Multi-factor dimensionality reduction applied to a large prospective investigation on gene–gene and gene–environment interactions

M.Manuguerra[1], G.Matullo[1,2,*], F.Veglia[1], H.Autrup[3], A.M.Dunning[4], S.Garte[5], E.Gormally[6], C.Malaveille[6], S.Guarrera[1], S.Polidoro[1], F.Saletta[1], M.Peluso[7], L.Airoldi[8], K.Overvad[9], O.Raaschou-Nielsen[10], F.Clavel-Chapelon[11], J.Linseisen[12], H.Boeing[13], D.Trichopoulos[14], A.Kalandidi[14], D.Palli[15], V.Krogh[16], R.Tumino[17], S.Panico[18], H.B.Bueno-De-Mesquita[19], P.H.Peeters[20], E.Lund[21], G.Pera[22], C.Martinez[23], P.Amiano[24], A.Barricarte[25], M.J.Tormo[26], J.R.Quiros[27], G.Berglund[28], L.Janzon[28], B.Jarvholm[29], N.E.Day[30], N.E.Allen[31], R.Saracci[6], R.Kaaks[6], P.Ferrari[6], E.Riboli[32,33] and P.Vineis[1,32,33]

[1]ISI Foundation, Torino, Italy, [2]Department of Genetics, Biology and Biochemistry, University of Turin, Turin, Italy, [3]Department of Environmental and Occupational Medicine, Aarhus Universitet, Aarhus, Denmark, [4]Department of Oncology, Strangeways Research Laboratory, University of Cambridge, Cambridge, UK, [5]Genetics Research Institute, Milan, Italy, [6]International Agency for Research on Cancer, Lyon, France, [7]Tuscany Cancer Institute, Cancer Risk Factor Branch, CSPO-Scientific Institute of Tuscany, Florence, Italy, [8]Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy, [9]Department of Clinical Epidemiology, Aalborg Hospital, Aarhus University Hospital, Aalborg, Denmark, [10]Institute of Cancer Epidemiology, Danish Cancer Society, Copenhagen, Denmark, [11]INSERM U521, Institut Gustave Roussy, Villejuif, France, [12]Division of Clinical Epidemiology, Deutsches Krebsforschungszentrum, Heidelberg, Germany, [13]German Institute of Human Nutrition, Potsdam-Rehbücke, Germany, [14]Department of Hygiene and Epidemiology, Medical School, University of Athens, Greece, [15]Molecular and Nutritional Epidemiology Unit, CSPO-Scientific Institute of Tuscany, Florence, Italy, [16]Department of Epidemiology, National Cancer Istitute, Milan, Italy, [17]Cancer Registry, Azienda Ospedaliera 'Civile MP Arezzo', Ragusa, Italy, [18]Dipartimento di Medicina Clinica e Sperimentale, Università Federico II, Naples, Italy, [19]Centre for Nutrition and Health, National Institute for Public Health and the Environment, Bilthoven, The Netherlands, [20]Department of Cancer Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center, Utrecht, The Netherlands, [21]Institute of Community Medicine, University of Tromso, Norway, [22]Department of Epidemiology, Catalan Institute of Oncology, Consejería de Sanidad y Servicios Sociales, Barcelona, Spain, [23]Andalusian School of Public Health, Granada, Spain, [24]Department of Public Health of Guipuzkoa, San Sebastian, Spain, [25]Public Health Institute, Navarra, Spain, [26]Consejería de Sanidad y Consumo, Murcia, Spain, [27]Public Health and Health Planning Directorate, Asturias, Spain, [28]Malmö Diet and Cancer Study, Lund University, Malmö, Sweden, [29]Department of Public Health and Clinical Medicine, University of Umeå, Sweden, [30]MRC Dunn Human Nutrition Unit, Cambridge, UK, [31]Cancer Research UK Epidemiology Unit, University of Oxford, UK, [32]Imperial College London, London, UK and [33]University of Torino, Italy

*To whom correspondence should be addressed. Section of Epidemiology, ISI Foundation (Institute for Scientific Interchange), Viale Settimio Severo, 65, 10133 Torino, Italy. Tel: +390116705601; Fax: +390112365601; Email: matullo@isiosf.isi.it

It is becoming increasingly evident that single-locus effects cannot explain complex multifactorial human diseases like cancer. We applied the multi-factor dimensionality reduction (MDR) method to a large cohort study on gene–environment and gene–gene interactions. The study (case-control nested in the EPIC cohort) was established to investigate molecular changes and genetic susceptibility in relation to air pollution and environmental tobacco smoke (ETS) in non-smokers. We have analyzed 757 controls and 409 cases with bladder cancer ($n = 124$), lung cancer ($n = 116$) and myeloid leukemia ($n = 169$). Thirty-six gene variants (DNA repair and metabolic genes) and three environmental exposure variables (measures of air pollution and ETS at home and at work) were analyzed. Interactions were assessed by prediction error percentage and cross-validation consistency (CVC) frequency. For lung cancer, the best model was given by a significant gene–environment association between the base excision repair (BER) XRCC1-Arg399Gln polymorphism, the double-strand break repair (DSBR) BRCA2-Asn372His polymorphism and the exposure variable 'distance from heavy traffic road', an indirect and robust indicator of air pollution (mean prediction error of 26%, $P < 0.001$, mean CVC of 6.60, $P = 0.02$). For bladder cancer, we found a significant 4-loci association between the BER APE1-Asp148Glu polymorphism, the DSBR RAD52-3′-untranslated region (3′-UTR) polymorphism and the metabolic gene polymorphisms COMT-Val158Met and MTHFR-677C > T (mean prediction error of 22%, $P < 0.001$, mean CVC consistency of 7.40, $P < 0.037$). For leukemia, a 3-loci model including RAD52-2259C > T, MnSOD-Ala9Val and CYP1A1-Ile462Val had a minimum prediction error of 31% ($P < 0.001$) and a maximum CVC of 4.40 ($P = 0.086$). The MDR method seems promising, because it provides a limited number of statistically stable interactions; however, the biological interpretation remains to be understood.

Abbreviations: MDR, multi-factor dimensionality reduction; CVC, cross-validation consistency; DSBR, double-strand break repair; ETS, environmental tobacco smoke; FPRP, false positive report probability.

Introduction

Environmental carcinogens contained in air pollution and environmental tobacco smoke (ETS) form mainly bulky DNA adducts and also generate interstrand cross-links and reactive oxygen species, which induce base damage, abasic sites, single- and double-strand breaks.

Many of the known environmental risk factors are metabolized in the organism to products that are either more carcinogenic or are detoxified. Genetic polymorphisms have been identified in many of these enzymes, and the biological consequence of some of such changes is an altered enzyme activity, which may influence the ratio between activation and deactivation, and thus the cancer risk (1). Enzymes involved in the biotransformation can be divided into phase 1 enzymes, involved in oxidative processes and phase 2 enzymes involved in the detoxification of the

primary compounds or their metabolites. Genetic polymorphisms in both phase 1 and phase 2 metabolic enzymes have been shown to influence both cancer risk and the DNA adduct levels (2). We have demonstrated previously that people with high levels of adducts at the time they were healthy had a subsequent higher risk of developing cancer (3).

Unrepaired damage can result in apoptosis or may lead to unregulated cell growth and cancer (4). Alternatively, the damage can be repaired at the DNA level enabling the cell to replicate as planned. Because of the importance of maintaining genomic integrity in the general and specialized functions of cells as well as in the prevention of carcinogenesis, genes coding for DNA repair molecules have been proposed as candidate cancer-susceptibility genes (5,6).

Studies to date indicate that variation in DNA repair and metabolic genes may influence cancer susceptibility (7,8); however, results are not fully consistent, and other potentially important polymorphisms in these and other genes have not been explored yet.

Based on the extensive reviews of the literature (8–11) several polymorphisms of DNA repair, metabolic and other genes potentially associated with lung cancer, bladder cancer and myeloid leukemia have been analyzed in the context of the GEN-AIR project, a case-control study nested in the EPIC investigation. The aim of the present paper is to apply the model-free multi-factor dimensionality reduction (MDR) method (12) to a large study whose primary goals were to investigate the role of air pollution and ETS in increasing the risk of cancer, and to investigate gene–environment and gene–gene interactions.

In fact, it is becoming increasingly evident that single-locus effects cannot explain complex multifactorial human diseases like cancer. Recent reviews and meta-analyses have demonstrated the difficulties associated with replicating main effect results. Thus, when the single polymorphism effect is not present alone or is not strong enough, the identification and characterization of susceptibility genes for cancer risk require the understanding of gene–gene interactions.

Materials and methods

Subjects

European Prospective Investigation into Cancer and Nutrition (EPIC) is a multicenter European study, coordinated by the International Agency for Research on Cancer (Lyon), in which more than 500 000 healthy volunteers were recruited in 10 European countries (France, Denmark, Germany, Greece, Italy, The Netherlands, Norway, Spain, Sweden and UK) corresponding to 23 recruitment centers. The cohort includes subjects of both genders, mostly in the age range 35–74 at recruitment. Recruitment took place between 1993 and 1998. Detailed dietary and lifestyle histories collected mainly through self-administered questionnaires, plus a 24 h dietary recall through person-to-person interview (in a 10% sample), anthropological measurements and a 30–40 ml blood sample are available. All questionnaire information is available in a computerized format. Signed informed consent forms were collected from all participants (except a sub-group of the Oxford cohort who gave consent on postal questionnaires).

Gen-Air is a case-control study nested within the EPIC cohort, aiming at studying the relationship between some types of cancer and air pollution or ETS. Cases are subjects with bladder, lung, oral, pharyngeal or laryngeal cancers and leukemia, all newly diagnosed after recruitment. Also deaths from respiratory diseases (COPD and emphysema) were identified and included. Only non-smokers or ex-smokers since at least 10 years have been included in Gen-Air. We have matched three controls per case for exposure assessment and the analysis of questionnaire data, and two controls per case for laboratory analyses. Matching criteria were gender, age (±5 years), smoking status (never or former smoker), country of recruitment, and follow-up time. Mean follow-up was 89 months (minimum 51 and maximum 123).

Gen-Air has been approved by the Ethical Committee of the International Agency for Research on Cancer, and by the local Ethical Committees of the 23 centres.

We have identified 4051 subjects (1074 cases and 2977 controls) in EPIC who met the Gen-Air protocol criteria. The distribution of cases by cancer site or cause of death was as follows: bladder cancer 227, lung cancer 271, oral, pharyngeal cancer 73, laryngeal cancer 58, leukemias 311, deaths from respiratory diseases 134. Of these subjects, 2410 had blood samples (846 cases and 1564 controls). The Malmo center has decided not to allow the use of their blood samples and the Umea center did not allow genetic analyses, but they participate in the rest of the project.

In the present study, we have analyzed 409 cases of bladder cancer ($n = 124$), lung cancer ($n = 116$), leukemia ($n = 169$) and 757 controls with available blood samples and successful DNA extraction and genotype analysis. The remaining endpoints were not analyzed as the methods used in this paper require a large number of subjects. Our approach did not keep the matching between cases and controls.

In addition to several gene variants (see below) we also included in the analyses three exposure variables, air pollution (as indirectly measured by the distance from a heavy traffic road) and exposure to ETS at work and at home. For details on these exposures and the ways they were measured see our previous publications (13,14).

DNA extraction and genotype analysis

The choice of the relevant polymorphisms (Table I) has been made on the basis of an extensive review of the literature (8–11) and encompasses genes involved directly or indirectly in DNA repair and metabolic genes.

DNA was extracted from 200–300 µl of buffy coats in Genova and Florence laboratories. DNA was isolated and purified as described in Peluso *et al.* (15).

Genotyping has been performed with Light Cycler at IARC by Dr Malaveille (genes: MnSOD, COMT, MPO, SULT-1A1), with Taqman at the ISI Foundation in Torino by Dr Matullo (XPD, XRCC1, PCNA, ERCC1, MGMT, OGG1) and at the University of Cambridge by Dr Dunning (BRCA1 and 2, NBS1, RAD51-52, XRCC2, LIG4, TP53), with a PCR–RFLP based assay at the University of Aarhus by Dr Autrup (GSTM1, GSTM3, GSTT1, GSTP, NQO1) and at the Genetics Research Institute in Milan by Dr Garte (NAT2, CYP1A1, CYP1B1, MTHFR).

MDR analysis

The common analytical tool to analyse gene–gene interactions under a multiplicative model is logistic regression analysis. Unfortunately, parametric and model-based statistical methods require specific hypotheses to be tested; in high-dimensional analysis such as in the investigation of all potential interactions, the number of hypotheses is inflated to a great extent. To address this issue it has been suggested that model-free, data-based exploratory methods are more flexible and powerful.

A MDR reduction method with related open-source software has been developed by Ritchie *et al.* (12) to detect and characterize high-order gene–gene and gene–environment interactions in studies with relatively small sample sizes.

The goal of MDR is to find the main factor and the combinations of $2, \ldots,$ N factors that are more frequently associated with case than with control status (adjusted for the ratio between them). The MDR method was applied as described previously (12,16–22). Briefly, to search for the best n-loci model (with $n = 1, \ldots,$ N), the dataset is randomly divided into 10 equal parts. A training set of 9/10 of the data are used to search for the best model, i.e. to classify each genotype combination as a high-risk or a low-risk pattern, depending on the number of cases and controls that present that combination. In Figure 1 each box represents a particular combination of genotypes for a 4-loci model (Figure 1A) and a 3-loci model (Figure 1B). If the box is dark, the combination is a high-risk pattern, i.e. we observe that the particular combination of genotypes for these polymorphisms is more frequently associated with the case status than with the control status. In Figure 1B the box representing the combination of the w/w genotype (w = wild-type, m = mutant) in CYP1A1-Ile462Val polymorphism, the w/m genotype in RAD52-2259C > T and the m/m genotype in MnSOD-Ala9Val is labelled as high-risk, although the number of cases is only 13 and that of controls is 22. The reason is that their ratio ($13/22 = 0.59$) is greater than the ratio of the number of cases to the number of controls for the leukemia dataset ($169/305 = 0.55$). The remaining 1/10 of the data are the testing set, used to control the goodness-of-fit of the model. This procedure is repeated 10 times, in order to use all the possible testing sets.

For each *n-loci* model the MDR method gives two scores, a mean prediction error percentage and the cross-validation consistency (CVC)

Table I. Gene variants included in the MDR model

Gene	Function[b]	Polymorphism	Genotype frequency (%)[a]		
			w/w	w/m	m/m
DNA repair					
ERCC2/XPD	NER	Asp312Asn	38.4	45.3	16.2
ERCC2/XPD	NER	Lys751Gln	35.7	46.1	18.2
PCNA	BER	6084 G > C (3′-UTR)	80	18.9	1.1
XRCC1	BER	Arg194Trp	87.6	12.2	0.3
XRCC1	BER	Pro206Pro	30.9	46	23.1
XRCC1	BER	Arg399Gln	44.8	43.6	11.6
XRCC3	DSBR	17893 A > G (IVS6–14)	51.1	40.1	8.9
XRCC3	DSBR	Thr241Met	34.6	49.8	15.6
APE1	BER	Asp148Glu	29.3	48.2	22.5
ERCC1	NER	Asn118Asn	35.6	46.8	17.6
MGMT	DRR	Leu84Phe	73.6	24.6	1.8
hOGG1	BER	Ser326Cys	62.6	32	5.4
BRCA1	DSBR	Pro871Leu	42.5	47.7	9.8
BRCA2	DSBR	Asn372His	51.6	41.3	7.1
NBS1	DSBR	Glu185Gln	49.2	40.2	10.6
RAD51	DSBR	135 G > C (5′-UTR)	87	12.7	0.3
RAD512	DSBR	172 G > T (5′-UTR)	33.8	48.7	17.5
RAD52	DSBR	2259 C > T (3′-UTR)	33.7	46.8	19.4
XRCC2	DSBR	Arg188His	83.2	16	0.9
LIG4	DSBR	Ala3Val	88.7	7.7	3.5
LIG45	DSBR	Thr9Ile	72.1	25	2.8
TP53	Cell cycle/apoptosis	Arg72Pro	57.5	36.5	5.9
Metabolic					
MnSOD	Oxidative scavenger	Ala9Val	25.2	52.7	22.1
NQO1	Oxidative scavenger	Pro187Ser	65	31.4	3.6
COMT	Phase 1	Val158Met	24.8	52.1	23.1
MPO	Phase 1	G > A SP1 site	60.3	35.4	4.2
SULT1A1	Phase 2	Arg213His	44.7	43.5	11.8
GSTM3	Phase 2	3 bp deletion (*A,*B)	68.7	28.4	3
GSTP1	Phase 2	Ile105Val	44.1	44.1	11.8
GSTP1	Phase 2	Ala114Val	83.9	15	1.1
CYP1A1	Phase 1	Ile462Val	83.6	15.4	1
CYP1B1	Phase 1	Val432Leu	16.4	48.6	35
MTHRF	Pholate/nucleotide	677 C > T	42.7	44.1	13.2
			w/w	**w/m + m/m**	
GSTT1	Phase 2	Gene deletion (*1, *2/*2)	75.5	24.5	
GSTM1	Phase 2	Gene deletion (*2/*2, *1)	55.6	44.4	
NAT2	Phase 2	Slow/rapid acetylator	42.7	57.3	

[a]w = Wild-type allele, m = mutant allele.
[b]NER = nucleotide excision repair, BER = base excision repair, DSBR = double-strand break repair, DRR = direct reversal repair.

frequency. The former is the proportion of subjects for whom an incorrect class prediction was made, while the latter is the number of times a particular combination of loci (model) is identified in each possible testing set. The best model is that with lower prediction error and maximum CVC. We repeated the complete analysis 10 times, using different random seeds to reduce the probability of biased results due to the chance divisions of the data in training and testing sets.

To evaluate the magnitude of the prediction error and the CVC, we permutated the status of cases and controls in the dataset and repeated the analysis 1000 times, obtaining for each *n*-loci model the prediction error and the CVC distributions under the null hypothesis of no association. Comparing the results and these distributions we obtained the *P*-values associated with each prediction error and CVC.

For each model, we computed the associated odds ratio (Table II). The reference group (unexposed) is formed by subjects presenting the combination of genotypes labelled as low-risk, while the exposed group is that with subjects that present the combination labelled as high-risk.

Limitations of MDR

Although MDR overcomes some of the limitations of the generalized linear model, there are also some other limitations to consider (12,19). First, MDR is computationally intensive for high-order models and would require genetic search algorithms to find solutions in a limited time. Second, MDR results are difficult to interpret. If a strong main effect is present, it is not simple to understand the importance of the other polymorphisms in the model. More generally, there is no information on the relations between the polymorphisms involved in each model; e.g. if the best model is a four-way interaction, it is not clear if it is the sum of two separate two-way interactions or a two-way interaction and two main effects, etc. Third, high dimensionality and a small dataset may lead to many multi-factor cells with either a small number of subjects or no subjects. It is not clear how the model sensitivity changes with the dimension of the dataset. For example, a source of noise is given by missing values. The proportion of missing values can vary among cases and controls and thus affect the results. In our dataset, the mean differences in percentage of missing values between cases and controls are <3%. Finally, MDR assumes that there is no genetic heterogeneity. If a group of cases are explained by a combination of loci different from the one that explains another group of cases, the MDR method can fail in finding the correct models.

Hypotheses tested

The main advantage of using a data-based statistical method like MDR is the ability to analyse a dataset without the need to hypothesize an a priori model (model free analysis). At the same time, it can be useful to test for some hypotheses, such as specifying the recessive or dominant model of inheritance. To simplify the analysis, we have searched for scenarios in which all the gene variants show a recessive or a dominant behaviour. In these analyses, we have assigned the genotypes with the highest frequency as wild-type for each polymorphism and used them as the reference group.

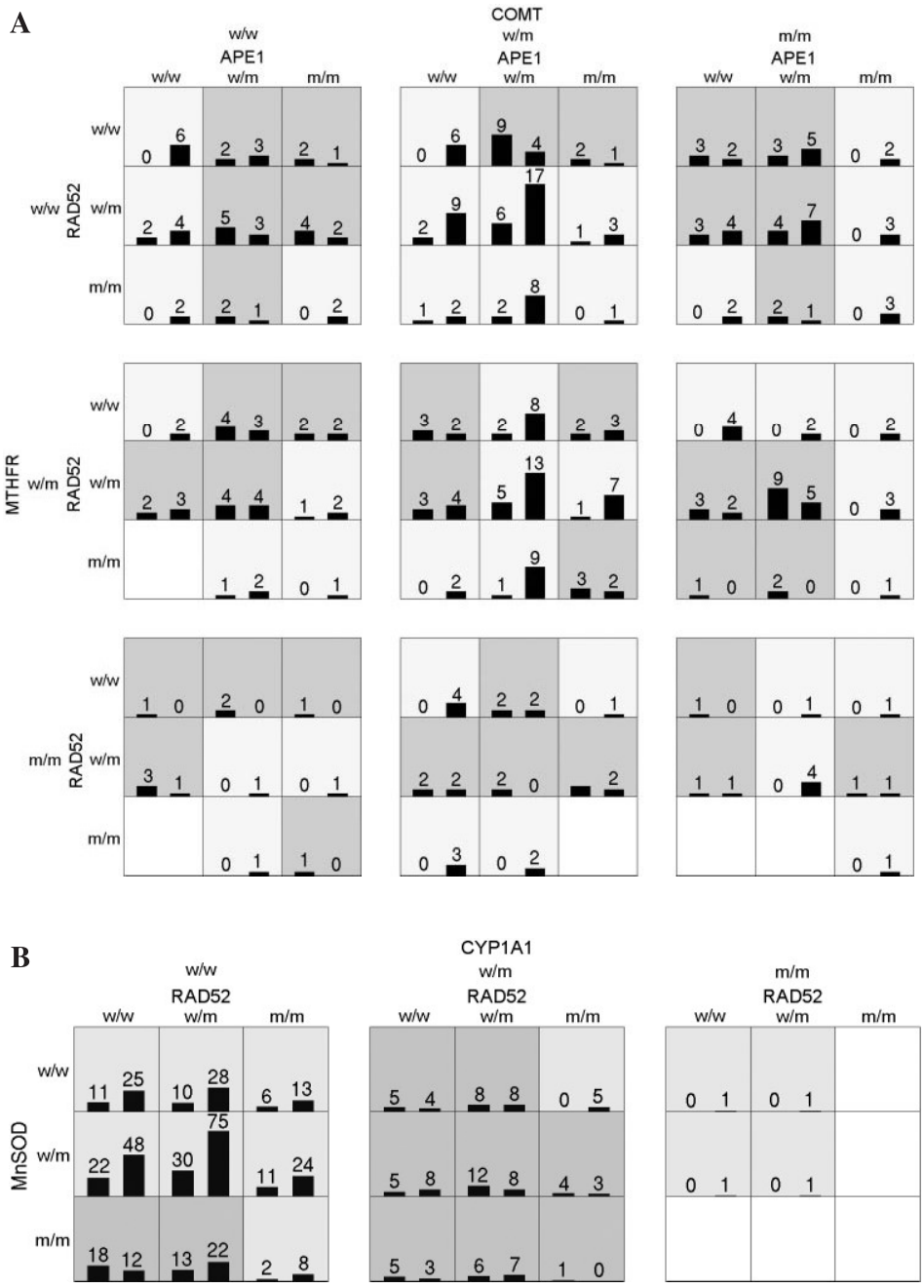

Fig. 1. Genotype combinations labelled as high-risk and low-risk for bladder cancer and leukemia. In each box the left and the right bars represent the number of cases and controls for each combination of genotypes. If the ratio of cases to controls exceeds a threshold (the ratio of total number of cases to total number of controls in the present analysis), the combination is labelled as high risk (dark boxes). (**A**) Bladder cancer. (**B**) Leukemia.

False positive report probability (FPRP)

Reports of associations between polymorphisms and complex diseases are greatly affected by the risk of being false positives. To estimate the magnitude of this probability we used the method described by Wacholder (23) and recently applied in some epidemiological studies (24–26). To compute the FPRP, we used the odds ratios from MDR models, in which we consider the given classification of high risk and low-risk genotype combinations, the *P*-values and the power to detect ORs of 2 and 4. The *P*-values are calculated from the comparison of the proportion of correct predictions of the model and the distribution of randomized datasets.

As there are many factors influencing the prior probability that the association between a genetic variant and a disease is real (27), the FPRP is commonly computed for prior probabilities ranging in a wide interval [1.E-5 to 0.25 in Wacholder (23), 0.001 to 0.50 in Hung (24) and Matullo (25,26)]. Considering the lack of information on the interactions between genes and environmental variables, in our study we have considered a wider interval for prior probabilities, with bottom and top values lower than those used in other studies, i.e. from 1.E-6 to 0.10.

Validation and interpretation of the results using information gain and interaction graphs

The results obtained with the MDR method were validated using an approach based on measures of information and entropy. Jakulin and Bratko (28) have provided a generalization of McGill's interaction information (29) to evaluate the information gain (IG) related to a class variable (e.g. case-control status) by merging several attributes together, over the information provided by the attributes independently.

Table II. Results of MDR for lung and bladder cancers and leukemia

Lung							
Model free				Prediction[a]	*P*-value	CVC	*P*-value
NQO1				0.68	0.015	9.60	0.030
NQO1	cyp1a1			0.70	0.002	5.50	0.131
road	XRCC1_28152	BRCA2		0.74	<0.001	6.60	0.023
XRCC3_18067	TP53	GSTP1_105	MTHFR	0.78	<0.001	2.30	0.537
Recessive model							
NQO1				0.68	0.040	8.80	0.224
NQO1	BRCA2			0.69	0.004	4.30	0.382
ERCC1	RAD52	NQO1		0.72	<0.001	5.80	0.087
ERCC1	BRCA2	RAD52	NQO1	0.73	0.001	4.20	0.096
Dominant model							
GSTM1				0.67	0.034	2.40	0.885
NQO1	cyp1a1			0.69	0.009	6.60	0.111
XPD_1	TP53	cyp1a1		0.70	0.002	2.60	0.517
BRCA2	TP53	GSTM3	cyp1a1	0.73	<0.01	2.30	0.380
Bladder							
Model free				Prediction[a]	*P*-value	CVC	*P*-value
GSTM3				0.45	1.000	2.40	0.946
LIG45	GSTM1			0.58	0.653	3.50	0.412
RAD52	COMT	MTHFR		0.70	<0.001	2.60	0.550
APE1	RAD52	COMT	MTHFR	0.78	<0.001	7.40	0.013–0.037
Recessive model							
MGMT				0.65	0.139	9.00	0.213
MGMT	TP53			0.65	0.063	0.90	1.000
APE1	LIG4	MTHFR		0.67	0.006	1.78	0.949
ETS	APE1	RAD52	COMT	0.69	<0.001	1.30	0.930
Dominant model							
MTHFR				0.50	1.000	2.90	0.943
COMT	MPO			0.66	0.026	3.50	0.493
ETS	XRCC2	GSTP1_114		0.63	0.061	1.50	0.941
ETS	XRCC3_2	NBS1	NQO1	0.72	<0.001	1.60	0.890
Leukemia							
Model free				Prediction[a]	*P*-value	CVC	*P*-value
NQO1				0.56	0.999	2.60	0.951
cyp1a1	nat2			0.65	0.029	4.00	0.203
RAD52	mnSOD	cyp1a1		0.69	<0.001	4.40	0.086
Road	APE1	RAD512	cyp1b1	0.73	<0.001	1.90	0.920
Recessive model							
NQO1				0.65	0.112	6.90	0.502
XRCC1_3	cyp1b1			0.66	0.016	4.90	0.345
XRCC1_3	NQO1	cyp1b1		0.67	0.005	2.90	0.545
ETS	XRCC1_3	mnSOD	MTHFR	0.56	0.950	4.20	0.120
Dominant model							
NQO1				0.65	0.032	5.80	0.292
cyp1a1	nat2			0.64	0.081	3.00	0.468
PCNA	cyp1a1	nat2		0.64	0.029	1.00	0.905
ETS	RAD512	GSTP1_114	cyp1a1	0.70	<0.001	3.22	0.170

The first *P*-value is for the proportion of subjects for whom a correct prediction was made; CVC is the number of times a particular combination of loci/variables (model) was identified in each possible testing set, with the corresponding *P*-value. The best model is the one with the lowest prediction error and maximum CVC. In the recessive model for lung cancer, the four-loci model has been regarded as the best model as it has lower false positive report probabilities than the three-loci model.
ETS = environmental tobacco smoke.
Road = distance from heavy traffic road.
[a]Proportion of correct predictions.

Let H(X) be the Shannon entropy of X (29). In the case with two attributes A, B and a class label C, the IG of A, B and C can be written as IG(A;B;C) = I(A;B|C) − I(A;B), where I(A;B|C) = H(A|C) + H(B|C) − H(A,B|C) and I(A;B) = H(A) + H(B) − H(A,B). I(A;B) is the mutual information between A and B, while I(A;B|C) is the conditional mutual information and measures the relationship between A and B in the context of C.

Generalizing from formulas in (29), Jakulin and Bratko define the k-way interaction information for $k = 3$, 4 to an arbitrary k as:

$$\mathrm{IG}(s) = -\sum_{\tau \in S} (-1)^{|S|-|\tau|} H(\tau) = I(S\backslash X|X) - I(S\backslash X), X \in S,$$

where $A=\{X_1, X_2, \ldots, X_n\}$ is a set of n attributes (e.g. the variables measured in the study) and $S \subseteq A$ is a subset of k attributes.

As the softwares used to implement these formulas do not permit missing values, their values have been randomly imputed 10 times respecting the genotype proportions in the data. The results reported are the mean values obtained from the analyses performed using each of the 10 datasets.

Computing the interaction information for all the possible combinations of k attributes chosen in a set of 40 variables (exposures and SNPs) is a time consuming process for $k > 3$ at the current speed of computers. To address this issue we used a two-step procedure following the indications given by Moore *et al.* (30). First a subset of more informative attributes to analyse is selected. To do this, the IG is estimated for each individual attribute and each pairwise combination of attributes. Pairs of attributes are sorted and those with the highest IG, or percentage of entropy removed, are selected for further consideration. The selection of the cut-off to exclude the attributes is made considering the distribution of the IG values for each pairwise combination. Usually, these distributions

Table III. False positive report probabilities and odds-ratios for the best models

	Odds-ratio (95% CI)	Power[a]	Expected OR = 2							Expected OR = 4					
			Prior probability							Prior probability					
			1E-01	1E-02	1E-03	1E-04	1E-05	1E-06		1E-01	1E-02	1E-03	1E-04	1E-05	1E-06
Lung model free															
road	6.22 (3.69–10.47)	0.001	**0.006**	**0.067**	0.421	0.879	0.986	0.999	0.116	**0.000**	**0.001**	**0.007**	**0.063**	0.404	0.871
XRCC1_28152	*P*-value = 7.90E-07														
BRCA2															
Lung recessive model															
ERCC1	6.55 (3.46–12.40)	0.004	**0.061**	0.418	0.879	0.986	0.999	1.000	0.137	**0.002**	**0.022**	**0.184**	0.693	0.957	0.996
BRCA2	*P*-value = 3.10E-05														
RAD52															
NQO1															
Bladder model free															
APE1	11.81 (6.99–19.95)	0.000002	**0.001**	**0.008**	**0.077**	0.456	0.894	0.988	0.003	**0.000**	**0.000**	**0.000**	**0.001**	**0.007**	**0.070**
RAD52	*P*-value = 2.00E-10														
COMT															
MTHFR															
Bladder model free															
RAD52	3.88 (2.51–6.00)	0.001	**0.000**	**0.000**	**0.000**	**0.004**	**0.036**	0.270	0.556	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.001**
mnSOD	*P*-value = 4.2E-10														
cyp1a1															

Each *P*-value is calculated from the proportion of correct predictions of the model and the distribution of randomized data sets. Bold type indicates the false positive report probabilities lower than 0.20.
[a]Estimation of the statistical power to detect the expected OR with a level equal to the reported *P*-value.

have a Gaussian-like shape, with only few interactions standing out from the crowd. In the second step, the IG is estimated for each attribute and each combination of two, three and four attributes using only the selected subset.

Measures of IG for main and interaction effects are useful to plot interaction graphs that can give indications to interpret the relationship between attributes. In the interaction graphs nodes and connections indicate the percentage of entropy in case-control status removed by each variable (main effect) and by each pairwise combination of attributes (interaction effect). A positive entropy indicates interaction while a negative entropy indicates redundancy.

Software

All MDR computations to find and test the models have been performed using the open-source java software MDR v.1.0.0rc1. FPRP results are calculated with the Excel spreadsheet distributed at http://jncicancer spectrum.oupjournals.org/jnci/content/vol96/issue6 (23). The formulas from Jakulin and Bratko have been implemented in R, v.2.2.1 (R Foundation for Statistical Computing, Vienna, Austria). The Orange software package (31) has been used to test the routines written in R and to build the interaction graphs.

Results

Table II shows for each cancer site and each number of loci/exposure variables evaluated (up to 4), the average CVC, the average prediction error and the associated *P*-values obtained from the randomized analysis of lung and bladder cancer, and leukemia datasets.

For lung cancer, the best result (model free analysis) is provided by a significant gene–environment interaction between XRCC1-Arg399Gln, BRCA2-Asn372His polymorphisms and the exposure variable 'heavy traffic road' (mean prediction error of 26%, $P < 0.001$, and mean CVC of 6.60, $P = 0.02$). The exposure variable represents distance from a heavy traffic road and is an indirect and robust indicator of air pollution. Moreover, a 4-loci interaction under a recessive model (ERCC1-Asn118Asn, BRCA2-Asn372His, RAD52-2259C > T and NQO1-Pro187Ser polymorphisms) showed also significant results (Table II, $P < 0.001$ for mean prediction error and $P = 0.096$ for CVC). The 4-loci interaction was regarded as a better model than the encapsulated 3-loci model (that showed comparable reliability for mean prediction error and CVC) because had lower FPRP values (Table III).

For bladder cancer, we found a significant 4-loci interaction between APE1-Asp148Glu, RAD52-2259C > T and the metabolic gene polymorphisms COMT-Val158Met and MTHFR-677C > T (mean prediction error of 22%, $P < 0.001$ and mean CVC of 7.40, $P = 0.013$–0.037).

For leukemia, a 3-loci interaction including RAD52-2259C > T, MnSOD-Ala9Val and CYP1A1-Ile462Val, had a minimum prediction error of 31% ($P < 0.001$) and a maximum CVC of 4.40 ($P = 0.086$). For both bladder cancer and leukemia, the recessive and dominant models did not yeld any significant results.

Table III reports the FPRP values calculated using the statistical power to detect ORs of 2 and 4 with α level equal to the observed P value. Results show an excellent reliability for the bladder 4-loci and leukemia 3-loci models, even assuming very low prior probabilities (<0.001). Figure 1A and B show the detailed models.

On the other hand, the two interaction models found for lung cancer seem to indicate false positive reports (as shown in Table III for OR = 2), since unrealistic prior probabilities are necessary to obtain FPRP values below 0.2.

In the two-step validation method, all the attributes involved in the MDR models for lung cancer, bladder cancer and leukemia were present in the more informative subset of variables selected by the IG method when used for main and interaction effects. In the second step, the lung cancer and leukemia models did not obtain high IG values, i.e. there were many models with other attributes (more than 10) reporting higher IGs. The bladder cancer model, on the other side, resulted to be the best model also for the IG method.

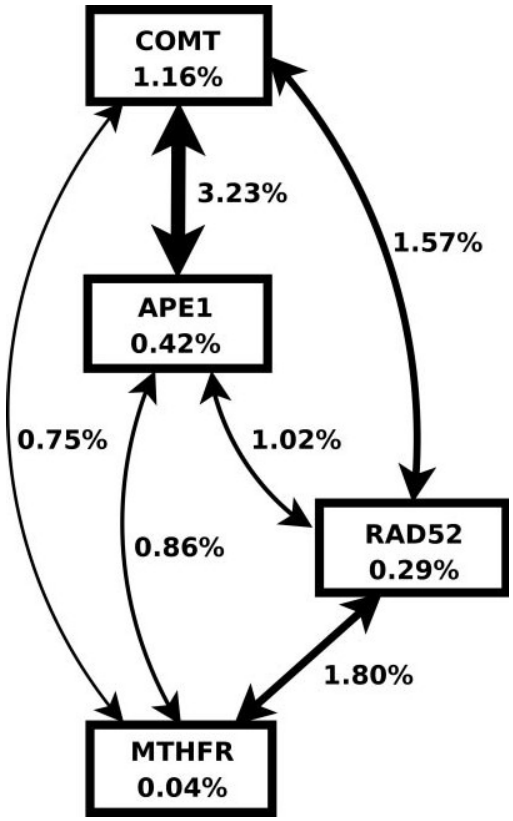

Fig. 2. Orange canvas interaction graph. Nodes and connections indicate the percentage of entropy in case-control status removed by each variable (main effect) and by each pairwise combination of attributes (interaction effect). The values are all positive, indicating a positive interaction between all the polymorphisms of the model.

The percentage of entropy removed by each variable and by each pairwise combination of attributes is shown in Figure 2. The values are all positive, indicating a positive interaction between all the polymorphisms included the model.

Discussion

In the present study, we investigated the possible combined contribution of genetic and environmental factors to determining complex cancer risk patterns through gene–gene and gene–environment interactions. We applied a robust and validated method to reduce the complexity of multi-factor analysis in a large prospective investigation on air pollution, ETS and cancer. The results were validated using a generalization to many attributes of the mutual information method, a well-established approach to evaluate interactions between two variables. The attributes selected in the lung cancer and leukemia models were factors in the subset of the more informative variables, while the MDR best model was confirmed by the IG method as the best 4-way interaction for bladder cancer (Figure 2). The patterns we have identified need further biological investigation for interpretation beyond their statistical significance. The biological plausibility of the interpretation is in fact a test of the viability of multidimension reduction methods.

The first consideration about the emerging patterns is that DNA repair genotypes seem to play an important role in increasing cancer risk. Amongst the genes that seem to be involved in the risk of lung and bladder cancer and of leukemia, RAD52 emerges repeatedly in our analysis, followed by XRCC1, BRCA2 and APE1. Although we investigated a limited number of DNA repair genes/polymorphisms, it seems that impaired base excision repair (APE1 and XRCC1) and double-strand break repair (RAD52 and BRCA2) could be common pathways leading to different forms of cancer. Unfortunately the literature on DNA repair gene variants and cancer risk is still scanty and unclear, reporting often contrasting results (8,32,33). The involvement of proximity to a heavy traffic road is the only environmental factor that interacts significantly with gene variants, and it is coherently associated with lung cancer (13). Distance from a heavy traffic road is an indirect and robust indicator of air pollution. In bladder cancer, APE1 and RAD52 DNA repair gene variants strongly interacted with some metabolic genes, namely COMT and MTHFR, in increasing the risk. APE1 is the major 5′ AP endonuclease of mammalian cells that is responsible for the incision of baseless lesions in DNA during BER. Although the *APE1*-Asp148Glu variant allele seems to have a small effect on AP endonuclease and DNA-binding activities (34) an increased risk of chromosomal aberration has been observed for this polymorphism (35). RAD52 is involved in DSBR through homologous recombination and facilitates strand exchange reaction catalysed by RAD51. Although the *RAD52* gene could be a good candidate due to the high number of aberrations and allelic losses in different chromosomes observed in bladder cancer cells, for the RAD52-2259C > T polymorphism insufficient information has been published for any conclusion about its functional consequences to be drawn.

COMT catalyses the methylation of various endobiotic and xenobiotic substances preventing quinone formation and redox cycling, and therefore might protect DNA from oxidative damage (1). A *G* to *A* transition, which results in amino acid change from valine to methionine at codon 108, leads to a lower COMT activity in a co-dominant manner (36). In a previous study we have analyzed the tumour biopsies of 45 patients with bladder cancer for p53 mutations by direct sequencing. All p53 mutations occurred in subjects with the COMT variant allele ($P = 0.03$) (37). Although, a recent study on bladder cancer did not find an association with the COMT variant (38), a role of this enzyme in bladder cancer is likely, as supported also by the concomitant use of a soluble isoform of COMT, calreticulin and gamma-synuclein as urinary markers to improve the diagnosis of bladder cancer (39). MTHFR is an enzyme that plays a key role in the metabolism of folate and in DNA synthesis, and common polymorphisms seem to affect its functional properties (40). The MTHFR-677 variant genotypes (CT or TT) are common in the population. The studies on MTHFR and bladder cancer risk are inconsistent, with negative results in three studies (41–43), while a complex pattern was apparent in another investigation, in which individuals carrying the variant genotypes (CT or TT) and reporting a low folate intake were at a significantly 3.5-fold increased risk of bladder cancer (95% CI: 1.6–6.5) (44). In contrast, individuals carrying a variant genotype and reporting a high folate intake were at a 1.4-fold increased risk (95% CI: 0.7–2.7), and wild-type homozygotes reporting a low folate intake were at only 1.6-fold increased risk (95% CI: 0.8–3.0). The interaction between genetic polymorphisms and folate intake

was significant on the multiplicative scale ($P = 0.01$) (44). The role of folate in inducing bladder cancer is biologically plausible, as suggested by the protective effect of fruit and vegetables on bladder cancer (45), but the hypothesis of an involvement of MTHFR is still uncertain.

Less clear is the lack of involvement (even in terms of interactions) of metabolic gene polymorphisms, such as NAT-2 and GSTM1, that are established genetic risk factors for bladder cancer as shown in recent large case-control studies (46). However, this difference may be due to the fact that our population includes only non-smokers. Similar considerations can also apply to the results on lung cancer and leukemia. In particular, for lung cancer we have not confirmed the involvement of XRCC1-Arg194Trp and of hOGG1-Ser326Cys polymorphisms, as recently suggested (24,32). However, we have identified a possible weak interaction of NQO1-Pro187Ser polymorphism in combination with ERCC1-Asn118Asn, BRCA2-Asn372His, RAD52-2259C > T polymorphisms; a positive association with lung cancer has been recently reported for NQO1-Pro187Ser polymorphism (47).

For leukemia, the involvement of a DSBR gene, i.e. RAD52, is plausible as suggested in different studies, since chromosomal translocations are very common genetic abnormalities in this disease (48,49), with evidence of loss of heterozygosity for a chromosomal segment including the RAD52 locus in T cell prolymphocytic leukaemia (50). Less clear is the involvement of a free radical scavenger such as MnSOD, although there are several studies reporting on down-regulation of this enzyme after treatment of leukaemic cells with different chemotherapeutic agents (51,52). However, no study investigated the relationship between leukaemia and MnSOD polymorphisms. On the other hand, CYP1A1 polymorphisms have been largely investigated in different types of leukaemias, showing positive associations in childhood leukaemia (53), and acute adult lymphoblastic leukaemia (54,55). A previously described interaction with the NAT-2 and GSTM1 genes (56) has not been confirmed in our study.

In summary, the method of multifactor-dimensionality reduction, even with the limitations previously discussed, seems promising. In fact, it provides a limited number of statistically stable associations among genes and between genes and the environment that have been validated by an alternative and well documented method. However, the lack of clear biological background knowledge hampers a clear interpretation of the findings.

Acknowledgements

The authors wish to thank to Angeline Andrew and Margaret Karagas for their useful comments. This paper was made possible by European Community grants to PV (5th Framework Programme, GENAIR investigation, grant QLK4-CT-1999-00927; 6th Framework Programme, ECNIS investigation, grant 513943) and a grant of the Compagnia di San Paolo to the ISI Foundation. All authors are independent from funders. Also, the work described in the paper was carried out with the financial support of: Europe Against Cancer Program of the European Commission (SANCO); Deutsche Krebshilfe; Deutsches Krebsforschungszentrum; German Federal Ministry of Education and Research; Danish Cancer Society; Health Research Fund (FIS) of the Spanish Ministry of Health; Spanish Regional Governments of Andalucia, Asturia, Basque Country, Murcia and Navarra; ISCIII, Red de Centros RCESP, C03/09, Spain; Cancer Research UK; Medical Research Council, United Kingdom; Stroke Association, United Kingdom; British Heart Foundation; Department of Health, United Kingdom; Food Standards Agency, United Kingdom; Wellcome Trust, United Kingdom; Greek Ministry of Health; Greek Ministry of Education; Italian Association for Research on Cancer (AIRC); Italian National Research Council; Dutch Ministry of Public Health, Welfare and Sports; World Cancer Research Fund; Swedish Cancer Society; Swedish Scientific Council; Regional Government of Skåne, Sweden; Norwegian Cancer Society; Research Council of Norway.

Conflict of Interest Statement: None declared.

References

1. Autrup,H. (2000) Genetic polymorphisms in human xenobiotica metabolizing enzymes as susceptibility factors in toxic response. *Mutat Res.*, **464**, 65–76.
2. Autrup,H. (2004) Gene-environment interaction in environmental carcinogenesis. In Nicolopou-Stamati *et al.* (eds), *Cancer as an Environmental Disease*. Kluwer Academic Publishers.
3. Peluso,M., Munnia,A., Hoek,G., Krzyzanowski,M., Veglia,F., Airoldi,L., Autrup,H., Dunning,A., Garte,S., Hainaut,P. *et al.* (2005) DNA adducts and lung cancer risk: a prospective study. *Cancer Res.*, **65**, 8042–8048.
4. Vispe,S., Yung,T.M., Ritchot,J., Serizawa,H. and Satoh,M.S. (2000) A cellular defense pathway regulating transcription through poly (ADP-ribosyl)ation in response to DNA damage. *Proc. Natl Acad. Sci. USA*, **97**, 9886–9891.
5. Squire,J.A., Whitmore,G.F. and Phillips,R.A. (1998) Genetic basis of cancer. In Tannock,I.F., Tannock,I. and Hill,R.P. (eds) *The Basic Science of Oncology* 3rd edn. McGraw-Hill Press.
6. Knudson,A.G.,Jr (1989) The genetic predisposition to cancer. *Birth Defects Orig. Artic. Ser.*, **25**, 15–27.
7. Friedberg,E.C., Walker,G.C. and Siede,W. (1995) *DNA Repair and Mutagenesis*. ASM Press Washington, DC.
8. Goode,E.L., Ulrich,C.M. and Potter,J.D. (2002) Polymorphisms in DNA repair genes and associations with cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, **11**, 1513–30.
9. Au,W.W., Navasumrit,P. and Ruchirawat,M. (2004) Use of biomarkers to characterize functions of polymorphic DNA repair genotypes. *Int. J. Hyg. Environ. Health*, **207**, 301–313.
10. Berwick,M., Matullo,G. and Vineis,P. (2002) Studies of DNA repair and human cancer: an update. In Wilson,S.H. and Suk,W.A., (eds) *Biomarkers of Environmentally Associated Disease: Technologies, Concepts and Perspectives*. Lewis publishers, CRC Press LLC, NY, pp. 83–107.
11. Berwick,M. and Vineis,P. (2000) Markers of DNA repair and susceptibility to cancer in humans: an epidemiologic review. *J. Natl Cancer Inst.*, **92**, 874–97.
12. Ritchie,M.D., Hahn,L.W., Roodi,N., Bailey,L.R., Dupont,W.D., Parl,F.F. and Moore,J.H. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, **69**, 138–147.
13. Vineis,P., Hoek,G., Krzyzanowski,M., Vigna-Taglianti,F., Veglia,F., Airoldi,L., Autrup,H., Dunning,A., Garte,S., Hainaut,P. *et al.* (2006) Air pollution and risk of lung cancer in a prospective study in Europe. *Int. J. Cancer*.
14. Vineis,P., Airoldi,L., Veglia,P., Olgiati,L., Pastorelli,R., Autrup,H., Dunning,A., Garte,S., Gormally,E., Hainaut,P. *et al.* (2005) Environmental tobacco smoke and risk of respiratory cancer and chronic obstructive pulmonary disease in former smokers and never smokers in the EPIC prospective study. *BMJ*, **330**, 277.
15. Peluso,M., Hainaut,P., Airoldi,L., Autrup,H., Dunning,A., Garte,S., Gormally,E., Malaveille,C., Matullo,G., Munnia,A. *et al.* (2005) Methodology of laboratory measurements in prospective studies on gene-environment interactions: the experience of GenAir. *Mutat Res.*, **574**, 92–104.
16. Hahn,L.W., Ritchie,M.D. and Moore,J.H. (2003) Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics*, **19**, 376–382.
17. Ritchie,M.D., Hahn,L.W. and Moore,J.H. (2003) Power of multifactor dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet. Epidemiol.*, **24**, 150–157.
18. Cho,Y.M., Ritchie,M.D., Moore,J.H., Park,J.Y., Lee,K.U., Shin,H.D., Lee,H.K. and Park,K.S. (2004) Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia*, **47**, 549–554.
19. Coffey,C.S., Hebert,P.R., Ritchie,M.D., Krumholz,H.M., Gaziano,J.M., Ridker,P.M., Brown,N.J., Vaughan,D.E. and Moore,J.H. (2004) An application of conditional logistic regression and multifactor

dimensionality reduction for detecting gene–gene interactions on risk of myocardial infarction: the importance of model validation. *BMC Bioinformatics*, **5**, 49.
20. Williams,S.M., Ritchie,M.D., Phillips,J.A.,3rd, Dawson,E., Prince,M., Dzhura,E., Willis,A., Semenya,A., Summar,M., White,B.C. *et al.* (2004) Multilocus analysis of hypertension: a hierarchical approach. *Hum. Hered.*, **57**, 28–38.
21. Ritchie,M.D. and Motsinger,A.A. (2005) Multifactor dimensionality reduction for detecting gene–gene and gene–environment interactions in pharmacogenomics studies. *Pharmacogenomics*, **6**, 823–834.
22. Martin,E.R., Ritchie,M.D., Hahn,L., Kang,S. and Moore,J.H. (2006) A novel method to identify gene–gene effects in nuclear families: the MDR-PDT. *Genet. Epidemiol.*, **30**, 111–123.
23. Wacholder,S., Chanock,S., Garcia-Closas,M., El Ghormli,L. and Rothman,N. (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl Cancer Inst.*, **96**, 434–442.
24. Hung,R.J., Brennan,P., Canzian,F., Szeszenia-Dabrowska,N., Zaridze,D., Lissowska,J., Rudnai,P., Fabianova,E., Mates,D., Foretova,L. *et al.* (2005) Large-scale investigation of base excision repair genetic polymorphisms and lung cancer risk in a multicenter study. *J. Natl Cancer Inst.*, **97**, 567–576.
25. Matullo,G., Dunning,A.M., Guarrera,S., Baynes,C., Polidoro,S., Garte,S., Autrup,H., Malaveille,C., Peluso,M., Airoldi,L. *et al.* (2005) DNA repair polymorphisms and cancer risk in non-smokers in a cohort study. *Carcinogenesis*.
26. Matullo,G., Guarrera,S., Sacerdote,C., Polidoro,S., Davico,L., Gamberini,S., Karagas,M., Casetta,G., Rolle,L., Piazza,A. *et al.* (2005) Polymorphisms/haplotypes in DNA repair genes and smoking: a bladder cancer case-control study. *Cancer Epidemiol. Biomarkers Prev.*, **14**, 2569–2578.
27. Matullo,G., Berwick,M. and Vineis,P. (2005) Gene–environment interactions: how many false positives? *J. Natl Cancer Inst.*, **97**, 550–551.
28. Jakulin,A. and Bratko,I. (2003) Analyzing attribute interactions. *Lect. Notes Artiff. Intell.*, **2838**, 229–240.
29. McGill,W.J. (1954) Multivariate information transmission. *Psychometrica*, **19**, 97–116.
30. Moore,J.H., Gilbert,J.C., Tsai,C.T., Chiang,F.T., Holden,T., Barney,N. and White,B.C. (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J. Theor. Biol.*, **241**, 252–261.
31. Demsar,J. and Zupan,B. (2004) Orange: from experimental machine learning to interactive data mining. *Omnipress*.
32. Hung,R.J., Hall,J., Brennan,P. and Boffetta,P. (2005) Genetic polymorphisms in the base excision repair pathway and cancer risk: a HuGE review. *Am. J. Epidemiol.*, **162**, 925–942.
33. Manuguerra,M., Saletta,F., Karagas,M.R., Berwick,M., Veglia,F., Vineis,P. and Matullo,G. (2006) XRCC3 and XPD/ERCC2 single nucleotide polymorphisms and the risk of cancer: a HuGE Review. *Am. J. Epidemiol.*, **164**, 297–302.
34. Hadi,M.Z., Coleman,M.A., Fidelis,K., Mohrenweiser,H.W. and Wilson,D.M.,3rd. (2000) Functional characterization of Ape1 variants identified in the human population. *Nucleic Acids Res.*, **28**, 3871–3879.
35. Au,W.W., Salama,S.A. and Sierra-Torres,C.H. (2003) Functional characterization of polymorphisms in DNA repair genes using cytogenetic challenge assays. *Environ. Health Perspect.*, **111**, 1843–1850.
36. Lotta,T., Vidgren,J., Tilgmann,C., Ulmanen,I., Melen,K., Julkunen,I. and Taskinen,J. (1995) Kinetics of human soluble and membrane-bound catechol O-methyltransferase: a revised mechanism and description of the thermolabile variant of the enzyme. *Biochemistry*, **34**, 4202–4210.
37. Martone,T., Vineis,P., Malaveille,C. and Terracini,B. (2000) Impact of polymorphisms in xeno(endo)biotic metabolism on pattern and frequency of p53 mutations in bladder cancer. *Mutat Res.*, **462**, 303–309.
38. Hung,R.J., Boffetta,P., Brennan,P., Malaveille,C., Gelatti,U., Placidi,D., Carta,A., Hautefeuille,A. and Porru,S. (2004) Genetic polymorphisms of MPO, COMT, MnSOD, NQO1, interactions with environmental exposures and bladder cancer risk. *Carcinogenesis*, **25**, 973–978.
39. Iwaki,H., Kageyama,S., Isono,T., Wakabayashi,Y., Okada,Y., Yoshimura,K., Terai,A., Arai,Y., Iwamura,H., Kawakita,M. *et al.* (2004) Diagnostic potential in bladder cancer of a panel of tumor markers (calreticulin, gamma-synuclein, and catechol-o-methyltransferase) identified by proteomic analysis. *Cancer Sci.*, **95**, 955–961.
40. Yamada,K., Chen,Z., Rozen,R. and Matthews,R.G. (2001) Effects of common polymorphisms on the properties of recombinant human methylenetetrahydrofolate reductase. *Proc. Natl Acad. Sci. USA*, **98**, 14853–14858.
41. Kimura,F., Florl,A.R., Steinhoff,C., Golka,K., Willers,R., Seifert,H.H. and Schulz,W.A. (2001) Polymorphic methyl group metabolism genes in patients with transitional cell carcinoma of the urinary bladder. *Mutat Res.*, **458**, 49–54.
42. Karagas,M.R., Park,S., Nelson,H.H., Andrew,A.S., Mott,L., Schned,A. and Kelsey,K.T. (2005) Methylenetetrahydrofolate reductase (MTHFR) variants and bladder cancer: a population-based case-control study. *Int. J. Hyg. Environ. Health*, **208**, 321–327.
43. Sanyal,S., Festa,F., Sakano,S., Zhang,Z., Steineck,G., Norming,U., Wijkstrom,H., Larsson,P., Kumar,R. and Hemminki,K. (2004) Polymorphisms in DNA repair and metabolic genes in bladder cancer. *Carcinogenesis*, **25**, 729–734.
44. Lin,J., Spitz,M.R., Wang,Y., Schabath,M.B., Gorlov,I.P., Hernandez,L.M., Pillow,P.C., Grossman,H.B. and Wu,X. (2004) Polymorphisms of folate metabolic genes and susceptibility to bladder cancer: a case-control study. *Carcinogenesis*, **25**, 1639–1647.
45. Peluso,M., Airoldi,L., Magagnotti,C., Fiorini,L., Munnia,A., Hautefeuille,A., Malaveille,C. and Vineis,P. (2000) White blood cell DNA adducts and fruit and vegetable consumption in bladder cancer. *Carcinogenesis*, **21**, 183–187.
46. Garcia-Closas,M., Malats,N., Silverman,D., Dosemeci,M., Kogevinas,M., Hein,D.W., Tardon,A., Serra,C., Carrato,A. and Garcia-Closas,R. (2005) NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. *Lancet*, **366**, 649–659.
47. Kiyohara,C., Yoshimasu,K., Takayama,K. and Nakanishi,Y. (2005) NQO1, MPO, and the risk of lung cancer: a HuGE review. *Genet. Med.*, **7**, 463–478.
48. Wu,G., Jiang,X., Lee,W.H. and Chen,P.L. (2003) Assembly of functional ALT-associated promyelocytic leukemia bodies requires Nijmegen Breakage Syndrome 1. *Cancer Res.*, **63**, 2589–2595.
49. Tennyson,R.B., Ebran,N., Herrera,A.E. and Lindsley,J.E. (2002) A novel selection system for chromosome translocations in *Saccharomyces cerevisiae*. *Genetics*, **160**, 1363–1373.
50. Salomon-Nguyen,F., Brizard,F., Le Coniat,M., Radford,I., Berger,R. and Brizard,A. (1998) Abnormalities of the short arm of chromosome 12 in T cell prolymphocytic leukemia. *Leukemia*, **12**, 972–975.
51. Gao,N., Rahmani,M., Shi,X., Dent,P. and Grant,S. (2006) Synergistic antileukemic interactions between 2-medroxyestradiol (2-ME) and histone deacetylase inhibitors involve Akt down-regulation and oxidative stress. *Blood*, **107**, 241–249.
52. Okada,N., Hirata,A., Murakami,Y., Shoji,M., Sakagami,H. and Fujisawa,S. (2005) Induction of cytotoxicity and apoptosis and inhibition of cyclooxygenase-2 gene expression by eugenol-related compounds. *Anticancer Res.*, **25**, 3263–3269.
53. Clavel,J., Bellec,S., Rebouissou,S., Menegaux,F., Feunteun,J., Bonaiti-Pellie,C., Baruchel,A., Kebaili,K., Lambilliotte,A., Leverger,G. *et al.* (2005) Childhood leukaemia, polymorphisms of metabolism enzyme genes, and interactions with maternal tobacco, coffee and alcohol consumption during pregnancy. *Eur. J. Cancer Prev.*, **14**, 531–540.
54. Gallegos-Arreola,M.P., Batista-Gonzalez,C.M., Delgado-Lamas,J.L., Figuera,L.E., Puebla-Perez,A.M., Arnaud-Lopez,L., Peralta-Leal,V., Ramirez-Jirano,L.J. and Zuniga-Gonzalez,G.M. (2004) Cytochrome P4501A1 polymorphism is associated with susceptibility to acute lymphoblastic leukemia in adult Mexican patients. *Blood Cells Mol. Dis.*, **33**, 326–329.
55. Joseph,T., Kusumakumary,P., Chacko,P., Abraham,A. and Radhakrishna Pillai,M. (2004) Genetic polymorphism of CYP1A1, CYP2D6, GSTM1 and GSTT1 and susceptibility to acute lymphoblastic leukaemia in Indian children. *Pediatr. Blood Cancer*, **43**, 560–567.
56. Krajinovic,M., Richer,C., Sinnett,H., Labuda,D. and Sinnett,D. (2000) Genetic polymorphisms of N-acetyltransferases 1 and 2 and gene-gene interaction in the susceptibility to childhood acute lymphoblastic leukemia. *Cancer Epidemiol. Biomarkers Prev.*, **9**, 557–562.