FEBS openbio

FEBS PRESS science publishing by scientists

METHOD

Development of a general logistic model for disease risk prediction using multiple SNPs

Cheng Long[1], Guanting Lv[2] and Xinmiao Fu[3]

1 West China Hospital of Sichuan University, Chengdu, Sichuan, China
2 Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China
3 Provincial University Key Laboratory of Cellular Stress Response and Metabolic Regulation, College of Life Sciences, Fujian Normal University, Fuzhou, Fujian, China

Keywords
disease risk prediction; GWASs; logistic regression; personalized medicine; precise medicine; SNP

Correspondence
X. Fu, Provincial University Key Laboratory of Cellular Stress Response and Metabolic Regulation, College of Life Sciences, Fujian Normal University, Fuzhou City, Fujian Province 350117, China
E-mail: xmfu@fjnu.edu.cn

(Received 16 June 2019, revised 8 August 2019, accepted 16 August 2019)

doi:10.1002/2211-5463.12722

Human diseases are usually linked to multiloci genetic alterations, including single-nucleotide polymorphisms (SNPs). Methods to use these SNPs for disease risk prediction (DRP) are of clinical interest. DRP algorithms explored by commercial companies to date have tended to be complex and led to controversial prediction results. Here, we present a general approach for establishing a logistic model-based DRP algorithm, in which multiple SNP risk factors from different publications are directly used. In particular, the coefficient β of each SNP is set as the natural logarithm of the reported odds ratio, and the constant coefficient β_0 is comprehensively determined by the coefficient and frequency of each SNP and the average disease risk in populations. Furthermore, homozygous SNP is considered a dummy variable, and the SNPs are updated (addition, deletion and modification) if necessary. Importantly, we validated this algorithm as a proof of concept: two patients with lung cancer were identified as the maximum risk cases from 57 Chinese individuals. Our logistic model-based DRP algorithm is apparently more intuitive and self-evident than the algorithms explored by commercial companies, and it may facilitate DRP commercialization in the era of personalized medicine.

Genome-wide association studies (GWASs) are increasingly uncovering the effect of genetic alterations in most of the common diseases [1] and are considerably accelerating the commercialization of personalized medicine [2]. One of the genetic alteration based applications for personalized medicine is disease risk prediction (DRP) [3–5], which has been reported to benefit individuals in terms of improving their lifestyle choices and facilitating preventive screening [3,4,6,7]. However, the discrepancy among DRPs of different direct-to-customer companies (e.g., 23andMe, Navigenics and deCODEme) has been widely reported [8–11] and substantially limits their commercialization. Such discrepancy has been mainly attributed to the difference in the selection of single-nucleotide polymorphism (SNP) risk factors and DRP algorithms. Although it is impossible for different DRP providers to select a palette of the same SNP risk factor markers for one type of disease, there is a high demand for the development of a general DRP algorithm.

Retrospectively, the algorithms of three DRP companies (23andMe, deCODEme and Navigenics) have notable differences in principle (as described in Data S1) and are complicated and compromised [9,10]. Whereas 23andMe and deCODEme transform the odds ratio (OR) of each SNP into a likelihood ratio, Navigenics transforms it into the relative risk. Thereafter, 23andMe multiplies the likelihood ratios of

Abbreviations
DRP, disease risk prediction; GWAS, genome-wide association study; OR, odds ratio; SNP, single-nucleotide polymorphism.

single genotypes by the average odds of the disease and converts these odds into risks, deCODEme multiplies the likelihood ratios by the average risk of disease, and Navigenics calculates the relative risks for all of the possible genotype combinations.

Apparently, these DRP algorithms are not suitable for generalization because of the following limitations [9,10]. First, all of them require the OR values of both homozygous and heterozygous SNPs. However, the OR values for many homozygous SNPs are not available in the existing GWAS literature, and this may limit the utility of high-risk SNPs whose OR values are available for only the heterozygous forms. Second, the DRP algorithm used by Navigenics requires a substantial amount of computer working memory [10]. Third, the DRP algorithm used by deCODEme is apparently not normalized; that is, the calculated risk might be larger than the upper limit value of 100% under certain circumstances. It is thus highly demanding to develop a more self-evident, robust and reliable algorithm for DRP.

Logistic regression has been widely used for risk factor identification and evaluation in general, [12] as well as in GWASs in particular. After a regression analysis of the experimental data, the generated logistic model containing certain risk factors can be applied for DRP [5,12–15]. However, the SNP risk factors related to a disease are usually identified in many studies by different groups of researchers; therefore, the generated logistic models in these studies are also different and would not be suitable for DRP directly. From this perspective, a method to build a general logistic model that integrates multiple SNP risk factors from multiple publications is in high demand for DRP. The algorithms explored by commercial DRP companies represent such efforts but apparently cannot meet the needs [8–11].

Here, we present a general approach to build a logistic model for DRP, in which multiple SNPs from multiple publications are integrated and can be promptly updated. Our model largely overcomes the aforementioned limitations of the currently used algorithms. Furthermore, we validated the model on lung cancer with Chinese individuals, illustrating its significance in preventive screening.

Materials and methods

Data collection from the literature

For each selected SNP, the original literature was downloaded, and the OR values for both homozygous (if available) and heterozygous SNPs were collected. SNPs with an OR <1.15 were omitted. Then, the frequency of each SNP in Chinese individuals and all individuals was extracted from the National Center for Biotechnology Information database by searching each SNP ID (for details, refer to Table 1), and the lifetime risk for lung cancer in Chinese men and women was assigned according to the Chinese Cancer Registry Annual Report 2010 [16]. These data were used by the algorithm for the regression analysis, as described in the Results and Discussion.

Ethics approval and consent to participate

The experiments in genotyping human SNPs were approved by the Ethics Committee of Third Hospital of Peking University. We confirm that our study is compliant with the ‘Guidance of the Ministry of Science and Technology of China for the Review and Approval of Human Genetic Resources’, and the study methodologies conformed to the standards set by the Declaration of Helsinki. The experiments were undertaken with the understanding and written consent of each subject.

Human subjects

Patients were normally subjected to different clinical tests, including peripheral blood-based clinical tests. The remaining blood (around at a volume of 1.5 mL) of each patient, accompanied with the information of only sex, age, smoking and clinical symptoms, was transferred by the hospital to us for DNA extraction and SNP genotyping, with the data being analyzed anonymously.

SNP genotyping

Genomic DNA was extracted from peripheral blood by using a QIAamp DNA Blood Mini Kit (Qiagen, Valencia, CA, USA) in accordance with the manufacturer’s instructions. The primers for SNP genotyping were designed using MassArray Assay Design v4.0 according to the sequence information flanking the SNP, were synthesized by Invitrogen and were validated by MassArray Analyzer Compact. The PCR, removal of free dNTP and sequencing by MassArray Analyzer Compact (Sequenom, San Diego, CA, USA) were performed according to the manufacturer’s instructions.

Statistics

Each sample was resequenced independently, and all of the SNP sequencing results were consistent between independent repeated experiments. Risk calculation was performed using Microsoft Excel (Microsoft, Redmond, WA, USA). A statistical analysis was performed in the MICROORIGIN software using the ANOVA algorithm at a significance level of 0.05.

Results and Discussion

Logistic model generated from a designed study does not meet the need for DRP on an increasing number of SNP-related diseases

Logistic regression has been widely used for risk factor identification and evaluation, and is generally presented as follows:

$$\mathrm{Logit}(P) = \ln(\frac{P}{1-P}) = \beta_0 + \beta_1 x_1 + \beta_i x_i \ldots + \beta_n x_n \quad (1)$$

where β_i and x_i are the coefficient and the logic value of each risk factor, respectively. Then, the disease risk P is expressed as follows:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_i x_i \ldots + \beta_n x_n)}}. \quad (2)$$

In a research study, researchers usually use a majority of their experimental data to estimate the coefficients in Eqns (1,2) by using statistical software through a maximum likelihood estimation algorithm [12]. Then, the remaining experimental data are subjected to the generated logistic model for validation [12]. Further, researchers can use the generalized model to predict the disease risk of new subjects. Apparently, this strategy is useful and effective for DRP based on the traditional risk factors [12].

In general, only one or a few novel SNP risk factors related to a disease are identified in a specific GWAS, in which the coefficients are determined and a specific logistic model is generated as described above. If the disease is linked to only these SNP risk factors, the logistic model generated from this GWAS is suitable for DRP directly. However, most human diseases are usually linked to a large number of SNP risk factors (sometimes up to 100 [17]), and each of them has a minor effect on the disease risk. In particular, these SNPs are identified one by one by different research groups and are thus timely reported in multiple publications. Accordingly, the logistic models generated in these studies are different and are not interchangeable for DRP directly.

Generation of the logistic model for DRP by directly using multiple SNPs from multiple publications

Now, the question is how to integrate the multiple SNPs from multiple publications to build a general model for DRP. Commercial DRP companies have developed algorithms to directly use the information of multiple SNPs from multiple publications for DRP by incorporating the OR value and the population frequency of each SNP, which are usually available in the GWAS literature and/or public databases. These algorithms, however, are complicated and lead to discrepant DRP results [8–11]. We thought to integrate the multiple SNP risk factors from multiple publications into a logistic model for DRP (as indicated by Eqn 2).

Regarding Eqn 2 for DRP, we need the coefficient β_i of each SNP and the constant coefficient β_0, as well as the value of x_i for each SNP that can be determined by the SNP genotyping of the subject. One way to determine β_i and β_0 is to resequence the known SNP risk factors in a large number of subjects and then perform a logistic regression analysis. This strategy, although used previously [5], does not make full use of the knowledge of the SNP risk factors that have been reported and thus has compromised cost-effectiveness. In particular, if a new SNP risk factor related to the disease is discovered, an additional clinical study needs to be performed for its incorporation into the logistic model.

In our approach, β_i in Eqn (2) is not estimated by a statistical analysis on the clinical data as performed by Ripatti *et al.* [5]. Rather, it is obtained directly by transforming the reported OR value of the SNP according to the logistic regression as follows:

$$\beta_i = \ln(\mathrm{OR}_i) \quad (3)$$

This is based on the definition of OR in a logistic regression model as follows:

$$\mathrm{OR}_i = \frac{\mathrm{odds}(x_i+1)}{\mathrm{odds}(x_i)} = \frac{\frac{P(x_i+1)}{1-P(x_i+1)}}{\frac{P(x_i)}{1-P(x_i)}} = \frac{e^{\beta_0+\beta_1 x_1+\beta_i x_i+\beta_i+\ldots+\beta_n x_n}}{e^{\beta_0+\beta_1 x_1+\beta_i x_i+\ldots+\beta_n x_n}} = e^{\beta_i} \quad (4)$$

where the OR value of a SNP risk factor is explained as the ratio of the relative disease risk of the person at risk to the relative disease risk of a person not at risk.

Next, β_0 can be determined as follows: if a SNP risk factor plays a role in the disease development, its contribution with respect to individuals should be considered the same as that with respect to populations. Then, Eqn 1 for an individual can be converted to the following form for populations:

$$\ln\left(\frac{P_0}{1-P_0}\right) = \beta_0 + \beta_1 f_1 + \beta_i f_i \ldots + \beta_n f_n \quad (5)$$

where the population frequency (f_i) of each SNP and the average population disease risk (P_0) are usually available in public databases; the coefficient (β_i) of each SNP has been already determined through Eqn 3. Then, β_0 can be calculated using Eqn 6 as follows:

$$\beta_0 = \ln\left(\frac{P_0}{1-P_0}\right) - (\beta_1 f_1 + \beta_i f_i \ldots + \beta_n f_n) \qquad (6)$$

Apparently, β_0 reflects the contribution of the unidentified and/or unselected SNP risk factors, as well as of the non-SNP risk factors, including the traditional risk factors (e.g., smoking and drinking). Importantly, if new critical SNPs are identified and need to be incorporated into our model, β_0 can be recalculated according to Eqn 6. In other words, our logistic model for DRP can be updated quickly if necessary.

In addition, the variance of P for each subject is collectively determined by the sum of the product of the variance of each β_i and the logic value x_i (according to Eqn 2), with the variance of each β_i being derived from the variance of OR for each SNP (according to Eqn 3). In this regard, OR with a smaller variance is favorable when multiple ORs are available for a specific SNP as reported in multiple publications.

Equations 1–6 represent a general approach to establishing a logistic model-based DRP algorithm by combining multiple SNPs from multiple publications, as summarized here. First, the SNP risk factor markers for a specific disease are selected from the existing GWAS literature, collecting their OR values and population frequencies. Second, the coefficient β_i of each SNP and the constant coefficient β_0 are determined using Eqns 3 and 6, respectively; therefore, Eqn 2 for DRP is ready. Third, the SNP risk factors for the subject are determined experimentally. Fourth, DRP is performed using Eqn 2. Note that both homozygous and heterozygous SNPs are suitable for DRP in our algorithm. If the OR value for the homozygous form of a SNP is available, it is considered a dummy variable and x_i is set as $\ln(\mathrm{OR}_{\text{homozygous}})/\ln(\mathrm{OR}_{\text{heterozygous}})$.

Collection of SNP markers for lung cancer risk prediction in Chinese individuals

Thus far, about 400 000 reference SNPs have been deposited for human genomes (refer to the National Center for Biotechnology Information SNP database: https://www.ncbi.nlm.nih.gov/snp/docs/RefSNP_about/). Of these, about 71 600 SNPs have been found to be associated with human diseases and traits, as reported in the NHGRI-EBI GWAS Catalog database [18,19] (https://www.ebi.ac.uk/gwas/). Lung cancer is considered a partially inherited complex disease [4], and about 80 SNPs have been reported to be associated with it [17]. Among these SNPs, a threshold value of 1.15 was set subjectively, and some genetically dependent SNPs were omitted. In the end, eight representative SNPs were selected for the lung cancer risk prediction in our study. The OR value and the population frequency for each of these eight SNPs in Chinese individuals were collected (Table 1).

Generation of algorithms for lung cancer risk prediction

Then, we determined β for each SNP (Table 1) according to the aforementioned rules. Given the lifetime risk for lung cancer in Chinese men and women being 5.62% and 2.56%, respectively, β_0 was calculated as −3.742 and −4.560 (for details, refer to Table 1). Together, the SNP-based risk prediction for lung cancer with Chinese men and women was finalized as follows:

$$P(\text{men}) = 1/(1 + e^{3.742-0.399x_1-0.604x_2-0.285x_3-0.262x_4-0.270x_5-0.166x_6-0.199x_7-0.148x_8}) \qquad (7)$$

$$P(\text{women}) = 1/(1 + e^{4.560-0.399x_1-0.604x_2-0.285x_3-0.262x_4-0.270x_5-0.166x_6-0.199x_7-0.148x_8}) \qquad (8)$$

Utilization of the algorithm for lung cancer risk prediction in Chinese individuals

To demonstrate the effectiveness of our logistic algorithm, we collected blood samples of 48 Chinese subjects (38 men and 10 women), genotyped the eight selected SNPs (for details, refer to Table S1) and then calculated the lung cancer risk of each subject by using Eqns (7,8). Remarkably, subject 17, who was diagnosed as the sole lung cancer patient among these 48 subjects and had the symptoms of poorly differentiated adenocarcinoma in the right lung and multiple metastases, had the highest absolute lifetime risk of 28.3% and a relative lifetime risk of 5.0 (Table 2). The absolute lifetime risk of this patient was significantly higher

Table 1. SNPs and parameters of logistic model for lifetime risk prediction of lung cancer. NA, not available.

SNP ID	Risk base	Frequency in Chinese individuals	Frequency in all individuals	OR of heterozygous SNPs	OR of homozygous SNPs	References	[a]β	[b]β_0 calculation
rs1820453	G	0.239	0.374	1.49	1.65	[20]	0.399	0.174
rs716274	G	0.25	0.438	1.83	2.96	[20]	0.604	0.294
rs9981861	G	0.125	0.332	1.33	[c]NA	[21]	0.285	0.036
rs16951095	C	0.805	0.917	1.3	NA	[22]	0.262	0.211
rs1051730	T	0.037	0.185	1.31	NA	[23]	0.270	0.010
rs402710	C	0.733	0.657	1.18	NA	[24]	0.166	0.121
rs2808630	G	0.22	0.211	1.22	NA	[25]	0.199	0.044
rs7626795	G	0.207	0.246	1.16	NA	[25]	0.148	0.031
Men	[c]Risk	β_0	Women	[c]Risk	β_0		Sum	0.921
	0.0562	−3.742		0.0256	−4.560			

[a] In the logistic model, coefficient β for each SNP was calculated as the natural logarithm of the OR value of the heterologous SNP on the basis of Eq 3. [b] For SNP without the OR value of the homozygous form, it was the product of β and the frequency in Chinese individuals (e.g., for rs9981861, 0.285 × 0.125 = 0.036). Otherwise, it was calculated as the sum of two products contributed, respectively, by the heterozygous and homozygous SNPs [e.g., for rs1820453, according to the Hardy-Weinberg equilibrium, the frequencies of heterozygous and homozygous SNPs in Chinese individuals were calculated as 2 × 0.239 × (1 − 0.239) = 0.364 and 0.239 × 0.239 = 0.057, respectively, and the total contribution was calculated as ln (1.49) × 0.364 + ln(1.65) × 0.057 = 0.174]. [c] The average lifetime risk for lung cancer in Chinese men and women was set as 5.62% and 2.56%, respectively, according to the Chinese Cancer Registry Annual Report 2010 [16].

than that of the remaining 47 persons ($P = 0.0000024$) and was significantly higher than that of the remaining 37 men ($P = 0.000015$). In addition, the average lifetime risk for the 38 men and the 10 women was 8.9% and 5.5%, respectively (Table 2); both of these values were not significantly different from the average disease risk of men and women ($P > 0.05$).

In another batch of genetic screening for nine subjects, we found that a female patient diagnosed with non-small-cell lung cancer had an absolute lifetime risk of 16.2% and a relative lifetime risk of 6.3 (Table 2), which was significantly higher than that of the remaining eight normal subjects. These results, although based on a test with a small group size and a limited number of positive patients, indicated that our algorithm was sensitive, to a certain degree, in screening the high-risk lung cancer patient.

Conclusions

In this study, we developed a general logistic model-based algorithm to calculate the disease risk by directly using multiple SNP risk factors from multiple publications, and tested this algorithm on Chinese individuals for lung cancer risk prediction. The logistic algorithm that we explored was apparently more intuitive and self-evident than the algorithms adopted by commercial DRP providers [7–11], although similar input parameters were used [9,10] (i.e., the average population risk of disease, the OR value and the genotype and/or allele frequencies of the SNPs).

Table 2. Lung cancer risk and clinical symptoms for 48 subjects.

ID	Absolute risk	Relative risk	Sex	Symptom
17	0.283	5.0	M	Poorly differentiated adenocarcinoma in the right lung, multiple metastases
23	0.174	3.1	M	Normal
27	0.174	3.1	M	Normal
25	0.152	2.7	M	Normal
26	0.145	2.6	M	Normal
16	0.141	2.5	M	Normal
2	0.138	2.5	M	After surgery for the right knee ligament reconstruction
32	0.122	2.2	M	Normal
22	0.119	2.1	M	Normal
24	0.119	2.1	M	Normal
19	0.094	1.7	M	Normal
4	0.092	1.6	M	Lumbar spinal stenosis
10	0.092	1.6	M	Lumbar disc herniation
11	0.092	1.6	M	Normal
30	0.087	1.5	M	Normal
28	0.086	3.4	F	Non-Hodgkin's lymphoma
15	0.084	1.5	M	Normal
43	0.081	1.4	M	Normal
18	0.081	3.2	F	Tubal pregnancy
33	0.078	1.4	M	Normal
8	0.077	1.4	M	Normal
31	0.077	3.0	F	Cholecystitis
9	0.076	1.3	M	Normal
34	0.073	1.3	M	Normal
1	0.072	1.3	M	Normal

Continued

Table 2. (Continued).

ID	Absolute risk	Relative risk	Sex	Symptom
47	0.072	1.3	M	Normal
38	0.071	1.3	M	Nasopharyngeal
21	0.069	2.7	F	Right breast invasive ductal carcinoma, appendicitis
14	0.067	2.6	F	Intrauterine pregnancy
48	0.064	1.1	M	Normal
40	0.063	1.1	M	Normal
12	0.061	1.1	M	Normal
29	0.058	2.3	F	After surgery for the right shoulder
42	0.056	1.0	M	Normal
20	0.054	1.0	M	Normal
39	0.053	0.9	M	Normal
37	0.044	0.8	M	Normal
6	0.042	0.8	M	Normal
3	0.040	0.7	M	Intertrochanteric fracture of the left femur
41	0.040	0.7	M	Normal
44	0.040	0.7	M	Normal
46	0.040	1.6	F	Normal
35	0.038	0.7	M	Normal
7	0.035	0.6	M	Normal
45	0.035	0.6	M	Normal
5	0.035	1.4	F	Abnormal ampulla of Vater (likely to be malignant)
13	0.022	0.8	F	Hemorrhoids and ovarian cysts, bronchial asthma
36	0.019	0.7	F	Osteoarthritis knees
N1	0.162	6.3	F	Non-small-cell lung cancer in the right lung
N2	0.040	0.7	M	Normal
N3	0.056	1.0	M	Normal
N4	0.081	1.4	M	Normal
N5	0.040	0.7	M	Normal
N6	0.035	0.6	M	Normal
N7	0.040	1.6	F	Normal
N8	0.072	1.3	M	Normal
N9	0.064	1.1	M	Normal

Our logistic algorithm had several advantages: (a) the coefficient β in the risk calculation equation directly reflected the contribution of each SNP and was simply determined by the OR value; (b) the SNP risk factor markers and the OR value were adjustable and/or promptly updated if necessary, such that β_0 could be recalculated using the new set of SNP markers; (c) the SNP risk factor markers, with or without the OR value for the homozygous forms, were all suitable for risk prediction; and (d) accordingly, the non-SNP risk factors (e.g., smoking and drinking) could also be included in our logistic model for DRP. Therefore, our study is of interest for the commercialization of DRP, given that similar logistic model-based algorithms can be established for common complex diseases (e.g., diabetes, cardiovascular disease and cancers) that have been reported to be linked quantitatively to multiple SNP risk factors [17].

Acknowledgements

This work was supported by National Natural Science Foundation of China Grants 31570778, 31770830 and 31972918.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

XF designed the research and established the model. GL performed genotyping experiments. XF and CL analyzed the data and wrote the manuscript.

References

1 Manolio TA, Brooks LD and Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* **118**, 1590–1605.

2 Hamburg MA and Collins FS (2010) The path to personalized medicine. *N Engl J Med* **363**, 301–304.

3 Wray NR, Goddard ME and Visscher PM (2008) Prediction of individual genetic risk of complex disease. *Curr Opin Genet Dev* **18**, 257–263.

4 Janssens AC and van Duijn CM (2010) An epidemiological perspective on the future of direct-to-consumer personal genome testing. *Investig Genet* **1**, 10.

5 Ripatti S, Tikkanen E, Orho-Melander M, Havulinna AS, Silander K, Sharma A, Guiducci C, Perola M, Jula A, Sinisalo J *et al.* (2010) A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet* **376**, 1393–1400.

6 Bloss CS, Schork NJ and Topol EJ (2014) Direct-to-consumer pharmacogenomic testing is associated with increased physician utilisation. *J Med Genet* **51**, 83–89.

7 Bloss CS, Topol EJ and Schork NJ (2012) Association of direct-to-consumer genome-wide disease risk estimates and self-reported disease. *Genet Epidemiol* **36**, 66–70.

8 Ng PC, Murray SS, Levy S and Venter JC (2009) An agenda for personalized medicine. *Nature* **461**, 724–726.

9 Kido T, Kawashima M, Nishino S, Swan M, Kamatani N and Butte AJ (2013) Systematic evaluation of personal genome services for Japanese individuals. *J Hum Genet* **58**, 734–741.

10 Kalf RR, Mihaescu R, Kundu S, de Knijff P, Green RC and Janssens AC (2014) Variations in predicted risks in personal genome testing for common complex diseases. *Genet Med* **16**, 85–91.

11 Kutz GD (2010) Direct-to-consumer genetic tests: misleading test results are further complicated by deceptive marketing and other questionable practices, T.U.S.G.A. Office, Editor.

12 Jiang J, Xue F and Xu T (2013) *Applied Medical Multivariate Statistics*. Science Press, Beijing.

13 Gail MH (2008) Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *J Natl Cancer Inst* **100**, 1037–1041.

14 Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ and Park JH (2013) Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat Genet* **45**, 400–405. 405 e1-3.

15 Spitz MR, Hong WK, Amos CI, Wu X, Schabath MB, Dong Q, Shete S and Etzel CJ (2007) A risk model for prediction of lung cancer. *J Natl Cancer Inst* **99**, 715–726.

16 Zhao P and Chen WQ (2011) Chinese Cancer Registry Annual Report 2010, National Office for Cancer Prevention and Control, National Central Cancer Registry, and Ministry of Health Disease Prevention and Control Bureau, Editors. Military Medical Science Press, Beijing.

17 NHGRI (2014) A Catalog of Published Genome-Wide Association Studies. The National Human Genome Research Institute, NIH.

18 Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L *et al*. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001–D1006.

19 Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E and Sollis E (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005–D1012.

20 Wu C, Xu B, Yuan P, Miao X, Liu Y, Guan Y, Yu D, Xu J, Zhang T, Shen H *et al*. (2010) Genome-wide interrogation identifies YAP1 variants associated with survival of small-cell lung cancer patients. *Cancer Res* **70**, 9721–9729.

21 Sato Y, Yamamoto N, Kunitoh H, Ohe Y, Minami H, Laird NM, Katori N, Saito Y, Ohnami S, Sakamoto H *et al*. (2011) Genome-wide association study on overall survival of advanced non-small cell lung cancer patients treated with carboplatin and paclitaxel. *J Thorac Oncol* **6**, 132–138.

22 Yoon KA, Park JH, Han J, Park S, Lee GK, Han JY, Zo JI, Kim J, Lee JE, Takahashi A *et al*. (2010) A genome-wide association study reveals susceptibility variants for non-small cell lung cancer in the Korean population. *Hum Mol Genet* **19**, 4948–4954.

23 Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, Yeager M *et al*. (2009) A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet* **85**, 679–691.

24 McKay JD, Hung RJ, Gaborieau V, Boffetta P, Chabrier A, Byrnes G, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J *et al*. (2008) Lung cancer susceptibility locus at 5p15.33. *Nat Genet* **40**: 1404–1406.

25 Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J *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.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1. DRP algorithms of commercial companies.

Table S1. SNP profiling results of 48 individuals.