EBioMedicine xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Multi-Omics Analysis Reveals a HIF Network and Hub Gene *EPAS1* Associated with Lung Adenocarcinoma

Zhaoxi Wang [a,1], Yongyue Wei [b,1], Ruyang Zhang [a,b,1], Li Su [a], Stephanie M. Gogarten [c], Geoffrey Liu [d], Paul Brennan [e], John K. Field [f], James D. McKay [e], Jolanta Lissowska [h], Beata Swiatkowska [i], Vladimir Janout [j], Ciprian Bolca [k], Milica Kontic [l], Ghislaine Scelo [e], David Zaridze [m], Cathy C. Laurie [c], Kimberly F. Doheny [n], Elizabeth K. Pugh [n], Beth A. Marosy [n], Kurt N. Hetrick [n], Xiangjun Xiao [o], Claudio Pikielny [o], Rayjean J. Hung [g], Christopher I. Amos [o], Xihong Lin [p], David C. Christiani [a,*]

[a] Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[b] Department of Epidemiology, Department of Biostatistics, School of Public Health, Nanjing Medical University, Nanjing, China
[c] Department of Biostatistics, School of Public Health, University of Washington, Seattle, WA, USA
[d] Princess Margaret Cancer Centre, Toronto, Canada
[e] Genetic Cancer Susceptibility group, International Agency for Research on Cancer, World Health Organization, Lyon, France
[f] Institute of Translational Medicine, University of Liverpool, Liverpool, United Kingdom
[g] Lunenfeld-Tanenbaum Research Institute, Sinai Health System, University of Toronto, Toronto, Canada
[h] Department of Cancer Epidemiology and Prevention, M. Sklodowska-Curie Institute - Oncology Center, Warsaw, Poland
[i] Nofer Institute of Occupational Medicine, Department of Environmental Epidemiology, Lodz, Poland
[j] Department of Epidemiology and Public Health, University of Ostrava, University of Olomouc, Olomouc, Czech Republic
[k] Thoracic Surgery Division, "Marius Nasta" National Institute of Pneumology, Bucharest, Romania
[l] Clinic of Pulmonology, Clinical Center of Serbia (KCS), Belgrade, Serbia
[m] Russian N.N. Blokhin Cancer Research Centre, Moscow, Russian Federation
[n] Center for Inherited Disease Research, Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[o] Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, NH, USA
[p] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

ARTICLE INFO

Article history:
Received 3 February 2018
Received in revised form 17 May 2018
Accepted 18 May 2018
Available online xxxx

Keywords:
Non-small cell lung cancer
Lung adenocarcinoma
Integrated analysis
Network analysis
Hypoxia-inducible factor

ABSTRACT

Recent technological advancements have permitted high-throughput measurement of the human genome, epigenome, metabolome, transcriptome, and proteome at the population level. We hypothesized that subsets of genes identified from omic studies might have closely related biological functions and thus might interact directly at the network level. Therefore, we conducted an integrative analysis of multi-omic datasets of non-small cell lung cancer (NSCLC) to search for association patterns beyond the genome and transcriptome. A large, complex, and robust gene network containing well-known lung cancer-related genes, including *EGFR* and *TERT*, was identified from combined gene lists for lung adenocarcinoma. Members of the hypoxia-inducible factor (HIF) gene family were at the center of this network. Subsequent sequencing of network hub genes within a subset of samples from the Transdisciplinary Research in Cancer of the Lung-International Lung Cancer Consortium (TRICL-ILCCO) consortium revealed a SNP (rs12614710) in *EPAS1* associated with NSCLC that reached genome-wide significance (OR = 1.50; 95% CI: 1.31–1.72; $p = 7.75 \times 10^{-9}$). Using imputed data, we found that this SNP remained significant in the entire TRICL-ILCCO consortium ($p = .03$). Additional functional studies are warranted to better understand interrelationships among genetic polymorphisms, DNA methylation status, and *EPAS1* expression.

1. Introduction

Lung cancer is the leading cause of cancer-related mortality worldwide for both men and women [1]. Although advances in cancer prevention, early detection, and treatment have been made in recent decades, the general prognosis for lung cancer remains poor. The high case–fatality ratio of lung cancer has been attributed to advanced

* Corresponding author at: Elkan Blout Professor of Environmental Genetics, Department of Environmental Health, Harvard TH Chan School of Public Health, 665 Huntington Avenue, Boston, MA 02115, USA.
E-mail address: dchris@hsph.harvard.edu (D.C. Christiani).
[1] These authors have equal contributions.

https://doi.org/10.1016/j.ebiom.2018.05.024

stage of disease at diagnosis, poor response to current therapies, and the aggressive biological nature of lung cancer. Non-small cell lung cancer (NSCLC) is the most common type of lung cancer accounting for about 85% of all lung cancers [2,3]. Histologic subtypes of NSCLC include adenocarcinoma, squamous cell carcinoma, and large cell carcinoma [4]. Further, previous studies have demonstrated that heritable factors are significantly important in lung cancer, independent of smoking history or exposure to environmental tobacco smoke [5,6].

More recently, genome-wide association studies (GWAS) have been used to identify multiple independent loci for most diseases, because GWAS can identify common disease susceptibility loci without prior knowledge of locus function or position [7,8]. Several GWAS have identified at least five loci associated with lung cancer risk—on chromosomes 3q28, 5p15, 6p21, 13q13.1 and 15q25—in populations with European ancestry [9–14]. Additional loci at 22q12 and 15q15 have been associated with lung cancer risk [15–17]. GWAS can define lung cancer-associated genomic loci with low to moderate effects, but cannot identify causal mutations given the complicated relationships among disease-associated loci.

Recent technological advances have permitted high-throughput measurement of the human genome, epigenome, metabolome, transcriptome, and proteome at the population level. Each study can offer complementary analyses of a certain biological function, and integrative multi-omics analyses are needed to uncover synergistic interactions [18]. However, because each omic study analyzes a different molecular layer, integrative analyses comparing top-ranked genes from different omic studies might not reveal much overlapping genes.

We hypothesized that there are subsets of genes identified from different omic studies that might have closely related biological functions and thus might directly interact at the network level. Therefore, it is possible to build network(s) with direct interactions among multiple molecular layers, characterized by higher network complexity and larger gene ratios, where network complexity is defined as the ratio of total number of connections between genes to number of genes within a network, and gene ratio is defined as proportion of genes within a network to total number of genes used to build a network. In addition, incorporating biological functionality from different molecular layers, such as RNA, proteome, and metabolome results, can boost the power of genetic mapping.

In this study, we conducted an integrative analysis of GWAS and transcriptomic profiling for NSCLC using network building based on an algorithm that searches for direct interactions from a high-quality, manually curated database of genetic and physical interactions. To evaluate the identified networks, we repeated network building from a large set of randomly generated gene lists for distributions of network complexity and gene ratio. We also used hub genes identified from significant networks for targeted sequencing and further validation in the Transdisciplinary Research in Cancer of the Lung-International Lung Cancer Consortium (TRICO-ILCCO) GWAS meta-analysis.

2. Materials and Methods

2.1. Study Population

This study was based on data derived from 1000 NSCLC cases and 1000 cancer-free controls, frequency-matched by age (±5 years), gender, and smoking status (by packyears) as previously described [19]. All cases were recruited at Massachusetts General Hospital (MGH) from 1992 to 2004, were > 18 years old, and had newly diagnosed, histologically confirmed primary NSCLC. Controls were healthy, non-blood-related family members and friends of patients with cancer or with cardiothoracic conditions undergoing surgery. Histological classification was done by two staff pulmonary pathologists at MGH according to the International Classification of Diseases for Oncology (ICD-O3). For histology analysis, the following codes were used: adenocarcinoma, 8140/3, 8250/3, 8260/3, 8310/3, 8480/3, and 8560/3; large cell carcinoma, 8012/3 and 8031/3; squamous cell carcinoma, 8070/3, 8071/3, 8072/3, and 8074/3; and other non-small cell carcinomas, 8010/3, 8020/3, 8021/3, 8032/3, and 8230/3. The Institutional Review Board of MGH and the Human Subjects Committee of the Harvard School of Public Health approved the study, and all participants signed consent forms.

2.2. GWAS Dataset

DNA was extracted from peripheral white blood cells using standard protocols and was genotyped using the Human610-Quad BeadChip (Illumina, San Diego, CA). Before association tests, we conducted a systematic quality evaluation of raw genotyping data according to a general quality control (QC) procedure described by Anderson et al. [20] Briefly, unqualified samples were excluded if they fit the following QC criteria: (i) overall genotype completion rates <95%; (ii) gender discrepancies; (iii) unexpected duplicates or probable relatives (based on pairwise identity by state value, PI_HAT in PLINK >0.185); or (iv) heterozygosity rates >6 standard deviations from the mean. Unqualified SNPs were excluded if they fit the following QC criteria: (i) overall genotype completion rates <95%; (ii) gender discrepancies; (iii) unexpected duplicates or probable relatives (based on pairwise identity by state value, PI_HAT in PLINK >0.185); (iv) heterozygosity rates >6 standard deviations from the mean; or (v) individuals were non-Caucasians (using the HapMap release 23, including JPT, CEPH, CEU, and YRI populations as reference). Unqualified SNPs were excluded if they fit the following QC criteria: (i) not mapped on autosomes; (ii) call rate < 95% in all GWAS samples; (iii) MAF < 0.01; or (iv) genotype distributions deviated from those expected by Hardy-Weinberg equilibrium ($p < 1.0 \times 10^{-6}$). After quality evaluation, we had a dataset of 984 cases and 970 controls with 543,697 autosomal SNPs for epistasis analysis.

2.3. Transcriptomic Profiling

FFPE tissues were obtained by surgical biopsy from patients with NSCLC and archived. Histopathologic sections were prepared from tumor and non-affected lung parenchyma tissue by manual microdissection of FFPE blocks. A pathologist who had no knowledge of the study outcome reviewed all tissue sections. Each specimen was evaluated for amount and quality of tumor cells and histologically classified using WHO criteria. Specimens with lower than 70% cancer cellularity were not included for transcriptomic profiling. Sectioned FFPE tissues were sent to Q^2 Solutions (formerly Expression Analysis Inc., Morrisville, NC) for RNA extraction, quality assessment, and transcriptomic profiling using whole genome-DASL assay [21]. The Whole-Genome DASL HT assay covered >47,000 annotated transcripts (Illumina, San Diego, CA) [22]. A total of 59 FFPE transcriptomic profiles were obained, including 39 tumor/non-involved tissues from adenocarcinomas, 16 tumor/non-involved tissues from squamous cell carcinomas, and 4 tumor/non-involved tissues from other types of lung carcer. Among them, there were 18 pairs of tumor and matched non-involved tissues of adenocarcinomas and 8 pairs of squamous cell carcinomas, which were used in the transcriptomic analysis.

2.4. External Transcriptomic Data

Two transcriptomic datasets of NSCLC were selected and raw data were downloaded from Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo): GSE10072 and GSE18842. GSE10072 focused on lung adenocarcinoma and included 180 paired adenocarcinoma and non-affected tissue samples [23]. GSE18842 included 91 samples of mainly squamous cell carcinoma [24]. GSE10072 and GSE18842 were generated from fresh snap-frozen samples from surgical resection and profiled on Affymetrix Human Genome U133 array (Affymetrix, Santa Clara, CA).

Please cite this article as: Wang Z, et al, Multi-Omics Analysis Reveals a HIF Network and Hub Gene *EPAS1* Associated with Lung Adenocarcinoma, EBioMedicine (2018), https://doi.org/10.1016/j.ebiom.2018.05.024