

## ORIGINAL ARTICLE

## Evidence for a genetical contribution to non-smoking-related lung cancer

Shamus R Carr,<sup>1</sup> Wallace Akerley,<sup>2</sup> Mia Hashibe,<sup>3</sup> Lisa A Cannon-Albright<sup>4,5</sup>

<sup>1</sup>Division of Thoracic Surgery, Department of Surgery, University of Maryland School of Medicine, Baltimore, Maryland, USA

<sup>2</sup>Division of Medical Oncology, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, Utah, USA

<sup>3</sup>Division of Public Health, Department of Family & Preventive Medicine, University of Utah School of Medicine, Salt Lake City, Utah, USA

<sup>4</sup>Division of Genetic Epidemiology, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, Utah, USA

<sup>5</sup>George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, Utah, USA

**Correspondence to**

Dr Shamus R Carr, Division of Thoracic Surgery, University of Maryland School of Medicine, 29 S Greene St, Suite 504, Baltimore, MD 21201, USA; scarr@mail.umaryland.edu

Received 1 December 2014

Revised 25 May 2015

Accepted 26 June 2015

Published Online First

15 July 2015

**ABSTRACT**

**Background** The majority of lung cancers are smoking-related, with environmental and genetical factors contributing. The interplay between environmental and genetical contributions in non-smoking-related lung cancers is less clear.

**Methods** We analysed a population-based computerised genealogy resource linked to a state-wide cancer registry of lung cancer cases (n=5544) for evidence of a genetical contribution to lung cancer predisposition in smoking (n=1747) and non-smoking cases (n=784). Statistical methods were used to test for significant excess relatedness of cases and estimate relative risk (RR) in close and distant relatives of lung cancer cases.

**Results** Significant excess relatedness was observed for all lung cancer cases ( $p < 0.001$ ) and for the subsets of smoking-related ( $p < 0.001$ ) and non-smoking-related ( $p < 0.001$ ) cases when all pairwise relationships were considered. Only the non-smoking-related subset of cases showed significant excess relatedness when close relationships were ignored ( $p = 0.020$ ). First-degree, second-degree, and fourth-degree relatives of non-smoking-related lung cancer cases had significantly elevated RR. An even higher elevated RR was observed for first-degree, second-degree, third-degree and fourth-degree relatives of smoking-related lung cancer cases.

**Conclusions** Non-smoking-related lung cancer cases show significant excess relatedness for close and distant relationships, providing strong evidence for a genetical contribution as well as an environmental contribution. Significant excess relatedness for only close family relationships in all lung cancer cases and in only smoking-related lung cancer cases implies environmental contribution. Additionally, the highest RR for lung cancer was observed in the relatives of smoking-related lung cancer, suggesting predisposition gene carriers who smoke are at highest risk for lung cancer. Screening and gene identification should focus on high-risk pedigrees.

**INTRODUCTION**

Lung cancer is the leading cause of cancer-related deaths worldwide.<sup>1</sup> In USA it was estimated to account for 159 260 deaths in 2014.<sup>1,2</sup> Overall survival from lung cancer at 5 years remains below 16% and has not changed in the past 30 years. Recently, low-dose screening for high-risk individuals has been shown to result in a 20% relative risk (RR) reduction in mortality from lung cancer.<sup>3</sup> Follow-up studies support this conclusion, lending support to lung cancer screening efforts. The major

**Key messages****What is the key question?**

- ▶ How much does genetics contribute to smoking-related and non-smoking-related lung cancer?

**What is the bottom line?**

- ▶ There is strong evidence for a genetical contribution in non-smoking-related lung cancer out to fourth degree relatives.

**Why read on?**

- ▶ The relative risk of developing lung cancer in smokers and non-smokers with a family history of lung cancer varies based upon genetical distance.

variables that define high-risk individuals are smoking and age.

While cigarette smoking is well established as the strongest risk factor for the development of lung cancer, environmental and genetical factors are suspected to play a role.<sup>4,5</sup> Studies have shown familial clustering of lung cancer cases, supporting evidence for a genetical contribution to the development of lung cancer.<sup>6–12</sup> These show an approximate twofold increase in lung cancer development in first-degree relatives. Similar results are observed in the Utah population.<sup>11,12</sup> This has led some to incorporate family history into lung cancer risk models.<sup>13</sup>

Presence of familial clustering in close relatives is supportive, but not sufficient, to prove a genetical contribution; the aetiology may be partly explained by shared environment and risk factors in close relatives, but may also include shared genetical predisposition. Non-genetical environmental exposures such as primary cigarette smoking, secondhand smoke, or shared occupations or behaviours can confound any evidence of genetical contribution to lung cancer, especially in close relatives.<sup>14–17</sup> Consequently, it remains difficult to show a clear genetical contribution to the risk of developing lung cancer. Evaluation of increased risk beyond first-degree relationships would provide greater weight to a genetical contribution; evaluation of familial clustering and risk in the absence of smoking would eliminate the largest confounding variable.

The unique Utah Population Database (UPDB) links Utah Cancer Registry (UCR) data from 1966



CrossMark

**To cite:** Carr SR, Akerley W, Hashibe M, et al. *Thorax* 2015;**70**:1033–1039.

with genealogy data representing the state of Utah from the mid-1800s, and with Utah death certificate (DC) data from 1904. This allowed us to analyse and examine evidence for a significant heritable contribution to predisposition to lung cancer in non-smoking lung cancer cases.

## DATA AND METHODS

### Utah population database

Originally created in the 1970s using family history data from the Family History Library of the Church of Jesus Christ of Latter-day Saints (Mormons), and since expanded using Utah vital statistics data, UPDB is a computerised genealogy of the Utah pioneers and their descendants.<sup>18</sup> The original genealogy included 1.6 million individuals and extended to six generations, it currently has over 7 million unique individuals; some pedigrees extend to 12 generations. A smaller set of 1.35 million individuals in the UPDB has at least 12 of their 14 immediate ancestors (both parents, all 4 grandparents, and at least 6 of their 8 great grandparents) in the genealogy. Individuals meeting these strict criteria were considered to have good genealogy content and were used for the genetical analyses to ensure individuals analysed had similar data available for relatives. All such individuals in the UPDB were assigned to cohorts based on 5-year birth year, sex and place of birth (Utah or not). This allowed selection of matched controls for cases, and allowed estimation of cohort-specific rates of lung cancer. Genealogy data in the UPDB has been record linked to various databases, including the UCR and the computerised Utah DCs.

### Utah Cancer Registry

The UCR, a founding member of the National Cancer Institute's SEER Program, has continuously collected data on every cancer occurring in the State of Utah since 1966. Only independent primary cancers are recorded in the UCR; cancer includes all in situ (except in situ cervical cancers) or malignant neoplasms (excluding basal and squamous cell carcinomas of the skin except in genital sites); over 98% follow-up is achieved.

Lung cancer cases were identified in the UCR based on primary site coding of the International Classification of Diseases for Oncology (C340–349; bronchus and lung), including histology codes 8000–9589 (which excludes leukaemias and lymphomas) in the UCR.

### Utah DCs: smoking status

Utah DCs from 1904 are computerised and record linked to the Utah genealogy data. Deaths from 1904 to 1956 were coded to International Classification of Disease revision 10; other deaths are coded to International Classification of Disease revisions 6–10, according to the year of death. All Utah DCs include primary cause of death, and many include contributing causes of death.

There are six separate distinctions for the tobacco contribution to death on a Utah DC. The category 'Non-user' identified non-smokers. Two categories, 'Was underlying cause of death' and 'Probably contributed to the cause of death' were used to identify smokers. The categories 'Did not contribute to cause of death', 'Unknown in relation to cause of death', and 'Unknown if user' were excluded from analysis.

### Genealogical index of familiarity test for excess relatedness

The genealogical index of familiarity (GIF) test for excess relatedness was designed specifically for use with the Utah genealogy and considers the average pairwise relatedness for a set of individuals.<sup>12</sup> The test is accomplished by comparing the

average pairwise relatedness of all pairs of individuals in a set (eg, all lung cancer cases) with their expected average pairwise relatedness in the Utah population, as estimated in 1000 sets of matched controls from the UPDB. The pairwise relatedness for a pair is measured as the Malécot's coefficient of kinship,<sup>19</sup> which measures the probability that randomly selected homologous genes from a common ancestor are shared identical by descent. The coefficient is based on the genetical distance between the pairs, through all common ancestors. The value is 1/2 for parent/offspring, 1/4 for siblings or grandparent/grandchild, and 1/8 for avunculars, for example. Unrelated individuals have a coefficient of 0. The GIF is multiplied by 10<sup>5</sup> for ease of presentation.

To calculate the significance of the GIF test we select 1000 sets of matched (5-year birth year, sex and place of birth) controls from the UPDB and calculate 1000 GIF statistics for these sets. The average pairwise relatedness for the cases is compared with the distribution of the 1000 control GIFs to ascertain the empirical significance of the GIF test. Matched controls for each lung cancer case and smoking subsets were selected from all individuals in the UPDB with good genealogy content. A GIF figure shows the contribution to the mean case GIF statistic (y axis) by the genetical distance of the pair (x axis).

The overall GIF test performs a test of the hypothesis of no excess familial clustering; the test considers all pairwise relationships. A disease with a strong environmental factor and no genetical contribution might show excess familial clustering, but this would only be exhibited in close relationships, for which individuals share environment and behaviour. For this reason we also consider the distant GIF (dGIF) test, which is similar to the overall GIF test, but ignores all pairwise relationships closer than first cousins (genetical distance 4), in cases and controls. A significant dGIF statistic shows evidence of excess relatedness among distant relationships, and provides strong support for a genetical contribution to the phenotype tested.

### Relative risks in relatives

Estimation of RR as the ratio of the observed number of affected relatives among a set of relatives to the expected number of affected relatives is the most common method to evaluate genetical contribution to phenotype. Cohort-specific rates of lung cancer from the UPDB were used to estimate the expected number of affected relatives. The cohort-specific lung cancer rate was estimated using all individuals with good genealogy content as the denominator and all lung cancer cases with good genealogy content as the numerator, for each cohort. The observed number of cases among a set of relatives was counted by cohort, excluding duplicates. The expected number of lung cancer cases was estimated by multiplying the number of

**Table 1** Classification of smoking status on a linked death certificate for Utah lung cancer cases

Tobacco-related category	All cases	Analysed cases
Probably contributed to death	8410	1235
Was underlying cause of death	3643	512
Did not contribute to cause of death	816	148
Unknown in relation to cause of death	992	184
Non-user	3248	784
Unknown if user	1866	307
Total	18 975	3170

**Table 2** Genealogical index of familiarity (GIF) analysis for excess relatedness for all lung cancer cases and for subsets of cases by smoking status from death certificate

Probands	n	Case GIF	Mean control GIF	GIF p Value	Case dGIF	Mean control dGIF	dGIF p Value
All lung cancers	5544	5.30	4.69	<0.001	4.23	4.11	0.057
Non-smoking	784	5.70	4.71	<0.001	4.62	4.11	0.020
Smoking	1747	5.67	4.74	<0.001	4.32	4.18	0.161

relatives in each cohort times the cohort-specific rate for lung cancer, then summing overall cohorts. RR was estimated for first-degree, second-degree, third-degree and fourth-degree relatives. The 95% CI for the RR is calculated as in Agresti.<sup>20</sup> A p value of less than 0.05 was considered significant for all statistical tests.

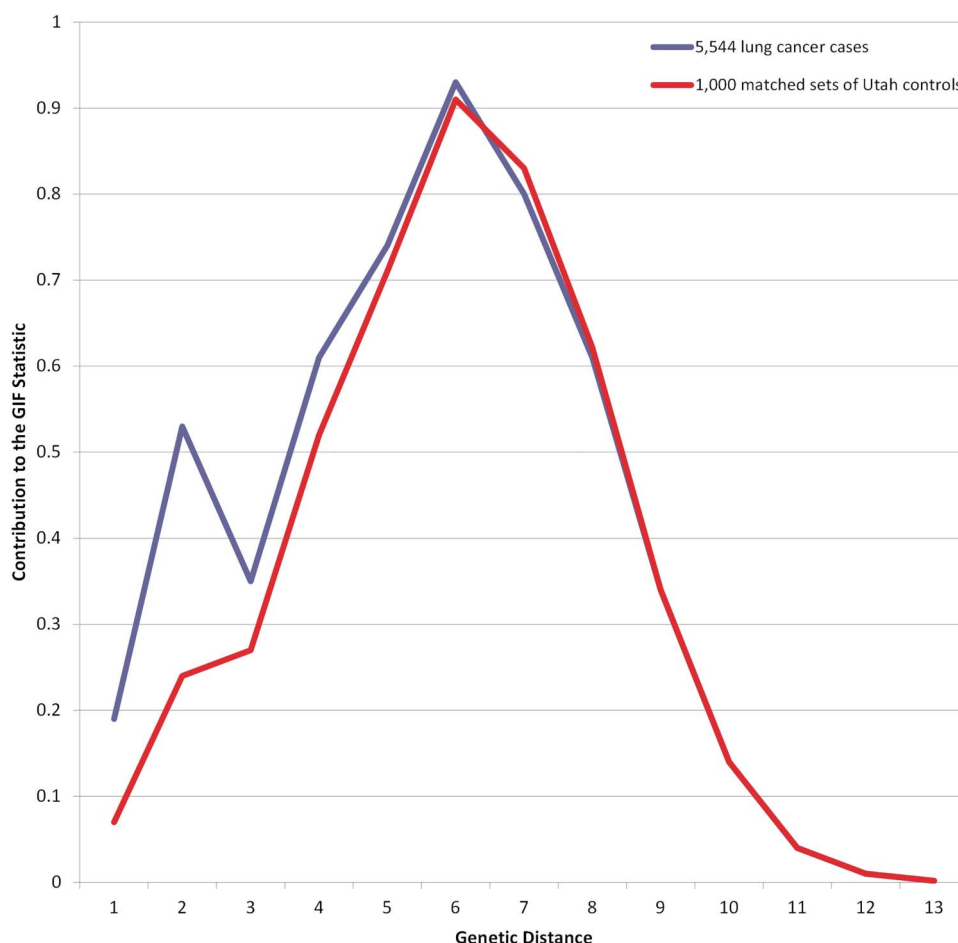
## RESULTS

Using strict criteria to ensure good genealogy content of at least 12 of 14 immediate ancestors (both parents, all 4 grandparents, and at least 6 of their 8 great grandparents) a smaller set of 1.35 million individuals in the UPDB<sup>18</sup> made up the cohort for analysis. There were 657 102 women and 692 207 men; 0.007%

are non-white; 1.20 million are born in Utah. These 1.35 million individuals were cross-referenced with the Utah DC and the UCR to identify 291 700 DCs and 93 870 cancer records. Table 1 shows the counts by tobacco contribution code for all deceased lung cancer cases in the UPDB (all cases), as well as for all deceased lung cancer cases with good genealogy content (analysed cases).

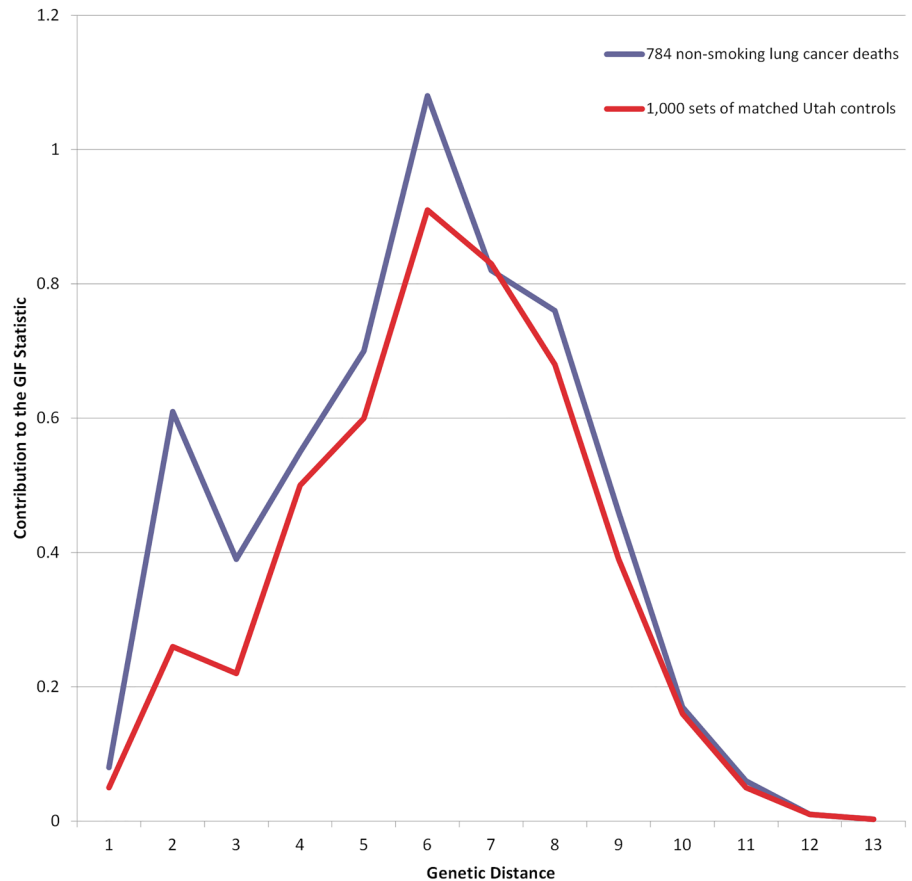
### Test for excess relatedness using the GIF method

Results for the GIF tests for excess relatedness among lung cancer cases are shown in table 2 which includes the set of individuals considered (Probands), the number of probands (n), the average relatedness of the probands (Case GIF), the average relatedness of the 1000 sets of controls (Mean control GIF), the empirical significance for the overall GIF (GIF p value), the average relatedness of the probands ignoring close relationships (Case dGIF), the average distant relatedness of the 1000 control sets (Mean control dGIF) and the empirical significance for the dGIF test (dGIF p value). Figure 1 shows the contribution to the GIF statistic by pairwise genetic distance for all lung cancer cases compared with 1000 sets of matched controls, allowing consideration of where the case and control relatedness distributions differ. All lung cancer cases show overall excess clustering over expected (Case GIF 5.30 vs Mean control GIF 4.69;  $p < 0.001$ ); however, as seen in figure 1, most of the excess relatedness is observed for genetical distance 1 (parent/offspring) and 2 (primarily siblings), relationships that share



**Figure 1** GIF analysis for all lung cancer cases compared with 1000 sets of matched controls. GIF, genealogical index of familiarity.

**Figure 2** GIF analysis for 784 non-smoking-related lung cancer deaths compared with 1000 sets of matched deceased Utah controls. GIF, genealogical index of familiarity.



common environment and genetics. The dGIF test for all lung cancers was not significant for all lung cancer cases (Case dGIF 4.23 vs Mean control dGIF 4.11;  $p=0.057$ ).

When only considering non-smoking lung cancer cases (table 2;  $n=784$ ), significant excess relatedness was observed for all relationships (Case GIF 5.70 vs Mean control GIF 4.71;  $p<0.001$ ), significant excess relatedness was also observed when close relationships were ignored (Case dGIF 4.62 vs Mean control dGIF 4.11;  $p=0.020$ ). Significant excess clustering was even observed when only relationships more distant than second cousins once removed (greater than genetic distance 7) were considered (data not shown;  $p=0.023$ ). Excess relatedness was observed for non-smoking lung cancer cases out to genetic distance 8, third cousins and beyond (figure 2).

When only smoking lung cancer cases were analysed (table 2;  $n=1747$ ), overall excess relatedness was significant for all relationships (Case GIF 5.67 vs Mean control GIF 4.74;  $p<0.001$ ), but when close relationships were ignored, no significant excess relatedness was seen (Case GIF 4.32 vs Mean control GIF 4.18; dGIF  $p=0.161$ ). Figure 3 shows the GIF results for the smoking-related lung cancer cases compared with 1000 sets of matched deceased Utah controls.

### Relative risks in relatives

Table 3 shows estimated RRs for lung cancer for first-degree to fourth-degree relatives of lung cancer cases, including the type of relative (Relative), the number of relatives ( $n$ ), the observed (obs) and expected (exp) number of relatives with lung cancer, the RR (95% CI), and the significance of the test for increased risk ( $p$  value). First-degree, second-degree, third-degree and fourth degree relatives of lung cancer cases are at a significantly increased risk of lung cancer (table 3). These estimates of RR in

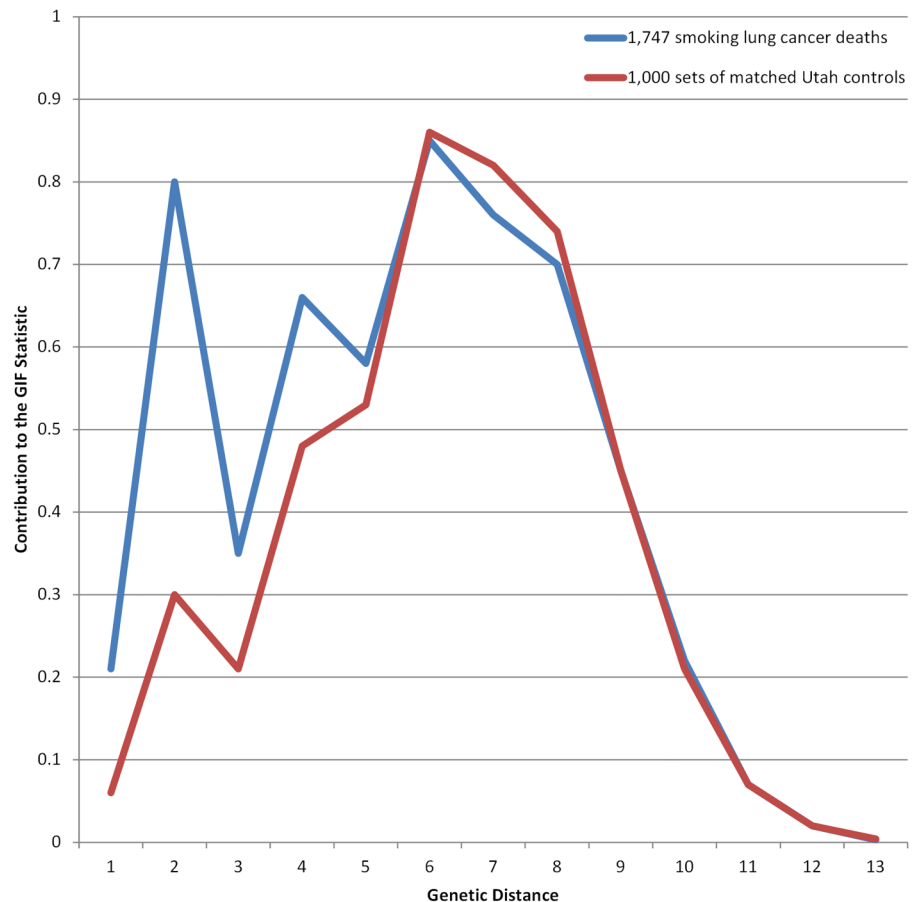
relatives of lung cancer cases showed similar results to the GIF analysis and support a contribution from genetical and environmental factors. However, degree of relationship and genetical distance do not exactly represent the same measurement. In figure 1, genetical distance 1 and 2 are primarily first-degree relationships, genetical distance 3 is primarily second-degree relationships, genetical distance 4 is primarily third-degree relationships, and genetical distance 5 is primarily fourth-degree relationships.

The effect of shared environment in lung cancer risk can be examined by estimating the RR for lung cancer in the spouses of lung cancer cases (table 4). RR for lung cancer in the spouses of all lung cancer cases was significantly elevated (RR=2.38; 95% CI 1.81 to 3.06). The RR for lung cancer in the spouses of smoking-related lung cancer cases was also significantly elevated and was even higher (RR=2.75; 95% CI 1.66 to 4.30). Spouses of non-smoking lung cancer cases were the only spouses who did not have significantly elevated risk for lung cancer (RR=1.24; 95% CI 0.50 to 2.55).

Table 5 shows the RR for non-smoking-related lung cancer in relatives of non-smoking-related lung cancer probands. For all degrees of relationship analysed, the RR of dying from non-smoking-related lung cancer was elevated, and RRs were significantly elevated for first-degree, second-degree and fourth degree relatives.

Table 6 shows the RR for smoking-related lung cancer among the relatives of smoking-related lung cancer probands. For all degrees of relationship analysed the RR of dying from smoking-related lung cancer was significantly elevated in the relatives of the smoking-related lung cancer cases. All smoking-related lung cancer RRs (table 6) were higher than the RR estimated for all lung cancer cases considered together (table 3).

**Figure 3** GIF analysis for 1747 smoking-related lung cancer deaths compared with 1000 sets of matched deceased Utah controls. GIF, genealogical index of familiarity.



## DISCUSSION

A number of publications since 1963<sup>21</sup> support a genetical contribution to lung cancer risk, similar to what has been shown for breast, colon and prostate cancers.<sup>22–24</sup> The confounding of shared genetics with shared environmental exposure (eg, smoking, occupation, childhood environment) makes separation of the genetical contribution to lung cancer challenging. Recognised comorbid conditions such as COPD are independent risk factors regardless of smoking status.<sup>25</sup> Additionally, unrecognised factors such as environmental tobacco smoke, radon, occupational exposures and diet may contribute to the development of lung cancer in smokers and non-smokers.<sup>5 17 26 27</sup> This unique study of the familial clustering of lung cancer using a population-based resource linking genealogy, a tumour registry and DC data begins to clarify the role of genetics in non-smoking-related lung cancer.

The GIF results clearly show significant excess relatedness for close and distant relationships for non-smoking-related lung cancer cases. Additionally, significantly increased RR for

non-smoking-related lung cancer cases among close and distant relatives is shown. These results are the first strong support for a genetical contribution to the subset of non-smoking-related lung cancer.

GIF analysis of the subset of smoking-related lung cancers also shows evidence for significant excess relatedness, but the dGIF results show excess familial relatedness observed was primarily due to close relationships; it was not observed when ignoring close relationships. The RR test differs from the GIF as it estimates risk for specific sets of relatives. The RR results for the smoking-related set of cases show significantly increased risks out to fourth-degree relatives, supporting a strong possibility that a genetical predisposition in some cases is enhanced by the additional risk factor of smoking.

Shared environmental exposure, specifically environmental tobacco smoke, has previously been associated with lung cancer.<sup>28</sup> In a meta-analysis an OR of 1.26 (95% CI 1.07 to

**Table 3** Estimates of relative risk (RR) for all lung cancers in relatives of all lung cancer cases

Relative	n	obs/exp	RR (95% CI)	p Value
First	44 755	767/351.9	2.18 (2.03 to 2.34)	<0.001
Second	122 158	763/574.9	1.33 (1.23 to 1.43)	<0.001
Third	314 362	1922/1696.9	1.13 (1.08 to 1.18)	<0.001
Fourth	641 360	3311/3160.0	1.05 (1.01 to 1.08)	0.0072

obs/exp, observed/expected.

**Table 4** Estimates of relative risk (RR) of lung cancer in the last spouse of cases with lung cancer

Probands	n (cases)	n (spouses)	obs/exp	Relative risk (95% CI)	p Value
All lung cancer	5554	2651	60/25.2	2.38 (1.81 to 3.06)	<0.001
Smoking cases	1741	683	19/6.9	2.75 (1.66 to 4.30)	<0.001
Non-smoking cases	784	476	7/5.7	1.24 (0.50 to 2.55)	0.34

obs/exp, observed/expected.



**Table 5** Relative risk estimates for non-smoking-related lung cancer among the relatives of 784 non-smoking-related lung cancer probands

Relative	n	obs/exp	Relative risk (95% CI)	p Value
First	2965	16/8.7	1.84 (1.05 to 2.99)	0.018
Second	6471	19/10.6	1.79 (1.08 to 2.79)	0.015
Third	18 029	49/46.0	1.07 (0.79 to 1.41)	0.66
Fourth	45 067	114/93.7	1.22 (1.00 to 1.46)	0.039

obs/exp, observed/expected.

1.47) was seen in non-smoking spouses of smokers and there was a linear relationship between lung cancer risk and the quantity and duration of exposure of smoking by the partner.<sup>26</sup> We demonstrate the RR for lung cancer in the spouses of smoking-related lung cancer cases (RR=2.75; 95% CI 1.66 to 4.30) is similar to the RR of smoking-related lung cancer in first-degree relatives of smoking-related lung cancer cases (RR=2.58; 95% CI 2.13 to 3.10), suggesting genetical risk factor(s) may be equivalent in strength to the smoking risk factor.

The possibility of a synergistic effect between smoking and genetics<sup>29</sup> is certainly likely. While the aim of this analysis was to explore evidence for a genetical contribution to non-smoking-related lung cancer, perhaps the most intriguing results are the RR estimates for smoking-related lung cancers. RR analysis differs from the GIF analysis as it only considers one relationship at a time, while the GIF analysis considers all pairs of cases. RR estimates for smoking-related lung cancer in the deceased relatives of smoking-related lung cancer cases were the highest RR observed (table 5), and were significantly elevated out to fourth-degree relatives. We hypothesise this may represent evidence for a gene-environment interaction in individuals who carry a predisposition variant, and who have further increased risk based on smoking behaviour. Genome-wide association studies support evidence for such a gene-environment interaction.<sup>30</sup> However, the effect of the genetical variant or familial mutation may be muted if the environmental exposures were absent in past generations.<sup>31</sup> Until the responsible predisposition genes are identified, and we can observe the outcomes in smoking and non-smoking predisposition carriers, we may not fully understand these combined effects.

In the most similar published study, Jonsson *et al*<sup>6</sup> evaluated the familial risk of lung cancer in the Icelandic population. In Iceland, as in Utah, there is a genealogical database and records are available for cause of death since 1955. A significantly increased risk of lung cancer was observed in first-degree, second-degree and third-degree relatives, with RRs of 2.69 (95% CI 2.20 to 3.23), 1.34 (95% CI 1.15 to 1.49) and 1.28 (95% CI 1.10 to 1.43), respectively. These RRs are similar to

**Table 6** Relative risk estimates for smoking-related-lung cancer among relatives of 1747 smoking-related probands

Relative	n	obs/exp	Relative risk (95% CI)	p Value
First	5747	114/44.1	2.58 (2.13 to 3.10)	<0.001
Second	12 105	77/45.8	1.68 (1.33 to 2.10)	<0.001
Third	32 104	265/198.4	1.34 (1.18 to 1.51)	<0.001
Fourth	76 600	417/362.0	1.15 (1.04 to 1.27)	0.0042

obs/exp, observed/expected.

those observed for all lung cancer cases in Utah for first-degree, second-degree and third-degree relatives, respectively: 2.18 (95% CI 2.03 to 2.34), 1.33 (95% CI 1.23 to 1.43) and 1.13 (95% CI 1.08 to 1.18). One limitation of the Icelandic study was that smoking information was only available for about 20% of their cohort, and direct adjustments for smoking was not possible. This is important when comparing it with this Utah study, as the prevalence of smoking in Iceland was over 20% in 2004,<sup>32</sup> which was considerably higher than Utah's estimate of 10%.<sup>33</sup>

There are limitations to analysis of a resource like the UPDB. These limitations include censorship, quality of DC data and lack of data for other known environmental risk factors such as secondhand smoke or occupation. Data for individuals who were not part of the original pioneer families or who do not have Utah vital statistics data would not be included in the genealogy and could therefore not be linked to cancer and DC data. Cancer data for individuals diagnosed out of state, or before 1966, would not be included. Tobacco use data was not available for lung cancer cases still alive, for those who died outside Utah, or for those whose attending physician at death did not complete the tobacco-related questions. While the DC has separate codes for the contribution of tobacco to the final cause of death, it relies on the physician to provide these data. These limitations apply to the entire UPDB, including cases and controls; they are assumed to be uniform and not associated with bias that would affect the hypothesis tests performed, but may reduce the power of these analyses.

While the population of Utah is very representative of the Mormon founding population from the mid-1800s, it does not represent an isolated population. The pioneer founders were a largely unrelated group of about 20 000 northern Europeans from England, Scotland, Wales, Denmark and Sweden.<sup>34</sup> Migration over ensuing generations contributed an additional 30 000–50 000 people to Utah. Recently, there has been a substantial influx of Hispanic, Polynesian and international religious migrants, but genealogies for these may not be complete in the UPDB. Utah has low to normal levels of inbreeding, compared with the rest of USA.<sup>35</sup> The population of Utah is recognised to be representative of USA and northern Europe; however, extrapolations to any other populations should not be made without additional analyses.

In this unique analysis, existing population-based data, with relationship data known for up to 12 generations, was linked to a state-wide cancer registry and DC data to explore the relationships between environment and genetics in the development of lung cancer. We show strong supporting evidence for a genetical contribution to non-smoking-related lung cancer while confirming the well-recognised evidence for familial clustering of smoking-related lung cancers. The failure of prior efforts to identify lung cancer predisposition genes of interest may in part be due to a focus on pedigrees with closely related, smoking-related lung cancers in whom a genetical predisposition gene is not always present. We propose screening and identification efforts for lung cancer predisposition genes should focus on the subset of high-risk non-smoking-related lung cancer pedigrees to lessen the noise from tobacco's contribution to lung cancer.

**Contributors** All authors contributed to the design. SRC and LAC-A were responsible for acquiring the data, interpretation of the data, drafting and editing the manuscript, while WA and MH additionally contributed in the editing of the manuscript. All authors have approved the final revised draft of the manuscript.

**Funding** This work was supported by the Utah Cancer Registry, which is funded by Contract No. HHSN261201000026C from the National Cancer Institute's SEER Program with additional support from the Utah State Department of Health and the

University of Utah. Partial support for all data sets within the Utah Population Database (UPDB) was provided by Huntsman Cancer Institute, University of Utah and the Huntsman Cancer Institute's Cancer Center Support grant, P30 CA42014 from National Cancer Institute. LAC-A acknowledges support from the University of Utah Center on Aging and the Huntsman Cancer Foundation.

**Competing interests** None declared.

**Ethics approval** Institutional Review Board.

**Provenance and peer review** Not commissioned; externally peer reviewed.

## REFERENCES

- 1 Ferlay J, Shin H-R, Bray F, *et al*. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer* 2010;127:2893–917.
- 2 Siegel R, Ma J, Zou Z, *et al*. Cancer statistics, 2014. *CA Cancer J Clin* 2014;64:9–29.
- 3 Aberle DR, Adams AM, Berg CD, *et al*, National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011;365:395–409.
- 4 Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers—a different disease. *Nat Rev Cancer* 2007;7:778–90.
- 5 Amabile J-C, Leuraud K, Vacquier B, *et al*. Multifactorial study of the risk of lung cancer among French uranium miners: radon, smoking and silicosis. *Health Phys* 2009;97:613–21.
- 6 Jonsson S, Thorsteinsdottir U, Gudbjartsson DF, *et al*. Familial risk of lung carcinoma in the Icelandic population. *JAMA* 2004;292:2977–83.
- 7 Cassidy A, Balsan J, Vesin A, *et al*. Cancer diagnosis in first-degree relatives and non-small cell lung cancer risk: results from a multi-centre case-control study in Europe. *Eur J Cancer* 2009;45:3047–53.
- 8 Nitadori J-I, Inoue M, Iwasaki M, *et al*. Association between lung cancer incidence and family history of lung cancer: data from a large-scale population-based cohort study, the JPHC study. *Chest* 2006;130:968–75.
- 9 Matakidou A, Eisen T, Houlston RS. Systematic review of the relationship between family history and lung cancer risk. *Br J Cancer* 2005;93:825–33.
- 10 Li X, Hemminki K. Inherited predisposition to early onset lung cancer according to histological type. *Int J Cancer* 2004;112:451–7.
- 11 Teerlink CC, Albright FS, Lins L, *et al*. A comprehensive survey of cancer risks in extended families. *Genet Med* 2012;14:107–14.
- 12 Cannon-Albright LA, Thomas A, Goldgar DE, *et al*. Familiality of cancer in Utah. *Cancer Res* 1994;54:2378–85.
- 13 Young RP, Hopkins RJ, Hay BA, *et al*. Lung cancer susceptibility model based on age, family history and genetic variants. *PLoS ONE* 2009;4:e5302.
- 14 Lorenzo Bermejo J, Hemminki K. Familial lung cancer and aggregation of smoking habits: a simulation of the effect of shared environmental factors on the familial risk of cancer. *Cancer Epidemiol Biomarkers Prev* 2005;14:1738–40.
- 15 Zhao Y, Wang S, Aunan K, *et al*. Air pollution and lung cancer risks in China—a meta-analysis. *Sci Total Environ* 2006;366:500–13.
- 16 Metayer C, Wang Z, Kleinerman RA, *et al*. Cooking oil fumes and risk of lung cancer in women in rural Gansu, China. *Lung Cancer* 2002;35:111–17.
- 17 Pallis AG, Syrigos KN. Lung cancer in never smokers: disease characteristics and risk factors. *Crit Rev Oncol Hematol* 2013;88:494–503.
- 18 Skolnick M. The Utah genealogical database: a resource for genetic epidemiology. *Banbury Rep* 1980;4:285–97.
- 19 Malécot G. *Les mathématiques de l'hérédité*. Paris: Masson & Cie, 1948.
- 20 Agresti A. *Categorical data analysis*. 2nd edn. New York: John Wiley & Sons, Inc., 2002.
- 21 Tokuhata GK, Lilienfeld AM. Familial aggregation of lung cancer among hospital patients. *Public Health Rep* 1963;78:277–83.
- 22 Miki Y, Swensen J, Shattuck-Eidens D, *et al*. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 1994;266:66–71.
- 23 Leppert M, Burt R, Hughes JP, *et al*. Genetic analysis of an inherited predisposition to colon cancer in a family with a variable number of adenomatous polyps. *N Engl J Med* 1990;322:904–8.
- 24 Tavtigian SV, Simard J, Teng DH, *et al*. A candidate prostate cancer susceptibility gene at chromosome 17p. *Nat Genet* 2001;27:172–80.
- 25 Young RP, Hopkins RJ, Christmas T, *et al*. COPD prevalence is increased in lung cancer, independent of age, sex and smoking history. *Eur Respir J* 2009;34:380–6.
- 26 Hackshaw AK, Law MR, Wald NJ. The accumulated evidence on lung cancer and environmental tobacco smoke. *BMJ* 1997;315:980–8.
- 27 Malhotra J, Sartori S, Brennan P, *et al*. Effect of occupational exposures on lung cancer susceptibility: a study of gene-environment interaction analysis. *Cancer Epidemiol Biomarkers Prev* 2015;24:570–9.
- 28 Trédaniel J, Boffetta P, Saracci R, *et al*. Exposure to environmental tobacco smoke and risk of lung cancer: the epidemiological evidence. *Eur Respir J* 1994;7:1877–88.
- 29 Rachtan J, Sokolowski A, Niepsuj S, *et al*. Familial lung cancer risk among women in Poland. *Lung Cancer* 2009;65:138–43.
- 30 Zhang R, Chu M, Zhao Y, *et al*. A genome-wide gene-environment interaction analysis for tobacco smoke and lung cancer susceptibility. *Carcinogenesis* 2014;35:1528–35.
- 31 Poirier JG, Brennan P, McKay JD, *et al*. Informed genome-wide association analysis with family history as a secondary phenotype identifies novel loci of lung cancer. *Genet Epidemiol* 2015;39:197–206.
- 32 Smoking habits 15–79 years 1987–2012 in Iceland. Statistics Iceland. 2014. <http://www.statice.is/Statistics/Health,-social-affairs-and-justice/Lifestyle-and-health> (accessed 1 Oct 2014).
- 33 Prevalence and Trends Data on Tobacco Use in Utah 2012. Behavioral Risk Factor Surveillance System. 2014. <http://apps.nccd.cdc.gov/brfss/display.asp?cat=TU&yr=2012&qkey=8161&state=UT> (accessed 1 Oct 2014).
- 34 Cannon-Albright LA, Farnham JM, Thomas A, *et al*. Identification and study of Utah pseudo-isolate populations-prospects for gene identification. *Am J Med Genet A* 2005;137A:269–75.
- 35 Jorde LB. Inbreeding in the Utah Mormons: an evaluation of estimates based on pedigrees, isonymy, and migration matrices. *Ann Hum Genet* 1989;53:339–55.