Use of pathway information in molecular epidemiology

Duncan C. Thomas,[1] David V. Conti,[1] James Baurley,[1] Frederik Nijhout,[2] Michael Reed[3] and Cornelia M. Ulrich[4]*

[1]Department of Preventive Medicine, University of Southern California, 1540 Alcazar St., CHP-220, Los Angeles, CA 90089-9011, USA
[2]Department of Biology, Duke University, Durham, NC, USA
[3]Department of Mathematics, Duke University, Durham, NC, USA
[4]Fred Hutchison Cancer Research Center, 1100 Fairview Avenue N., PO Box 19024, Seattle, WA 98109-1024, USA
**Correspondence to:* Tel: +1 323 442 1218; Fax: +1 323 442 2349; E-mail: dthomas@usc.edu

Date received (in revised form): 23rd June 2009

Abstract

Candidate gene studies are generally motivated by some form of pathway reasoning in the selection of genes to be studied, but seldom has the logic of the approach been carried through to the analysis. Marginal effects of polymorphisms in the selected genes, and occasionally pairwise gene–gene or gene–environment interactions, are often presented, but a unified approach to modelling the entire pathway has been lacking. In this review, a variety of approaches to this problem is considered, focusing on hypothesis-driven rather than purely exploratory methods. Empirical modelling strategies are based on hierarchical models that allow prior knowledge about the structure of the pathway and the various reactions to be included as 'prior covariates'. By contrast, mechanistic models aim to describe the reactions through a system of differential equations with rate parameters that can vary between individuals, based on their genotypes. Some ways of combining the two approaches are suggested and Bayesian model averaging methods for dealing with uncertainty about the true model form in either framework is discussed. Biomarker measurements can be incorporated into such analyses, and two-phase sampling designs stratified on some combination of disease, genes and exposures can be an efficient way of obtaining data that would be too expensive or difficult to obtain on a full candidate gene sample. The review concludes with some thoughts about potential uses of pathways in genome-wide association studies.

Keywords: *colorectal cancer, complex diseases, folate, gene–environment interactions, gene–gene interactions*

Introduction

Molecular epidemiology has advanced from testing associations of disease with single polymorphisms, to exhaustive examination of all polymorphisms in a candidate gene using haplotype tagging single nucleotide polymorphisms (SNPs), to studying increasing numbers of candidate genes simultaneously. Often, gene–environment and gene–gene interactions are considered at the same time. As the number of main effects and interactions proliferate, there is a growing need for a more systematic approach to model development.[1]

In recognition of this need, the American Association for Cancer Research held a special conference[2] in May 2007, bringing together experts in epidemiology, genetics, statistics, computational biology, systems biology, toxicology, bioinformatics and other fields to discuss various multidisciplinary approaches to this problem.

A broad range of exploratory methods have been developed recently for identifying interactions, such as neural nets, classification and regression trees, multi-factor dimension reduction, random forests, hierarchical clustering, etc.[3–7] Our focus here, however, is instead on hypothesis-driven

methods based on prior understanding about the structure of biological pathways postulated to be relevant to a particular disease. Our primary purpose is to contrast mechanistic and empirical methods and explore ways of combining the two.

The folate pathway as an example

Folate metabolism provides a rich example to illustrate these challenges. Folate has been implicated in colorectal cancer,[8] coronary heart disease[9] and neural tube defects,[10,11] among other conditions. Several steps in the metabolism of folate could be involved in these various diseases (Figure 1) and could have quite different effects. The pathway is complex, involving 19 enzymes or carrier proteins, with various feedback loops and two main cycles, the folate and the methionine cycles. The former is involved in pyrimidine synthesis through the action of thymidylate synthase (TS), potentially leading to uracil misincorporation into DNA and subsequent DNA damage and repair or misrepair. The latter is involved in DNA methylation through the conversion of S-adenosyl methionine (SAM) to S-adenosyl homocysteine (SAH) by DNA-methyltransferase (DNMT). These two mechanisms in particular have been suggested as important links between folate and carcinogenesis, although other possibilities include purine synthesis (via the aminoimidazole-carboxamide ribonucleotide transferase [AICART] reaction) and homocysteine itself. Because polymorphisms that tend to increase one of these effects may decrease others, their effects on disease endpoints can be quite different, depending on which part of the pathway is more important. A detailed mathematical model for this system has been developed by Nijhout *et al.*[12,13] and Reed *et al.*,[14,15] based on the equilibrium solution to a set of linked ordinary differential equations for Michaelis–Menten kinetics and implemented in software available at http://metabolism.math.duke.edu/.

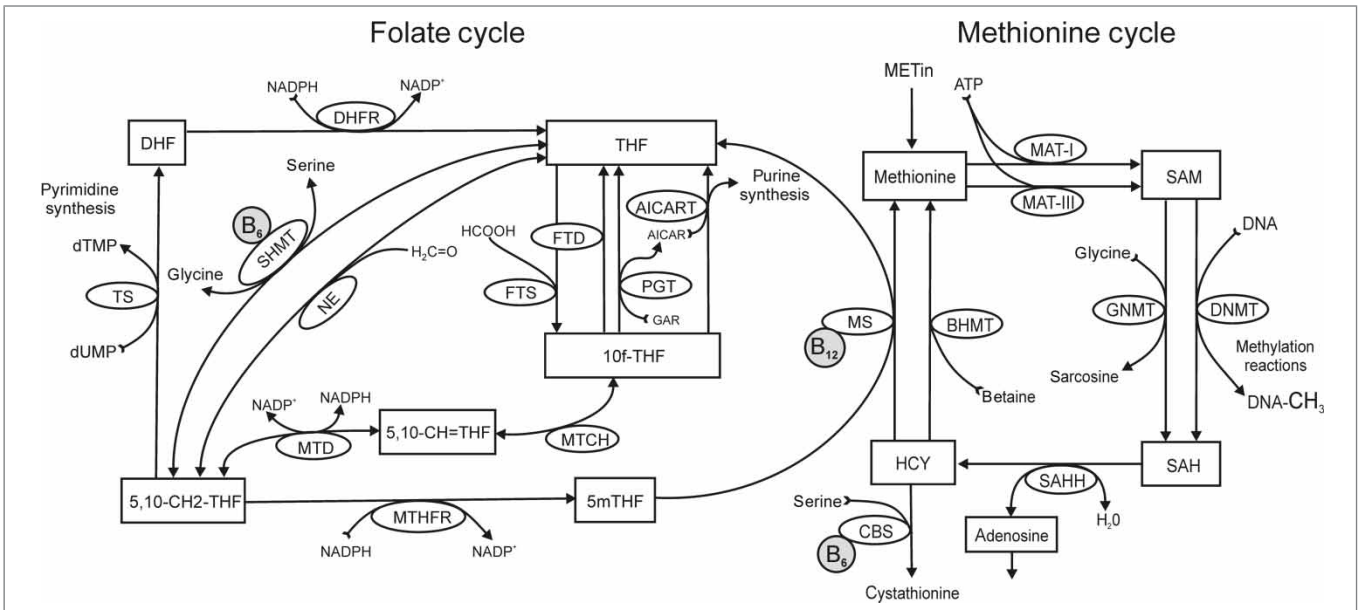

Figure 1. Biochemical diagram of folate metabolism (reproduced with permission from Reed *et al.*[14]).
AICART, aminoimidazolecarboxamide ribonucleotide transferase; BHMT, betaine-homocysteine methyltransferase; CBS, cystathionine b-synthase; DHFR, dihydrofolate reductase; DNMT, DNA-methyltransferase; dTMP, thymidine monophosphate; FTD, 10-formyltetrahydrofolate dehydrogenase; FTS, 10-formyltetrahydrofolate synthase; GAR, glycinamide ribotid; G-NMT, glycine N-methyltransferase; HCOOH; formic acid; $H_2C{=}O$, formaldehyde; HCY, homocysteine; MAT, methionine adenosyl transferase; MS, methionine synthase; MTCH, 5,10-methylenetetrahydrofolate cyclohydrolase; MTD, 5,10-methylenetetrahydrofolate dehydrogenase; MTHFR, 5,10-methylenetetrahydrofolate reductase; NE, nav-enzymatic; PGT, phosphoribosyl glycinamide transferase; SAH, S-adenosylhomocysteine; SAHH, SAH hydrolase; SAM, S-adenosylmethionine; SHMT, serine hydroxymethyltransferase; THF, tetrahydrofolate; 5m-THF, 5-methylTHF; 5,10-CH_2-THF, 5,10-methyleneTHF; 10f-THF, 10-formylTHF; TS, thymidylate synthase.

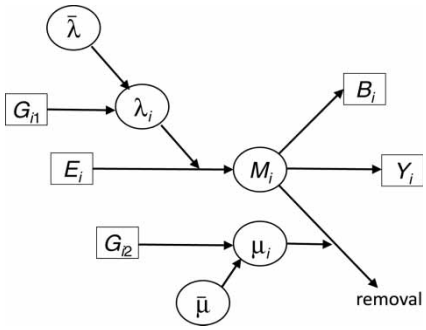

Figure 2. Schematic representation of simplified one-compartment model.

To illustrate the various approaches, we simulated some typical data in the form that might be available from a molecular epidemiology study — specifically data on genetic variants, various environmental exposures, a disease outcome or clinical trait, and, possibly, biomarker measurements on some or all subjects. We began with a population of 10,000 individuals with randomly generated values of intracellular folate E_1 (the total tetrahydrofolate [THF] concentration in the six compartments forming the closed loop shown on the left-hand side of Figure 1), methionine intake E_2 (METin, log-normally distributed) and 14 of the key genes G shown in Figure 1. For each gene, a person-specific value of the corresponding V_{max} was sampled from log-normal distributions with genotype-specific geometric means (GMs = 0.6, 0.8, or 1.1 times the overall GM) and common geometric standard deviations (GSD = 1.1) and K_m appropriate for that enzyme (see Table 1 in Reed *et al.*[14] for reference values for V_{max} and K_m for each gene). The differential equations were then evaluated to determine the steady-state solutions for ten intermediate metabolite concentrations and 14 reaction rates for each individual, based on their specific environmental variables and enzyme activity rates. The probability of disease was calculated under a logistic model for each of four scenarios for the causal biological mechanism — homocysteine concentration, the rate of DNA methylation reactions and the rates of purine and pyrimidine synthesis — and a binary disease indicator Y was sampled with the corresponding probability. Only the data on (Y,E,G) were retained from the first 500 cases and 500 controls for the first level of the epidemiological analysis. For some analyses, we also simulated biomarkers[16] on stratified subsamples of these subjects, as will be described later. Various summaries of the correlations among the (X,E,G) values for the remaining 9,000 subjects were deposited into what we shall call the 'external database' for use in constructing priors, as described below (no individual Y data were used for this purpose).

Table 1 shows the univariate associations of each gene with disease under each assumption about the causal risk factor. In these simulations, only one of these was taken as causal at a time, each scaled with the relative risk coefficient $\beta = 2.0$ per standard deviation of the respective risk factor. When homocysteine concentration was taken as the causal factor, the strongest association was with genetic variation in the cystathionine b-synthase (*CBS*) and S-adenosylhomocysteine hydrolase (*SAHH*) genes. The remaining three columns relate to various reaction rates as causal mechanisms. For pyrimidine synthesis (characterised here by the TS reaction rate), the strongest influence was seen for genetic variation in *TS* and the 5,10-methyleneTHF dehydrogenase (*MTD*) gene. For purine synthesis (reflected in the AICART reaction rate), the strongest associations were with genetic variation in the phosphoribosyl glycinamide transferase (*PGT*) gene and somewhat weaker for *MTD* and 5,10-methyleneTHF cyclohydrolase (*MTCH*) and serine hydroxymethyltransferase (*SHMT*) genes; interestingly, the disease risk is not particularly related to the *AICART* genotype itself. When DNA methylation (reflected by the DNMT reaction rate) was assumed to be causal, none of the genetic associations were as strong as for the other three causal mechanisms, the strongest being with the 5,10-methyleneTHF reductase (*MTHFR*) gene, *SAHH* and *MTD*. Genetic variation in *DNMT* was not explicitly simulated, but the reaction rates for this enzyme were identical to those for methionine adenosyl transferase (*MAT-II*) and SAHH, reflecting a rate-limiting step. Thus, genetic variation in *MAT-II* had no effect on risk, the reaction rate being driven entirely by *SAHH*. Other rate-limited

Table 1. Marginal odds ratios (ORs) for the association of each gene with disease under various choices of reaction rates or intermediate metabolite concentrations as the causal risk factor (ORs are expressed relative to the low enzyme activity rate genotype)

Genes	Simulated causal intermediate variable (β = 2 per SD)			
	Homocysteine concentration	Pyrimidine synthesis (*TS*)	Purine synthesis (*AICART*)	DNA methylation (*DNMT*)
1. DHFR	1.012	0.963	0.978	0.988
2. TS	0.910***	0.437***	1.129***	1.103***
3. MTD	0.753***	2.451***	0.540***	1.659***
4. MTCH	0.793***	1.805***	0.532***	1.372***
5. PGT	1.059	0.863**	0.200***	0.950
6. AICART	1.044	0.989	0.972	0.963
7. FTD	0.969	1.009	0.713***	1.077*
8. FTS	1.048	0.899***	1.709***	0.957
9. SHMT	1.256***	0.558***	0.592***	0.639***
10. MTHFR	1.298***	1.073*	1.153***	0.573***
11. MS	1.197***	0.790***	0.736***	0.815***
12. SAHH	0.428***	1.097**	1.108**	0.564***
13. CBS	2.753***	1.073*	1.028	0.925*
14. MAT-II	0.998	1.013	1.014	0.999
Exposures				
1. Intracellular folate	0.790***	1.783***	1.961***	1.543***
2. Methionine intake	3.819***	1.226***	1.112**	1.342***

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

AICART, aminoimidazolecarboxamide ribonucleotide transferase; CBS, cystathionine b-synthase; DHFR, dihydrofolate reductase; FTD, 10-formyltetrahydrofolate dehydrogenase; FTS, 10-formyltetrahydrofolate synthase; MAT, methionine adenosyl transferase; MS, methionine synthase; MTCH, 5,10-methylenetetrahydrofolate cyclohydrolase; MTD, 5,10-methylenetetrahydrofolate dehydrogenase; MTHFR, 5,10-methylenetetrahydrofolate reductase; PGT, phosphoribosyl glycinamide transferase; SAHH, S-adenosyl-homocysteine hydrolase; SHMT, serine hydroxymethyltransferase; TS, thymidylate synthase.

combinations included dihydrofolate reductase (*DHFR*) with TS, MTD with MTCH, and PGT with AICART. Methionine intake was the strongest environmental exposure factor for the simulation with homocysteine as the causal mechanism, whereas intracellular folate had a stronger effect under the other three mechanisms.

Mechanistic vs empirical models

For the four highlighted simulations, we also conducted multiple logistic regressions in a stepwise manner, offering methionine, folate, the 14 genotypes and all 91 pairwise G × G and 28 G × E interactions (Table 2). These are difficult to interpret, however, owing to the large numbers of comparisons and unstable regression coefficients, particularly in the models that include interaction terms. In an attempt to gain greater insight into mechanisms, attention will now be turned to more pathway-driven modelling approaches, based on hierarchical or mechanistic models. The former extend the standard logistic models summarised in Table 2 by the addition of 'prior covariates'

Table 2. Multiple stepwise logistic regression models, including only main effects or main effects and G × G/G × E interaction terms for four different choices of the causal variable (gene names are given in Table 1; E_1 = intracellular folate concentration; E_2 = methionine intake)

Simulated causal risk factor							
Homocysteine concentration		Pyrimidine synthesis (*TS*)		Purine synthesis (*AICART*)		DNA methylation (*DNMT*)	
G, E	***G, E, G × E, G × G***	***G, E***	***G, E, G × E, G × G***	***G, E***	***G, E, G × E, G × G***	***G, E***	***G, E, G × E, G × G***
E_1^{--}	E_1^{--}	E_1^{+++}	E_1^{+++}	E_1^{+++}	E_1^{+++}	E_1^{+++}	E_1^{+++}
E_2^{+++}	E_2^{+++}	E_2^{+}	G_2^{---}	G_3^{--}	G_5^{--}	E_2^{++}	G_3^{+++}
G_3^{---}	G_3^{-}	G_2^{---}	G_3^{+++}	G_4^{---}	G_8^{+++}	G_3^{+++}	G_9^{---}
G_4^{-}	G_{12}^{--}	G_3^{+++}	G_4^{+++}	G_5^{---}	G_9^{---}	G_4^{+++}	G_{10}^{---}
G_9^{+++}	G_{13}^{+++}	G_4^{+++}	G_9^{-}	G_7^{---}	$G_1 \times G_{10}^{-}$	G_8^{-}	$E_1 \times G_1^{+}$
G_{10}^{+++}	$G_1 \times G_{14}^{--}$	G_8^{-}	$E_2 \times G_2^{-}$	G_8^{+++}	$G_3 \times G_5^{+++}$	G_9^{---}	$E_2 \times G_2^{--}$
G_{11}^{+}	$G_2 \times G_9^{-}$	G_9^{---}	$E_2 \times G_6^{-}$	G_9^{--}	$G_4 \times G_7^{+}$	G_{10}^{---}	$E_2 \times G_{12}^{--}$
G_{12}^{---}	$G_3 \times G_{14}^{++}$		$G_2 \times G_{13}^{--}$	G_{11}^{---}	$G_4 \times G_{11}^{+++}$	G_{12}^{---}	$G_2 \times G_5^{-}$
G_{13}^{+++}	$G_4 \times G_{13}^{++}$		$G_7 \times G_{14}^{-}$	G_{12}^{++}	$G_5 \times G_7^{+}$		$G_3 \times G_8^{+}$
	$G_8 \times G_{10}^{---}$		$G_8 \times G_9^{++}$		$G_5 \times G_{12}^{-}$		$G_4 \times G_5^{--}$
	$G_8 \times G_{12}^{++}$		$G_{11} \times G_{14}^{++}$				$G_5 \times G_6^{+}$
	$G_9 \times G_{11}^{--}$						$G_5 \times G_{12}^{+++}$
	$G_9 \times G_{12}^{-}$						$G_{11} \times G_{13}^{+}$
	$G_{12} \times G_{13}^{+}$						

$^{+}p < 0.05$; $^{++}p < 0.01$; $^{+++}p < 0.001$ for positive associations; $^{-}$, $^{--}$, $^{---}$ denote corresponding levels of significance for negative associations.
AICART, aminoimidazolecarboxamide ribonucleotide transferase; DNMT, DNA methyltransferase; TS, thymidylate synthase.

incorporating knowledge about the relative risk coefficients predicted by the pathway. The latter attempt to model the pathways explicitly, using simplified versions of physiologically based pharmacokinetic (PBPK) models, thereby requiring stronger assumptions about reaction dynamics and population distributions of rate parameters.

Hierarchical models for disease–pathway associations

In the first level, the epidemiological data are fitted using a conventional 'empirical' model for the main effects and interactions among the various input genotypes and exposures, here denoted generically as $\mathbf{X} = (X_{ip})_{p=1\ldots P} = (E, G, G \times G, G \times E, G \times G \times E, \ldots)$; for example, a logistic regression model of the form

$$\text{logit}\ \Pr(Y_i = 1|\mathrm{X}_i) = \beta_0 + \sum_{p=1}^{P} \beta_p X_{ip} \qquad (1)$$

the sum being taken over the range of terms included in the $\mathbf{X}$ vector. Note that all possible effects of some predetermined complexity (eg all main effects and two-way, or perhaps higher order, interactions possibly limited to subsets relevant to the hypothesised pathway structure) are included, rather than using some form of model selection, as was done in the stepwise analyses summarised in Table 2.

In the second-level model, each of the regression coefficients from Eq. (1) is in turn regressed on a vector $\mathbf{Z}_p = (Z_{pv})_{v=1\ldots V}$ of 'prior covariates' describing characteristics of the corresponding terms in X; for example,

$$\beta_p \sim N(\pi_0 + \sum_{v=1}^{v} \pi_v Z_{pv}, \sigma^2) \qquad (2)$$

There are many possibilities for what could be included in the set of prior covariates, ranging from indicator variables for which of several pathways each gene might act in,[17] *in silico* predictions of the functional significance of polymorphisms in each gene,[18,19] or genomic annotation from formal ontologies.[20] Summaries of the effects of genes on expression levels ('genetical genomics') or of associations of genes with relevant biomarkers might also be used as prior covariates. Rebbeck *et al.*[21] provide a good review of available tools that could be used for constructing prior covariates.

Alternatively, one could model the variances, for example:

$$\beta_p \sim N\left(\pi' Z_p, \sigma^2 \exp(\varphi' Z_p)\right) \qquad (3)$$

For example, suppose the $\mathbf{X}$ vector comprised effects for different polymorphisms within each gene and one had some prior predictors of the effects of each polymorphism (eg *in silico* predictions of functional effects or evolutionary conservation) and other predictors of the general effects of genes (eg their roles in different pathways or the number of other genes that they are connected to in a pathway). Then, it might be appropriate to include the former in the $\boldsymbol{\pi}'\mathbf{Z}$ part of the model for the means, and the latter in the $\boldsymbol{\varphi}'\mathbf{Z}$ part of the model for the variances.

So far, the second-level models have assumed an independence prior for each of the regression coefficients; but now, suppose we have some prior information about the relationships *among* the genes, such as might come from networks inferred from gene co-expression data. Let $\mathbf{A} = (A_{pq})_{p,q=1..P}$ denote a matrix of prior connectivities between pairs of genes — for example, taking the value 1 if the two are adjacent (connected) in some network or 0 otherwise. Then, one might replace the independence prior of Eq. (2) by a multivariate prior of the form:

$$\beta \sim N_P(\pi' Z, \sigma^2 (I - \rho A)^{-1})$$

This is known as the conditional autoregressive model, and is widely used in spatial statistics.[22] Sample WinBUGS code to implement these and other models described below are available in an online supplement.

In applications to the folate simulation, we tried two variants of this model. First, we considered three prior covariates in $\mathbf{Z}$: an indicator for whether a gene is involved in the methionine cycle; whether it is involved in the folate cycle; and the number of other genes it is connected to in the entire network (a measure of the extent to which it might have a critical role as a 'hub' gene). The $\mathbf{A}$ matrix was specified in terms of whether a pair of genes had a metabolite in common, either as substrate or product.

Table 3 summarises the results of several models, including these three prior covariates in the means or variance model, as well as the connectivities in the covariance model. As would be expected, in the zero mean model, all the significant parameter estimates were shrunk towards zero because of the large number of genes with no true effect in the model. In general, none of the prior covariates significantly predicted the means. The estimates of the βs in all these models were much closer to the simple maximum likelihood estimates (the first column), however, and their standard errors were generally somewhat smaller, indicating the 'borrowing of strength' from each other. In the model with covariates in the prior variances, however, the number of connections for each gene was significantly associated with the variance. In the final model, with correlations between genes being given by indicators for whether they were connected in the graph, the posterior distribution for the parameter ρ is constrained by the requirement that the covariance matrix be positive definite, but showed strong evidence of gene–gene correlations following the pattern given by the connectivities in Figure 1. The generally

Table 3. Summary of hierarchical modelling fits (parameter estimates [SEs]) for selected genetic effects (β_G), prior covariates ($\mathbf{Z}'\pi$) and prior correlations ($\sigma^2\mathbf{A}$) for simulation with homocysteine concentration as the causal variable

	No prior	$N(0,\sigma^2)$	$N(Z\pi,\sigma^2)$	$N(0,\sigma^2e^{Z'\psi})$	$N(Z'\pi,\sigma^2e^{Z'\psi})$	$N(Z'\pi,\sigma^2(I-\rho A)^{-1})$
Genetic main effects						
G_3: *MTD*	−0.370 (0.113)	−0.341 (0.111)	−0.352 (0.109)	−0.327 (0.112)	−0.340 (0.109)	−0.106 (0.114)
G_4: *MTCH*	−0.258 (0.133)	−0.229 (0.123)	−0.245 (0.124)	−0.216 (0.120)	−0.207 (0.117)	−0.304 (0.127)
G_9: *SHMT*	0.300 (0.121)	0.282 (0.112)	0.272 (0.112)	0.155 (0.109)	0.336 (0.079)	0.128 (0.112)
G_{10}: *MTHFR*	0.335 (0.116)	0.301 (0.110)	0.313 (0.113)	0.293 (0.110)	0.245 (0.116)	0.198 (0.114)
G_{11}: *MS*	0.240 (0.112)	0.206 (0.106)	0.219 (0.114)	0.106 (0.095)	0.190 (0.102)	0.060 (0.112)
G_{12}: *SAHH*	−0.809 (0.131)	−0.735 (0.126)	−0.760 (0.135)	−0.752 (0.128)	−0.760 (0.125)	−0.648 (0.127)
G_{13}: *CBS*	1.492 (0.134)	1.360 (0.123)	1.417 (0.129)	1.400 (0.129)	1.419 (0.133)	1.149 (0.123)
Prior covariates						
π_1: folate (mean)			−0.020 (0.638)		−0.668 (0.192)	−0.641 (0.453)
π_2: methionine (mean)			0.115 (0.480)		−0.266 (0.151)	0.203 (0.377)
π_3: connections (mean)			0.000 (0.092)		0.113 (0.027)	0.022 (0.030)
ψ_1: folate (variance)				0.171 (0.337)	−0.312 (1.536)	
ψ_2: methionine (variance)				−0.181 (0.327)	−0.913 (1.259)	
ψ_3: connections (variance)				0.472 (0.185)	0.901 (0.390)	
Posterior variances and correlations						
$\sigma_\beta = SD(\beta\|\mathbf{Z})$		0.492 (0.109)	0.658 (0.143)	1.175 (0.491)	2.362 (2.714)	0.877 (0.285)
$\sigma_\pi = SD(\pi)$			0.864 (0.385)		0.858 (0.356)	0.889 (0.403)
$\sigma_\psi = SD(\psi)$				0.342 (0.059)	1.350 (1.010)	
$\rho = corr(\beta\|\mathbf{A})$						0.597 (0.170)

weak effects of prior covariates in these models may simply reflect the crudeness of these classifications. Below, we will revisit these models with more informative covariates based on the quantitative predictions of the differential equations model.

Mechanistic models

Whereas hierarchical models are generally applicable whenever one has external information about the genes and exposures available, in some circumstances the dynamics of the underlying biological process may be well enough understood to support mechanistic modelling. These are typically based on systems of ordinary differential equations (ODEs) describing each of the intermediate nodes in a graphical model as deterministic quantities given by their various inputs (exposures or previous substrates) with reaction rates determined by genotypes (Figure 2). For example, in a sequence $j = 1, \ldots, J$ of linear kinetic steps, with conversion from metabolite M_j to M_{j+1} at rate λ_j and removal at rate μ_j, the instantaneous concentration is given by the differential equation:

$$\frac{dM_j}{dt} = (\lambda_{j-1}M_{j-1} - (\lambda_j + \mu_j)M_j \qquad (4)$$

leading to the equilibrium solution for the final metabolite M_J as:

$$M_J(E,\mathbf{G}) = E \times \prod_{j=1}^{J}\left(\frac{\lambda_{j-1}(G_{j-1}^{(\lambda)})}{(\lambda_j(G_j^{(\lambda)}) + \mu_j(G_j^{(\mu)})}\right)$$

where X_0 denotes the concentration of exposure E. This predicted equilibrium concentration of the final metabolite in the graph is then treated as a covariate in a logistic model for the risk of disease:

$$\text{logit}[\Pr(Y_i = 1)] = \beta_0 + \beta_1 M_J(E_i, \mathbf{G}_i)$$

If sufficient external knowledge about the genotype-specific reaction rates is available, these could be treated as fixed constants, but more likely they would need to be estimated jointly with the βs in the disease model using a non-linear fitting program. More sophisticated non-linear models are possible — for example, incorporating Michaelis–Menten kinetics by replacing each of the λM terms in Eq. (4) by expressions of the form:

$$\frac{V_{max}^{\lambda j}(G_j)M_j}{M_j + K_m^{\lambda j}(G_j)}$$

and similarly for the μM terms. The resulting equilibrium solutions for $M_J(E,\mathbf{G})$ are now more complex solutions to a polynomial equation. For example, with only a single intermediate metabolite with one activation rate λ and one detoxification rate μ, the solution becomes:

$$M = E\frac{\lambda/\mu}{1 + \left((1/K_m^{\lambda}) - (\lambda/\mu)/K_m^{\mu}\right)E}$$

where $\lambda = V_{max}^{\lambda}(G_1)/K_m^{\lambda}$ and $\mu = V_{max}^{\mu}(G_2)/K_m^{\mu}$ denote the low-dose slopes of the two reactions. These solutions can be either upwardly or downwardly curvilinear in E, depending on whether the term in parentheses is positive or negative (basically, whether the creation of the intermediate exceeds the rate at which it can be removed). For the fitted values in the application below (third block of Table 4), the dose–response relationship for $M|E$ is upwardly curved for all genotype combinations (not shown).

A more realistic and more flexible model would allow for stochastic variation in the reaction rates λ_{ij} and μ_{ij} for each individual i conditional on their genotypes G_{ij}; for example, $\lambda_{ij} \sim LN(\bar{\lambda}_j(G_{ij}), \sigma_j^2)$ and likewise for μ_{ij}[23] or similarly for their corresponding V_{max} and K_m.[24] The population genotype-specific rates are, in turn, assumed to have log-normal prior distributions $\bar{\lambda}_j(g) \sim LN(\bar{\bar{\lambda}}_j, \omega_j^2)$ (and similarly for the μs), with vague priors on the population means $\bar{\bar{\lambda}}_j$, inter-individual variances σ_j^2 and between-genotype variances ω_j^2. The individual data might be further supplemented by available biomarker measurements B_{ij} of either the enzyme activity levels or intermediate metabolite concentrations, modelled as $B_{ij} \sim LN(\lambda_{ij},\omega^2)$ and $B_{ij} \sim LN(M_{ij},\omega^2)$, respectively.

The WinBUGS software[25] has an add-in called PKBUGS,[26] which implements a Bayesian analysis of population pharmacokinetic parameters.[27–31] More complex models can, in principle, be fitted using the add-in WBDIFF (http://www.winbugs-development.org.uk/wbdiff.html), which allows user-specified differential equations as nodes in a Bayesian graphical model.

To illustrate the approach, we consider a highly simplified model with only a single intermediate metabolite M (homocysteine). We assume this is created at linear rate λ determined by *SAHH* and removed at linear rate μ determined by *CBS*. The ratios of λ and μ between genotypes are estimated jointly with β. The first two lines of Table 4 provide the results of fitting the linear kinetics model, with and without inter-individual variability in the two rate parameters. Although, of course, many other genes are involved in the simulated model, the estimated homocysteine concentrations M are strongly predictive of disease, and both genes have highly significant effects on their respective rates. Allowing additional random variability in these rates slightly increased the population average genetic effects. For the Michaelis–Menten models, we allowed the V_{max}s to depend on genotype, while keeping the K_ms fixed. Not all the parameters can be independently estimated, but only the ratios μ_0/λ_0 and K_m^{μ}/K_m^{λ}, along

Table 4. Results of Markov chain Monte Carlo fitting of single-compartment models with homocysteine as an unobserved intermediate metabolite, created at a rate depending on *SAHH* (λ) and removed at a rate depending on *CBS* (μ), applied to the simulation taking homocysteine concentration as the causal variable

Model	Ln(β)	λ		μ	
		Mean	SD	Mean	SD
Linear:					
λ, μ fixed	1.63 (0.13)	0.152 (0.030)	0	0.226 (0.031)	0
λ, μ random	1.55 (0.13)	0.178 (0.029)	0.079 (0.024)	0.256 (0.037)	0.082 (0.021)
Michaelis–Menten:					
λ, μ fixed: $\ln(\mu_0/\lambda_0)$		0	0	0.765 (0.037)	0
$\ln[V_{max}(1)/V_{max}(0)]$	2.77 (0.13)	0.058 (0.010)	0	0.086 (0.010)	0
$\ln(K_m^{\mu}/K_m^{\lambda})$		0	0	−0.743 (0.042)	0
γs random: $\ln(\mu_0/\lambda_0)$		0	0	1.022 (0.002)	0.007 (0.001)
$\ln[V_{max}(1)/V_{max}(0)]$	2.99 (0.06)	0.061 (0.011)	0.023 (0.005)	0.091 (0.001)	0.003 (0.001)
$\ln(K_m^{\mu}/K_m^{\lambda})$		0	0	−1.008 (0.001)	0.005 (0.001)
Stochastic differential equations:					
$N = 10$ fixed	1.22 (0.07)	0.209 (0.016)	0	0.273 (0.014)	0
$N \sim \Gamma(100,1)$: $\hat{N} = 116$	3.16 (0.06)	0.161 (0.011)	0	0.215 0.010)	0

CBS, cystathionine b-synthase; SAHH, S-adenosylhomocysteine.

with the genetic rate ratios λ_1/λ_0 and μ_1/μ_0. Allowing the V_{max}s and K_ms to vary between subjects leads to some instability, but did not substantially alter the population mean parameter estimates. Adding in biomarker measurements B_i as surrogates for M_i for even a subset of subjects, as described below, substantially improved the precision of estimation of all the model parameters (results not shown).

Combining mechanistic and statistical models

Such an approach is likely to be impractical for complex looped pathways like folate, however. In this case, one might use the results of a preliminary exploratory or hierarchical model to simplify the pathway to a few key rate-limiting steps, so as to yield a simpler unidirectional model for which the differential equation steady-state solutions can be obtained in closed form.

Rather than taking $M(E,G)$ as a deterministic node in the mechanistic modelling approach described above, a fully Bayesian treatment would use *stochastic* differential equations to derive $\Pr(M|E,G)$. For example, suppose one postulated that the rate of change dM/dt depends on the rate at which it is created as a constant rate $\lambda(G_1)E$ and the rate at which it is removed at rate $\mu(G_2)M$. (Of course, the exposures E could be time dependent, in which case one would be interested in the long-term average of M rather than its steady state, but in most epidemiological studies there is little information available on short-term variation in exposures, so the following discussion is limited to the case of time-constant exposures.) Consider a discrete number of molecules and let $p_m(t) = \Pr(M = m|T = t)$. Then, the resulting stochastic differential equation becomes:

$$\frac{dp_m}{dt} = -(\lambda E + \mu m)p_m + \lambda E p_{m-1} + \mu(m+1)p_{m+1}$$

The solution turns out to be simply a Poisson distribution for m with mean $E(m) = \lambda E/\mu$. This

suggests as a distribution for continuous metabolite concentrations M in some volume of size N:

$$p(M) = Ne^{-\lambda EN/\mu}(\lambda EN/\mu)^{NM}/\Gamma(NM+1)$$

where N now controls the dispersion of the distribution. More complex solutions for Michaelis–Menten kinetics with a finite number of binding sites have been provided by Kou *et al.*,[32] who showed that the classical solutions still held in expectation, but other properties — like the distribution of waiting times in various binding states — were different, appearing to demonstrate a non-Markov memory phenomenon, particularly at high substrate concentrations. Further stochastic variability arises from fluctuations in binding affinity due to continual changes in enzyme conformation.[33]

To illustrate the general idea, we fitted this simplified version of the model, treating λ and μ as fixed genotype-specific population values, yielding the estimates shown in the last line of Table 4. The dispersion parameter N cannot be estimated, but the results for other parameters are relatively insensitive to this choice; the results in Table 4 are based on either a fixed value $N = 10$ or using an informative $\Gamma(100,1)$ prior; as N gets very large, the estimates converge to those in the first line for linear kinetics with fixed genotype-specific λ and μ.

For more complex models, for which analytic solution of the differential equations may be intractable, the technique of approximate Bayesian computation[34] may be helpful. The basic idea is, at each Markov chain Monte Carlo cycle, to simulate data from the differential equations model using the current and proposed estimates of model parameters and evaluate the 'closeness' of the simulated data to the observed data in terms of some simple statistics. This is then used to decide whether to accept or reject the proposed new estimates, rather than having to compute the likelihood itself.

A simpler approach uses the output of a PBPK simulation model as prior covariates in a hierarchical model. Let $Z_{ge} = E[M(G_g,E_e)]$ denote the predicted steady-state concentrations of the final metabolite from a differential equations model for a particular combination of genes and/or exposures (thus, $Z_{gg'}$ might represent the predicted effect of a G × G interaction between genes g and g'). As discussed above, other Zs could comprise variances of predicted Ms across a range of input values as a measure of the sensitivity of the output to variation in that particular combination of inputs. Z_{ge} could also be a vector of several different predicted metabolite concentrations if there were multiple hypotheses about which was the most aetiologically relevant.

For the folate application, the Z matrix was obtained by correlating the simulated intermediate phenotypes v (reaction rates or metabolite concentrations) with the 14 genotypes, 91 G × G and 28 G × E interaction terms. The resulting correlation coefficients for the four simulated causal variables were then used as a vector of *in silico* prior covariates $Z_p = (Z_{pv})_{v=1..4}$ for the relative risk coefficients β_p. The full set of correlations Z_{pv} across all ten metabolites and nine non-redundant velocities were also used to compute an adjacency matrix as $A_{pq} = \text{corr}_v(Z_{pv}, Z_{qv})$, representing the extent to which a pair of genes had similar effects across the whole range of intermediate phenotypes. The effects of these *in silico* covariates (Table 5) were substantially stronger than for the simple indicator variables illustrated earlier. In each simulation, the prior covariate corresponding to the causal variable was the strongest predictor of the genetic main effects.

Designs incorporating biomarkers

Ultimately, it may be helpful to incorporate various markers of the internal workings of a postulated pathway, perhaps in the form of biomarker measurements of intermediate metabolites, external bioinformatic knowledge about the structure and parameters of the network, or toxicological assays of the biological effects of the agents under study. For example, in a multi-city study of air pollution, we are applying stored particulate samples from each city to cell cultures with a range of genes experimentally knocked down to assess their genotype-specific biological activities. We will then incorporate these measurements directly into the analysis of G × E interactions in epidemiological data.[35] See Thomas,[1] Thomas *et al.*,[2] Conti *et al.*[20]

Table 5. Summary of hierarchical modelling fits for selected genetic effects (β_G), prior covariates ($Z'\pi$) and prior standard deviations (σ_β and σ_π) for simulation with different intermediates as the causal variable, using the Z matrix derived from independent data from the same simulation model (see text). Bolded entries have posterior credibility intervals that exclude zero

	Simulated causal variable			
	Homocysteine concentration	**Pyrimidine synthesis (*TS*)**	**Purine synthesis (*AICART*)**	**DNA methylation (*DNMT*)**
Genetic main effects				
G_2: *TS*	0.06 (0.10)	**−0.97 (0.11)**	0.15 (0.11)	0.12 (0.10)
G_3: *MTD*	**−0.35 (0.11)**	**1.09 (0.12)**	**−0.81 (0.12)**	**0.59 (0.11)**
G_4: *MTCH*	−0.26 (0.14)	**0.60 (0.13)**	**−0.68 (0.14)**	**0.35 (0.12)**
G_5: *PGT*	0.01 (0.05)	0.05 (0.18)	**−1.90 (0.19)**	−0.04 (0.16)
G_7: *FTS*	0.13 (0.12)	0.06 (0.14)	**−0.62 (0.13)**	0.11 (0.12)
G_8: *FTD*	−0.12 (0.13)	**−0.34 (0.14)**	**0.59 (0.13)**	**−0.26 (0.12)**
G_9: *SHMT*	**0.29 (0.12)**	**−0.72 (0.10)**	**−0.53 (0.12)**	**−0.43 (0.10)**
G_{10}: *MTHFR*	**0.34 (0.10)**	−0.04 (0.10)	0.13 (0.11)	**−0.70 (0.11)**
G_{11}: *MS*	**0.23 (0.10)**	−0.18 (0.10)	**−0.47 (0.11)**	−0.13 (0.10)
G_{12}: *SAHH*	**−0.78 (0.13)**	0.22 (0.13)	**0.34 (0.14)**	**−0.54 (0.12)**
G_{13}: *CBS*	**1.43 (0.13)**	0.07 (0.12)	0.06 (0.12)	0.01 (0.11)
Prior covariates				
π_1: homocysteine	**1.65 (0.81)**	−0.16 (0.63)	−0.02 (0.70)	0.00 (0.57)
π_2: v*TS*	−0.18 (0.57)	**1.45 (0.61)**	0.08 (0.62)	0.17 (0.54)
π_3: v*AICART*	0.11 (0.59)	−0.35 (0.61)	**2.19 (0.81)**	−0.28 (0.55)
π_4: v*DNMT*	−0.01 (0.69)	0.27 (0.69)	−0.01 (0.70)	1.01 (0.73)
Posterior standard deviations (SDs)				
$\sigma_\beta = \mathrm{SD}(\beta\|\mathbf{Z})$	0.49 (0.12)	0.48 (0.11)	0.52 (0.11)	0.47 (0.11)
$\sigma_\pi = \mathrm{SD}(\pi)$	1.14 (0.48)	1.14 (0.47)	1.33 (0.58)	0.99 (0.40)

AICART, aminoimidazolecarboxamide ribonucleotide transferase; CBS, cystathionine b-reductase; DNMT, DNA methyltransferase; FTD, 10-formyltetrahydrofolate dehydrogenase; FTS, 10-formyltetrahydrofolate synthase; MS, methionine synthase; MTCH, 5,10-methylenetetrahydrofolate cyclohydrolase; MTD, 5,10-methylenetetrahydrofolate dehydrogenase; MTHFR, 5,10-methylenetetrahydrofolate reductase; PGT, phosphoribosyl glycinamide transferase; SAHH, S-adenosylhomocysteine; SHMT, serine hydroxymethyltransferase; TS, thymidylate synthase.

and Parl *et al.*[36] for further discussion about approaches to incorporating biomarkers and other forms of biological knowledge into pathway-driven analyses.

Typically biomarker measurements are difficult to obtain and are only feasible to collect on a subset of a large epidemiological study. While one might consider using a simple random sample for this purpose, greater efficiency can often be obtained by stratified sampling. Suppose the parent study is a case-control study with exposure information and DNA already obtained. One might then consider sampling on the basis of some combination of disease status, exposure and the genotypes of one or more genes thought to be particularly important for the intermediate phenotype(s) for which biomarkers are to be obtained. The optimal design would require knowledge of the true model (which, of course, is

Table 6. Estimated log relative risk per unit change of true long-term homocysteine concentrations, treated as a latent variable in a single compartment linear-kinetics model; data simulated assuming homocysteine is the causal variable. The simulated coefficients are 2.0 for homocysteine and 0 for TS

Sampling scheme	Subsample size	Biomarker(s) measured		
		Homocysteine	TS enzyme	Both
Random	80	1.92 (0.21) –	– 2.71 (0.43)	2.68 (0.46) −0.04 (0.20)
	200	1.82 (0.15) –	– 3.26 (0.54)	2.28 (0.21) −0.04 (0.14)
Stratified by G, E, and Y	$8 \times 10 = 80$	1.77 (0.25) –	– 2.47 (0.64)	2.62 (0.94) −0.05 (0.28)
	$8 \times 25 = 200$	1.82 (0.16) –	– 2.93 (0.39)	2.03 (0.20) −0.14 (0.14)

unknown), but a balanced design, selecting the subsample so as to obtain equal numbers in the various strata defined by disease and predictors is often nearly optimal.[37–39] The analysis can then be conducted by full maximum likelihood, integrating the biomarkers for unmeasured subjects over their distribution (given the available genotype, exposure and disease data) or by some form of multiple imputation, quasi-likelihood[40] or MCMC methods. Here, the interest is not in the association of disease with the biomarker B itself, but rather with the unobserved intermediate phenotype M it is a surrogate for. The disease model is thus of the form $\Pr(Y|M)$, with a latent process model for $\Pr(M|G,E)$ and a measurement model for $\Pr(B|M)$.

Again, using the folate simulation as the example, we simulated biomarkers for samples of ten or 25 individuals selected at random from each of the eight cells defined by disease status, the *MTHFR* genotype and high or low folate intake. A measurement B of either homocysteine concentration or the TS enzyme activity level was assumed to be normally distributed around their simulated equilibrium concentrations with standard deviations 10 per cent of that the true long-term average concentrations.

These data were analysed within a conventional measurement error framework[41,42] by treating the true long-term average values of homocysteine or TS activity as a latent variable X in a model given by the following equations:

$$\text{logit}\,\Pr(Y=1) = \beta_0 + \beta_1 M$$

$$M \sim N(\mathrm{X}'\alpha, \sigma^2) \text{ where } \mathrm{X} = (\mathrm{G}, \mathrm{E}, \ldots)$$

$$B \sim N(M, \tau^2)$$

For joint analyses of homocysteine and TS activity measurements, $\mathbf{M}$ and $\mathbf{B}$ were assumed to be bivariate normally distributed with $\mathbf{M} \sim N_2(\mathrm{X}'\mathrm{A}, \Sigma)$ and $\mathbf{B} \sim N_2(\mathbf{M}, \mathbf{T})$, and Y as having a multiple logistic dependence on $\mathbf{M}$. Only the main effects of the 14 genes and two environmental factors were included in $\mathbf{X}$ for this analysis. While the model can be fitted by maximum likelihood, it is convenient to use MCMC methods, which more readily deal with arbitrary patterns of missing $\mathbf{B}$ data. Thus, it is not essential for the different biomarkers to be measured on the same subset of subjects, but some overlap is needed to estimate the covariances Σ_{12} and T_{12}. More complex mechanistic models could, of course, be used in place of the regression model $\mathbf{M}|\mathbf{X}$. For this model to be identifiable, however, it is essential that distinct biomarkers be available for each of the intermediate phenotypes included in the disease model.

Estimates of the effects of both homocysteine and TS enzyme activity were highly significant in univariate analyses, even though the simulated

causal variable is homocysteine. In bivariate analyses, however, the TS effect became non-significant, owing to the strong positive correlation ($r_\Sigma = 0.45$; 95 per cent confidence interval [CI] 0.21, 0.71) between the residuals of M, while correlation between the residuals of the measurement errors was not significant ($r_T = 0.34$; 95% CI -0.12, $+0.63$). Although the standard errors varied strongly with subsample size, stratified sampling did not seem to improve the precision of the estimates. The reason for this appears to be that the biased sampling is not properly allowed for in the Bayesian analysis. Further work is needed to explore whether incorporating the sampling fractions into a conditional likelihood would yield more efficient estimators in the stratified designs.

Dealing with reverse causation: Mendelian randomisation

The foregoing development assumes that the biomarker measurement B or the underlying phenotype M of which it is a measurement is not affected by the disease process. While this may be a reasonable assumption in a cohort or nested case-control study where biomarker measurements are made on stored specimens obtained at entry to the cohort rather than after the disease has already occurred, it is a well known problem (known as 'reverse causation') in case-control studies. In this situation, one might want to restrict biomarker measurements only to controls and use marginal likelihood or imputation to deal with the unmeasured biomarkers for cases. Alternatively, one might consider using case measurements in a model that includes terms for differential error in the measurement model, $\Pr(B|M,Y)$.

These ideas have been formalised in literature known as 'Mendelian randomisation' (MR),[43–47] sometimes referred to as 'Mendelian deconfounding'.[48] Here, the focus of attention is not the genes themselves, but intermediate phenotypes (M) as risk factors for disease. The genes that influence M are treated as 'instrumental variables' (IVs)[49–54] in an analysis that indirectly infers the $M-Y$ relationship from separate analyses of the $G-M$ and $G-Y$ relationships. The appeal of the approach is that uncontrolled confounding and reverse causation are less likely to distort these relationships than they are to distort the $M-Y$ relationship if studied directly. In essence, the idea of imputing M values using G as an IV in a regression of Y on $E(M|G)$ is a form of MR argument. Nevertheless, the approach is not without its pitfalls,[55–58] both as a means of testing the null hypothesis of no causal connection between M and Y and as a means of estimating the magnitude of its effect. Particularly key is the assumption that the effect of G on Y is mediated solely through M. For complex pathways, the simple MR approach is unlikely to be of much help, but the idea of using samples free of reverse causation to learn about parts of the model from biomarker measurements and incorporating these into the analysis of a latent variable model is promising.

To illustrate these methods, consider the scenario where homocysteine is the causal variable for disease. The logistic regression of disease directly on homocysteine yields a logRR coefficient β of 2.57 (SE 0.22) per SD change of homocysteine (Table 7). This estimate is, however, potentially subject to confounding and reverse causation, and indeed in this simulation we generated an upward bias in $B|M$ of 50 per cent of the SD of M, which produced a substantial overestimate of the simulated $\beta = 2$. An MR estimate could in principle be obtained by using any of the genes in the pathway as an IV, *MTHRF* being the most widely studied. The regression of homocysteine on *MTHFR* yields a regression coefficient of $\alpha = 0.216$ (0.079) and a logistic regression of disease on *MTHFR* yields a regression coefficient of $\gamma = 0.112$ (0.142), to produce an MR estimate of $\beta = \gamma/\alpha = 0.52$ (0.68). Since *MTHRF* is only a relatively weak predictor of homocysteine concentrations in this simulation, however, it is a poor instrumental variable, as reflected in the large SE of the ratio estimate. Several other genes, exposures and interactions have much stronger effects on both homocysteine and disease risk — notably, *SAHH* and *CBS*, which yield significant MR estimates, 1.27 (0.33) and 1.09 (0.20), respectively. These differences between

Table 7. Mendelian randomisation estimates of the effect of homocysteine on disease risk

Analysis	α in B|G	γ in Y|G	β in Y|B
Direct: *Y*|*B*	—	—	2.57 (0.22)
Mendelian randomisation			
MTHFR	0.216 (0.079)	0.112 (0.142)	0.52 (0.68)
SAHH	−0.633 (0.088)	−0.801 (0.175)	1.27 (0.33)
CBS	0.917 (0.074)	0.995 (0.166)	1.09 (0.20)
Single imputation		**δ in** B|G,Y	
$E(B\|G)$	$R^2 = 0.43$	—	1.32 (0.16)
$E(B\|G,Y)\|_{Y=0}$	$R^2 = 0.71$	1.33 (0.05)	1.28 (0.20)
$E(B\|G,Y=0)$	$R^2 = 0.43$	—	1.31 (0.20)
Joint Bayesian			
	—	−0.04 (0.95)	1.92 (0.15)

estimates using different IVs and their underestimation of the simulated β suggest that simple Mendelian randomisation is inadequate to deal with complex pathways.

A stepwise multiple regression model for $\widehat{M} = E(B|G)$ included 13 main effects and G × G interactions and attained an R^2 of 0.43. Treating these predicted homocysteine concentrations as the covariate yielded a single imputation estimate of the log RR for disease of 1.32 (0.16), only slightly less precise than that from the logistic regression of disease directly on the measured values. While robust to uncontrolled confounding, this approach is not robust to reverse causation or misspecification of the prediction model; for example, it fails to include any exposure effects, which we have excluded to avoid distortion by reverse causation. More importantly, it also assumes that the entire effect of the predictors is mediated through homocysteine; this is true for this simulation, but is unlikely to be in practice. While not quite as downwardly biased as the Mendelian randomisation estimates (resulting from the improved prediction of *B*|G), the incompleteness of the model has still produced some underestimation.

Since we have simulated the case where the biomarker measurements are distorted by disease status, one might consider one of two alternative single imputation analyses. If both cases and controls have biomarker measurements available, one might include disease status in a model for $\widehat{M} = E(B|\mathbf{G}, Y) = \boldsymbol{\alpha}'\mathbf{G} + \delta Y$, and then set $Y = 0$ in the fitted regression in order to estimate the pre-disease values for the cases. Alternatively, one could fit the model for $\widehat{M} = E(B|\mathbf{G})$ using data *only* from controls and then apply the fitted model to *all* subjects, cases and controls. In either case, one would use only the *predicted* values for all subjects, not the *actual* biomarker measurements for those having them. In these simulated data, these approaches yield log RR estimates of 1.28 (0.20) and 1.31 (0.20), respectively. Either of these approaches avoids the circularity of using disease status to predict $B|\mathbf{G}$, Y and then using it again in the regression of Y on $\widehat{M} = E(B|\mathbf{G}, Y)$. While the first approach uses more of the data, it requires a stronger assumption that the effect of Y on B is correctly specified, including possible interactions with **G**. In this simulation, the estimate of δ is 1.33 (0.06), substantially biased away from the simulated value of 0.50 because it includes some of the causal effect of X on Y. A fully Bayesian analysis jointly estimates the bias term δY in the full model $p_\alpha(M|E,G)p_\beta(Y|M)p_{\gamma,\delta}(B|M,Y)$. In this simulation, the fully

Bayesian analysis yielded an estimate of $\beta = 2.95$ (0.22) and $\delta = -0.02$ (1.02). Obviously, δ is so poorly estimated and β so overestimated that this approach appears to suffer from problems of identifiability that require further investigation.

In the Colon Cancer Family Registries,[59] we have pre-disease biospecimens on several hundred relatives of probands who were initially unaffected and subsequently became cases themselves. In a currently ongoing substudy of biomarkers for the folate pathway, it will be possible to use these samples to estimate the effect of reverse causation directly. Of course, it would have been even more informative to have both pre- and post-diagnostic biomarker measurements on incident cases to model reverse causation more accurately.

Incorporating external information: Ontologies

There are now numerous databases available that catalogue various types of genomic information. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is perhaps the most familiar of these for knowledge about the structure of pathways and the parameters of each step therein. Others include the Gene Ontology, Biomolecular Interaction Network Database, Reactome, PANTHER, Ingenuity Pathway Analysis, BioCARTA, GATHER, DAVID and the Human Protein Reference Database, (see, for example, Meier and Gehring,[60] Thomas *et al.*[61] and Werner[62] for reviews). Literature mining is emerging as another tool for this purpose,[63] although potentially biased by the vagaries of research and publication trends. Such repositories form part of a system for organising knowledge known as an 'ontology'.[64] Representation of our knowledge via an ontology may provide a more useful and broadly informative platform to generate system-wide hypotheses about how variation in human genes ultimately impacts on organism-level phenotypes via the underlying pathway or complex system. Since the biological and environmental knowledge relevant to most diseases spans many research fields, each with specific theories guiding ongoing research, expertise across the entire system by one individual scientist is limited. While the information that contributes to each knowledge domain may contain uncertainties and sources of error stemming from the underlying experiments and studies, biases in the selection of genes and pathways chosen to be included and lack of comparability across terms and databases, an ontology as a whole can generate hypotheses and links across research disciplines that may only arise when information is integrated from several disciplines across the entire span of suspected disease aetiology. An ontology should not be taken as the truth, but rather as the current representation of knowledge that can, and should, be updated as new findings arise and hypotheses are tested. Evaluation of the accuracy of ontologies is an active research area.

In our folate simulation, we considered three prior covariates for Z in Table 3. The creation of these priors followed directly from the network representation given in Figure 1, obtained from a previously published article representing one research group's interpretation of the folate pathway.[14] An ontology, such as Gene Ontology (GO), has the potential advantage of allowing for the construction of prior covariates across a range of biological mechanisms. For example, a very refined biological process captured by the GO term *folic acid and derivative biosynthetic process* indicates two genes (*MTCH* and *MS*) from our example set of genes. A more general term, *methionine biosynthetic process*, identifies three genes (*MTCH*, *MTHFR* and *MS*). Finally, a broad process, such as *one-carbon compound metabolic process*, identifies five genes (*SAHH*, *DHFR*, *MAT-II*, *MTCH* and *SHMT*). Since an ontology has a hierarchical structure in a easily computable format, one may consider more quantitative approaches in generating prior covariates, such as the distance between two genes in the ontology. Across the full range of 184 GO terms involving one or more of these 14 genes, positively correlated sets include (*MTHFR*, *MTCH*, *MS*), (*MTD*, *CBS*), (*FTD, DHFR*) and (*AICART, TS*), while *PGT* and *MTCH* are negatively correlated. Figure 3 represents these correlations using a complete agglomerative clustering.

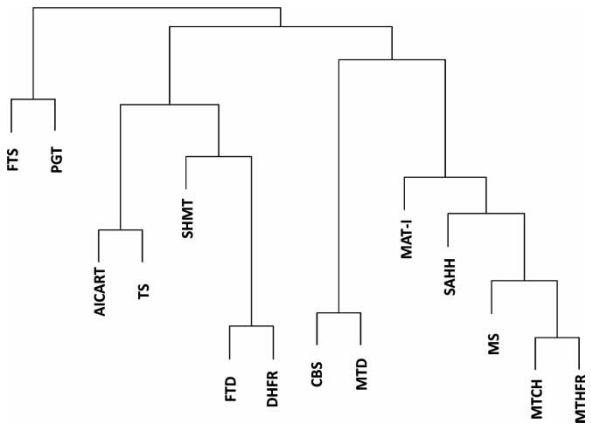

Figure 3. Hierarchical clustering of folate genes based on 184 GO terms.

Although both approaches to building prior covariates, via either the visual interpretation of a network or the use of Gene Ontology, use knowledge of biological mechanisms, they lack a formal link of these mechanisms to disease risk or organism-level phenotypes. Such links may be critical when generating hypotheses or informing statistical analyses using biological mechanisms. Many publicly available ontologies provide a vast amount of structural information on various biological processes, but interpretation or weighting of the importance of those processes in relation to specific phenotypes will only come when ontologies from biological domains are linked to ontologies characterising phenotypes. As one example, Thomas *et al.*[61] created a novel ontological representation linking smoking-related phenotypes and response to smoking cessation treatments with the underlying biological mechanisms, mainly nicotine metabolism. Most of the ontological concepts created for this specific ontology were created using concept definitions from existing ontologies, such as SOPHARM and Gene Ontology. This ontology was used in Conti *et al.*[20] to demonstrate the use in pathway analysis as a systematic way of eliciting priors for a hierarchical model. Specifically, the ontology was used to generate quantitative priors to reduce the space of potential models and to inform subsequent analysis via a Bayesian model selection approach.

Dealing with uncertainty in pathway structure

A more general question is how to deal with model uncertainty in any of these modelling strategies. The general hierarchical modelling strategy was first extended by Conti *et al.*[65] to deal with uncertainty about the set of main effects and interactions to be included in X using stochastic search variable selection.[66] Specifically, they replaced the second-level model by a pair of models, a logistic regression for the probability that $\beta_p = 0$ and a linear regression of the form of Eq. (2) for the expected values of the coefficient, given that it was not zero. In turn, the pair of second-level models inform the probability that any given term will be included in the model at the current iteration of the stochastic search. Thus, over the course of MCMC iterations, variables are entered and removed, and one can then estimate the posterior probability or Bayes factor (1) for each factor or possible model (2), for whether each factor has a non-zero β averaging over the set of other variables in the model, or (3) the posterior mean of each β, given that it is non-zero. Other alternatives include the Lasso prior,[67] which requires only a single hyperparameter to accomplish both shrinkage and variable selection in a natural way, and the elastic net,[68] which combines the Lasso and normal priors and can be implemented in a hierarchical fashion combining variable selection at lower levels (eg among SNPs within a pathway) and shrinkage at higher levels (eg between genes within a pathway or between pathways) (Chen *et al.* Presented at the Eastern North American Region Meeting of the Biometric Society; San Antonio, TX: February 2009).

In an analysis, utilising the methods described by Conti *et al.*,[20] of the simulated data when homocysteine is the causal variable (Table 5, first column) and incorporating an exchangeable prior structure in which all genes are treated equally (ie intercept only in the prior covariate matrix, Z), the posterior probabilities of including the two modestly significant genes *TS* and *FTD* are 0.57 and 0.48, respectively. By contrast, when the prior covariate matrix is derived from the 'external database' from the

simulation model and is thus more informative of the underlying mechanism, these posterior probabilities change to 0.84 and 0.14, respectively. These changes in the posterior probabilities of inclusion reflect the covariate values for these genes in relation to homocysteine concentration and the AICART reaction velocity (the two prior covariates with the largest estimated second-level effects). In the case of *TS*, the velocities for these covariates are large, resulting in an increase in the posterior probability of inclusion. By contrast, for *FTD* these values are much smaller and there is a subsequent decrease.

For mechanistic models, the 'topology' of the model Λ and the corresponding vector of model parameters θ_Λ are treated as unknown quantities, about which we might have some general prior knowledge in the form of the 'ontology' *Z*. In the microarray analysis world, Bayesian network analysis has emerged as a powerful technique for inferring the structure of a complex network of genes.[69] Might such a technique prove helpful for epidemiological analysis?

One promising approach is 'logic regression', which considers a set of tree-structured models relating measurable inputs (genes and exposures) to a disease trait through a network of unobserved intermediate nodes representing logical operators (AND, OR, XOR etc).[70] To allow for uncertainty about model form, a MCMC method is used to update the structure of the graphical model by adding, deleting, moving or changing the types of the intermediate nodes.[71] Although appealing as a way of representing the biochemical pathways, logic regression does not exploit any external information about the form of network. It also treats all intermediate nodes as binary, so it is more suitable for modelling regulatory than metabolic pathways where the intermediate nodes would represent continuous metabolite concentrations.

To overcome some of these difficulties, we relaxed the restriction to binary nodes, parameterising the model as:

$$M_j = \theta_{j1} M_{p_{j1}} + \theta_{n2} M_{p_{j2}} + (1 - \theta_{j1} - \theta_{j2}) M_{p_{j1}} M_{p_{j2}} \quad (5)$$

When both input nodes (the 'parents' $p_{j=}$ [p_{j1}, p_{j2}]) are binary, various combinations of θs will yield the full range of possible logical operators (eg AND = [0,0], OR = [1,1]), but this framework allows great flexibility in modelling interactions between continuous nodes, while remaining identifiable. The *M*s are treated as deterministic nodes, so the final metabolite concentration M_J (E,G;Λ,θ) can be calculated via a simple recursion. The disease risk is assumed to have a logistic dependence on M_J. Prior knowledge about the topology can be incorporated by use of a measure of similarity of each fitted network to the postulated true network (eg the proportion of connections in the true graph which are represented in the fitted one, minus the number of connections in the fitted graph which are not represented in the true one). In the spirit of Monte Carlo logic regression, the topology of the graph is modified by proposing to add or delete nodes or to move a connection between them using the Metropolis–Hastings algorithm.[72] Finally, the model parameters are updated conditional on the current model form. By post-processing the resulting set of graphs, various kinds of inference can be drawn, such as the posterior probability that a given input appears in the fitted graphs, that a pair of inputs is represented by a node in the graph, or the marginal effect of any input or combination of inputs on the disease risk. In small simulations, we demonstrated that the model could correctly identify the true network structure (or logically equivalent ones) and estimate the parameters well, while not identifying any incorrect models. In an application to data on ten candidate genes from the Children's Health Study, we were able to replicate the interactions found by a purely exploratory technique[73] and identified several alternative networks with comparable Bayes factors.

The folate pathway poses difficulties for mechanistic modelling because it is not a directed acyclic graph (DAG); although each arrow in Figure 1 is directed, the graph contains numerous cycles (feedback loops), making direct computation of probabilities difficult. In some instances, such cycles can be treated as single composite nodes with complex deterministic or stochastic laws, thereby rendering the remainder of the graph acyclic, but when there

are many interconnected cycles, as in the folate pathway, such decomposition may be difficult or impossible to identify. Might it be possible, however, to identify a simpler DAG that captures the key behaviour of the network? Since any DAG would be an oversimplification and there could be many such DAGs that provide a reasonable approximation, the problem of model uncertainty is important.

A further extension of the Baurley *et al.* approach to the folate simulation will now be summarised. As in their approach, we assume that each node has exactly two inputs, but now distinguish three basic types of nodes, $G \times G$, $G \times M$ (or $G \times E$) and $M \times M$. $G \times G$ nodes are treated as logical operators, yielding a binary output as high or low risk. $G \times M$ and $G \times E$ nodes represent intermediate metabolite concentrations, treated as continuous variables with deterministic values given by Michaelis–Menten kinetics with rate parameters $V_{max}(G)$ and K_m. $M \times M$ nodes are regression expressions yielding a continuous output variable with the mean parameterised as in Eq. (5). Disease risk is assumed to have a logistic dependence on one or more of the Zs. Finally, each measured biomarker B is assumed to be log-normally distributed around one of the Ms, with some measurement error variance. Rather than treating the intermediate nodes as deterministic, the likelihood of the entire graph is now calculated by peeling over possible states of all the intermediate nodes.

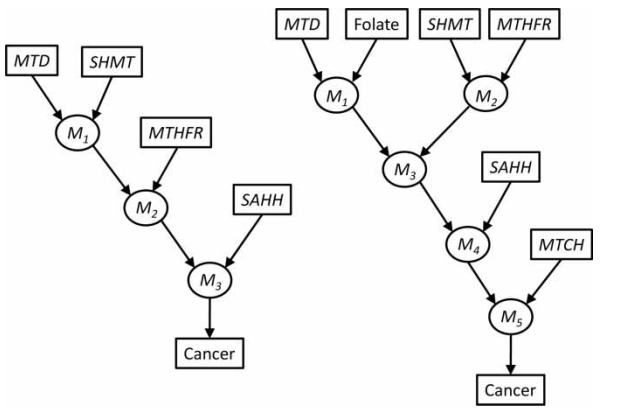

Figure 4. Top-ranking topologies without incorporating priors: left, gene only; right, genes and exposures. With no priors, the two topologies have posterior probabilities 3.9 per cent and 2.3 per cent, respectively. Using a topology derived by hierarchical clustering of the A matrix from simulated data, the top-ranked gene-only topology was identical to that shown on the left, with posterior probability of 9.5 per cent. Using the GO topology shown in Figure 3, the same genes were included, but reordered as (((*MTHR*, *SAHH*), *MTD*), *SHMT*)with a posterior probability of 6.4 per cent.

Figure 4 shows the topologies discovered by the MCMC search. The largest Bayes factors are obtained when using no prior topologies. With a prior topology, essentially the same networks are found, with somewhat different Bayes factors.

Pathways in a genome-wide context

Genome-wide association studies (GWAS) are generally seen as 'agnostic' — the antithesis of hypothesis-driven pathway-based studies. Aside from the daunting computational challenge, their primary goal is, after all, the discovery of novel genetic associations, possibly in genes with unknown function or even with genomic variation in 'gene desert' regions not known to harbour genes. How, then, could one hope to incorporate prior knowledge in a GWAS? The response has generally been to wait until the GWAS has been completed (after a multi-stage scan and independent replication) and then conduct various *in vitro* functional studies of the novel associations before attempting any pathway modelling.

The idea of incorporating prior knowledge from genomic annotation databases or other sources as a way of improving the power of a genome-wide scan for discovery has, however, been suggested by several authors. Roeder *et al.*,[74] Saccone *et al.*,[75] Wakefield[76–78] and Whittemore[79] introduced variants of a weighted false discovery rate, while Lewinger *et al.*[80] and Chen and Witte[81] described hierarchical modelling approaches for this purpose. These could be applied at any stage of a GWAS to improve the prioritisation of variants to be taken forward to the next stage. For example, Sebastiani *et al.*[82] used a Bayesian test to incorporate external information for prioritising SNP associations from the first stage of a GWAS using pooled DNA, to be subsequently tested using individual genotyping. Roeder *et al.*[74] originally

suggested the idea of exploiting external information in the context of using a prior linkage scan to focus attention in regions of the genome more likely to harbour causal variants, but subsequent authors have noted that various other types of information, such as linkage disequilibrium, functional characterisation or evolutionary conservation, could be included as predictors. An advantage of hierarchical modelling is that multiple sources can be readily incorporated in a flexible regression framework, whereas the weighted FDR requires *a priori* choice of a specific weighting scheme.

A recent trend has been the incorporation of pathway inference in genome-wide association scans,[75,83–89] borrowing ideas from the extensive literature on network analysis of gene expression array data.[90,91] Currently, the most widely used tool for this purpose is gene set enrichment analysis,[92] which in GWAS applications aims to test whether groups of genes in a common pathway tend to rank higher in significance. Several published applications have yielded novel insights using this approach,[93–96] although others have found that no specific pathway outranks the most significant single markers,[89,97,98] suggesting that the approach may not be ideal for all complex diseases. Many other empirical approaches have been used in the gene-expression field, including Bayesian network analysis,[69,99,100] neural networks,[101] support vector machines[102] and a variety of other techniques from the fields of bioinformatics, computational or systems biology and machine learning.[103–111] Most of these are empirical, although in the sense of trying to reconstruct the unknown network structure from observational data, rather than using a known network to analyse the observational data. It is less obvious how such methods could be applied to mining single-marker associations from a GWAS, but they could be helpful in mining G × G interactions. Even simple analyses of GWAS data can be computationally demanding, particularly if all possible G × G interactions are to be included, and analyses incorporating pathway information is likely to be even more daunting. Recent developments in computational algorithms for searching high-dimensional spaces and parallel cluster computing implementations may, however, make this feasible.

Recently, several authors[112–116] have undertaken analyses of the association of genome-wide expression data with genome-wide SNP genotypes in search of patterns of genetic control that would identify *cis*- and *trans*-activating factors and master regulatory regions. Ultimately, one could foresee using networks inferred from gene expression directly as priors in a hierarchical modelling analysis for GWAS data, or a joint analysis of the two phenotypes, but this has yet to be attempted. Other novel technologies, such as whole-genome sequencing, metabolomics, proteomics and so on may provide other types of data that will inform pathway-based analysis on a genome-wide scale.

Conclusions

As in any other form of statistical modelling, the analyst should be cautious in interpretation. An pointed out by Jansen:[117]

> 'So, the modeling of the interplay of many genes — which is the aim of complex systems biology — is not without danger. Any model can be wrong (almost by definition), *but particularly complex (overparameterized) models have much flexibility to hide their lack of biological relevance*' [emphasis added].

A good fit to a particular model does not, of course, establish the truth of the model. Instead, the value of models, whether descriptive or mechanistic, lies in their ability to organise a range of hypotheses into a systematic framework in which simpler models can be tested against more complex alternatives. The usefulness of the Armitage–Doll[118] multistage model of carcinogenesis, for example, lies not in our belief that it is a completely accurate description of the process, but rather in its ability to distinguish whether a carcinogen appears to act early or late in the process or at more than one stage. Similarly, the importance of the Moolgavkar–Knudson two-stage clonal-expansion model[119] lies in its ability to test whether a carcinogen acts as an 'initiator' (ie on the mutation rates) or a 'promoter' (ie on proliferation rates). Such inferences can be valuable, even if the model itself

is an incomplete description of the process, as must always be the case.

Although mechanistic models do make some testable predictions about such things as the shape of the dose–response relationship and the modifying effects of time-related variables, testing such patterns against epidemiological data tends to provide only weak evidence in support of the alternative models, and only within the context of all the other assumptions involved. Generally, comparisons of alternative models (or specific sub-models) can only be accomplished by direct fitting. Visualisation of the fit to complex epidemiological datasets can be challenging. Any mechanistic interpretations of model fits should therefore consider carefully the robustness of these conclusions to possible misspecification of other parts of the model.

Acknowledgments

This work was supported in part by NIH grants R01-CA92562, P50-ES07048, R01-CA112237 and U01-ES015090 (D.C.T., D.V.C., J.B.), R01-CA105437, R01-CA105145, R01-CA59045 (C.M.U.) and NSF grants DMS-0616710 and DMS-0109872 (F.N., M.R.). The authors are particularly grateful to Wei Liang and Fan Yang for programming support.

References

1. Thomas, D.C. (2005), 'The need for a comprehensive approach to complex pathways in molecular epidemiology', *Cancer Epidemiol. Biomarkers Prev.* Vol. 14, pp. 557–559.
2. Thomas, D.C., Baurley, J.W., Brown, E.E., Figueiredo, J. *et al.* (2008), 'Approaches to complex pathways in molecular epidemiology: Summary of an AACR special conference', *Cancer Res.* Vol. 68, pp. 10028–10030.
3. Cook, N.R., Zee, R.Y. and Ridker, P.M. (2004), 'Tree and spline based association analysis of gene-gene interaction models for ischemic stroke', *Stat. Med.* Vol. 23, pp. 1439–1453.
4. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R. *et al.* (2001), 'Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer', *Am. J. Hum. Genet.* Vol. 69, pp. 138–147.
5. Hoh, J. and Ott, J. (2003), 'Mathematical multi-locus approaches to localizing complex human trait genes', *Nat. Rev. Genet.* Vol. 4, pp. 701–709.
6. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q. *et al.* (1999), 'Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation', *Proc. Natl. Acad. Sci. USA* Vol. 96, pp. 2907–2912.
7. Tahri-Daizadeh, N., Tregouet, D.A., Nicaud, V., Manuel, N. *et al.* (2003), 'Automated detection of informative combined effects in genetic association studies of complex traits', *Genome Res.* Vol. 13, pp. 1952–1960.
8. Potter, J.D. (1999), 'Colorectal cancer: Molecules and populations', *J. Natl. Cancer Inst.* Vol. 91, pp. 916–932.
9. Frosst, P., Blom, H.J., Milos, R., Goyette, P. *et al.* (1995), 'A candidate genetic risk factor for vascular disease: A common mutation in methylenetetrahydrofolate reductase', *Nat. Genet.* Vol. 10, pp. 111–3.
10. Ulrich, C.M. and Potter, J.D. (2006), 'Folate supplementation: Too much of a good thing?', *Cancer Epidemiol. Biomarkers Prev.* Vol. 15, pp. 189–93.
11. Molloy, A.M., Brody, L.C., Mills, J.L., Scott, J.M. *et al.* (2009), 'The search for genetic polymorphisms in the homocysteine/folate pathway that contribute to the etiology of human neural tube defects', *Birth Defects Res. A Clin. Mol. Teratol.* Vol. 85, pp. 285–94.
12. Nijhout, H.F., Reed, M.C., Budu, P. and Ulrich, C.M. (2004), 'A mathematical model of the folate cycle: New insights into folate homeostasis', *J Biol. Chem.* Vol. 279, pp. 55008–16.
13. Nijhout, H.F., Reed, M.C. and Ulrich, C.M. (2008), 'Mathematical models of folate-mediated one-carbon metabolism', *Vitam. Horm.* Vol. 79, pp. 45–82.
14. Reed, M.C., Nijhout, H.F., Neuhouser, M.L., Gregory, J.F., 3rd. *et al.* (2006), 'A mathematical model gives insights into nutritional and genetic aspects of folate-mediated one-carbon metabolism', *J. Nutr.* Vol. 136, pp. 2653–61.
15. Reed, M.C., Thomas, R.L., Pavisic, J., James, S.J. *et al.* (2008), 'A mathematical model of glutathione metabolism', *Theor Biol. Med. Model*, Vol. 5, p. 8.
16. Ulrich, C.M., Neuhouser, M., Liu, A.Y., Boynton, A. *et al.* (2008), 'Mathematical modeling of folate metabolism: Predicted effects of genetic polymorphisms on mechanisms and biomarkers relevant to carcinogenesis', *Cancer Epidemiol. Biomarkers Prev.*, Vol. 17, pp. 1822–31.
17. Hung, R.J., Brennan, P., Malaveille, C., Porru, S. *et al.* (2004), 'Using hierarchical modeling in genetic association studies with multiple markers: Application to a case-control study of bladder cancer', *Cancer Epidemiol. Biomarkers Prev.* Vol. 13, pp. 1013–1021.
18. Capanu, M., Orlow, I., Berwick, M., Hummer, A.J. *et al.* (2008), 'The use of hierarchical models for estimating relative risks of individual genetic variants: An application to a study of melanoma', *Stat. Med.* Vol. 27, pp. 1973–1992.
19. Hung, R.J., Baragatti, M., Thomas, D., McKay, J. *et al.* (2007), 'Inherited predisposition of lung cancer: A hierarchical modeling approach to DNA repair and cell cycle control pathways', *Cancer Epidemiol. Biomarkers Prev.* Vol. 16, pp. 2736–2744.
20. Conti, D.V., Lewinger, J.P., Swan, G.E., Tyndale, R.F. *et al.* (2009), 'Using ontologies in hierarchical modeling of genes and exposures in biologic pathways', in: Swans, G.E. (ed.), *Phenotypes and Endophenotypes: Foundations for Genetic Studies of Nicotine Use and Dependence*, NCI Tobocco Control Monographs, Bethesda, MD, pp. 539–584.
21. Rebbeck, T.R., Spitz, M. and Wu, X. (2004), 'Assessing the function of genetic variants in candidate gene association studies', *Nat. Rev. Genet.* Vol. 5, pp. 589–597.
22. Besag, J., York, J. and Mollie, A. (1991), 'Bayesian image restoration with two applications in spatial statistics (with discussion)', *Ann. Inst. Statist. Math.* Vol. 43, pp. 1–59.
23. Cortessis, V. and Thomas, D.C. (2003), 'Toxicokinetic genetics: An approach to gene–environment and gene–gene interactions in complex metabolic pathways', in: Bird, P., Boffetta, P., Buffler, P. and Rice, J. (eds), *Mechanistic Considerations in the Molecular Epidemiology of Cancer*, IARC Scientific Publications, Lyon, France, pp. 127–150.
24. Du, L., Conti, D.V. and Thomas, D.C. (2006), 'Physiologically-based pharmacokinetic modeling platform for genetic and exposure effects in metabolic pathways', *Genet. Epidemiol.* Vol. 29, p. 234.
25. Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000), 'Winbugs — A Bayesian modelling framework: Concepts, structure, and extensibility', *Stat. Comput.* Vol. 10, pp. 325–337.
26. Lunn, D.J., Best, N., Thomas, A., Wakefield, J. and Spiegelhalter, D. (2002), 'Bayesian analysis of population PK/PD models: General concepts and software', *J. Pharmacokinet. Pharmacodyn* Vol. 29, No. 3, pp. 271–307.
27. Racine-Poon, A. and Wakefield, J. (1998), 'Statistical methods for population pharmacokinetic modelling', *Stat. Meth. Med. Res.* Vol. 7, pp. 63–84.

28. Bois, F.Y. (2001), 'Applications of population approaches in toxicology', *Toxicol. Lett.* Vol. 120, pp. 385–394.
29. Bennett, J.E. and Wakefield, J.C. (1996), 'A comparison of a Bayesian population method with two methods as implemented in commercially available software', *J. Pharmacokinet. Biopharm.* Vol. 24, pp. 403–432.
30. Wakefield, J. (1996), 'Bayesian individualization via sampling-based methods', *J. Pharmacokinet. Biopharm.* Vol. 24, pp. 103–131.
31. Best, N.G., Tan, K.K., Gilks, W.R. and Spiegelhalter, D.J. (1995), 'Estimation of population pharmacokinetics using the Gibbs sampler', *J. Pharmacokinet. Biopharm.* Vol. 23, pp. 407–435.
32. Kou, S.C., Cherayil, B.J., Min, W., English, B.P. *et al.* (2005), 'Single-molecule Michaelis-Menten equations', *J. Phys. Chem. B* Vol. 109, pp. 19068–19081.
33. English, B.P., Min, W., van Oijen, A.M., Lee, K.T. *et al.* (2006), 'Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited', *Nat. Chem. Biol.* Vol. 2, pp. 87–94.
34. Marjoram, P., Molitor, J., Plagnol, V. and Tavare, S. (2003), 'Markov chain Monte Carlo without likelihoods', *Proc. Natl. Acad. Sci. USA* Vol. 100, pp. 15324–15328.
35. Thomas, D.C. (2007), 'Using gene-environment interactions to dissect the effects of complex mixtures', *J. Expo. Sci. Environ. Epidemiol.* Vol. 17 (Suppl. 2), pp. S71–S74.
36. Parl, F., Crooke, P., Conti, D.V. and Thomas, D.C. (2008), 'Pathway-based methods in molecular cancer epidemiology', in: Rebbeck, T.R., Ambrosone, C.B. and Shields, P.G. (eds), *Fundamentals of Molecular Epidemiology*, Informa Healthcare, New York, NY, pp. 189–204.
37. Spiegelman, D., Carroll, R.J. and Kipnis, V. (2001), 'Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument', *Stat. Med.* Vol. 20, pp. 139–160.
38. Holcroft, C.A. and Spiegelman, D. (1999), 'Design of validation studies for estimating the odds ratio of exposure-disease relationships when exposure is misclassified', *Biometrics* Vol. 55, pp. 1193–1201.
39. Thomas, D.C. (2007), 'Multistage sampling for latent variable models', *Lifetime Data Anal.* Vol. 13, pp. 565–581.
40. Breslow, N.E. and Chatterjee, N. (1999), 'Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis', *Appl. Statist.* Vol. 48, pp. 457–468.
41. Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006), *Measurement Error in Nonlinear models: A Modern Perspective* (2nd edn.), Chapman and Hall CRC Press, London, UK.
42. Thomas, D.C., Stram, D. and Dwyer, J. (1993), 'Exposure measurement error: Influence on exposure-disease relationships and methods of correction', *Annu. Rev. Publ. Health* Vol. 14, pp. 69–93.
43. Davey Smith, G. and Ebrahim, S. (2003), '"Mendelian randomization": Can genetic epidemiology contribute to understanding environmental determinants of disease?', *Int. J. Epidemiol.* Vol. 32, pp. 1–22.
44. Davey Smith, G. and Ebrahim, S. (2004), 'Mendelian randomization: Prospects, potentials, and limitations', *Int. J. Epidemiol.* Vol. 33, pp. 30–42.
45. Davey Smith, G. and Ebrahim, S. (2005), 'What can Mendelian randomisation tell us about modifiable behavioural and environmental exposures?', *BMJ* Vol. 330, pp. 1076–1079.
46. Lewis, S.J. and Davey Smith, G. (2005), 'Alcohol, aldh2, and esophageal cancer: A meta-analysis which illustrates the potentials and limitations of a Mendelian randomization approach', *Cancer Epidemiol. Biomarkers Prev.* Vol. 14, pp. 1967–1971.
47. Thompson, J.R., Minelli, C., Abrams, K.R., Tobin, M.D. *et al.* (2005), 'Meta-analysis of genetic studies using Mendelian randomization — A multivariate approach', *Stat. Med.* Vol. 24, pp. 2241–2254.
48. Tobin, M.D., Minelli, C., Burton, P.R. and Thompson, J.R. (2004), 'Commentary: Development of Mendelian randomization: From hypothesis test to "Mendelian deconfounding"', *Int. J. Epidemiol.* Vol. 33, pp. 26–29.
49. Glynn, R.J. (2006), 'Commentary. Genes as instruments for evaluation of markers and causes', *Int. J. Epidemiol.* Vol. 35, pp. 932–934.
50. Hernan, M.A. and Robins, J.M. (2006), 'Instruments for causal inference: An epidemiologist's dream?', *Epidemiology* Vol. 17, pp. 360–372.
51. Brookhart, M.A., Wang, P.S., Solomon, D.H. and Schneeweiss, S. (2006), 'Instrumental variable analysis of secondary pharmacoepidemiologic data', *Epidemiology* Vol. 17, pp. 373–374.
52. Buzas, J.S. and Stefanski, L.A. (1996), 'Instrumental variable estimation in generalized linear measurement error models', *J. Am. Stat. Assoc.* Vol. 91, pp. 999–1006.
53. Greenland, S. (2000), 'An introduction to instrumental variables for epidemiologists', *Int. J. Epidemiol.* Vol. 29, p. 1102.
54. Martens, E.P., Pestman, W.R., de Boer, A., Belitser, S.V. *et al.* (2006), 'Instrumental variables: Application and limitations', *Epidemiology* Vol. 17, pp. 260–267.
55. Didelez, V. and Sheehan, N. (2007), 'Mendelian randomization as an instrumental variable approach to causal inference', *Stat. Meth. Med. Res.* Vol. 16, pp. 309–330.
56. Nitsch, D., Molokhia, M., Smeeth, L., DeStavola, B.L. *et al.* (2006), 'Limits to causal inference based on Mendelian randomization: A comparison with randomized controlled trials', *Am. J. Epidemiol.* Vol. 163, pp. 397–403.
57. Bautista, L.E., Smeeth, L., Hingorani, A.D. and Casas, J.P. (2006), 'Estimation of bias in nongenetic observational studies using "Mendelian triangulation"', *Ann. Epidemiol.* Vol. 16, pp. 675–680.
58. Thomas, D.C. and Conti, D.V. (2004), 'Commentary. The concept of "Mendelian randomization"', *Int. J. Epidemiol.* Vol. 33, pp. 21–25.
59. Newcomb, P.A., Baron, J., Cotterchio, M., Gallinger, S. *et al.* (2007), 'Colon cancer family registry: An international resource for studies of the genetic epidemiology of colon cancer', *Cancer Epidemiol. Biomarkers Prev.* Vol. 16, pp. 2331–2343.
60. Meier, S. and Gehring, C. (2008), 'A guide to the integrated application of on-line data mining tools for the inference of gene functions at the systems level', *Biotechnol. J.* Vol. 3, pp. 1375–1387.
61. Thomas, P.D., Mi, H., Swan, G.E., Lerman, C. *et al.* (2009), 'A systems biology network model for genetic association studies of nicotine addiction and treatment', *Pharmacogenet. Genomics* Vol. 19, pp. 538–551.
62. Werner, T. (2008), 'Bioinformatics applications for pathway analysis of microarray data', *Curr. Opin. Biotechnol.* Vol. 19, pp. 50–54.
63. Jensen, L.J., Saric, J. and Bork, P. (2006), 'Literature mining for the biologist: From information retrieval to biological discovery', *Nat. Rev. Genet.* Vol. 7, pp. 119–129.
64. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D. *et al.* (2000), 'Gene ontology: Tool for the unification of biology', *Nat. Genet.* Vol. 25, pp. 25–29.
65. Conti, D.V., Cortessis, V., Molitor, J. and Thomas, D.C. (2003), 'Bayesian modeling of complex metabolic pathways', *Hum. Hered.* Vol. 56, pp. 83–93.
66. George, E.I. and McCulloch, R.E. (1993), 'Variable selection via Gibbs sampling', *J. Am. Stat. Assoc.* Vol. 88, pp. 881–889.
67. Park, T. and Casella, G. (2008), 'The Bayesian lasso', *J. Am. Stat. Assoc.* Vol. 103, pp. 681–686.
68. Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *J. R. Stat. Soc. Ser. B* Vol. 67, pp. 301–320.
69. Friedman, N. (2004), 'Inferring cellular networks using probabilistic graphical models', *Science* Vol. 303, pp. 799–805.
70. Ruczinski, I., Kooperberg, C. and LeBlanc, M.L. (2004), 'Exploring interactions in high-dimensional genomic data: An overview of logic regression, with applications', *J. Multivar. Anal.* Vol. 90, pp. 178–195.
71. Kooperberg, C. and Ruczinski, I. (2005), 'Identifying interacting SNPs using Monte Carlo logic regression', *Genet. Epidemiol.* Vol. 28, pp. 157–170.
72. Hastings, W. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* Vol. 57, pp. 97–109.
73. Millstein, J., Conti, D.V., Gilliland, F.D. and Gauderman, W.J. (2006), 'A testing framework for identifying susceptibility genes in the presence of epistasis', *Am. J. Hum. Genet.* Vol. 78, pp. 15–27.
74. Roeder, K., Devlin, B. and Wasserman, L. (2007), 'Improving power in genome-wide association studies: Weights tip the scale', *Genet. Epidemiol.* Vol. 31, pp. 741–747.
75. Saccone, S.F., Saccone, N.L., Swan, G.E., Madden, P.A. *et al.* (2008), 'Systematic biological prioritization after a genome-wide association

study: An application to nicotine dependence', *Bioinformatics* Vol. 24, pp. 1805–1811.
76. Wakefield, J. (2008), 'Bayes factors for genome-wide association studies: Comparison with p-values', *Genet. Epidemiol.* Vol. 33, pp. 79–86.
77. Wakefield, J. (2008), 'Reporting and interpretation in genome-wide association studies', *Int. J. Epidemiol.* Vol. 37, pp. 641–653.
78. Wakefield, J. (2007), 'A Bayesian measure of the probability of false discovery in genetic epidemiology studies', *Am. J. Hum. Genet.* Vol. 81, pp. 208–227.
79. Whittemore, A.S. (2007), 'A Bayesian false discovery rate for multiple testing', *J. Appl. Statist.* Vol. 34, pp. 1–9.
80. Lewinger, J.P., Conti, D.V., Baurley, J.W., Triche, T.J. *et al.* (2007), 'Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation', *Genet. Epidemiol.* Vol. 31, pp. 871–882.
81. Chen, G.K. and Witte, J.S. (2007), 'Enriching the analysis of genome-wide association studies with hierarchical modeling', *Am. J. Hum. Genet.* Vol. 81, pp. 397–404.
82. Sebastiani, P., Zhao, Z., Abad-Grau, M.M., Riva, A. *et al.* (2008), 'A hierarchical and modular approach to the discovery of robust associations in genome-wide association studies from pooled DNA samples', *BMC Genet.* Vol. 9, p. 6.
83. Wang, K., Li, M. and Bucan, M. (2007), 'Pathway-based approaches for analysis of genomewide association studies', *Am. J. Hum. Genet.* Vol. 81, pp. 1278–1283.
84. Elbers, C.C., van Eijk, K.R., Franke, L., Mulder, F. *et al.* (2009), 'Using genome-wide pathway analysis to unravel the etiology of complex diseases', *Genet. Epidemiol.* Vol. 33, pp. 419–431.
85. Chasman, D.I. (2008), 'On the utility of gene set methods in genome-wide association studies of quantitative traits', *Genet. Epidemiol.* Vol. 32, pp. 658–668.
86. Holden, M., Deng, S., Wojnowski, L. and Kulle, B. (2008), 'Gsea-SNP: Applying gene set enrichment analysis to SNP data from genome-wide association studies', *Bioinformatics* Vol. 24, pp. 2784–2785.
87. Bush, W.S., Dudek, S.M. and Ritchie, M.D. (2009), 'Biofilter: A knowledge-integration system for the multi-locus analysis of genome-wide association studies', *Pac. Symp. Biocomput.* pp. 368–379.
88. Rajagopalan, D. and Agarwal, P. (2005), 'Inferring pathways from gene lists using a literature-derived network of biological relationships', *Bioinformatics* Vol. 21, pp. 788–793.
89. Hong, M.G., Pawitan, Y., Magnusson, P.K. and Prince, J.A. (2009), 'Strategies and issues in the detection of pathway enrichment in genome-wide association studies', *Hum. Genet.* (Epub ahead of print).
90. Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A. *et al.* (2003), 'Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes', *Nat. Genet.* Vol. 34, pp. 267–273.
91. Pan, W. (2005), 'Incorporating biological information as a prior in an empirical Bayes approach to analyzing microarray data', *Stat. Appl. Genet. Mol. Biol.* Vol. 4, Art. 12.
92. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S. *et al.* (2005), 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Stat. Appl. Genet. Mol. Biol.* Vol. 4, Art. 12, *Proc. Natl. Acad. Sci. USA* Vol. 102, pp. 15545–15550.
93. Lesnick, T.G., Papapetropoulos, S., Mash, D.C., Ffrench–Mullen, J. *et al.* (2007), 'A genomic pathway approach to a complex disease: Axon guidance and Parkinson disease'. *PLoS Genet.* Vol. 3, p. e98.
94. Baranzini, S.E., Galwey, N.W., Wang, J., Khankhanian, P. *et al.* (2009), 'Pathway and network-based analysis of genome-wide association studies in multiple sclerosis', *Hum. Mol. Genet.* Vol. 18, pp. 2078–2090.
95. Torkamani, A., Topol, E.J. and Schork, N.J. (2008), 'Pathway analysis of seven common diseases assessed by genome-wide association', *Genomics* Vol. 92, pp. 265–272.
96. Vink, J.M., Smit, A.B., de Geus, E.J., Sullivan, P. *et al.* (2009), 'Genome-wide association study of smoking initiation and current smoking', *Am. J. Hum. Genet.* Vol. 84, pp. 367–379.
97. Perry, J.R., McCarthy, M.I., Hattersley, A.T., Zeggini, E. *et al.* (2009), 'Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach', *Diabetes* Vol. 58, pp. 1463–1467.
98. Kasperaviciute, D., Weale, M.E., Shianna, K.V., Banks, G.T. *et al.* (2007), 'Large-scale pathways-based association study in amyotrophic lateral sclerosis', *Brain* Vol. 130, pp. 2292–2301.
99. Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000), 'Using Bayesian networks to analyze expression data', *J. Comput. Biol.* Vol. 7, pp. 601–620.
100. Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J. *et al.* (2004), 'Advances to Bayesian network inference for generating causal networks from observational biological data', *Bioinformatics* Vol. 20, pp. 3594–3603.
101. Ritchie, M.D., White, B.C., Parker, J.S., Hahn, C.W. *et al.* (2003), 'Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases', *BMC Bioinformatics* Vol. 4, p. 28.
102. Byvatov, E. and Schneider, G. (2003), 'Support vector machine applications in bioinformatics', *Appl. Bioinformatics* Vol. 2, pp. 67–77.
103. Schafer, J. and Strimmer, K. (2005), 'An empirical Bayes approach to inferring large-scale gene association networks', *Bioinformatics* Vol. 21, pp. 754–764.
104. Wu, C.C., Huang, H.C., Juan, H.F. and Chen, S.T. (2004), 'Genenetwork: An interactive tool for reconstruction of genetic networks using microarray data', *Bioinformatics* Vol. 20, pp. 3691–3693.
105. Franke, L., van Bakel, H., Fokkens, L., de Jong, E.D. *et al.* (2006), 'Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes', *Am. J. Hum. Genet.* Vol. 78, pp. 1011–1025.
106. Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U. *et al.* (2005), 'Reverse engineering of regulatory networks in human b cells', *Nat. Genet.* Vol. 37, pp. 382–390.
107. Kim, T.H. and Ren, B. (2006), 'Genome-wide analysis of protein-DNA interactions', *Annu. Rev. Genom. Hum. Genet.* Vol. 7, pp. 81–102.
108. Tu, Z., Wang, L., Arbeitman, M.N., Chen, T. *et al.* (2006), 'An integrative approach for causal gene identification and gene regulatory pathway inference', *Bioinformatics* Vol. 22, pp. e489–e496.
109. Yu, H., Zhu, X., Greenbaum, D. *et al.* (2004), 'Topnet: A tool for comparing biological sub-networks, correlating protein properties with topological statistics', *Nucleic Acids Res.* Vol. 32, pp. 328–337.
110. Blais, A. and Dynlacht, B.D. (2005), 'Constructing transcriptional regulatory networks', *Genes Dev.* Vol. 19, pp. 1499–1511.
111. Xie, Y., Pan, W., Jeong, K.S. and Khodursky, A. (2007), 'Incorporating prior information via shrinkage: A combined analysis of genome-wide location data and gene expression data', *Stat. Med.* Vol. 26, pp. 2258–2275.
112. Dixon, A.L., Liang, L., Moffatt, M.F., Chen, W. *et al.* (2007), 'A genome-wide association study of global gene expression', *Nat. Genet.* Vol. 39, pp. 1202–1207.
113. Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J. *et al.* (2005), 'Genome-wide associations of gene expression variation in humans', *PLoS Genet.* Vol. 1, p. e78.
114. Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L. *et al.* (2004), 'Genetic analysis of genome-wide variation in human gene expression', *Nature* Vol. 430, pp. 743–747.
115. Cheung, V.G., Spielman, R.S., Ewens, K.G., Weber, T.M. *et al.* (2005), 'Mapping determinants of human gene expression by regional and genome-wide association', *Nature* Vol. 437, pp. 1365–1369.
116. Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M. *et al.* (2003), 'Natural variation in human gene expression assessed in lymphoblastoid cells', *Nat. Genet.*, Vol. 33, pp. 422–425.
117. Jansen, R.C. (2003), 'Studying complex biological systems using multifactorial perturbation', *Nat. Rev. Genet.* Vol. 4, pp. 145–151.
118. Armitage, P. and Doll, R. (1954), 'The age distribution of cancer and multi-stage theory of carcinogenesis', *Br. J. Cancer* Vol. 8, No. 1, pp. 1–12.
119. Moolgavkar, S. and Knudson, A. (1981), 'Mutation and cancer: A model for human carcinogenesis', *J.N.C.I.*, Vol. 66, pp. 1037–1052.