Carcinogenesis vol.31 no.3 pp.462–465, 2010
doi:10.1093/carcin/bgp315
Advance Access publication December 18, 2009

Genome-wide association study in discordant sibships identifies multiple inherited susceptibility alleles linked to lung cancer

Antonella Galvan, Felicia S.Falvella, Elisa Frullanti, Monica Spinola, Matteo Incarbone[1], Mario Nosotti[2], Luigi Santambrogio[2], Barbara Conti, Ugo Pastorino, Anna Gonzalez-Neira[3] and Tommaso A.Dragani*

Fondazione IRCCS, Istituto Nazionale Tumori, 20133 Milan, Italy, [1]Istituto Clinico Humanitas, 20089 Rozzano, Italy, [2]Fondazione IRCCS, Ospedale Maggiore Policlinico, Università degli Studi di Milano, 20122 Milan, Italy and [3]Genotyping Unit (CeGen), Centro Nacional de Investigaciones Oncologicas, E-28029 Madrid, Spain

*To whom correspondence should be addressed. Department of Predictive and Preventive Medicine, Fondazione IRCCS, Istituto Nazionale Tumori, Via G. Venezian 1, 20133 Milan, Italy. Tel: +39 0223902642; Fax: +39 0223902764; Email: tommaso.dragani@istitutotumori.mi.it

We analyzed a series of young (median age = 52 years) non-smoker lung cancer patients and their unaffected siblings as controls, using a genome-wide 620 901 single-nucleotide polymorphism (SNP) array analysis and a case–control DNA pooling approach. We identified 82 putatively associated SNPs that were retested by individual genotyping followed by use of the sib transmission disequilibrium test, pointing to 36 SNPs associated with lung cancer risk in the discordant sibs series. Analysis of these 36 SNPs in a polygenic model characterized by additive and interchangeable effects of rare alleles revealed a highly statistically significant dosage-dependent association between risk allele carrier status and proportion of cancer cases. Replication of the same 36 SNPs in a population-based series confirmed the association with lung cancer for three SNPs, suggesting that phenocopies and genetic heterogeneity can play a major role in the complex genetics of lung cancer risk in the general population.

Introduction

In addition to the major role of cigarette smoking, inherited predisposition of a complex nature appears to play a role in lung cancer risk. Segregation analyses in lung cancer families (1) and genetic linkage studies in mouse models (2) support the polygenic nature of inherited susceptibility to lung cancer (3). However, there is currently no model that adequately explains the complex genetics of lung cancer predisposition and results from linkage and/or association studies are contrasting. For example, a single major genetic locus linked to lung cancer risk in pedigrees with multiple lung cancer members and mapping on chromosome 6q23–25 (4) has not been detected in genome-wide association studies of population-based case–control series (5–9).

Herein, we carried out a genome-wide analysis in a family-based series consisting of young non-smoker lung cancer cases and their unaffected sibs as controls in order to avoid the problem of population stratification. Our findings suggest that lung cancer risk in non-smokers has an inherited susceptibility component that may override the strong role of the smoking habit (10). A few loci associated with lung cancer risk in non-smokers were also confirmed in an independent population-based series.

Subjects and methods

Series and DNA pool construction

The family-based case–control series consisted of 80 Italian lung cancer patients and their healthy sibs recruited with the help of Associazione Marta Nurizzo, Brugherio, Italy (http://www.martalive.org/foreign.htm), on a voluntary basis, with recruitment criteria consisting in the non-smoking status and young age (<60 years) of lung cancer cases, although 14 cases were >60 years and 5 cases were ex-smokers (Table I).

The population-based case–control series consisted of pathologically documented lung adenocarcinoma (ADCA) patients and unrelated healthy individuals enrolled at Istituto Nazionale Tumori (Milan, Italy) from 1980 to 2007. Controls were recruited among blood donors or subjects participating in a computed tomography screening for lung cancer prevention (11) (Table I). Study protocols were approved by the institute ethics committee and written informed consent was obtained from each subject for the use of his/her biological material.

Genomic DNA was extracted from peripheral blood by standard methods and quantified using Picogreen dsDNA Quantitation Kit (Invitrogen, Carlsbad, CA). DNA of 80 discordant sibships was used to generate two pools (cases or controls) containing 30 ng of each DNA sample.

Single-nucleotide polymorphism markers and genotyping

Genome-wide genotyping was carried out in DNA pools prepared from cases and controls of the family-based series. DNA (200 ng per sample) was hybridized using the Infinium II Assay Human610-Quad BeadChip on the Sentrix BeadChip platform (Illumina, San Diego, CA), which allows analysis of 620 901 genetic markers chosen from the International HapMap release 23. For each DNA pool, single-nucleotide polymorphism (SNP) array analysis was carried out in quadruplicate to verify genotype reproducibility and to estimate technical variability. Data were obtained as intensity signals, which were used to determine the allele frequencies of each SNP and to reconstruct the number of chromosomes carrying each of the two possible alleles.

Individual samples were genotyped using MassARRAY (Sequenom, San Diego, CA). Multiplex polymerase chain reaction assays were designed using Sequenom SpectroDESIGNER software by entering the sequence containing the SNP site and 100 bp of flanking sequence on either side of the SNP. SNPs were grouped into multiplexes according to the mass of the extension product over the SNP site. Polymerase chain reaction was carried out in 384-well reaction plates in a volume of 5 μl using 2.5 ng genomic DNA. The extension products were spotted onto a 384-well spectroCHIP before analysis by MALDI-TOF mass spectrometry.

Statistical analysis

Differences between lung cancer cases and their sib controls in allelic frequencies assessed in SNP array hybridization were analyzed using random variance *t*-statistics (12) and BRB ArrayTools (http://linus.nci.nih.gov/BRB-ArrayTools.html). Differences in chromosome counts between cases and controls were tested by Fisher's exact test or by chi-square analysis when the normal approximation was appropriate. The correlation of the allelic frequencies between SNP array and individual genotypes was expressed as a Pearson's coefficient. Association analyses were carried out using PLINK software (13), which included analysis of Hardy–Weinberg equilibrium, family-based transmission disequilibrium test (14) and population-based association analyses between disease status and genotype/allelotype. A generalized linear model with binomial errors was used to test the relationship between genetic susceptibility score and proportion of lung cancer cases; the mean values of genetic susceptibility scores were also analyzed using the Kruskal–Wallis test. The age was downcoded to binary dummy variables (age in decades), which were used as covariates in logistic analyses. Linkage disequilibrium between SNP markers was evaluated using JLIN program (15).

Results

The family-based series points to multiple SNPs associated with lung cancer risk

The genome-wide SNP array analysis conducted in quadruplicate on DNA pools of either 80 lung cancer cases or matched healthy sib controls allowed the screening of 620 901 SNPs, with data obtained as intensity signals for each of the two SNP alleles. Analysis of allelic frequencies of each SNP from case and control DNA pools, deleting SNPs whose minor allele frequency in both cases and controls was <0.1, revealed 659 SNPs with parametric *P*-values $\leq 1.0 \times 10^{-7}$ (equivalent at a false discovery rate $P = 0.0008$). For these 659 SNPs, we estimated the number of chromosomes in cases and controls using a

Abbreviations: ADCA, adenocarcinoma; SNP, single-nucleotide polymorphism.

2×2 contingency table analysis and obtained 82 SNPs putatively associated with disease at $P \leq 0.001$. In analysis of all 82 SNPs in individual cases and healthy sib controls by MassARRAY, three SNPs failed polymerase chain reaction or MassEXTEND primer design, two SNPs failed genotyping, one SNP was monomorphic and one SNP showed highly significant deviation from the Hardy–Weinberg equilibrium, reducing the number of markers to 75 SNPs.

Correlation analysis of the minor allele frequencies estimated in cases and controls either in DNA pools by SNP array analysis or in individual samples by MassARRAY for the 75 SNPs associated with lung cancer demonstrated the reliability of the pooling approach ($r = 0.78$, $P < 2.2 \times 10^{-16}$, Figure 1).

Single-point analysis using the sib transmission disequilibrium test indicated that 36 of the 75 genotyped SNPs were significantly associated with disease status (Table II). The strongest associations were observed for SNPs rs11833102 mapping in the carboxypeptidase M (*CPM*) gene on chromosomes 12, rs17120323 in the sarcoglycan zeta (*SGCZ*) gene on chromosome 8, rs12445758 within cadherin 13, H-cadherin (heart) (*CDH13*) gene on chromosome 16 and rs325702 mapping in the cyclic nucleotide-gated channel alpha 4 (*CNGA4*) gene on chromosome 11.

The 36 SNPs that were statistically associated with lung cancer in the discordant sibships analysis were replicated in a population-based lung ADCA case–control series. None of them showed significant deviation from the Hardy–Weinberg equilibrium, except for rs11074274 ($P = 0.005$, in cases only). Logistic analysis adjusted for sex, age and smoking habit indicated that three SNPs (rs1261411, rs4330610 and rs3019885) were significantly associated in the population-based series ($P < 0.05$) (Table III). Comparison of the cases with early tumor onset (age up to 60 years; $n = 198$) versus the whole controls confirmed the association of the SNP rs3019885 on chromosome 8 ($P = 0.029$) and detected the association of the SNP rs16937762 on chromosome 9 ($P = 0.024$).

The polygenic model of additive and interchangeable effects of the rare alleles explains lung cancer risk in discordant sibships

In the family-based series, we tested our recently proposed polygenic model, characterized by additive and interchangeable effects of the rare alleles (6), for the interpretation of individual risk of lung cancer.

Table I. Characteristics of patients with lung cancer and their healthy controls in the family-based and population-based case–control series

Subject characteristics	Discordant sibs series		Population-based series	
	Controls	Cases	Controls	Cases
No. of subjects	80	80	516	482
Median age (range)[a]	51 (31–73)	52 (31–80)	59 (31–77)	63 (34–78)
Gender				
Male	29	19	391	361
Female	51	61	125	121
Smoker status				
Never	54	75	25	69
Ever	26	5	485	400
Histological type				
ADCA	NA	47	NA	482
NSCLC	NA	30	NA	0
SCLC	NA	3	NA	0

NSCLC, non-small cell lung carcinoma; SCLC, small-cell lung carcinoma; NA, not applicable.
[a]Age in years.

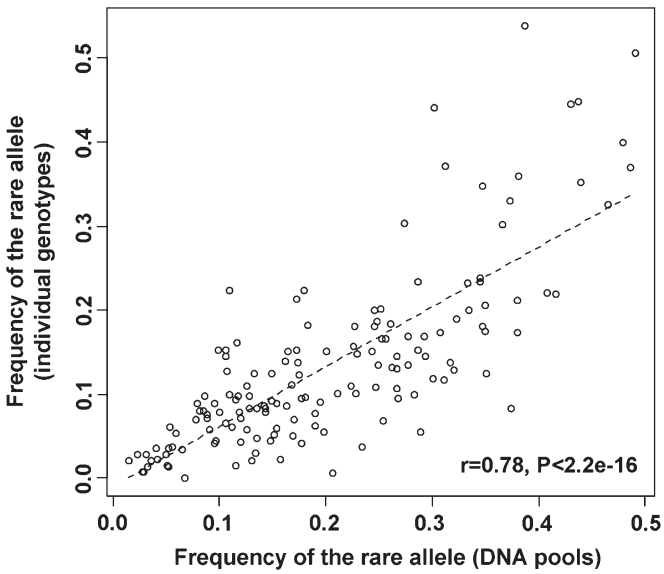

Fig. 1. Correlation between SNP frequencies measured by SNP array analysis of DNA pools and frequencies measured by genotyping of individual samples. Plotted data represent frequencies of the rare allele of 75 SNPs putatively associated with lung cancer risk.

Table II. SNPs showing statistically significant association with lung cancer risk in the discordant sibs series

SNP	Chromosome	Position (Mb)	Gene	*P*-value[a]	OR[b]	95% CI
rs639739	1	4.29		0.0124	0.5	0.24–1.06
rs12748434	1	52.11	*NRD1*	0.058	2.30	0.78–6.81
rs1261411	1	56.79	*PPAP2B*	0.0114	1.65	1.01–2.69[c]
rs2765529	1	119.38	*WARS2*	0.0482	0.66	0.39–1.14
rs6676647	1	119.57		0.0233	0.64	0.37–1.09
rs10931664	2	195.53		0.058	0.57	0.24–1.36
rs721377	3	14.41		0.0126	0.33	0.12–0.95[c]
rs9813644	3	49.89		0.0254	3.63	0.74–17.80
rs1918071	3	54.95	*CACNA2D3*	0.072	0.55	0.26–1.14
rs1456196	3	119.14		0.0348	1.68	0.78–3.60
rs12648320	4	92.54		0.0196	0.54	0.30–0.97[c]
rs28475332	4	188.67		0.0201	0.41	0.16–1.02
rs7713580	5	41.93		0.096	1.56	0.68–3.60
rs16889292	6	78.43		0.0339	0.14	0.02–1.12
rs12663498	6	151.05	*PLEKHG1*	0.055	0.49	0.27–0.91
rs17160175	7	31.47		0.033	1.86	0.92–3.76
rs11773530	7	31.48		0.0164	1.95	0.97–3.93
rs4330610	7	85.18		0.059	0.36	0.09–1.39
rs17120323	8	14.72	*SGCZ*	0.0011[d]	1.94	1.11–3.38[c]
rs3019885	8	118.09		0.052	0.69	0.42–1.13
rs12342234	9	13.32		0.0126	1.83	0.88–3.81
rs16937762	9	19.75	*SLC24A2*	0.0196	0.35	0.11–1.11
rs12001157	9	71.29	*APBA1*	0.090	1.63	0.78–3.42
rs325702	11	6.22	*CNGA4*	0.0045	2.41	1.06–5.49[c]
rs820900	11	38.15		0.0114	0.31	0.10–0.99[c]
rs10842402	12	24.78		0.061	0.63	0.33–1.19
rs11833102	12	67.61	*CPM*	0.0006[d]	2.44	1.21–4.94[c]
rs9544359	13	76.21		0.0254	1.91	0.93–3.94
rs1958226	14	81.27		0.052	0.55	0.24–1.26
rs11074274	15	92.75	*MCTP2*	0.0348	2.54	0.87–7.43
rs12445758	16	82.2	*CDH13*	0.0016[d]	1.93	1.16–3.22[c]
rs790097	17	69.12		0.096	0.48	0.16–1.45
rs4426464	19	1.72	*ONECUT3*	0.0067	2.57	1.17–5.65[c]
rs755032	20	23.9		0.0067	2.04	0.91–4.57
rs2516542	22	19.69		0.0076	3.1	1.18–8.11[c]
rs4823153	22	42.63		0.0046	0.51	0.27–0.97

CI, confidence interval.
[a]DFAM procedure in PLINK toolset, nominal *P*-values. SNPs sorted by chromosome and position.
[b]Based on allelic test for association.
[c]$P < 0.05$, logistic regression procedure in PLINK toolset, based on allelic test for association, i.e. rare allele versus common allele.
[d]$P < 0.05$ by 20 000 permutations of the whole series (75 SNPs).

Table III. SNPs showing statistically significant population-based association with lung ADCA risk

SNP	Chromosome	Position (Mb)	Gene	Rare/common allele	Frequency of the rare allele (controls/cases)	OR[a]	95% CI[a]	*P*-value[a]
rs1261411	1	56.79	*PPAP2B*	A/G	0.39/0.44	1.23	1.02–1.49	0.033
rs4330610	7	85.18		T/C	0.05/0.03	0.49	0.29–0.84	0.009
rs3019885	8	118.09	*SLC30A8*	G/T	0.46/0.50	1.25	1.03–1.51	0.021

CI, confidence interval.
[a]Logistic regression procedure in PLINK toolset, adjusted for sex, age in decades and smoking habit; *P*-values obtained by allelic test for association.

Analysis included a total of 35 SNPs, reflecting the exclusion of SNP rs17160175 due to its high linkage disequilibrium with rs11773530 ($D' = 1.0$, $r^2 = 0.97$), and 151 subjects, with five controls and four cases removed from the dataset because >80% genotypes were missing.

To test the model, a score of +1 or −1 was attributed to the rare allele of each SNP based on its association with increased or decreased lung cancer risk, respectively. For each subject, the sum of the scores for all 35 SNPs was obtained as a general estimator of the individual genetic risk. The average estimator was -1.6 ± 0.3 (mean ± standard error) in controls and 2.0 ± 0.3 in cases, respectively ($P = 2.0 \times 10^{-11}$, Kruskal–Wallis test), and the proportion of lung cancer cases increased at higher genetic susceptibility scores (Figure 2; $P = 5.9 \times 10^{-9}$).

Applying the same polygenic model in the population-based series, by using the three confirmed SNPs (Table III), we found that the average estimator was 1.25 ± 0.03 in controls and 1.40 ± 0.03 in cases, respectively ($P = 0.0019$, Kruskal–Wallis test).

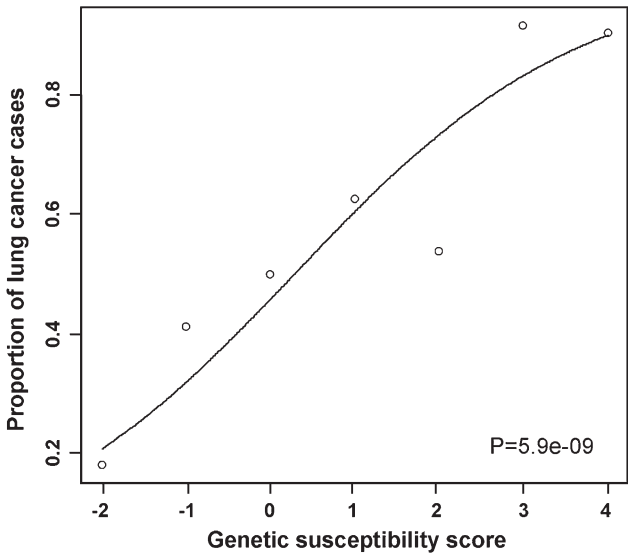

Fig. 2. A polygenic inheritance model with additive and interchangeable effects of rare alleles at lung cancer modifier loci explains the individual risk of lung cancer in the family-based series. Scatterplot shows the proportions of subjects that are cases as a function of genetic susceptibility score and the fitted line.

Discussion

Our use of a sibling-based study design detected loci statistically associated with lung cancer risk, even with a limited number of subjects. Notwithstanding the poor feasibility of the recruitment of healthy sibs controls to carry out family-based genome-wide association studies in lung cancer, due to late age at cancer onset for most of the cases, the possible benefits resulting from the appropriate matching of cases and controls may justify the effort.

An important advantage of the discordant sibs design is the possibility to exclude the potential for bias due to population stratification, which is common in population-based studies (14). Indeed, cases and controls derive from the same pedigree whose DNA differences may lie in genetic polymorphisms putatively responsible for the disease status. The effect (lung cancer risk estimation) detected by the discordant sib pair design (1:1 case:control ratio) and sib transmission disequilibrium test is due to the combined presence of linkage and association (14).

Starting from a genome-wide association studies of DNA pools, we identified 36 SNPs that showed significant linkage/association in the family-based series (Table I). Individual genotyping confirmed the robustness of our pooling approach (Figure 1), demonstrating that this method produces reliable results and is time- and cost-effective. Of these 36 genetic markers, 13 mapped within genes. The most significantly associated SNPs ($P \leq 0.0045$), i.e. rs11833102, rs17120323, rs12445758 and rs325702, mapped in carboxypeptidase M (*CPM*), sarcoglycan zeta (*SGCZ*), cadherin 13, H-cadherin (heart) (*CDH13*) and cyclic nucleotide-gated channel alpha 4 (*CNGA4*) genes, respectively. Overexpression of CPM was recently reported to correlate negatively with disease survival in human lung ADCA patients (16), and aberrant methylation of the *CDH13* gene was observed in lung ADCA (17). Thus, our findings point to the relevance of genetic components in the modulation of individual lung cancer risk in non-smokers.

Interestingly, we found that one of the associated SNPs (rs12663498, $P = 0.055$) maps on 6q25.1, the same locus previously linked to lung cancer risk in pedigrees with multiple lung cancer members (4). The SNP maps within the pleckstrin homology domain-containing family G (with RhoGef domain) member 1 (*PLEKHG1*) gene, which lies 2 Mb from RGS17, the major candidate gene for the familial lung cancer susceptibility locus (18).

Our present findings in non-smokers discordant sib pairs did not confirm the previously reported population-based association of lung cancer risk with the chromosome 15q25 nicotinic receptor locus (7–9,11). These results are supported by our recent meta-analysis showing that this locus is not associated with lung cancer risk in >1000 never-smoker cases and >1800 controls (19).

Another aspect that we should take into consideration is the role of genetic heterogeneity in the predisposition to cancer. Indeed, independent loci may modulate the risk of sporadic and of familial cancer, as the model of breast cancer susceptibility demonstrated (20,21). Also, we should consider the great impact of the major environmental risk factor, i.e. smoking habit, and the difficulty in separating the genetic and environmental contributions to lung cancer risk. Indeed, a study in monozygotic and dizygotic twins showed that the possible sharing of the same environmental risk factors may play a major role in lung cancer risk (22).

The application of the previously proposed polygenic model of interchangeable and additive effects of rare alleles (6) to the 35 SNPs associated in the discordant sib-based series showed a highly statistically significant association between the genetic susceptibility score and the proportion of cancer cases (Figure 2), supporting the plausibility of this type of polygenic model.

Our single-point analysis confirmed in the population-based series only 3 of 36 SNPs that were statistically associated in the

family-based series (Tables II and III). Since the results of the family-based series are not biased by population structure and most of the detected SNPs presumably represent real associations, the scarce effects of the same SNPs in the population-based series rest in either the existence of significant population admixture, masking real associations, or the existence of phenocopies and a high degree of genetic heterogeneity in the general population. In the latter case, a model of 'private' genetic epidemiology (23) may account for the genetic effects detected in lung cancer families.

Funding

Associazione and Fondazione Italiana Ricerca Cancro; Fondo Investimenti Ricerca di Base, Italy; CEGEN (Spanish National Genotyping Centre), Centro Nacional de Investigaciones Oncologicas Node, Spain.

Acknowledgements

The authors thank Associazione Marta Nurizzo for their help in recruiting the non-smoker family-based series and Harvard-Partners Center for Genetics and Genomics Genotyping Facility, Cambridge, MA, for custom genotyping by MassARRAY.

Conflict of Interest Statement: None declared.

References

1. Xu,H. *et al*. (2005) Complex segregation analysis reveals a multigene model for lung cancer. *Hum. Genet.*, **116**, 121–127.
2. Dragani,T.A. (2003) 10 Years of mouse cancer modifier loci: human relevance. *Cancer Res.*, **63**, 3011–3018.
3. Dragani,T.A. *et al*. (1996) A polygenic model of inherited predisposition to cancer. *FASEB J.*, **10**, 865–870.
4. Bailey-Wilson,J.E. *et al*. (2004) A major lung cancer susceptibility locus maps to chromosome 6q23-25. *Am. J. Hum. Genet.*, **75**, 460–474.
5. Spinola,M. *et al*. (2006) Association of the PDCD5 locus with lung cancer risk and prognosis in smokers. *J. Clin. Oncol.*, **24**, 1672–1678.
6. Galvan,A. *et al*. (2008) A polygenic model with common variants may predict lung adenocarcinoma risk in humans. *Int. J. Cancer*, **123**, 2327–2330.
7. Amos,C.I. *et al*. (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.*, **40**, 616–622.
8. Hung,R.J. *et al*. (2008) A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*, **452**, 633–637.
9. Thorgeirsson,T.E. *et al*. (2008) A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, **452**, 638–642.
10. Matakidou,A. *et al*. (2005) Systematic review of the relationship between family history and lung cancer risk. *Br. J. Cancer*, **93**, 825–833.
11. Falvella,F.S. *et al*. (2009) Transcription deregulation at the 15q25 locus in association with lung adenocarcinoma risk. *Clin. Cancer Res.*, **15**, 1837–1842.
12. Wright,G.W. *et al*. (2003) A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, **19**, 2448–2455.
13. Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
14. Spielman,R.S. *et al*. (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.*, **62**, 450–458.
15. Carter,K.W. *et al*. (2006) JLIN: a java based linkage disequilibrium plotter. *BMC Bioinformatics*, **7**, 60.
16. Tsakiris,I. *et al*. (2008) The presence of carboxypeptidase-M in tumour cells signifies epidermal growth factor receptor expression in lung adenocarcinomas: the coexistence predicts a poor prognosis regardless of EGFR levels. *J. Cancer Res. Clin. Oncol.*, **134**, 439–451.
17. Kubo,T. *et al*. (2009) DNA methylation in small lung adenocarcinoma with bronchioloalveolar carcinoma components. *Lung Cancer*, **65**, 328–332.
18. You,M. *et al*. (2009) Fine mapping of chromosome 6q23-25 region in familial lung cancer families reveals RGS17 as a likely candidate gene. *Clin. Cancer Res.*, **15**, 2666–2674.
19. Galvan,A. *et al*. (2009) Nicotine dependence may link the 15q25 locus to lung cancer risk. *Carcinogenesis,* 10.1093/carcin/bgp282.
20. Hunter,D.J. *et al*. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.*, **39**, 870–874.
21. Pharoah,P.D. *et al*. (2002) Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.*, **31**, 33–36.
22. Lichtenstein,P. *et al*. (2000) Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.*, **343**, 78–85.
23. Ioannidis,J.P. (2006) Commentary: grading the credibility of molecular evidence for complex diseases. *Int. J. Epidemiol.*, **35**, 572–577.

Received August 25, 2009; revised November 6, 2009; accepted December 9, 2009