BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 4 2014, pages 514–522
doi:10.1093/bioinformatics/btt708

Gene expression

Advance Access publication December 5, 2013

Chemical structure informing statistical hypothesis testing in metabolomics

Hongjie Zhu[1,2,*,†] and Man Luo[3,4,†]

[1]Department of Biostatistics and Programming, Sanofi, Bridgewater, NJ 08807, USA, [2]Department of Psychiatry and Behavioral Sciences, Duke University, Durham, NC 27710, USA, [3]Exploratory Clinical & Translational Research, Bristol-Myers Squibb, Princeton, NJ 08543, USA and [4]Center for Human Health Assessment, The Hamner Institutes for Health Sciences, Durham, NC 27709, USA

Associate Editor: John Hancock

ABSTRACT

Motivation: Metabolomics has been shown as an effective tool to study various biological and biomedical phenotypes, whereas interrogating the inherently noisy metabolite concentration data with limited sample size remains a major challenge. Accumulating evidence suggests that metabolites' structures are relevant to their bioactivities.

Results: We present a new strategy to boost the statistical power of hypothesis testing in metabolomics by incorporating quantitative molecular descriptors for each metabolite. The strategy selects potentially informative summary molecular descriptors and outputs chemical structure-informed false discovery rates. The effectiveness of the proposed strategy is demonstrated by both simulation studies and a real application. In a metabolomic study on Alzheimer's disease, the posterior inclusion probability for summary molecular descriptors reaches 0.97. By incorporating the structure data, our approach uniquely identifies multiple Alzheimer's disease signatures, which are consistent with existing evidence. These results evidently suggest the value of the proposed approach for metabolomic hypothesis-testing problems.

Availability and implementation: A code package implementing the strategy is freely available at https://github.com/HongjieZhu/CIMA.git.

Contact: hongjie.zhu@sanofi.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 25, 2013; revised on November 11, 2013; accepted on December 2, 2013

1 INTRODUCTION

Over the last decades, major advances in analytical chemistry have resulted in the emergence of metabolomics. It enables simultaneous quantifying of hundreds to thousands of metabolites present in a set of biological samples, e.g. plasma, urine and cerebrospinal fluid, to report on metabolic pattern associated with conditions of interest, such as disease or drug exposure (Kaddurah-Daouk *et al*., 2008; Patti *et al*., 2012). The identities and concentrations of metabolites, or the so-called metabotype, reflect net interactions between gene and environment, providing information that can possibly bridge a gap between genotype and phenotype.

Attesting to this belief, metabolomics has been widely used to understand disease pathogenesis and drug effects, as well as to predict variation in drug response, among many other applications (Corona *et al*., 2012; Kaddurah-Daouk *et al*., 2008; Mamas *et al*., 2011). These studies typically involve identifying 'metabolomic signatures' among the measured metabolites: examples are metabolites that are influenced by an environmental stimulus, e.g. drug treatment, and metabolites that are associated with a phenotype of interest, e.g. a disease status or drug response. Common metabolomics data analysis practice uses conventional statistical inference tools, such as Student's *t*-tests and regression techniques, to identify the signatures. These methods, as well as many multivariate chemometrics and statistics tools (Korman *et al*., 2012; Lindon *et al*., 2007) that are frequently used, essentially treat metabolites as individual variables instead of biological entities, of which rich prior knowledge has been accumulated and is accessible from literatures and/or databases. A variety of methodologies have been developed recently to take advantages of such information to facilitate statistical inference on omic data: some incorporate network relationships among the biological entities (Li and Li, 2008; Wei and Pan, 2008), based on a general assumption that the connected ones on the network tend to share similar behavior, whereas others perform selection at the level of pathways (Ramanan *et al*., 2012; Zhu and Li, 2011), acknowledging the modularity of biological networks.

A distinctive attribute of metabolites is that each of them can be characterized by its unique chemical features. These features can be quantified by a variety of molecular descriptors (MD) (Todeschini and Consonni, 2009), which transform chemical structures into numbers (Supplementary Fig. S1). Different types of MDs can be obtained from existing software, e.g. Dragon (Talete SRL, Milan, Italy), ranging from simple molecular properties, e.g. molecular weight and polarity, to elaborative spatial formulations, e.g. 3D autocorrelations and atom pairs, to complex molecular fingerprints, e.g. Daylight (http://www.daylight.com) and UNITY 2D fingerprints (http://www.tripos.com). It has been acknowledged that many physicochemical or pharmacological properties of metabolites could be revealed by their structure information. For instance, medicinal chemists use the structure–activity relationships (Carbó-Dorca *et al*., 2001) identified from a series of compounds targeting a specific binding protein to design new drugs with better bioactivity (Patani and LaVoie, 1996); toxicologists predict a specific toxicity profile of a

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

compound based on its similar key structural moiety with known toxins (Cronin and Dearden, 1995; Nelson, 2001). It has also been observed that the levels of metabolites sharing the same or similar key chemical structures are more likely to be affected together under certain environmental stimulus. For example, serotonin (5-HT) and dopamine (DA) are both monoamine neurotransmitters, structurally similar to each other, but belong to different pathways. Under the stimulus of 3,4-methylenedioxymethamphetamine, which interacts with both 5-HT and DA receptors, the levels of these two metabolites are both upregulated in the brain, leading to the hallucinogenic effect (Capela *et al*., 2009).

Because the function and bioactivity of metabolites are closely tied to their structure, we hypothesize that integrating the structure information into the analysis of metabolite concentration data can significantly improve the discovery of metabolomic signatures. To evaluate this hypothesis, we build a regression normal mixture (RNM) model that can merge information from the two data sources, metabolites' concentration data and their chemical structure data, into the estimation of false discovery rate (FDR), which is nowadays routinely used to control for multiple testing in metabolomics studies. As an illustrative example, suppose a total of m metabolites are measured for a cohort of patients and healthy controls to evaluate the following *null* hypotheses: H_i: level of the i-th metabolite is not different between patients and controls, $i = 1, \dots, m$. Depending on the actual statistical method being used to test the hypotheses, different types of summary statistics can be derived for each metabolite, e.g. a t-statistic from a parametric two-sample t-test, or a sum of ranks from a non-parametric Wilcoxon rank sum test, along with a P-value. To control for multiple testing, the summary test statistics can be fitted to a mixture model (Efron and Tibshirani, 2002). The model estimates for each metabolite a Bayesian posterior probability of its *null* hypothesis being true, based on which various types of FDRs can be calculated. A standard mixture model assumes the same prior belief/probability of each metabolite being *null*. Variations have been proposed to incorporate additional information into the specification of the prior probabilities. One example is the class of mixtures-of-experts models (McLachlan and Peel, 2000), in which the prior probabilities are modeled as functions of some covariates. Wei and Pan (2008) used a spatially correlated mixture model that incorporates network relationships among genes into the specification of prior probabilities for microarray studies. The RNM model in our strategy uses several summary molecular descriptors (SMD) to estimate metabolite-specific prior probabilities. We further perform variable selection on the SMDs, which can suggest how likely they are informative for a particular metabolomic study. In general, the selected structural information and the test results based on the concentration data are jointly used to output chemical structure-informed FDR for the evaluation of the importance of metabolites.

Based on the RNM model, we propose an integrated chemical structure-informed metabolomic hypothesis-testing strategy (Fig. 1). Simulation studies demonstrate the effectiveness of the RNM model both in identifying informative SMD(s) and in improving the identification of metabolomic signatures with the SMDs. As proof-of-concept, we applied the proposed strategy to a metabolomic study of Alzheimer's disease (AD). A high posterior inclusion probability of the first SMD being selected and the comprehensive biological evidence that supports the identified AD signatures unique to our approach both suggest the potential benefit of incorporating structure information into the analysis of metabolomic data.

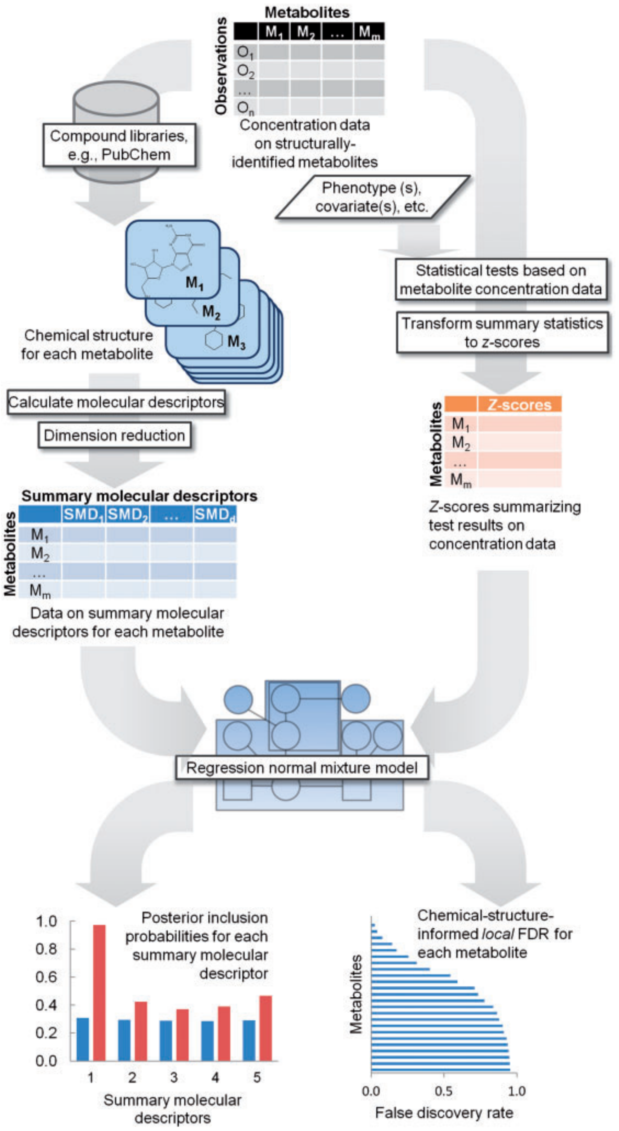
Metabolites
M1 M2 ... Mm
Observations
O1
O2
...
On
Concentration data on structurally-identified metabolites
Compound libraries, e.g., PubChem
Phenotype (s), covariate(s), etc.
M1
M2
M3
Chemical structure for each metabolite
Statistical tests based on metabolite concentration data
Transform summary statistics to z-scores
Calculate molecular descriptors
Dimension reduction
Z-scores
Metabolites
M1
M2
...
Mm
Z-scores summarizing test results on concentration data
Summary molecular descriptors
SMD1 SMD2 ... SMDd
Metabolites
M1
M2
...
Mm
Data on summary molecular descriptors for each metabolite
Regression normal mixture model
1.0
0.8
0.6
0.4
0.2
0.0
Posterior inclusion probabilities for each summary molecular descriptor
1 2 3 4 5
Summary molecular descriptors
Chemical-structure-informed local FDR for each metabolite
Metabolites
0.0 0.5 1.0
False discovery rate

Fig. 1. Illustration of the chemical structure-informed metabolomic hypothesis-testing strategy

2 METHODS

We begin this section with an introduction of quantifying and summarizing the structure information of metabolites with the SMDs, and then discuss the RNM model that incorporates the SMDs into metabolomic hypothesis testing. We conclude this section by a summary of the proposed strategy illustrated in Figure 1.

2.1 Obtaining and summarizing molecular descriptors

To quantify the structures of metabolites, we first retrieve their canonical simplified molecular-input line-entry system (SMILES) strings from online compound libraries, e.g. PubChem (Wang *et al.*, 2009), and then calculate a wide range of MDs for each metabolite from Dragon Software (version 5.5). Given the relatively high dimensionality of the MD data and strong correlations among some MDs, we use principal component analysis (PCA) to generate lower-dimensional projections that can retain major variance within the MD data. Because the MDs are typically of different scales, we standardize them before running PCA. The top d principal components that capture the largest variance are retained as SMDs.

2.2 The regression normal mixture model

The RNM model uses the SMDs to facilitate the estimation of FDR for metabolomic hypothesis testing, which used to be evaluated with the summary test statistics derived from metabolite concentration data alone. Many research problems in metabolomics fall into the hypothesis-testing class. Examples include testing which metabolites are influenced by a drug treatment, or which metabolites are associated with a clinical phenotype. In such cases, each metabolite under study is assumed to be in either of the following two states: the *null* hypothesis is true or it is not true (*non-null*). We refer to them as *null* and *non-null* metabolites, respectively.

The RNM model adopts a finite mixture model framework (McLachlan and Peel, 2000), with which Efron and Tibshirani (2002) developed empirical Bayesian approaches to deal with large-scale simultaneous inference problems in microarray studies. McLachlan *et al.* (2006) proposed a simple parametric implementation, namely, a two-component standard normal mixture (SNM) model, based on normal transformation to the original test statistics of metabolites. Specifically, a z-score can be calculated for a metabolite i from its P-value: $z_i = \Phi^{-1}(1-p_i)$, where Φ represents the cumulative standard normal distribution. Consistent with the P-value, the z-score for a metabolite reflects its statistical significance. Therefore, it is reasonable to assume that the z-scores of the *null* and *non-null* metabolites follow distinct distributions. We use an indicator variable C_i to denote the states of metabolites: $C_i = 0$ ($C_i = 1$) corresponds to that the *null* (*non-null*) hypothesis holds for the i-th metabolite. Denoting the prior probability of a metabolite being *null* (*non-null*) as π_0 (π_1), i.e. $Pr(C_i = 0) = \pi_0$ ($Pr[C_i = 1] = \pi_1$), we can represent the density of the z-score for the i-th metabolite using an SNM model with two components corresponding to the two states:

$$f(z_i) = \pi_0 f_0(z_i) + \pi_1 f_1(z_i) \tag{1}$$

where f_0 and f_1 represent the *null* and *non-null* densities of z-score, respectively. The posterior probability of the metabolite i being *null* given its z-score is as follows:

$$\tau_0(z_i) = Pr(C_i = 0|z_i) = \frac{\pi_0 f_0(z_i)}{f(z_i)} = 1 - \frac{\pi_1 f_1(z_i)}{f(z_i)}$$

This posterior probability is the *local* FDR defined by Efron and Tibshirani (2002), which has been a widely used variant of Benjamini and Hochberg (1995) *tail-area* FDR. The *tail-area* FDR is a *global* measure of FDR for a given rejection region, e.g. $Z \in [\lambda, +\infty)$, where Z is the random z-score and λ is a given cutoff. Corresponding to the i-th metabolite, the *tail-area* FDR for rejecting $Z \geq z_i$ has a posterior probability interpretation of $Pr(null|Z \geq z_i)$, which can be shown equal to $E_f[\tau_0(Z)|Z \geq z_i]$ (Efron, 2005), where E_f indicates expectation with respect to $f(Z)$. Accordingly, the *tail-area* FDR corresponding to the i-th metabolite can be estimated by the average (estimated) *local* FDR of the metabolites whose z-scores are not less than z_i. In the usual situation where $\tau_0(Z)$ decreases as z-score increases, the *tail-area* FDR will be less than the *local* FDR. Besides, the *local* FDR has an advantage in evaluating individual behavior of metabolites (Efron, 2005).

In contrast to the conventional SNM model that assumes the same prior probabilities of being *null* to all the metabolites, the RNM model fits metabolite-specific prior probabilities based on the SMDs that summarize the chemical structures of all the metabolites under study. Corresponding to the SNM model in (1), a two-component RNM model for $f(z_i)$ is as follows:

$$f(z_i) = \pi_{i,0} f_0(z_i) + \pi_{i,1} f_1(z_i) \tag{2}$$

where $\pi_{i,0}$ and $\pi_{i,1}$ represent the metabolite-specific prior probabilities for metabolite i. We estimate them with the SMDs through a logit link model:

$$logit(\pi_{i,1}) = \log(\frac{\pi_{i,1}}{\pi_{i,0}}) = \beta_0 + \sum_{j=1}^{d} \beta_j D_{i,j}$$

where D is the design matrix of the SMDs with $D_{i,j}$ representing the value of the j-th SMD for the metabolite i, and β_0 and $\beta = (\beta_1, \ldots, \beta_d)$ are unknown regression intercept and coefficient parameters, respectively. Based on (2), the posterior probability of the metabolite i being *null* is

$$Pr(C_i = 0|D, z_i) = \frac{\pi_{i,0} f_0(z_i)}{f(z_i)} = 1 - \frac{\pi_{i,1} f_1(z_i)}{f(z_i)}$$

which we call the chemical structure-informed *local* FDR. Because the prior probabilities are now unique to each individual metabolite, it is no longer meaningful to calculate the corresponding *tail-area* FDR.

Two-sided tests are frequently used in discovery studies. In such cases, the *non-null* metabolites can be further divided into two substates denoted here as negative *non-null* and positive *non-null*. Applying the two-component RNM model essentially assumes the same effect of an SMD to the prior probabilities of being positive and negative *non-null*, which may not be conceivable. To relief this assumption, it is intuitive to distinguish the two *non-null* substates in the modeling. The indicator variable C_i now takes three distinct values: $C_i = 0$ for *null*, $C_i = 1$ for negative *non-null* and $C_i = 2$ for positive *non-null*. The calculation of z-score should also reflect the direction of the original test results. For instance, when two-sample t-tests are used to evaluate which metabolites are different between patients and controls, the z-scores can be defined as $z_i = \text{sgn}(t_i)\Phi^{-1}(1 - p_i/2)$, where t_i is t-statistic derived for the i-th metabolite. With this definition, either an extreme negative or extreme positive z-score indicates a departure from the *null* hypothesis. We use a three-component RNM model for the density of such z-scores:

$$f(z_i) = \pi_{i,0} f_0(z_i) + \pi_{i,1} f_1(z_i) + \pi_{i,2} f_2(z_i)$$

or more specifically in our application,

$$f(z_i) = \pi_{i,0}\phi(z_i; 0, \sigma_0^2) + \pi_{i,1}\phi(z_i; \mu_1, \sigma_1^2) + \pi_{i,2}\phi(z_i; \mu_2, \sigma_2^2)$$

where the last two components correspond to the negative and positive *non-null*, respectively, with $\mu_1 < 0$ and $\mu_2 > 0$. Accordingly, we estimate the log-ratios of prior probabilities between each *non-null* state and *null* state with the SMDs using the following link models:

$$\begin{aligned} \log\left(\tfrac{\pi_{i,1}}{\pi_{i,0}}\right) &= \beta_{1,0} + \sum_{j=1}^{d} \beta_{1,j} D_{i,j} \\ \log\left(\tfrac{\pi_{i,2}}{\pi_{i,0}}\right) &= \beta_{2,0} + \sum_{j=1}^{d} \beta_{2,j} D_{i,j} \end{aligned} \tag{3}$$

The chemical structure-informed *local* FDRs for the i-th metabolite are as follows: $1 - \pi_1 f_1(z_i)/f(z_i)$ for claiming it being negative *non-null* and $1 - \pi_2 f_2(z_i)/f(z_i)$ for claiming it being positive *non-null*.

Because all the SMDs may not be informative, we further impose random effect stochastic search variable selection (Meuwissen and Goddard, 2004) on them to inform their relevance to the prior probabilities for the *null* and *non-null* states of metabolites. To save space, we only elaborate this for the three-component RNM model, and it is a trivial task to modify for the two-component model. We assume the intercept and coefficient parameters for all the SMDs are independent, and let $I_{k,j}$

be an indicator for the j-th SMD being selected ($I_{k,j} = 1$) or not ($I_{k,j} = 0$) in the k-th model in (3). The following mixture prior for $\beta_{k,j}, k = 1, 2; j = 1, \ldots, d$ is used:

$$f(\beta_{k,j}|I_{k,j}, \sigma_\beta^2) = (1 - I_{k,j})\phi(\beta_{k,j}; 0, g\sigma_\beta^2) + I_{k,j}\phi(\beta_{k,j}; 0, \sigma_\beta^2)$$

where g is fixed to be a small positive number (0.001 in our study), $\sigma_\beta^2 \sim \text{Uniform}(0, 20)$; $I_{k,j}|p_I \sim \text{Bernoulli}(p_I)$; $p_I \sim \text{Uniform}(0, 1)$; and $\beta_{k,0} \sim N(0, 10^2)$, for $k = 1, 2$. Under these specifications, the marginal prior inclusion probability of each SMD being selected equals to 0.5 (Supplementary Note S1).

We mainly follow Wei and Pan (2008)'s prior specifications and use vague priors for the rest parameters in the RNM model: $\mu_1 \sim N(0, 10^6)I(a, 0)$, a truncated normal distribution between $a = \min_i z_i$ and 0; $\mu_2 \sim N(0, 10^6)I(0, b)$, $b = \max_i z_i$; and $\sigma_l^2 \sim \text{InverseGamma}(0.1, 0.1)$, for $l = 0, 1, 2$. The same specifications are also adopted for a three-component SNM model applied in the simulation and the AD studies.

Supplementary Figure S2 shows a graphical representation of the three-component RNM model. The full specification of the model is given in Supplementary Note S2. The RNM model can be readily implemented in WinBUGS (Lunn *et al.*, 2000) and fitted with Markov chain Monte Carlo (MCMC) algorithms. To assess the convergence of the MCMC process and determine the number of burn-ins, we initiate two chains, one with all the SMDs initially selected and the other with none selected, and examine the trace plots and the potential scale reduction factors (Gelman and Rubin, 1992) for μ_ls, σ_ls, σ_β, $I_{k,j}$s and C_is.

As primary indices of interest, the posterior inclusion probabilities of the SMDs can be estimated from the MCMC sample mean of $I_{k,j}$s. The estimated inclusion probabilities for an SMD measure how often the SMD is selected into the modeling of the prior probabilities for the states of metabolites, and thus reflects how much the input data, namely, the SMD data and metabolites' z-scores, favors the relevance of the SMD to the states of metabolites. The posterior probabilities of a metabolite being in *null*/*non-null* states can be estimated from the MCMC sample mean of C_is, from which its corresponding *local* FDRs can be estimated as described earlier in the text.

2.3 The chemical structure-informed hypothesis-testing strategy

Figure 1 illustrates the workflow of the entire strategy. To address a research question of interest, one applies appropriate conventional analysis, e.g. linear regression, to metabolite concentration data and other data as needed, e.g. phenotypes, covariates. The analysis gives test statistics and P-values for metabolites under study, which can be transformed to z-scores (Section 2.2). In parallel, one can obtain chemical structures of metabolites by searching their names in the chemical libraries, and then calculate and summarize the MDs for all the metabolites (Section 2.1). The RNM model takes the two sources of data as inputs and generates two outputs: the chemical structure-informed FDRs of the metabolites and the posterior inclusion probabilities of the SMDs. A high posterior inclusion probability (e.g. >0.8) indicates the SMD's relevance to the states of metabolites, and the chemical structure-informed FDR is likely to outperform the conventional estimation of FDR without incorporating the structural information.

3 RESULTS

We evaluated the proposed chemical structure-informed hypothesis-testing strategy at two levels. First, we performed simulation studies to examine the effectiveness of the RNM model on identifying potentially informative SMD(s) and *non-null* metabolites. Second, we applied the strategy to a real study seeking new metabolomic signatures for AD.

3.1 Simulation studies

We simulated diverse scenarios to evaluate the RNM model that incorporates the SMDs into the estimation of *local* FDR by varying the following factors: the total number of metabolites, the percentages of *non-null* metabolites, the total number of SMDs and the number and effect sizes of informative SMDs. We examined the posterior probabilities of the informative SMD(s) being selected and compared them with those of the non-informative ones. We then performed Receiver-Operating Characteristics curve (ROC) analysis to evaluate its performance on identifying *non-null* metabolites.

3.1.1 Simulation setup Settings for different simulation scenarios are shown in Table 1. In each scenario, the null/non-null state of the i-th metabolite was generated based on the following values: $S_{1,i} = \sum_{j=1}^{d} \beta_{1,j}D_{i,j} + e_{1,i}$ and $S_{2,i} = \sum_{j=1}^{d} \beta_{2,j}D_{i,j} + e_{2,i}$. $D_{i,j}$s were independently and randomly sampled from $N(0, 1)$, and D was then orthogonally transformed to mimic SMDs produced by PCA; $e_{1,i}$ and $e_{2,i}$ are random noises sampled from $N(0, 1)$, which introduce uncertainties in determining the null/non-null states of metabolites. The first m_1 (m_2) metabolites having the largest S_1 (S_2) values were assigned to be negative (positive) non-null. Occasionally, there were metabolites assigned to both states. Such metabolites were assigned to the larger one between S_1 and S_2, and the search continued until all the m_1 and m_2 metabolites were selected. All the rest of metabolites were null. Given the null/non-null states of metabolites, z_i was sampled from

$$f(z_i|C_i) = \begin{cases} \phi(z_i; 0, 1) & \text{if } C_i = 0 \ (\textit{null}) \\ \phi(z_i; -1, 1) & \text{if } C_i = 1 \ (\text{negative } \textit{non} - \textit{null}) \\ \phi(z_i; 1, 1) & \text{if } C_i = 2 \ (\text{positive } \textit{on} - \textit{null}) \end{cases}$$

Applying the RNM model to the simulated SMDs and z-scores, one can estimate the posterior probabilities of each metabolite being in different states. By varying cutoffs on the posterior probabilities, we performed ROC analysis for distinguishing each pair of states and derived the area under the curve (AUC) values, the main index for method evaluation.

To better assess the potential benefit from incorporating structure information with the RNM model, we compared its AUC scores with the optimal AUCs that can be achieved with the z-scores alone. Specifically, for discriminating any pair of states, l_1 versus l_2, the Neyman–Pearson lemma establishes that the true likelihood ratio, $f(z_i|C_i = l_1)/f(z_i|C_i = l_2)$, or any monotone increasing function of it, maximizes the height of ROC all the way along the curve (thus the optimal AUC is achieved) (Eguchi and Copas, 2002). The optimal AUC values of the likelihood ratios can be determined analytically (Faraggi and Reiser, 2002), and are 0.76 for *null* versus either *non-null* state given the current *null* and *non-null* densities of z-score.

3.1.2 Simulation results We ran each simulation scenario 100 times. Inference was made on 10 000 MCMC samples after 20 000 burn-ins with a thinning rate of 10. The potential scale

Table 1. Results of simulation studies

Scenario	Number of metabolites	Class of *non-null* versus *null*	Percentage of *non-null* metabolites	Number of SMDs	Effects of SMDs	Posterior inclusion probabilities of SMDs	AUC
1	$m = 200$	Negative	$m_1/m = 10\%$	5	$\beta_1 = [1, 0, 0, 0, 0]$	[0.51, 0.38, 0.38, 0.38, 0.38]	.84 (.07)
		Positive	$m_2/m = 10\%$		$\beta_2 = [-1, 0, 0, 0, 0]$	[0.51, 0.38, 0.37, 0.39, 0.38]	.84 (.07)
2	$m = 200$	Negative	$m_1/m = 10\%$	5	$\beta_1 = [2, 0, 0, 0, 0]$	[0.63, 0.37, 0.38, 0.38, 0.37]	.91 (.07)
		Positive	$m_2/m = 10\%$		$\beta_2 = [-2, 0, 0, 0, 0]$	[0.64, 0.37, 0.38, 0.38,0.38]	.92 (.06)
3	$m = 200$	Negative	$m_1/m = 10\%$	1	$\beta_1 = [1]$	[0.56]	.87 (.05)
		Positive	$m_2/m = 10\%$		$\beta_2 = [-1]$	[0.56]	.88 (.06)
4	$m = 200$	Negative	$m_1/m = 10\%$	5	$\beta_1 = [1, -1, 0, 0, 0]$	[0.52, 0.49, 0.40, 0.41, 0.40]	.86 (.07)
		Positive	$m_2/m = 10\%$		$\beta_2 = [-1, 1, 0, 0, 0]$	[0.51, 0.50, 0.39, 0.40, 0.40]	.87 (.06)
5	$m = 400$	Negative	$m_1/m = 10\%$	5	$\beta_1 = [1, 0, 0, 0, 0]$	[0.68, 0.35, 0.35, 0.37, 0.36]	.86 (.05)
		Positive	$m_2/m = 10\%$		$\beta_2 = [-1, 0, 0, 0, 0]$	[0.64, 0.33, 0.34, 0.33, 0.33]	.86 (.06)
6	$m = 200$	Negative	$m_1/m = 20\%$	5	$\beta_1 = [1, 0, 0, 0, 0]$	[0.72, 0.42, 0.41, 0.42, 0.43]	.85 (.06)
		Positive	$m_2/m = 5\%$		$\beta_2 = [-1, 0, 0, 0, 0]$	[0.43, 0.34, 0.34, 0.35, 0.34]	.83 (.08)
7	$m = 200$	Negative	$m_1/m = 10\%$	5	$\beta_1 = [0, 0, 0, 0, 0]$	[0.36, 0.35, 0.35, 0.36, 0.36]	.65 (.07)
		Positive	$m_2/m = 10\%$		$\beta_2 = [0, 0, 0, 0, 0]$	[0.35, 0.35, 0.35, 0.35, 0.36]	.66 (.06)

Note: The last two columns give average values (over 100 simulations) of the posterior inclusion probabilities of SMDs and the AUC values for distinguishing *non-null* from *null* metabolites. The standard deviations of AUCs are given in parentheses. For comparison, the AUCs that can be achieved with the true likelihood ratios based on *z*-scores alone are 0.76 for *null* versus either *non-null* state.

reduction factors for the parameters under examination are all below 1.1 (results not shown), indicating that there is no suggested evidence of lack of convergence. Simulation results are given in Table 1.

Each run of Scenario 1 simulates 200 metabolites, and the percentages of positive and negative *non-null* metabolites are both 10%. One of the five SMDs is related to the *null*/*non-null* states of metabolites. Results show that the posterior inclusion probabilities for the informative SMD are 0.1 over those of the non-informative ones. Correspondingly, the AUCs of incorporating structure information with the RNM model are clearly higher than those of the true likelihood ratio on *z*-scores. It is not surprising to see that the advantage of incorporating structure information further increases with the following changes to the simulation settings: increasing the effect size of the informative SMD (Scenario 2), reducing the number of non-informative SMDs (Scenario 3), increasing the number of informative SMDs (Scenario 4) or increasing the total number of metabolites (Scenario 5). In all of these studies, the informative SMD(s) has notably higher posterior inclusion probabilities than the non-informative ones, and the difference is enlarged with increased effect size of the informative SMD (Scenario 2) and/or increased number of metabolites (Scenario 5). In Scenario 6, the percentages of negative and positive *non-null* become 20% and 5%, respectively. By comparing Scenario 6 with Scenario 1, one can see that the posterior inclusion probability for the informative SMD (and its difference from those of the non-informative ones) also varies positively with the percentage of *non-null* metabolites. When none of the SMDs is informative (Scenario 7), the posterior inclusion probabilities for the SMDs are all around $0.35 \sim 0.36$. Correspondingly, incorporating these SMDs does not show advantage in identifying *non-null* metabolites. In metabolomics studies, metabolites are often found correlated. To examine the behavior of the RNM model in this situation, we further divided the 200 metabolites into 10 groups in Scenario 3, and let pairwise correlations of the SMDs ($D_{i,j}$'s) for metabolites within each group be 0.5 (compound symmetric structure). Additionally, to simulate correlation among metabolites that is not related to structural similarity, we adopted the same correlation structure for the random noises ($e_{1,i}$ and $e_{2,i}$). Results show that the AUCs of the RNM model remain the same (0.87 and 0.88), indicating the model is robust to correlation among metabolites. Finally, to examine whether the proposed strategy will inflate the type I error, we simulated a scenario where there are no true *non-null* metabolites. We used the same simulation settings for Scenario 3, except that only the 160 *null* metabolites were analyzed with (the RNM model) and without (the SNM model) incorporating structure information. Supplementary Figure S3 gives the distribution of the estimated posterior probabilities of being *non-null* for the 160 metabolites over 100 runs. Compared with the SNM model, the RNM model in general gives lower posterior probabilities for the metabolites being *non-null*. In other words, in the case of no *non-null* metabolites, incorporating molecular structure information with the proposed approach leads to higher estimated FDR than the conventional approach.

In general, the simulation studies demonstrate that posterior inclusion probabilities for the SMDs can help identify informative SMD(s) and indicate whether a particular study can benefit from adopting the proposed approach. When informative SMDs exist, the RNM model is able to make use of such information to improve the discrimination between *null* and *non-null* metabolites.

3.2 Metabolomic signatures for Alzheimer's disease

We then applied the proposed approach to a clinical metabolomic study of AD (Kaddurah-Daouk *et al.*, 2013; Motsinger-Reif *et al.*, 2013). The AD is a well-known neurodegenerative disorder and a leading cause of dementia with currently no effective cure or preventive therapy (http://www.alz.org). Our goal is to identify metabolites that have different levels between AD patients and cognitively normal (CN) participants, which may lead to new biomarkers and provide novel diagnostic and therapeutic insights. Cerebrospinal fluid samples from 40 AD and 38 CN participants were profiled with two platforms, a liquid chromatography electrochemical array and a gas chromatography time-of-flight mass spectrometer. A total of 121 metabolites was structurally identified (Supplementary Table S1). The *null* hypothesis for a metabolite is that there is no difference in its level between AD and CN groups, denoted by $AD = CN$, whereas the positive *non-null* (negative *non-null*) is that its level is higher (lower) in the AD group than CN, denoted by $AD > CN$ ($AD < CN$). Levels of metabolites were adjusted for the use of two AD treatment drugs (binary variables) by building a linear regression model for each metabolite, and the residuals are compared between AD and CN with Wilcoxon rank sum tests. A z-score was calculated for each metabolite from its P-value and Hodges–Lehmann estimation of the difference between AD and CN subjects. To implement the proposed strategy, we calculated 882 MDs from Dragon software, which involve (but not limited to) ring descriptors, topological indices, walk and path counts, connectivity indices and geometrical descriptors. The MDs with zero variance for the metabolites under study were removed. One from each pair of highly correlated MDs (correlation coefficient ≥ 0.95) was removed. After these steps, 202 MDs remained (Supplementary Table S2), and the top five SMDs were retained. They capture 53.5% of the total variance and the first SMD captures 21.4%. The SMDs and the z-scores derived earlier in the text were then fitted to a three-component RNM model to estimate *local* FDR of metabolites as described in Section 2. For comparison purpose, we also applied a three-component SNM model not incorporating the structure information to the z-scores, which has the same specification for the *null*/*non-null* densities for z_i and prior distributions for the mean and variance parameters as the RNM model.

Results show that the posterior inclusion probability for the first SMD in the modeling of the prior probability ratio between $AD > CN$ and $AD = CN$ reaches 0.97, and is much higher than those for the rest SMDs (Fig. 2). This suggests that the structure data are highly likely to be informative for the identification of metabolites that are increased in AD. The top 10 MDs with the highest contribution to this SMD are given in Supplementary Table S3. Table 2 provides results for the metabolites that are given the lowest (≤ 0.05) chemical structure-informed *local* FDR. A histogram for the posterior probabilities of being *non-null* for all the metabolites is available in Supplementary Figure S4. The average posterior probabilities for $AD < CN$ and $AD > CN$ among all the metabolites given by the proposed strategy are 0.03 and 0.40, respectively, lower than those (0.09 and 0.52) given by the conventional SNM model not incorporating structure information. Compared with the conventional approach, our approach promotes multiple metabolites that are mapped to two key neurotransmitter pathways: the purine and tryptophan metabolism pathways (Fig. 3). Also promoted are three carbohydrates and the pseudouridine.

In the purine pathway, guanosine (GR) and inosine (IN) are promoted by the proposed approaches, whereas xanthosine (XANTH) ranks highest by both the proposed and conventional approaches. As shown in Figure 3a, cyclic adenosine monophosphate (cAMP) and cyclic guanosine monophosphate (cGMP) are hydrolyzed by phosphodiesterase, whose expression and activity has been shown upregulated in AD patients (Domek-Lopacinska and Strosznajder, 2010), leading to decreased cGMP and cAMP and increased GR, IN, and XANTH. The decreased level of cGMP and cAMP, two important secondary messengers, is one of the key reasons for early AD symptoms, such as memory loss and poor judgment.

The proposed approach also highlights the abnormalities of tryptophan (TRP), 5-hydroxytryptophan (5-HTP), 5-hydroxyindoleacetic acid (5-HIAA), kynurenine (KYN) and indole-3-acetic acid (I-3-AA) in AD, which are all involved in the TRP metabolism pathway. The mechanism underlying the increased concentration of 5-HTP and 5-HIAA may involve the upregulated monoamine oxidase-A (MAO-A) activity in AD, which has shown strong association with increased serotonin deamination (Kumagae *et al.*, 1991), and thus the level of metabolites in the 5-HIAA branch. Recent studies show a markedly increased

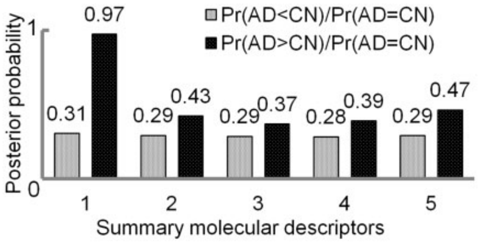

Fig. 2. Posterior inclusion probabilities for the selection of SMDs in the modeling of prior probability ratios between *non-null* and *null* states of metabolites in the AD study

Table 2. Metabolomic differences between AD and CN participants that are given the lowest ($\leq$0.05) chemical structure-informed local FDR

Metabolites	*Non-null* state	SNM-based		Chemical structure-informed	
		Local FDR	Rank	*Local* FDR	Rank
Xanthosine	$AD > CN$	0.11	1	0.00	1
Inosine	$AD > CN$	0.18	12	0.01	2
5-hydroxyindoleacetic acid	$AD > CN$	0.17	10	0.01	3
Guanosine	$AD > CN$	0.36	32	0.02	4
Vanillylmandelic acid	$AD > CN$	0.14	6	0.02	5
Indole-3-acetic acid	$AD > CN$	0.19	13	0.02	6
Glutathione	$AD > CN$	0.11	2	0.03	7
Kynurenine	$AD > CN$	0.16	9	0.03	8
Tryptophan	$AD > CN$	0.35	31	0.03	9
Sucrose	$AD > CN$	0.29	23	0.04	10
Inulobiose	$AD > CN$	0.27	20	0.04	11
Pseudouridine	$AD > CN$	0.50	55	0.05	12
Maltose	$AD > CN$	0.51	56	0.05	13

Note: Also provided are the *local* FDRs estimated by the SNM model not incorporating structure information.

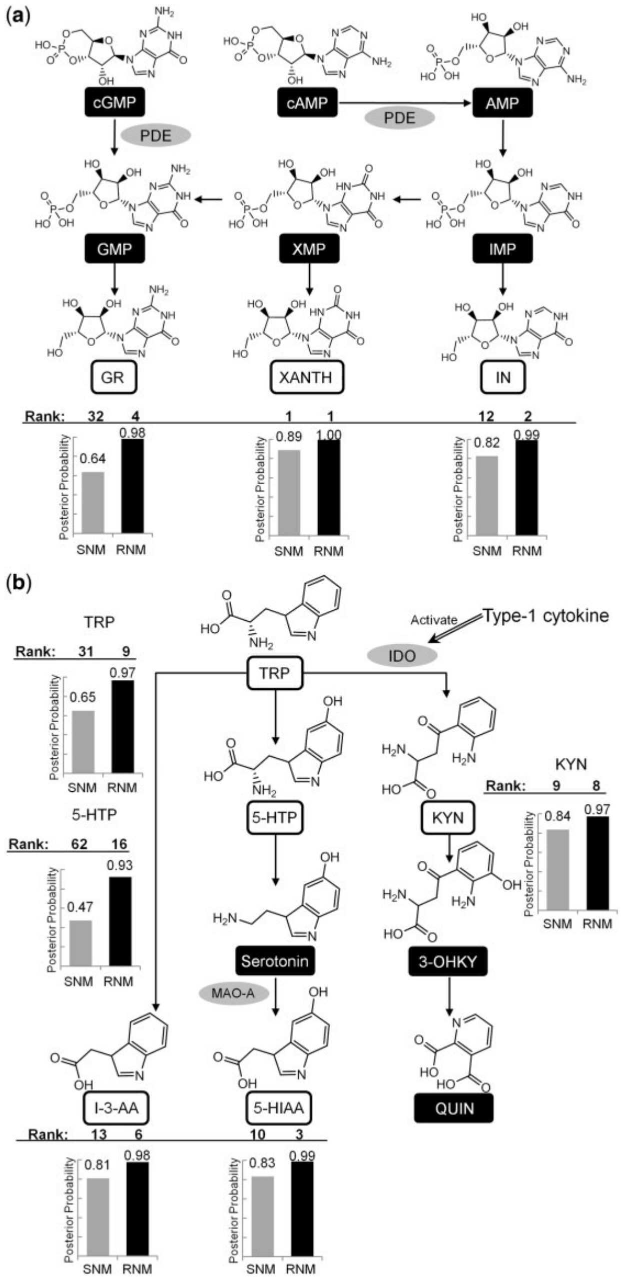

Fig. 3. Metabolites on the purine (**a**) and tryptophan (**b**) metabolism pathways that were found elevated in AD. Dark metabolites are not measured. The bar plots give the posterior probabilities of the metabolites being higher in AD along with their ranks given by the corresponding approaches

concentration of monocyte chemoattractant protein-1, a type 1 cytokine, in AD patients (Zhang *et al.*, 2013), which induces indoleamine-pyrrole 2,3-dioxygenase (IDO) activation and thus upregulates the catabolism of TRP into KYN, 3-hydroxykynurenine (3-OHKY) and quinolinic acid (QUIN). The increased level of I-3-AA might be another pathogenesis of AD due to its cytotoxic effect to neurons: the administration of I-3-AA in pregnant mice has been found to induce neuroepithelium apoptosis and decrease neuron formation in the fetuses (Furukawa *et al.*, 2007). Our findings provide further evidence for the involvement of the tryptophan pathway in elucidating the mechanisms of AD.

Maltose, inulobiose and sucrose are a group of carbohydrates highlighted by the proposed approach. Excess intake of these natural or artificial sweeteners has been shown in mouse models to cause insulin resistance and metabolic alterations (Cao *et al.*, 2007;Carvalho *et al.*, 2012), which play important roles in the exacerbation of oxidative stress, mitochondrial abnormalities and increased amyloid β protein levels in the brain. The correlation between high-carbohydrate diets and AD has been also reported in human studies (Henderson, 2004).

Pseudouridine (Ψ) is the oxidized form of urinary nucleosides (Charette and Gray, 2000). Oxidative damage of RNA plays a critical role in the mechanisms of neurodegenerative disorders, including AD (Shan *et al.*, 2007). The potential mechanism of Ψ-induced brain damage might arise from the additional hydrogen bond in its structure compared with uridine, leading to increased risk of hydrogen bonding with the phosphate of its own or adjacent nucleotides. This structural change on RNA might cause an incorrect translation and thus decreased protein production and function in the brain. In addition, an increased level of pseudouridine in urine samples of AD patients has been reported recently (Lee *et al.*, 2007).

In summary, top metabolomic signatures identified by the proposed approach highlight four mechanisms likely to play critical roles in the etiology of AD. They are either evidenced by *in vivo* animal studies and/or supported by observations in human patients. On one hand, these results reflect the heterogeneous and multifactorial nature of AD; on the other, they show the advantage of the proposed chemical structure-informed hypothesis-testing strategy.

4 DISCUSSION

In this article, we have motivated and presented a new approach featured as chemical structure-informed metabolomic hypothesis testing. Simulation studies suggest that when some of the structure data is relevant to the *null*/*non-null* states of metabolites, the new approach shows clear advantage on the identification of *non-null* metabolites. By implementing the proposed approach to identify metabolomic signatures for AD, we find an SMD being selected with high posterior probability and underscore several pathways or groups of metabolites as potential AD signatures that are well supported by the existing knowledge on the molecular mechanisms of AD. Although further studies are needed to validate the new findings, we think that the joint evidence strongly attests to the advantages of the proposed approach in this study.

The proposed chemical structure-informed approach differs in multiple aspects from the class of network-based approaches mentioned earlier in Section 1, which may make the former more preferable for metabolomics studies. First, owing to their limited analytical capability, many if not the majority of the current metabolomics studies only measure a limited number of structurally identified metabolites. This lack of coverage of metabolic networks typically makes it practically infeasible to apply the network-based approaches. In contrast, the proposed approach is readily applicable to these studies. Additionally, the network-based approaches may miss some coregulation relationships among metabolites that share the same or similar key chemical structures but are not close on the networks, e.g. the coregulation relationship between 5-HT and DA discussed in Section 1. In another aspect, the common assumption made by the network-based approaches that metabolites connected on the networks tend to share similar behavior is often realized by encouraging similarities among parameters/coefficients corresponding to connected variables on the networks. On the contrary, the proposed approach does not simply assume metabolites sharing similar structure would behave similarly; instead, it performs selection on the structure data, so that the most relevant SMDs have more weight in determining the states of metabolites.

Different from the standard mixture models, the RNM model in our approach regresses prior probabilities for metabolites being in different states on several SMDs, where metabolites may be regarded as individual 'observations'. Simulation studies show that increasing the number of metabolites can improve the estimation of the effects of SMDs and thus the states of metabolites. Because SMDs are generated from PCA, the percentage of total variance captured by the top ones is typically considered for choosing the number of SMDs. In practice, however, model complexity is another issue that should be considered, especially when the number of metabolites is limited. In the AD study, we take a conservative number of five SMDs and the first SMD is found most likely to be informative.

The current modeling of prior probabilities for the *null* and *non-null* states of metabolites in (3) considers the states of metabolites as a nominal response variable. However, it is arguably more intuitive to treat it as an ordinal response and use the corresponding link models, e.g. the cumulative link models (Agresti, 2002). Some of them, e.g. the proportional odds models, assume each SMD has the same effect in the modeling of different prior probability ratios. It can be seen that this assumption does not stand in the AD study. The generalized proportional odds logit models allow different effects of the same SMD (Agresti, 2002). Specifically, the ratios of prior probabilities on the left side of (3) become $\pi_{i,1}/(\pi_{i,0}+\pi_{i,2})$ and $\pi_{i,2}/(\pi_{i,0}+\pi_{i,1})$. This way of modeling was implemented in both simulations and the AD study, and it gives similar results (not shown). One can also modify the prior specification for the regression coefficients in (3) to jointly select one SMD in both link models, or to incorporate prior beliefs on the effects of SMDs.

Although the AD study shows favorable evidence for the proposed approach, it is not our intention to claim that it is prescriptive for studies of all different phenotypes. Instead, to inform whether the structure information should be incorporated, the RNM model is accompanied by a selection on the SMDs through stochastic search variable selection. Via MCMC methods, this addition also introduces a Bayesian model averaging (Clyde, 1999) for the estimation of the states of metabolites, whose advantage in estimation accuracy has been well documented (Hoeting *et al.*,1999; Raftery and Zheng, 2003). Besides, the proposed approach is tested with metabolites from two analytical platforms, which only cover a portion of the entire metabolome. Evaluation should be extended to more studies and different classes of metabolites.

Finally, the proposed approach focuses on the general metabolomic hypothesis-testing problem, whereas incorporating structure information into other common metabolomic inference practices, such as multivariate predictive model construction and network inference, is warranted for further research.

ACKNOWLEDGEMENTS

The authors thank the associate editor and the three anonymous referees for their constructive comments, which have greatly helped improve the article. The authors thank Dr Rima Kaddurah-Daouk and the Pharmacometabolomics Research Network for permission to use the metabolomics data. The authors thank Drs Zhaobang Zeng and Hua Zhou, whose comments helped us clarify the approach presented here.

Conflict of Interest: none declared.

REFERENCES

Agresti,A. (2002) *Categorical Data Analysis*. 2nd edn. Wiley, Hoboken, NJ.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, **57**, 289–300.

Cao,D. *et al.* (2007) Intake of sucrose-sweetened water induces insulin resistance and exacerbates memory deficits and amyloidosis in a transgenic mouse model of Alzheimer disease. *J. Biol. Chem.*, **282**, 36275–36282.

Capela,J.P. *et al.* (2009) Molecular and cellular mechanisms of ecstasy-induced neurotoxicity: an overview. *Mol. Neurobiol.*, **39**, 210–271.

Carbó-Dorca,R. *et al.* (2001) *Fundamentals of Molecular Similarity*. Kluwer Academic/Plenum Publishers, New York.

Carvalho,C. *et al.* (2012) Metabolic alterations induced by sucrose intake and Alzheimer's disease promote similar brain mitochondrial abnormalities. *Diabetes*, **61**, 1234–1242.

Charette,M. and Gray,M.W. (2000) Pseudouridine in RNA: what, where, how, and why. *IUBMB Life*, **49**, 341–351.

Clyde,M.A. (1999) Bayesian model averaging and model search strategies. *Bayesian Stat.*, **6**, 157–185.

Corona,G. *et al.* (2012) Pharmaco-metabolomics: an emerging "omics" tool for the personalization of anticancer treatments and identification of new valuable therapeutic targets. *J. Cell. Physiol.*, **227**, 2827–2831.

Cronin,M.T.D. and Dearden,J.C. (1995) Qsar in toxicology.3. prediction of chronic toxicities. *Quant. Struct. –Act. Relatsh.*, **14**, 329–334.

Domek-Lopacinska,K.U. and Strosznajder,J.B. (2010) Cyclic GMP and nitric oxide synthase in aging and Alzheimer's disease. *Mol. Neurobiol.*, **41**, 129–137.

Efron,B. (2005) Local false discovery rates. In: *Technical Report*. Department of Statistics, Stanford University, Stanford, CA.

Efron,B. and Tibshirani,R. (2002) Empirical bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.*, **23**, 70–86.

Eguchi,S. and Copas,J. (2002) A class of logistic-type discriminant functions. *Biometrika*, **89**, 1–22.

Faraggi,D. and Reiser,B. (2002) Estimation of the area under the ROC curve. *Stat. Med.*, **21**, 3093–3106.

Furukawa,S. *et al.* (2007) Indole-3-acetic acid induces microencephaly in mouse fetuses. *Exp. Toxicol. Pathol.*, **59**, 43–52.

Gelman,A. and Rubin,D. (1992) Inference from iterative simulation using multiple sequences. *Stat. Sci.*, **7**, 457–511.

Henderson,S.T. (2004) High carbohydrate diets and Alzheimer's disease. *Med. Hypotheses*, **62**, 689–700.

Hoeting,J.A. *et al.* (1999) Bayesian model averaging: a tutorial. *Stat. Sci.*, **14**, 382–401.

Kaddurah-Daouk,R. *et al.* (2008) Metabolomics: a global biochemical approach to drug response and disease. *Ann. Rev. Pharmacol. Toxicol.*, **48**, 653–683.

Kaddurah-Daouk,R. *et al.* (2013) Alterations in metabolic pathways and networks in Alzheimer's disease. *Transl. Psychiatry*, **3**, e244.

Korman,A. *et al.* (2012) Statistical methods in metabolomics. *Methods Mol. Biol.*, **856**, 381–413.

Kumagae,Y. *et al*. (1991) Deamination of norepinephrine, dopamine, and serotonin by type a monoamine oxidase in discrete regions of the rat brain and inhibition by RS-8359. *Jpn J. Pharmacol*., **55**, 121–128.

Lee,S.H. *et al*. (2007) Increased urinary level of oxidized nucleosides in patients with mild-to-moderate Alzheimer's disease. *Clin. Biochem*., **40**, 936–938.

Li,C.Y. and Li,H.Z. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**, 1175–1182.

Lindon,J. *et al*. (2007) *The Handbook of Metabonomics and Metabolomics*. Elsevier, Amsterdam and Oxford.

Lunn,D.J. *et al*. (2000) Winbugs—a bayesian modelling framework: Concepts, structure, and extensibility. *Stat. Comput*., **10**, 325–337.

Mamas,M. *et al*. (2011) The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Arch. Toxicol*., **85**, 5–17.

McLachlan,G. and Peel,D. (2000) *Finite Mixture Models*. Wiley, New York.

McLachlan,G.J. *et al*. (2006) A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, **22**, 1608–1615.

Meuwissen,T.H. and Goddard,M.E. (2004) Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet. Sel. Evol*., **36**, 261–279.

Motsinger-Reif,A. *et al*. (2013) Comparing metabolomic and pathologic biomarkers alone and in combination for discriminating Alzheimer's disease from normal cognitive aging. *Acta Neuropathol. Commun*., **1**, 28.

Nelson,S.D. (2001) Structure toxicity relationships - how useful are they in predicting toxicities of new drugs?. *Adv. Exp. Med. Biol*., **500**, 33–43.

Patani,G.A. and LaVoie,E.J. (1996) Bioisosterism: a rational approach in drug design. *Chem. Rev*., **96**, 3147–3176.

Patti,G.J. *et al*. (2012) Innovation: metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol*., **13**, 263–269.

Raftery,A.E. and Zheng,Y.Y. (2003) Discussion: performance of bayesian model averaging. *J. Am. Stat. Assoc*., **98**, 931–938.

Ramanan,V.K. *et al*. (2012) Pathway analysis of genomic data: concepts, methods and prospects for future development. *Trends Genet*., **28**, 323–332.

Shan,X. *et al*. (2007) Messenger RNA oxidation is an early event preceding cell death and causes reduced protein expression. *FASEB J*., **21**, 2753–2764.

Todeschini,R. and Consonni,V. (2009) *Molecular Descriptors for Chemoinformatics*. 2nd edn. Wiley, Weinheim, Germany.

Wang,Y. *et al*. (2009) Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res*., **37**, W623–W633.

Wei,P. and Pan,W. (2008) Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics*, **24**, 404–411.

Zhang,R. *et al*. (2013) Systemic immune system alterations in early stages of Alzheimer's disease. *J. Neuroimmunol*., **256**, 38–42.

Zhu,H. and Li,L. (2011) Biological pathway selection through nonlinear dimension reduction. *Biostatistics*, **12**, 429–444.