Genome-wide analysis of runs of homozygosity identifies new susceptibility regions of lung cancer in Han Chinese

Cheng Wang, Zhengfeng Xu, Guangfu Jin, Zhibin Hu, Juncheng Dai, Hongxia Ma, Yue Jiang, Lingmin Hu, Minjie Chu, Songyu Cao, Hongbing Shen

Department of Epidemiology and Biostatistics and Ministry of Education (MOE) Key Lab for Modern Toxicology, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu 210029, China;
Center of Prenatal Diagnosis, Nanjing Maternal and Child Health Hospital of Nanjing Medical University, Nanjing, Jiangsu 210004, China;
State Key Laboratory of Reproductive Medicine, Nanjing Medical University, Nanjing, Jiangsu 210029, China;
Section of Clinical Epidemiology, Jiangsu Key Laboratory of Cancer Biomarkers, Prevention and Treatment, Cancer Center, Nanjing Medical University, Nanjing, Jiangsu 210029, China.

Received 31 January 2013, Revised 27 February 2013, Accepted 14 March 2013, Epub 25 April 2013

Abstract

Runs of homozygosity (ROHs) are a class of important but poorly studied genomic variations and may be involved in individual susceptibility to diseases. To better understand ROH and its relationship with lung cancer, we performed a genome-wide ROH analysis of a subset of a previous genome-wide case-control study (1,473 cases and 1,962 controls) in a Han Chinese population. ROHs were classified into two classes, based on lengths, intermediate and long ROHs, to evaluate their association with lung cancer risk using existing genome-wide single nucleotide polymorphism (SNP) data. We found that the overall level of intermediate ROHs was significantly associated with a decreased risk of lung cancer (odds ratio = 0.63; 95% confidence interval: 0.51-0.77; \( P = 4.78 \times 10^{-6} \)), while the long ROHs seemed to be a risk factor of lung cancer. We also identified one ROH region at 14q23.1 that was consistently associated with lung cancer risk in the study. These results indicated that ROHs may be a new class of variation which may be associated with lung cancer risk, and genetic variants at 14q23.1 may be involved in the development of lung cancer.

Keywords: lung cancer, runs of homozygosity (ROHs), genome-wide association study

INTRODUCTION

Lung cancer is the leading cause of cancer death worldwide; its incidence and mortality rates have been increasing rapidly in the last three decades[1]. Although lung cancer is largely caused by tobacco smoking, there is accumulating evidence for the role of genetic factors in the occurrence of the disease[2]. Until now,
several genome-wide association studies (GWAS) have captured a few polymorphic variations at 3q28 (TP63), 5p15.33 (TERC/CLPTM1L), 5q32 (PPP2R2B-STK32A-DPYSL3), 6p21.32 (BTN21), 6p21.33 (BAT3/MSH5), 9p21 (CDKN2A,CDKN2B), 10p14 (GATA3), 12p13.33 (RAD52), 15q25.1 (CHRNA5/CHRNA3/CHRNB4), 13q12.12 (MPEG-TNFRSF19), 17q24.2 (BPTF), 20q13.2 (CYP24A1) and 22q12.2 (MTMR-HORMAD2-LIF) that are associated with lung cancer risk [3-11]. However, these findings are not sufficient to explain the heritability of lung cancer. Thus, further studies are needed to discover more regions/variants related to lung cancer risk.

A run of homozygosity (ROH) is defined as a continuous or uninterrupted stretch of a genomic sequence without heterozygosity in the diploid state. In general, very long ROHs are the by-product of recent inbreeding or chromosome abnormality, whereas the derivation of relatively shorter ROHs is still disputable. It is also hypothesized that ROH may be a result of linkage disequilibrium (LD) [12,13]. However, increased LD in the vicinity of a given variant is neither necessary nor sufficient for a series of variants to be included in an ROH [14]. Emerging evidence has supported that ROH may represent a novel type of independent characteristic of the genome [15]. With the development of the genotyping technology of single nucleotide polymorphism (SNP), it has become more feasible to carry out ROH studies. Several studies have reported that ROHs are widely but not randomly distributed in the outbred human genomes [12,14,16], and have been implicated in multiple complex diseases [12,17-24]. In the current study, we attempted to clarify the association of ROHs with the development of lung cancer by using a case-control study including 1,473 cases and 1,962 cancer-free controls of Han Chinese.

**SUBJECTS AND METHODS**

### Subjects

Demographic and clinical information is summarized in **Table 1** and has been described elsewhere [9]. Subjects which were included in this study consisted of 1,473 cases and 1,962 controls. The cases that were histopathologically or cytologically confirmed as lung cancer were recruited from local hospitals. All cancer-free control subjects were selected from individuals receiving routine physical examinations at local hospitals or those participating in our community-based screening of non-communicable diseases. All subjects were unrelated ethnic Han Chinese. At recruitment, informed consent was obtained from each subject, and this study protocol was approved by the local institutional review boards of authors’ affiliated institutions.

**Table 1** Summary of Class A and B ROHs at 14q23.1 and its association with lung cancer risk

<table>
<thead>
<tr>
<th>Variable</th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (Mean±SD)</td>
<td>60.08±10.30</td>
<td>59.35±9.74</td>
</tr>
<tr>
<td>Gender (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1057</td>
<td>1214</td>
</tr>
<tr>
<td>Female</td>
<td>416</td>
<td>748</td>
</tr>
<tr>
<td>Smoking status (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current smokers</td>
<td>741</td>
<td>636</td>
</tr>
<tr>
<td>Former smokers</td>
<td>168</td>
<td>83</td>
</tr>
<tr>
<td>Never smokers</td>
<td>504</td>
<td>1243</td>
</tr>
<tr>
<td>Smoking levels (Mean± S.D.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ 25 (Pack-years)</td>
<td>254</td>
<td>327</td>
</tr>
<tr>
<td>&gt; 25 (Pack-years)</td>
<td>655</td>
<td>392</td>
</tr>
<tr>
<td>Histology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Squamous cell carcinoma</td>
<td>421</td>
<td></td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>896</td>
<td></td>
</tr>
<tr>
<td>Small cell carcinoma</td>
<td>129</td>
<td></td>
</tr>
<tr>
<td>Others *</td>
<td>27</td>
<td></td>
</tr>
</tbody>
</table>

*Other includes large cell lung cancer and mixed cell carcinoma.
removed. Finally, 591,370 SNPs from 1,473 cases and 1,962 controls were used for ROH detection.

ROHs were determined using the command “--homozyg” implemented in PLINK v1.07[25], by setting the minimum length of ROH at 0.50 Mb, the minimum number of SNPs per ROH at 50 and gap threshold between two ROHs at 0.10 Mb. The minimum length of ROH was selected in order to exclude some short copy number variations (CNVs) and false ROHs formed by chance[16]. A ROH was broken into two if a gap was found more than 100 kb between adjacent homozygous SNPs. To show a full view of ROH burden, we set the minimum number of SNPs in an ROH at 50, 60, 70, 80, 90 and 100 and performed detection, respectively.

**Statistical analysis**

For each individual, $F_{\text{ROH}}$, defined as the proportion of the autosomal genome in ROHs above a specified length threshold (the total length of all their ROHs in the autosome divided by the total SNP-mappable autosomal distance)[16,21], was used as a predictor of case-control status in ROH description and burden analysis.

Mclust from the mclust package (v.4) in R was used to run unsupervised Gaussian fitting of the ROH length distribution. According to Pemberton et al.[26], we divided ROHs into two groups (Class A and B as one group, and Class C as another) depending on their

**Fig. 1** Distribution of runs of homozygosity (ROHs) at different single nucleotide polymorphism (SNP) thresholds. X axis represents $F_{\text{ROH}}$ of each individual. Y axis represents the number of individuals with specific $F_{\text{ROH}}$. 
length. For each group, overlapping pools between individuals were defined in case and control separately via a program coded by R based on the algorithm from command “-homozyg-group” implemented in PLINK v1.07, which calculated overlapping ROH number in each SNPs and considered regions with peak overlapping ROH number as pools. Considering that there might be difference between ROH structure in cases and controls, pools identified either in case or controls were involved in subsequent analysis. While analyzing, the status of each pool was coded as 0 for no ROH, 1 for class A or class B ROH, and 2 for class C ROH (dummy variables). Pools with frequency greater than 20% in either the case or control groups were considered to be “hotspot” and were further analyzed in the study.

The relationship of \( F_{ROH} \) with lung cancer risk was analyzed by using logistic regression to assess the burden of ROHs on lung cancer. The association of each hotspot with lung cancer risk was also evaluated by using logistic regression with adjustment for age, gender and pack-years of smoking as top principle component. Population structure was evaluated by principal component analysis (PCA) using SNPs in the software package EIGENSTRAT 3.0\textsuperscript{21,27}. R v2.15.1 was used for general statistical analysis\textsuperscript{21}.

**RESULTS**

\( F_{ROH} \) of ROHs was defined as a SNP number of more than 50 approximates to a normal distribution except for some individuals with extremely large \( F_{ROH} \) (Fig. 1A to 1F). To extensively evaluate ROHs across the entire genome, we set 50 as a default value to define ROHs in subsequent analysis. \( F_{ROH} \) distribution suggested that there were multiple components (Fig. 1A); therefore, we classified ROHs into three classes (class A: 500Kb-689.346Kb, class B: 689.346Kb-1548.887Kb, and class C: over 1548.887Kb) according to the cluster method based on Gaussian mixture model (see Method).

We evaluated the ROH burden on lung cancer using \( F_{ROH} \). As shown in Table 2, overall moderate ROH level (\( F_{ROH} < 0.353 \)) was significantly associated with a decreased risk of lung cancer (OR = 1.03, 95% CI: 0.99-1.07). In contrast, long \( F_{ROH} \) levels was significantly associated with an increased risk of lung cancer (OR = 1.13, 95% CI: 1.01-1.26, \( P = 0.030 \)). As individuals were divided into 4 ROH levels according to the quartile of \( F_{ROH} \) in controls, logistic regression analysis showed that the high levels of moderate ROH were consistently associated with a decreased risk of lung cancer (trend OR = 0.85, 95% CI: 0.79-0.91, \( P = 3.33 \times 10^{-5} \)) as compared to low levels while the association was not observed with the high levels of long ROH (trend OR = 1.06, 95% CI: 0.99-1.14, \( P = 0.08 \)).

To identify specific regions associated with lung cancer risk, we performed a genetic association analysis of ROH hotspots. Totally 3,288 and 3,657 ROH hotspots detected from the case and control groups, respectively, were separately evaluated for an association.

**Table 2** Distribution of \( F_{ROH} \) for overall ROHs between lung cancer cases and controls and its association with lung cancer risk

<table>
<thead>
<tr>
<th>ROH group</th>
<th>Case/Control</th>
<th>OR (95%CI)</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \leq 4.78% )</td>
<td>399/459</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4.78%-5.01%</td>
<td>397/462</td>
<td>0.90 (0.72-1.12)</td>
<td>3.29 \times 10^{-1}</td>
</tr>
<tr>
<td>5.01%-5.26%</td>
<td>341/517</td>
<td>0.72 (0.57-0.89)</td>
<td>2.75 \times 10^{-2}</td>
</tr>
<tr>
<td>&gt; 5.26%</td>
<td>336/523</td>
<td>0.63 (0.51-0.78)</td>
<td>3.33 \times 10^{-2}</td>
</tr>
<tr>
<td>TREND</td>
<td>1,473/1,962</td>
<td>0.85 (0.79-0.91)</td>
<td>4.46 \times 10^{-8}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ROH group</th>
<th>Case/Control</th>
<th>OR (95%CI)</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \leq 0.155% )</td>
<td>322/456</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0.155%-0.246%</td>
<td>348/511</td>
<td>0.92 (0.73-1.15)</td>
<td>4.52 \times 10^{-1}</td>
</tr>
<tr>
<td>0.246%-0.353%</td>
<td>384/474</td>
<td>1.17 (0.93-1.46)</td>
<td>1.73 \times 10^{-3}</td>
</tr>
<tr>
<td>&gt; 0.353%</td>
<td>384/475</td>
<td>1.13 (0.90-1.41)</td>
<td>2.86 \times 10^{-3}</td>
</tr>
<tr>
<td>TREND</td>
<td>1,473/1,962</td>
<td>1.06 (0.99-1.14)</td>
<td>8.55 \times 10^{-6}</td>
</tr>
</tbody>
</table>

**Table 3** Summary of Class A and B ROHs at 14q23.1 and its association with lung cancer risk

<table>
<thead>
<tr>
<th>chr14:58952891-58963188</th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROH</td>
<td>ROH free</td>
<td>ROH</td>
</tr>
<tr>
<td>chr14:59160264-59160264</td>
<td>281</td>
<td>1192</td>
</tr>
<tr>
<td>chr14:59121306-59143688</td>
<td>288</td>
<td>1185</td>
</tr>
<tr>
<td>chr14:5916443-59116443</td>
<td>287</td>
<td>1186</td>
</tr>
<tr>
<td>chr14:59145071-59145071</td>
<td>286</td>
<td>1187</td>
</tr>
<tr>
<td>chr14:59098920-59098920</td>
<td>279</td>
<td>1194</td>
</tr>
<tr>
<td>chr14:59163968-59163968</td>
<td>246</td>
<td>1227</td>
</tr>
<tr>
<td>chr14:59174889-59174889</td>
<td>252</td>
<td>1221</td>
</tr>
</tbody>
</table>

\*Position of pools were based on Hg17 (Human May 2004 Assembly). ROH: Runs of homozygosity.

ROH: Runs of homozygosity.
tion with lung cancer risk. Several intermediate ROH pools at 14q23.1 were found to be significantly associated with lung cancer risk (PFDR < 0.05, OR < 1) after multiple test correction (FDR) (Table 3), and interestingly, long ROH pools at the same location exerted reverse effects on lung cancer (P < 0.05, OR > 1).

DISCUSSION

SNPs are considered as the major source of genetic diversity in humans and have been extensively implicated in multiple diseases and traits. Although GWAS has successfully established the link between SNPs and phenotypes, the identified loci can only explain a small fraction of the risk of diseases or the variance of traits. In addition to SNPs, other types of genetic variants may also contribute to the individual risk of disease either as causal variants or as proxies for causal variants. In the current study, we performed a genome-wide survey on ROHs, a mysterious type of genomic variant, utilizing SNP data from genome SNP scanning chips, evaluated the overall ROH levels on lung cancer and conducted genome-wide association of ROHs with lung cancer risk. We found a significantly decreased overall ROH level among lung cancer cases and identified a ROH region at 14q23.1 that was consistently associated with lung cancer risk. This study has made an important effort to investigate the role of ROHs in a specific disease and provided a proof-of-principle approach that can be used in further ROH studies of other diseases using existing GWAS data.

In this study, $F_{\text{ROH}}$ of ROHs approximated to a normal distribution, which is consistent with the results reported in previous studies \[21,26\]. According to the right-skewed distribution of $F_{\text{ROH}}$ for overall ROHs, the boundaries of each component were defined at 689,346 kb and 1,548,887 kb, similar to the cutoffs (0.5 Mb and 1.5 Mb) reported by McQuillan et al. \[16\]. Moreover, we found that the similar $F_{\text{ROH}}$ distributions of class A and class B ROHs, which, however, were significantly different from that of class C. Class C ROHs, representing long ROHs, are very rare in Han Chinese and may be caused by recent inbreeding \[16,26,29\]. According to the simulation analyses in previous studies \[26-29\], a large sample size (e.g., 12,000-65,000) is needed for an adequate statistical power as detecting the associations of long ROHs with phenotypes. In the current study, we paid attention primarily to class A and B ROHs that have a relatively high frequency. These ROHs are probably the consequence of ancestry inbreeding and positive selection and act as a stable genomic structure across populations \[16,26\]. Intriguingly, we obtained consistent results that long ROHs (class C), accompanied by some deleterious recessive variants passed down through recent inbreeding, contributed to the development of cancer. Furthermore, we found a protective effect of intermediate ROHs.

On the basis of candidate gene/region approach, ROHs have been reported to be associated with several diseases, including schizophrenia, late-onset Alzheimer’s disease, cancer and etc. \[22,31-33\]. In the current study, we identified several pools associated with lung cancer risk in the region (chr14:59,835,003-60,105,136) at 14q23.1 including genes DAAM1, GRP135, c14orf149, JKAMP and RTN1. Gain of copy numbers on this region was associated with shorter lung cancer survival \[34\]. The phenomenon of ROH increase in this region may indicate the presence of a number of deleterious heterozygosity mutation, suggesting that it is worthy of doing some further studies in this region.

DAAM1 is one important participant in the planar cell polarity (PCP) pathway, which is one of the key sub-pathways of the Wnt signaling pathway. Bounded to both DVL and Rho, DAAM1 is thought to function as a scaffolding protein and mediates Wnt-induced DBL-Rho complex formation \[35\]. DAAM1 is also considered as a form in homology (FH) protein that is expressed in complementary patterns in lungs \[36\] and DAAM1 protein is enriched at the apical surface of airways \[37\]. However, there was no further data and research on the relationship between DAAM1 and the lungs. Recently, two studies found that DAAM1 was involved in regulating heart and kidney morphogenesis \[38,39\]. In our study, the protective region in 14q23.1 included DAAM1 upstream position and some parts of DAAM1, within which there were some eQTL SNPs of DAAM1 that may regulate the expression of DAAM1.

JKAMP, also called JNK1-associated membrane protein (JAMP), was reported to be associated with JNK1. Its association with JNK1 outcompetes JNK1 association with MKP5, resulting in increased and prolonged JNK1 activity following stress \[40\]. JNK-deficient mice exhibited delayed epithelial development in the lungs \[41\]. Furthermore, one study indicated that JNK was activated in a subset of NSCLC biopsy samples and promoted ontogenesis of the bronchial epithelium \[42\].

RTN1 was regarded as neuroendocrine-specific proteins and identified by screening the expression library of the small-cell lung cancer (SCLC) NCI-H82 cell line with antibodies to the previously identified proteins \[43\]. The gene can produce 3 different transcripts with identical 3-prime ends but unique amino termini. The B transcript was found only in the NCI-H82 cell lines, while A and C transcripts were found in 18 different SCLC lines but not in any of the 11 non-endocrine NSCLCs \[44\].
In summary, our results indicated that the overall intermediate ROH levels were low in lung cancer cases and may be implicated with susceptibility to lung cancer. We also identified one intermediate ROH region at 14q23.1 that was associated with the risk of lung cancer in Han Chinese population by using genome-wide ROH association analysis. It is important to further elucidate the genetic determinants of lung cancer. However, the exact mechanisms through which this region leads to the development of lung cancer are still unclear and need to be further investigated. Nevertheless, this study represents a pioneering effort to explore the role of intermediate ROHs in a specific disease on a genome-wide scale.

References

[25] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-
[26] Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. Genomic patterns of homozygos-