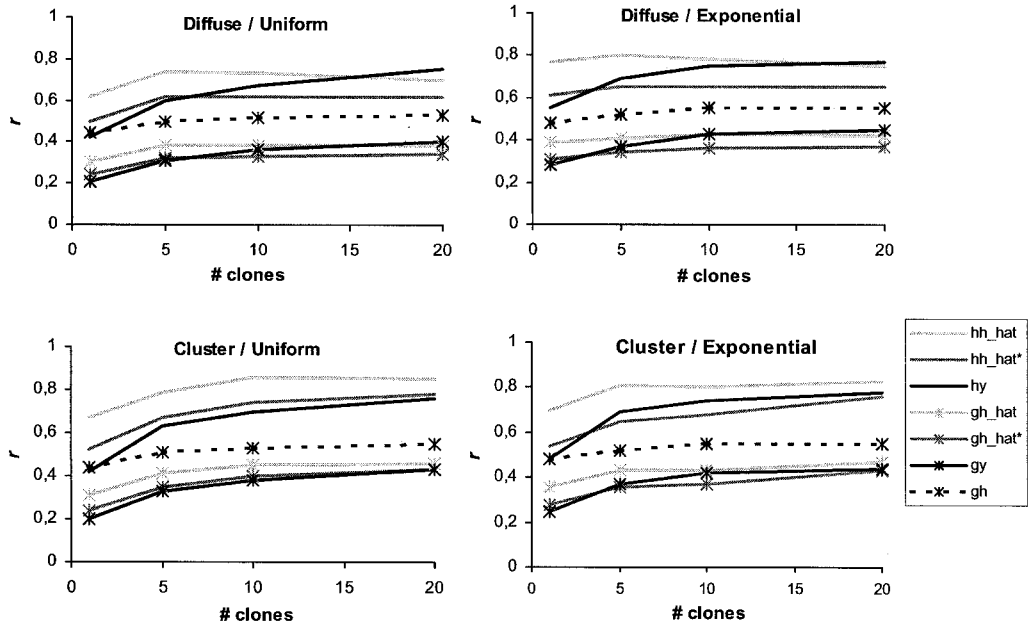

FIGURE 3.—Effect of the gene expression scenario on correlations (r) between true liability and different variables (plain lines): estimated liability (light gray, labeled hh_hat), estimated liability when deleting the cDNA clones that affect true liability from the data set (dark gray, labeled hh_hat*), and phenotype (black, labeled hy). Correlations between genotypic value and different variables (starred lines) are true liability (dashed black lines, labeled gh), estimated liability (solid light gray, labeled gh_hat), estimated liability when the genes affecting true liability are removed (solid dark gray, labeled gh_hat*), and phenotype (solid black, labeled gy). Results correspond to QTL effect = 1 SD and QTL frequencies 0.25/0.50/0.25, averaged over 500 simulation replicates.

the actual causal genes. A main issue, then, is to evaluate the chance of identifying a gene whose expression affects liability. Figure 2a shows that this depends mostly on the number of cDNA clones actually involved in h; we did not find any influence of the QTL effect or of the QTL allele frequencies, so results were averaged over effects and frequencies. If liability is monogenic, the probability that the cDNA clone is included in at least one of the PLS components (supergenes) varies between 60%, when weights are uniformly distributed, and 80%, when an exponential distribution is used. This simply reflects the fact that weights are, on average, larger in the exponential than in the uniform scenario. The distribution of weights did not seem to affect the results appreciably when liability was polygenic. Nevertheless, clustering of gene effects increased the number of true significant cDNAs identified. This is a consequence of a higher number of total significant cDNAs in PLS in clustered compared to diffuse scenarios (Figure 2b). In fact, the percentage of true causal cDNAs among significant ones was similar in diffuse and in clustered scenarios. There is clear evidence that coexpression can be strong in both humans and Drosophila (CARON *et al.* 2001; ARBEITMAN *et al.* 2002). All in all, on average, only between two and four real causal clones are identified as significant when 20 expression levels are actually involved in the liability (see Figure 2a). This may explain why discordant sets of cDNA clones are identified in different microarray experiments, *e.g.*, when discriminating between estrogen receptor-positive and -negative breast cancers (GRUVBERGER *et al.* 2001; KHAN *et al.* 2001; WEST *et al.* 2001; PÉREZ-ENCISO and TENENHAUS 2003).

A positive association was found between the number of cDNA clones in h and the number of clones included in $\hat{h}$ (Figure 2b), and the association was even stronger when causal cDNAs were coexpressed in a cluster. However, as the number of clones in h increased, the relative effect of each clone is expected to decrease, especially in a diffuse scenario, and so does the power of PLS for identifying each effect. This may explain the lack of linearity of the association in the diffuse scenario. The number of PLS components fitted was also affected by the actual number of cDNA clones in the liability: only one PLS component was retained in ~95% of replicates when the true liability was monogenic ($n_g = 1$). This does not mean that the PLS component consisted of a single cDNA, as the number of genes in the PLS component was ~4 (Figure 2b). A second component was significant when liability was polygenic in 10–20% of replicates. An exponential distribution for the weights $\boldsymbol{\omega}$ increased the percentage with two PLS components fitted, but this percentage was roughly the same for 5, 10, or 20 genes within this scenario. This may reflect the fact that PLS is computed such that the number of components is minimized.

The results shown in Figure 2a do not imply that the liability was estimated poorly. On the contrary, the correlation between estimated and true liabilities was ~0.80 over a wide range of parameters. Consider

TABLE 2

Main results

			QTL frequencies											
			0.25/0.50/0.25						0.50/0.00/0.50					
			P value QTL[d]		% QTL max[e]		% SNP1 max[f]		*P* value QTL		% QTL max		% SNP1 max	
Scenario[a]	a[b]	n_g[c]	$\hat{h}$	y	$\hat{h}$	y	$\hat{h}$	y	$\hat{h}$	y	$\hat{h}$	y	$\hat{h}$	y
Diffuse/uniform	0.5	1	0.24	0.37	38	18	19	7	0.18	0.27	54	29	28	18
		5	0.21	0.29	38	30	20	9	0.08	0.11	69	54	20	14
		10	0.18	0.22	44	34	13	13	0.06	0.07	71	70	27	24
	1.0	1	0.16	0.20	66	40	26	18	0.08	0.10	77	65	40	20
		5	0.03	0.05	88	72	30	21	0.02	0.01	94	93	43	37
		10	0.02	0.03	89	81	29	24	0.01	3×10^{-3}	94	95	42	42
	1.5	1	0.09	0.11	77	62	37	18	0.10	0.10	71	69	44	30
		5	0.03	0.02	90	88	39	28	0.01	10^{-3}	93	98	40	42
		10	9×10^{-3}	5×10^{-3}	95	96	31	32	0.01	5×10^{-4}	96	98	47	47
Diffuse/exponential	1.0	1	0.07	0.12	80	63	24	18	0.04	0.07	88	81	43	38
		5	0.02	0.03	88	85	24	24	3×10^{-3}	3×10^{-3}	98	97	47	47
		10	8×10^{-3}	0.01	95	93	35	30	8×10^{-4}	2×10^{-4}	98	98	43	43
Cluster/uniform	1.0	5	0.02	0.06	90	72	29	23	7×10^{-3}	6×10^{-3}	97	93	42	35
		10	7×10^{-3}	0.01	95	85	36	27	3×10^{-3}	10^{-3}	99	98	45	41
Cluster/ exponential	1.0	5	0.01	0.03	92	84	26	23	2×10^{-3}	3×10^{-3}	97	97	48	47
		10	0.01	0.01	94	93	27	24	2×10^{-4}	3×10^{-4}	100	99	51	47

[a] See Table 1 for description of each scenario.
[b] QTL effect in SD units.
[c] Number of cDNA levels in true liability.
[d] Mean ANOVA *P* value using estimated liability ($\hat{h}$) or phenotype (y).
[e] Percentage of replicates when maximum statistics, using estimated liability ($\hat{h}$) or phenotype (y), coincided with QTL position.
[f] Percentage of replicates when maximum statistics, using estimated liability ($\hat{h}$) or phenotype (y), coincided with closest SNP when QTL genotype was not included in the region scan.

Figure 3, where the correlation between true and estimated liability is labeled "hh_hat" (the plain light gray line). This correlation was independent of the QTL effect (results not shown), but it increased slightly as the number of cDNAs in the true liability increased. Figure 3 also shows that the advantage of using microarray data over simply the phenotypes was inversely related to the number of cDNAs in the true liability and that it was maximum when liability was monogenic (compare the lines labeled hh_hat *vs.* "hy"). Interestingly, a clustered scenario was more favorable than a diffuse scenario, especially when weights were uniformly distributed.

An implicit assumption of our model is that the clones that influence liability have been spotted in the microarray. If this were not so, the possible advantage of using microarray data would be reduced, although it would seldom be nil because of possible correlations of expression between spotted and nonspotted genes. We evaluated this possibility by removing from the data set all cDNA clone data involved in h and carrying out the PLS analysis subsequently. It can be seen (Figure 3, dark gray lines labeled "hh_hat*") that the correlation between liability and its estimate is still high, but below that with the phenotype when liability is polygenic. A clustered scenario makes the loss in accuracy smaller when causal cDNAs are not spotted. Figure 3 also depicts the correlations between genotypic values and liability, its estimate, or the phenotype (starred lines). The dashed black line, labeled "gh" (correlation between genotype and true liability) sets the maximum correlation that can be expected. The trends of correlations with true liability or genotype were similar (compare plain and starred lines). Again, as the number of cDNAs in liability increased, the value of phenotypic information relative to that of estimated liability increased. Clustering, instead, favored the usefulness of the estimated liability.

Table 2 presents the ANOVA *P* values obtained using either $\hat{h}$ or phenotype (y), as well as the percentage of replicates where the significance was maximum at the QTL position or at the closest SNP (SNP1) when the QTL position was not included in the genome region scan. These percentages somewhat reflect the confidence that we can have in the estimated QTL position with and without microarray information. Not all cases analyzed are reported to facilitate legibility. First, note that estimated liability performed relatively better than phenotype at intermediate rather than at extreme QTL

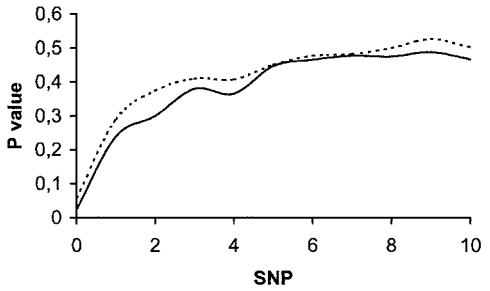

FIGURE 4.—Example of profile of the *P* values: ANOVA *P* value with estimated liability (—) and ANOVA *P* value with phenotype (---). Cluster/uniform scenario with QTL effect = 1 SD, five cDNA clones in true liability, and QTL frequencies equal to 0.25/0.50/0.25, averaged over 500 simulation replicates, is shown.

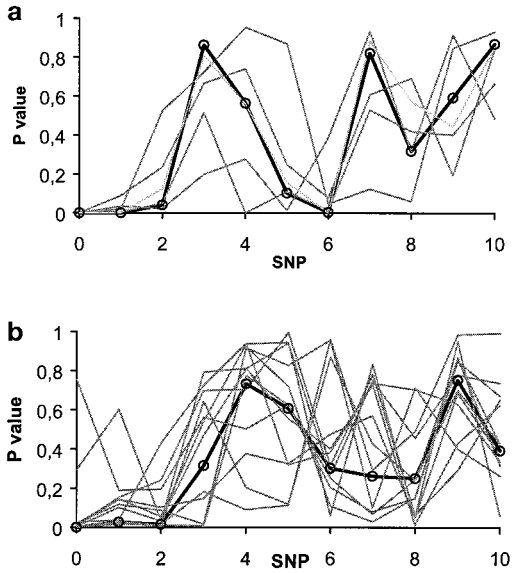

FIGURE 5.—*P*-value profile of a single replicate, for the estimated liability (thick circled line) and each of its individual cDNA components (thin gray lines), QTL effect = 1 SD in the diffuse/uniform scenario. (a) The true number of cDNA clones in the liability is 1; (b) the number of cDNA clones is 10. The QTL is in position 0.

frequencies. Other things being equal, microarray data will be more useful if liability is monogenic, as could be expected from results in Figure 3. Similarly, the percentage of replicates where the position of maximum significance coincided with the QTL position was comparatively better with $\hat{h}$ than with y at small QTL effects. Clustering and an exponential scenario favored using $\hat{h}$ over only the phenotype when significance was low.

Figure 4 displays an illustration of the performance of the test over the interval considered when we use the phenotype y (dashed line) or the estimated liability with PLS $\hat{h}$ (solid line). On average, the point of maximum significance (minimum *P* value) coincides with the QTL position (position zero), and over the entire interval the power is larger when using microarray data than when using the disease status (phenotype) only. Although the average test was maximum at the QTL position, see Table 2 for the percentage of replicates when this actually happened. Association studies are well known for the difficulty in obtaining replicable results (EMAHAZION *et al.* 2001), which is due, in part, to the wide variability that disequilibrium exhibits (NORDBORG and TAVARÉ 2002). Note that when the causal (QTL) mutation was not genotyped, the closest SNP coincided with the maximum statistic in <50% of the replicates for most of the cases studied (Table 2).

The purpose of this article was not to assess extensively the impact of disequilibrium variability on cDNA-QTL studies. However, the study illustrates that a PLS-estimated liability will normally have a more stable behavior than any of its components (*i.e.*, cDNA measurements here) taken individually. For instance, Figure 5 displays results for two individual replicates in the diffuse/uniform scenario where the estimated liability and all of its individual components are shown when the true liability is monogenic (Figure 5a) or polygenic (Figure 5b). The variability of the estimated liability was lower than that of any of its individual components, with the trend increasing as the number of cDNA clones in the liability increases. Moreover, the noise will increase significantly when all expression levels are analyzed separately; recall that a typical microarray experiment consists of thousands of measurements. This means that it may be very difficult to interpret all sets of QTL profiles when cDNA levels are analyzed individually.

The four scenarios considered in this work are idealized representations of a variety of possible gene interaction networks. The diffuse/uniform case corresponds to the simplest scenario and may be viewed as a "null hypothesis" model. In the diffuse/exponential model we study the effect of unequal gene contributions, which is perhaps a more realistic assumption. Overall, it seems that gene clustering is far more relevant than the fact of having unequal weights (*e.g.*, Figures 2 and 3). There is ample evidence of coexpression of large clusters of genes (CARON *et al.* 2001; ARBEITMAN *et al.* 2002), but in the context of this work we are interested in coexpressed genes that are causal as well. If coexpressed genes are not causal, the number of genes in the PLS components will increase (Figure 2a), but not the number of causal significant genes.

There is, thus far, no empirical evidence about the genetic basis of gene expression on a genome-wide scale in outbred populations, although some experiments concerning QTL analysis in crosses between inbred stocks have begun to appear, notably that of BREM *et al.* (2002) in yeast but also in mice and maize (SCHADT *et al.* 2003). BREM *et al.* (2002) found that a good percentage (~80%) of expression levels were controlled by more than one gene, probably by at least five genes. In a few cases, a single genome region controlled several expression

levels, ranging from 7 to 87 levels. In summary, they found a variety of gene architectures affecting expression levels, as was the case in Schadt *et al.* (2003). We can say, in light of our simulation study, that a polygenic basis will be one of the main challenges for interpreting QTL expression data; it will increase the number of significant cDNA clones but it will be less likely that the true causal cDNAs are among those that are significant (Figure 2). Clustering will enhance this phenomenon.

A final word of caution should be said. Throughout, we have assumed that the expression levels are measured without error or at least measured with the same precision as the disease is diagnosed. However, this is not necessarily true because of technical problems in the microarray devices, rapid changes in mRNA concentrations, or imperfect conversion into cDNA. All these phenomena will hamper the usefulness of microarray data but it is difficult to quantify their effect at this stage of knowledge. There are specific statistical techniques for dealing with the problem of regressors measured with errors that can prove to be valuable in this setting (Nguyen *et al.* 2002; Suh and Schaferb 2002).

CONCLUSION

In the almost complete absence of real data that combine marker and gene expression data, a fundamental problem is how to simulate a "realistic," or at least plausible, data set reflecting as much as possible the actual complexity of correlation between expression levels. Here we propose a novel alternative that is attractive for several reasons. First, we simulate conditionally on real expression data, including the missing data pattern. Second, we also allow for random variation by choosing each time a different subset of causal cDNAs subject to a series of constraints (*e.g.*, equal expected frequency of disease and healthy individuals). Third, we allow for a complex genetic basis, in the sense that there are environmental influences (no one-to-one correspondence between genotype and trait exists) and a nonlinear relationship between trait and genotypic effects.

It might well be that the results presented here represent the worst-case scenario in using microarray data to help in mapping complex trait genes, as the underlying liability was constructed arbitrarily, although using the true variation in expression levels and respecting some constraints like that of matching the frequency of cases and controls. It is also likely that future technologies will allow us to measure expression levels in an increasing number of individuals, leading to much more information than can be obtained at present. In any case, our simulation study suggests the following: (1) the relative usefulness of microarray data increases as both the number of expression levels in the underlying liability and the QTL effect decreases, but increases with gene expression clustering; (2) some sort of data reduction is necessary to smooth out the wide variability apparent in each expression level when taken individually; and (3) it is unlikely that the cDNAs identified as significant in PLS (or in similar data reduction techniques) are the truly responsible cDNA clones, especially as the number of cDNAs in the liability increases. This can occur even when there is a very high correlation between true liability and its estimate. This corresponds with the cautionary remark made some years ago (Lander 1999): correlations or associations found with microarray experiments should not be viewed as cause-effect relationships. The same caution is in order when interpreting similar experiments yielding distinct results. For instance, the experiments may declare different sets of genes as "significant" when discriminating disease subtypes, but such genes may not be the causal ones.

We thank Bruce Walsh and the referees for suggestions and A. Miller for making his subroutines available to the public. Work was funded by grants to D.G. (National Research Institute CGP-United States Department of Agriculture 99-35205-8162 and National Science Foundation DEB-0089742; United States) and to M.P.E. (*Action en Bioinformatique*; France). This research started while M.P.E. and M.A.T. were visiting scientists at the University of Wisconsin-Madison.

LITERATURE CITED

Alter, O., P. O. Brown and D. Botstein, 2000 Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA **97:** 10101–10106.

Arbeitman, M. N., E. E. Furlong, F. Imam, E. Johnson, B. H. Null *et al.*, 2002 Gene expression during the life cycle of Drosophila melanogaster. Science **297:** 2270–2275.

Brem, R. B., G. Yvert, R. Clinton and L. Kruglyak, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science **296:** 752–755.

Caron, H., B. van Schaik, M. van der Mee, F. Baas, G. Riggins *et al.*, 2001 The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science **291:** 1289–1292.

Dumas, P., Y. Sun, G. Corbeil, S. Tremblay, Z. Pausova *et al.*, 2000 Mapping of quantitative trait loci (QTL) of differential stress gene expression in rat recombinant inbred strains. J. Hypertens. **18:** 545–551.

Eaves, I. A., L. S. Wicker, G. Ghandour, P. A. Lyons, L. B. Peterson *et al.*, 2002 Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes. Genome Res. **12:** 232–243.

Emahazion, T., L. Feuk, M. Jobs, S. L. Sawyer, D. Fredman *et al.*, 2001 SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. Trends Genet. **17:** 407–413.

Esposito-Vinci, V., and M. Tenenhaus, 2001 PLS logistic regression, pp. 117–130 in *PLS and Related Methods, Proceedings of the PLS01 International Symposium*, edited by V. Esposito-Vinci, C. Lauro, A. Morineau and M. Tenenhaus. CSIA-CERESTA, Paris.

Frank, I. E., and J. H. Friedman, 1993 A statistical view of some chemometrics regression tools. Technometrics **35:** 109–135.

Gianola, D., M. Perez-Enciso and M. A. Toro, 2003 On marker-assisted prediction of genetic value: beyond the ridge. Genetics **163:** 347–365.

Gruvberger, S., M. Ringner, Y. Chen, S. Panavally, L. H. Saal *et al.*, 2001 Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res. **61:** 5979–5984.

Hastie, T., R. Tibshirani and J. H. Friedman, 2001 *The Elements of Statistical Learning*. Springer Verlag, New York.

Holter, N. S., M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar *et al.*, 2000 Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc. Natl. Acad. Sci. USA **97:** 8409–8414.

Holter, N. S., A. Maritan, M. Cieplak, N. V. Fedoroff and J. R.

Banavar, 2001 Dynamic modeling of gene expression data. Proc. Natl. Acad. Sci. USA **98:** 1693–1698.

Hosmer, D. W., and S. Lemeshow, 2000 *Applied Logistic Regression*. John Wiley & Sons, New York.

Jansen, R. C., and J. Nap, 2001 Genetical genomics: the added value from segregation. Trends Genet. **17:** 388–391.

Khan, J., J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi *et al.*, 2001 Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. **7:** 673–679.

Knudsen, S., 2002 *A Biologist's Guide to Analysis of DNA Microarray Data*. John Wiley & Sons, New York.

Lander, E. S., 1999 Array of hope. Nat. Genet. **21:** 3–4.

McPeek, M. S., and A. Strahs, 1999 Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine scale genetic mapping. Am. J. Hum. Genet. **65:** 858–875.

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

Nguyen, D. V., A. B. Arpat, N. Wang and R. J. Carroll, 2002 DNA microarray experiments: biological and technological aspects. Biometrics **58:** 701–717.

Nordborg, M., and S. Tavaré, 2002 Linkage disequilibrium: what history has to tell us. Trends Genet. **18:** 83–90.

Pérez-Enciso, M., and M. Tenenhaus, 2003 Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. Hum. Genet. **112:** 581–592.

Perou, C. M., T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey *et al.*, 2000 Molecular portraits of human breast tumours. Nature **406:** 747–752.

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **422:** 297–302.

Sorlie, T., C. M. Perou, R. Tibshirani, T. Aas, S. Geisler *et al.*, 2001 Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA **98:** 10869–10874.

Storey, J., and R. Tibshirani, 2003 SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays, pp. 320–346 in *The Analysis of Gene Expression Data: Methods and Software*, edited by G. Parmigiani, E. Garrett, R. Irizarry and S. Zeger. Springer Verlag, New York.

Suh, E. Y., and D. W. Schaferb, 2002 Semiparametric maximum likelihood for nonlinear regression with measurement errors. Biometrics **58:** 448–453.

Tenenhaus, M., 1998 *La Régression PLS*. Editions Technip, Paris.

Umetrics, 2001 *SIMCA-P9*. Umetrics, Umea, Sweden.

West, M., C. Blanchette, H. Dressman, E. Huang, S. Ishida *et al.*, 2001 Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl. Acad. Sci. USA **98:** 11462–11467.

Wold, S., H. Martens and H. Wold, 1983 The multivariate calibration problem in chemistry solved by the PLS method, pp. 286–293 in *Proceedings of the Conference on Matrix Pencils*, edited by A. Ruhe and B. Kagstrom. Springer Verlag, Heidelberg, Germany.

Communicating editor: J. B. Walsh