Original Article

Identification of heritable rare variants associated with early-stage lung adenocarcinoma risk

Rui Fu[1,2#], Jia-Tao Zhang[3#], Rong-Rong Chen[4], Hong Li[5], Zai-Xian Tai[4], Hao-Xiang Lin[4], Jian Su[2], Xiang-Peng Chu[2], Chao Zhang[1,2], Zhen-Bin Qiu[2], Zi-Hao Chen[1,2], Wen-Fang Tang[2,6], Song Dong[2], Xue-Ning Yang[2], Guo-Qing Zhang[5], Guo-Ping Zhao[5], Yi-Long Wu[2], Wen-Zhao Zhong[1,2]

[1]School of Medicine, South China University of Technology, Guangzhou, China; [2]Guangdong Lung Cancer Institute, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China; [3]The Second School of Clinical Medicine, Southern Medical University, Guangzhou, China; [4]GenePlus-Beijing Institute, Beijing, China; [5]Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China; [6]Department of Cardiothoracic Surgery, Zhongshan City People's Hospital, Zhongshan, China

Contributions: (I) Conception and design: R Fu, JT Zhang, GQ Zhang, GP Zhao, YL Wu, WZ Zhong; (II) Administrative support: GQ Zhang, GP Zhao, YL Wu, WZ Zhong; (III) Provision of study materials or patients: R Fu, JT Zhang, J Su, XP Chu, C Zhang, ZB Qiu, ZH Chen, WF Tang, S Dong, XN Yang, WZ Zhong; (IV) Collection and assembly of data: R Fu, JT Zhang, RR Chen, H Li, ZX Tai, HX Lin; (V) Data analysis and interpretation: R Fu, JT Zhang, RR Chen, H Li, ZX Tai, HX Lin; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Wen-Zhao Zhong. Guangdong Lung Cancer Institute, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China; School of Medicine, South China University of Technology, Guangzhou, China. Email: 13609777314@163.com.

Background: In East Asia, the number of patients with adenocarcinoma, especially those presenting with ground-glass nodules (GGNs), is gradually increasing. Family aggregation of pulmonary GGNs is not uncommon; however, genetic predisposition in these patients remains poorly understood and identification of genes involved in the cause of these early-stage lung cancers might contribute to understanding of the underlying mechanisms and potential prevention strategies.

Methods: Fifty patients with early-stage lung adenocarcinoma (LUAD) presenting as GGNs and a first-degree family history of lung cancer (FHLC) from 34 independent families were enrolled into this study. Germline mutations of these patients were analyzed with whole exome sequencing (WES) and compared with age- and sex-matched 39 patients with sporadic lung cancer and 689 local healthy people. We used a stepwise variant filtering strategy, gene-based burden testing, and enrichment analysis to investigate rare but potentially pathogenic heritable mutations. Somatic tumor mutations were analyzed to consolidate germline findings.

Results: In total, 1,571 single nucleotide variants (SNVs) and 238 frameshifts with a minor allele frequency (MAF) <0.01, which were rare, recurrent, and potentially damaging candidates, were finally identified through the filtering in the GGN cohort. Pathway analysis showed the extracellular matrix to be the top dysregulated pathway. Gene-based burden testing of these highly disruptive risk-conferring heritable variants showed that *MSH5* [odds ratio (OR), 9.28, 95% confidence interval (CI): 2.49–35.87], *MMP9* (OR, 8.11, 95% CI: 2.22–28.43), and *CYP2D6* (OR, 8.09, 95% CI: 2.68–24.92) were significantly enriched in our cohort ($P<0.05$). The number of rare damaging germline variants in non-smoking patients was significantly higher than that of smoking-affected patients (Spearman's $\rho=-0.39$, $P=0.02$).

Conclusions: Heritable, potentially deleterious, and rare candidate variants of *MSH5*, *MMP9* and *CYP2D6* were significantly associated with early-stage LUAD presenting with GGNs. Nonsmoking patients likely have a higher genetic predisposition to this type of cancer than smoking-affected patients. These results have extended our understanding of the underlying mechanisms of early-stage LUAD.

Keywords: Lung cancer; adenocarcinoma; genetic predisposition; ground-glass nodule (GGN); germline mutation

Submitted Sep 24, 2021. Accepted for publication Mar 20, 2022.
doi: 10.21037/tlcr-21-789
View this article at: https://dx.doi.org/10.21037/tlcr-21-789

Introduction

Lung cancer is a malignancy with the highest morbidity and mortality worldwide (1). During the past two decades, the proportions of patients with adenocarcinoma, women, nonsmokers, and patients with a family history of malignant tumors has significantly increased in China (2). In the United States, the incidence of lung cancer was also higher in young females than in young males and the changes in epidemiological trends had not been explained fully by sex differences in smoking behaviors or in outdoor air pollution exposure (3). Although tobacco smoking is the major etiological component in lung cancer, an inherited predisposition might act independently or in concert with smoking (4). Therefore, susceptibility to lung adenocarcinoma (LUAD) needs further study.

In the past twenty years, the promotion of low-dose computed tomography (CT) has increased the detection rate of pulmonary ground-glass nodules (GGNs) (5). Compared with lung cancers presenting as solid nodules, those presenting as GGNs are characterized by inert growth and better prognosis. The pathological diagnosis is possibly pre-invasive or early-stage LUAD, including atypical adenomatous hyperplasia, adenocarcinoma *in situ* (AIS), minimally invasive adenocarcinoma (MIA), or invasive adenocarcinoma (IAC) (6,7). Few trials have studied the familial genetic susceptibility of early-stage LUAD, and potentially high-risking heritable variants in pre-invasive and invasive LUAD remain largely unknown.

Previous studies have shown that some damaging germline mutations could lead to LUAD familial aggregation (8-10). Familial LUAD is more likely to carry germline epidermal growth factor receptor (*EGFR*) mutations along with other cancer predisposition mutations, and potential genetic modifiers might contribute to somatic mutation (11-13). Nevertheless, reported damaging mutations, for example germline *EGFR* mutation, explain only a small proportion of patients with LUAD and familial aggregation and did not specially cover the population with GGN (13,14). Dozens of susceptibility loci implicated in lung cancer have been identified in genome-wide association studies (GWAS) (15). However, they could only explain a limited proportion of the genetic component of lung cancer pathogenesis with modest odds ratios (ORs) (1.1–1.4) (16,17). This has also been referred to as missing heritability and is due in part to the fact that GWAS focuses on common alleles [minor allele frequency (MAF) >0.05]. In brief, many genetic studies had limited genetic explanations for LUAD or did not focus on early-stage LUAD manifesting as GGNs.

In contrast, previous studies have reported that rare and deleterious variants with MAF <0.01 and modest-to-high effect sizes may have an important role in the etiology of complex traits and can explain missing heritability, which cannot be explained by common variants (18-21). Some low-frequency coding variants at lung cancer risk loci evaluated by exome sequencing were proven to be associated with lung carcinogenesis (22). Selecting whole exome sequences of specific individuals with extreme phenotypes, such as those with a family history, is an economical approach in identifying rare causal variants in targeted loci (18). Therefore, in our study, we sequenced select cases of pre-invasive and invasive LUAD manifesting as GGNs in patients with a first-degree family history of lung cancer (FHLC) to reveal rare and potential inheritable carcinogenetic variants among Eastern Asian patients with early-stage LUAD (*Figure 1*). The following article was presented in accordance with the STROBE reporting checklist (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-21-789/rc).

Methods

Study design and population

This study enrolled 50 patients with a first-degree relative family history of histologically confirmed lung cancer from 2019 to 2020. All patients were pathologically diagnosed as pre-invasive or IAC manifesting as GGNs on CT. Peripheral blood and tumor samples were collected for whole exome sequencing (WES) and further analysis. In addition, blood samples from age- and sex-matched

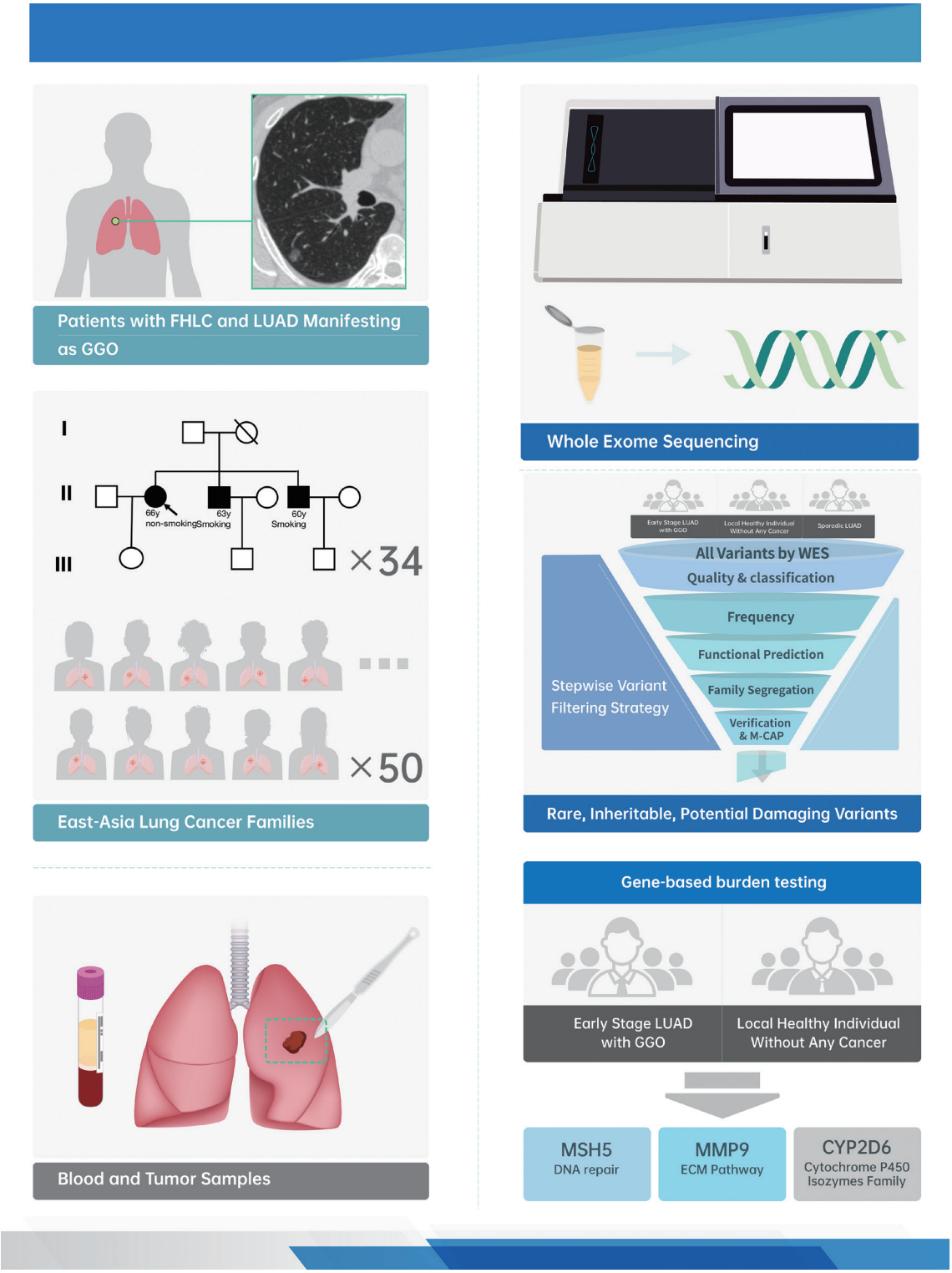

Figure 1 General view of the study. FHLC, first-degree family history of lung cancer; LUAD, lung adenocarcinoma; GGO, ground-glass opacity; WES, whole exome sequencing; ECM, extracellular matrix; M-CAP, Mendelian Clinically Applicable Pathogenicity.

39 patients with sporadic LUAD (without FHLC) and 678 local healthy people were collected retrospectively. Considering the impact of secondhand smoking in the family, we divided all the families into non-smoking families (no smoker in the family) and smoking families (at least one person in the family who lived with the patient and had a smoking history). The patients who lived in smoking families or had smoking habits themselves were defined as smoking-affected patients. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Institutional Review Board of Guangdong Provincial People's Hospital (No. GDREC2019523H), and written informed consent was obtained from all individual participants.

Library preparation, capture enrichment, exome sequencing, and variant identification

Serial peripheral blood (2–4 mL) was sampled and collected in ethylenediaminetetraacetic acid vacutainer tubes (BD Biosciences, Franklin Lakes, NJ, USA). Buffy coat DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA, USA). Forty-four patients with GGNs had tumor tissues available for somatic mutation analysis, and frozen tissue or serial sections from formalin-fixed paraffin-embedded tumor tissues were used for tumor genomic DNA extraction using the QIAamp DNA mini kit (Qiagen). DNA concentration was measured using a Qubit fluorometer and the Qubit dsDNA High Sensitivity Assay Kit (Invitrogen, Carlsbad, CA, USA). Before library construction, 1 μg of buffy coat DNA was sheared to 300 bp fragments using a Covaris S2 ultrasonicator (Covaris, Woburn, MA, USA). Following sonication, 200 ng of DNA from each sample was used for library construction. Samples underwent 2 enzymatic steps followed by the NEBNext Ultra II End Repair/dA-Tailing Module (New England Biolabs, Ipswich, MA, USA). Successful adapter ligation was confirmed with an 8-cycle polymerase chain reaction (PCR) using KAPA HiFi HotStart ReadyMix (KAPA Biosystems, Wilmington, MA, USA) with PCR primers containing a custom-synthesized barcode sequence (10 bp), which was used as a unique sample identifier. The adapter-ligated and indexed DNA fragments from 1–2 libraries were mixed in equal amounts to obtain a single pool containing 4.5 μg of DNA. DNA libraries of the peripheral blood were hybridized to the xGen Exome Research Panel (v1, Integrated DNA Technologies, Skokie, IL, USA) according to the manufacturer's instructions. Each peripheral blood DNA sample was sequenced on the Geneplus-2000 sequencing platform (Geneplus, Beijing, China) using paired-end reads according to the manufacturer's instructions. A mean depth of 211 for the germline WES and 401 for somatic WES were used. Low-quality reads and reads containing adaptor sequences were removed. Clean data were mapped to the human reference genome HG19 using Burrows-Wheeler Aligner software (23) (BWA, version 0.7.10). The best practices to call SNPs and Indels were followed using the Genome Analysis Toolkit (24).

Variant annotation and filtering

Germline variants were annotated for mutation types, transcripts, and allele frequencies of the healthy population in the public database using the variant effector predictor tool (VEP) (25). To identify the most possible rare damaging candidate variants, we filtered variants before analysis by filtering out non-functional variants; keeping variants with allele frequencies <0.01 from all populations and East Asian populations of ExAC (26), 1,000 G (27) and gnomAD (26) databases, and assessing the allele frequency in 678 healthy Chinese individuals; keeping variants predicted as damaging and deleterious by PolyPhen-2 (28) and SIFT (29); and keeping variants meeting the family segregation rule (i.e., all the patients tested in the same family must carry the variant). We also predicted the clinical pathogenicity of variants using the M-CAP Score, which dismisses variants with an uncertain significance (30). Somatic variants were filtered to exclude synonymous variants, known germline variants in the patient, and variants that occur at a population frequency of >1% in the Exome Sequencing Project (31).

Pathway enrichment analysis

Germline variants associated with biochemical metabolic pathways and signal transduction pathways were analyzed using pathway enrichment analyses based on the Kyoto Encyclopedia of Genes and Genome (KEGG), Gene Ontology (GO) and Reactome Pathway database using the clusterProfiler (32), DOSE (33), and ReactomePA (34) packages, respectively. All these enrichment analyses used a hypergeometric model and the Benjamini and Hochberg model to adjust q-values to the estimated false discovery rate. Statistical significance was established at an adjusted P value of <0.05.

Gene-based burden testing

Gene-based burden testing was performed for the targeted genes (genotype present or genotype absent) in the case subjects (34 unrelated index patients from 34 families) and the sporadic LUAD cohort compared to the healthy cohort. We calculated ORs with 95% confidence intervals (CIs) using Fisher's exact tests and corrected the P values for multiple testing by applying the Benjamini and Hochberg approach against the total number of genes in the test. Statistical significance was established at and adjusted P value of <0.05.

Statistical analysis

Spearman's rank correlation coefficient was used to compare the clinicopathological characteristics with germline variants. Statistical significance was established at P<0.05 (two-sided). Pathway enrichment analyses and gene-based burden testing used the Benjamini and Hochberg model to adjust q-values to the estimated false discovery rate. Statistical significance was established at an adjusted P value of <0.05 (two-sided).

Results

Patient clinical information

In total, 50 patients with GGNs from 34 independent families were recruited for this study (Figure S1). Fourteen families with two familial members and one family with three individuals enrolled with available samples. Most patients were female (n=29) and non-smokers (n=40), and the mean age at GGN diagnosis was 51 (range, 30–75) years. All cases presented as GGNs on CT and were pathologically confirmed as pre-invasive or IAC (Table S1). Among the 50 patients, 44 had tumor tissues available for somatic mutation analysis, except patient G0002, G0104, G0004, G0008, G0120, and G0121, as their lesions were too small and had to be used entirely for pathological diagnosis with no surplus available for WES. For the 39 patients with sporadic LUAD, most of them were female (n=27) and non-smokers (n=33), and the mean age at lung cancer diagnosis was 53 (range, 22–79) years (Table S2). For the healthy cohort, most people were female (n=436) and the mean age was 48 (range, 17–67) years.

Inheritable carcinogenetic variants of patients with GGNs

A total of 435,980 germline single nucleotide variants (SNVs) and 119,189 indels were identified by WES, with a mean of 82,880 SNVs [standard deviation (SD), 48,259; range, 44,366–156,617] and 14,460 indels (SD, 11,387; range, 6,229–34,072) for each patient from the GGN cohort. The variants were further filtered using a stepwise filtering strategy covering read quality and mutation classifications, including frameshift, missense, splicing, and stop gain. SNVs and indels with MAFs >0.01 in any of the ExAC, 1,000 G, or gnomAD databases, and an internal exome data cohort of local healthy individuals were filtered out. Furthermore, 3,786 SNVs and 440 frameshifts were predicted as potentially damaging or deleterious through PolyPhen and SIFT, and we identified 2,325 SNVs and 238 frameshifts meeting the family segregation criteria. Finally, we manually checked the allele frequency in the Allele Frequency Aggregator database to exclude variants with MAF >0.01 in the Asian population. As most of SNVs were missense mutations, we used M-CAP (30), a pathogenicity classifier for rare missense variants in the human genome with a high sensitivity to dismiss variants of uncertain significance, using >0.025 as the pathogenicity threshold. Finally, we retained 1,571 SNVs and 238 frameshifts, which were defined as rare, recurrent, and potentially pathogenic candidates (*Figure 2*). With the same filtering steps except family segregation criteria, the sporadic lung cancer cohort had 2,391 SNVs and 342 frameshifts left, while the 678 healthy controls had 32,329 SNVs and 4,643 frameshifts left (Table S3).

The KEGG pathway analysis of the 1,571 filtered SNVs and 238 filtered frameshifts indicated that "focal adhesion" and "extracellular matrix (ECM)-receptor interaction" were significantly enriched in the mutated genes (*Figure 3A*). The GO enrichment analysis showed that the top 3 dysregulated biological processes were associated with "ECM," "extracellular structure organization," and "collagen-containing ECM" (*Figure 3B*). The Reactome pathway analysis demonstrated that "degradation of the ECM" and "ECM proteoglycans" were among the top dysregulated pathways (*Figure 3C*). Collectively, these suggested that rare, potentially damaging, and inheritable variants associated with the ECM are possibly related to the risk of early-stage LUAD risk (adjusted P value <0.05).

We further examined the distribution of the 1,571 filtered SNVs and 238 filtered frameshifts from 34 families. The number of variants varied remarkably among different families (median 40; range, 15–90). We analyzed the correlations between the number of filtered variants and clinicopathological characteristics using Spearman's rank coefficient of correlation (*Figure 4A*). As expected, the

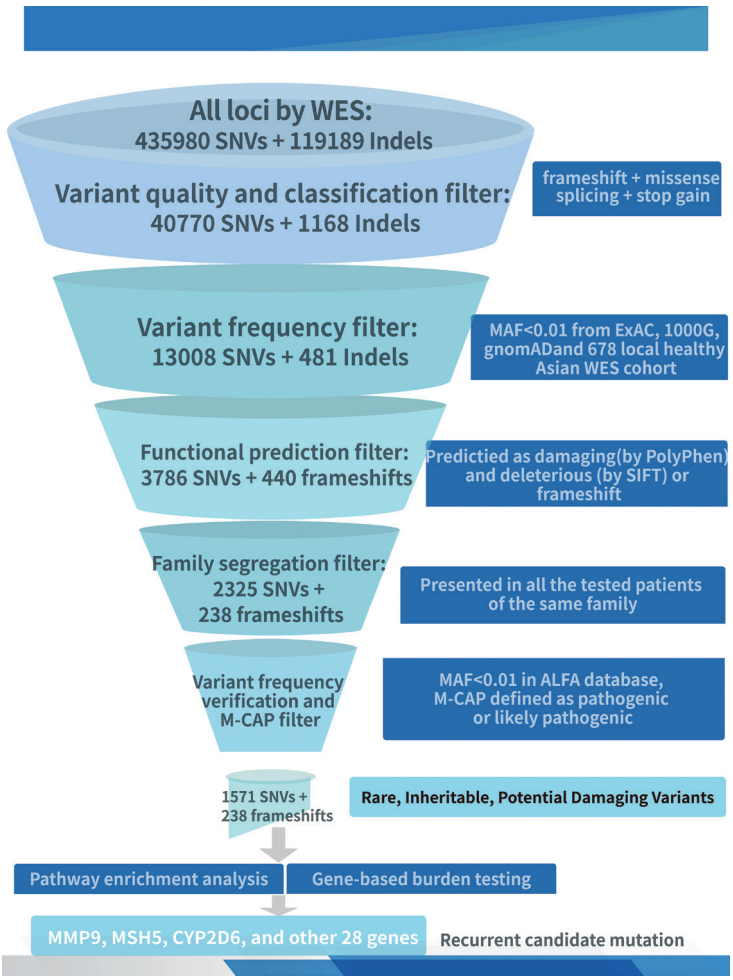

Figure 2 Workflow and annotation pipeline for the identification of candidate variants. WES, whole exome sequencing; SNV, single nucleotide variant; MAF, minor allele frequency; M-CAP, Mendelian Clinically Applicable Pathogenicity; SIFT, Sorting Intolerant From Tolerant.

pure or mixed GGN subtype demonstrated on CT was significantly associated with the pathological diagnosis (Spearman's ρ=0.52, P<0.001). This was consistent with the proposition that the solid component of GGNs can predict the invasiveness of early-stage LUAD (35). Interestingly, we found that the number of variants was significantly associated with smoking history (Spearman's ρ=−0.39, P=0.02), with fewer variants in smoking families and more in non-smoking families. This suggests that many more innate genetic predisposition factors are needed for non-smoking patients to lead to pre-invasive and invasive LUAD manifesting as GGNs.

While examining the somatic mutation signatures, we noticed there were more C>A mutations and less C>T mutations in the smoking patients (P=0.0044 and P=0.0042, respectively; *Figure 4B*). In contrast, the mutation signatures were similar in AIS, MIA, and IAC (*Figure 4C*). The median number of somatic mutations was 28 (range: 4 to 88). As expected, AIS had fewer somatic mutations than IAC (medium 31 *vs.* 47, P=0.00278, *Figure 4D*). *EGFR* mutations were the most prominent and significant variations, followed by those in *MED12*, *FOXA2*, *OR1S1* (n=6), *ERBB2*, *POFUT2*, *TGFB1*, *TP53*, and *SP4* (n=5) (*Figure 4E*). Actionable *EGFR* mutations were found in 7 of the 18 patients with pure GGNs and 11 of the 26 patients with mixed GGNs or solid tumors (38.89% *vs.* 57.69%, P=0.36) (available online: https://cdn.amegroups.cn/static/public/tlcr-21-789-1.pdf).

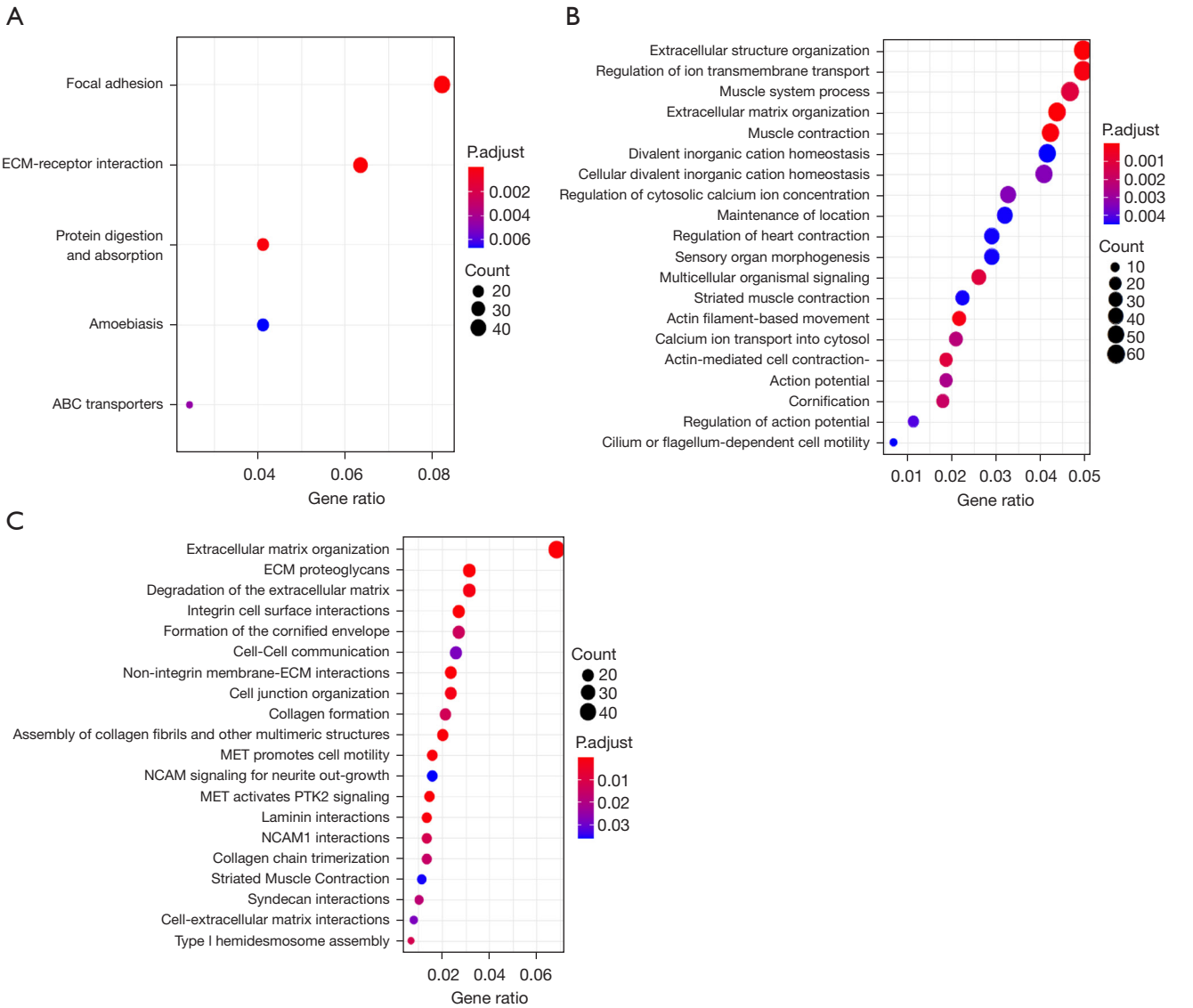

Figure 3 Pathway analysis of the filtered 1,571 SNVs and 238 frameshifts indicated enrichment of mutations in ECM pathway related genes. (A) KEGG pathway analysis; (B) GO enrichment analysis; (C) Reactome pathways analysis. ECM, extracellular matrix; SNV, single nucleotide variant; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genome; ABC, ATP-binding cassette.

Recurrent predisposition germline variants in adenocarcinoma families

Of the 1,571 filtered SNVs and 238 filtered frameshifts, 35 SNVs and 10 frameshifts were present in ≥2 families (Table S4). When aggregating the variant data at the gene level, there were 338 SNVs and 49 frameshifts in 192 genes presenting in ≥2 families, and 79 SNVs and 10 frameshifts in 31 genes in ≥3 families (*Figure 4F*). Gene-based burden testing showed that *MMP9* (OR, 8.11; 95% CI: 2.22–28.43), *MSH5* (OR, 9.28; 95% CI: 2.49–35.87), and *CYP2D6* (OR, 8.09; 95% CI: 2.68–24.92) were significantly enriched in our cohort (*Table 1*).

MMP9 encodes matrix metallopeptidase 9 (707 amino acids), spans 7.6 kb, and contains 13 exons. MMP9 plays an essential role in local proteolysis of the ECM and in leukocyte migration (protein ID: P14780). We identified 3 rare missense variants in *MMP9* (G615W, T246I, C373W) (*Table 1*, available online: https://cdn.amegroups.cn/static/

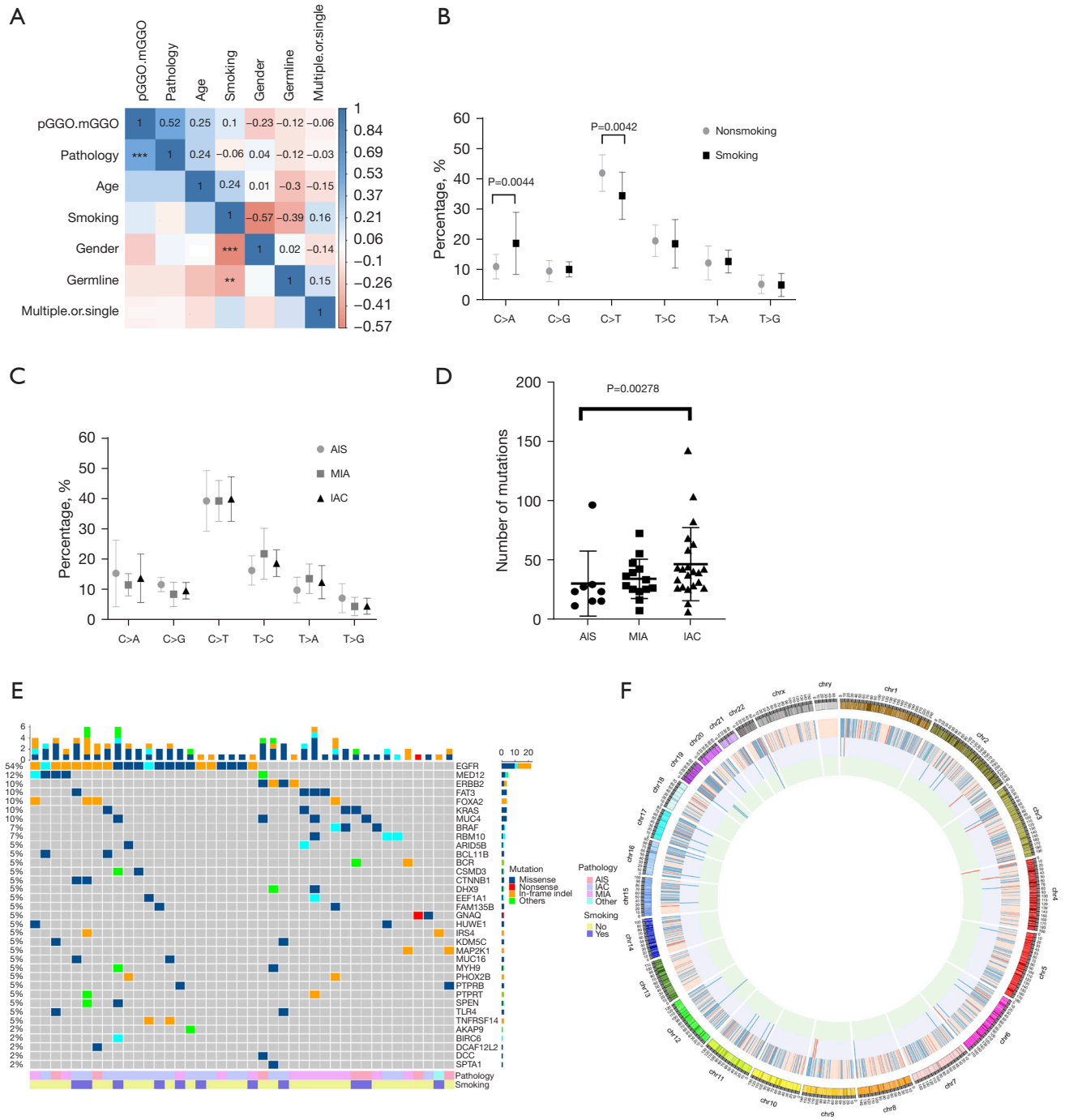

Figure 4 Signature of germline and somatic mutations. (A) Correlation of mutation numbers with clinicopathological characteristics showed the association of smoking and the number of germline mutations. ***, P<0.001; *, P<0.05. (B) Somatic mutation signature of patients with or without smoking; (C) Somatic mutation signature of different pathologic types; (D) Number of somatic mutations in different pathologic types early-stage pulmonary adenocarcinoma; (E) Landscape of somatic driver mutations; (F) Distribution of the filtered variants. Chr, chromosome; inner circle, presented in ≥3 families; 2nd inner circle, presented in ≥2 families; 3rd inner circle, presented in ≥1 family. pGGO, pure ground-glass opacity; mGGO, mixed ground-glass opacity; AIS, adenocarcinoma in situ; MIA, minimally invasive adenocarcinoma; IAC, invasive adenocarcinoma.

public/tlcr-21-789-2.pdf). Since the over-expression of MMPs can destroy the basement membrane, tumor cells or their accompanying stromal cells bearing MMPs are better able to penetrate endothelial basement membranes and become invasive (36). The expression level of *MMP9* increases from AIS to IAC, especially in the non-invasive phase, suggesting that increased expression of *MMP9* occurs before non-invasive lesions become invasive tumors (37).

CYP2D6 encodes cytochrome P450 2D6 (446 amino acids, spans 4 kb, and contains 9 exons. The R441H substitution is located on exon 9 and affects a highly evolutionarily conserved site in the crystal structure of CYP2D6 (protein ID: Q9Y512) (Figure S2). This variant had an M-CAP score of 0.690 (available online: https://cdn.amegroups.cn/static/public/tlcr-21-789-2.pdf). The M-CAP score of the other 2 variants (R474W, Y355C) was 0.095 and 0.082 respectively, which indicates that they are possibly pathogenic as well.

MSH5 encodes mutS homolog 5 (834 amino acids), which contains 25 exons and functions in the DNA mismatch repair pathway. Notably, GWAS have identified susceptibility loci for lung carcinogenesis by GWAS in this gene (22,38,39). The A685T substitution in the *MSH5* is highly evolutionarily conserved (protein ID: O43196) (Figure S3). It was identified in 2 non-smoking female patients with multiple GGNs diagnosed with pre-invasive LUAD from family 32, and 1 non-smoking female patient with a GGN diagnosed as invasive LUAD from family 18 (*Table 1*). The M-CAP scores of A685T and the other variant (R287H) were 0.089 and 0.128, respectively (available online: https://cdn.amegroups.cn/static/public/tlcr-21-789-2.pdf).

Considering the function of *MSH5* in the DNA repair pathway (40,41), we further explored other DNA repair genes besides *MSH5* in our cohort. We found rare recurrent germline mutations in *APEX1*, *FANCM*, *MNAT1*, *MSH4*, *PNKP*, and *RAD54L* (Table S5). As most DNA repair genes serve as tumor suppressors, we further queried whether patients with rare germline mutations in DNA repair genes had somatic mutations in these genes as well. Indeed, patients from families 9, 17, 31, 38, 40, and 42 had both somatic and germline mutations in DNA repair genes (Table S5).

Discussion

The number of patients with LUAD presenting as GGNs is gradually increasing in East Asian populations, and their genetic predisposition remains unclear. This study analyzed the germline variants of patients with pre-invasive or invasive LUAD presenting as GGNs as well as patients with FHLC and sporadic LUAD and East-Asian healthy people without cancer, using a stepwise variant filtering strategy. Using WES data and gene-based burden testing, we identified rare, heritable, and potentially pathogenic candidates in early-stage LUAD.

Pathway enrichment analyses showed that germline variants in genes associated with the ECM may contribute to the carcinogenesis of LUAD presenting as GGNs, especially the *MMP9*. Numerous studies have demonstrated the crucial role of different stromal components during cancer development and metastasis (42). Genetic and epigenetic changes, such as aberrant promoter methylation or aberrant miRNA expression, lead to misexpression of collagens, laminins, proteoglycans, proteases, and integrins in the tumor microenvironment (43). Changes in biomechanical properties of the ECM are involved in the development of cancer (44). Focal adhesion complexes, as an adaptor linking the ECM to the actomyosin cytoskeleton, can help cells perceive environmental external forces and lead to many functional consequences (36). From hyperplasia and carcinoma *in situ*, to invasive lesions, oncogenic transformation involves a series of genetic and epigenetic changes, including genetic mutations and expression changes of different ECM adhesion receptors and growth factor receptors. Moreover, it modifies the ability of tumor cells to sense and respond to external forces and mechanical properties of the ECM (44). *MMP9* is an important ECM enzyme. External carcinogens could induce production of MMP9 and epithelial-mesenchymal transition progression in lung cancer by activating the Shp2/ERK1/2/JNK/Smad2/3 signaling pathways (45). With age or certain diseases, MMPs may be deregulated at genetic or post-genetic levels and destabilize the ECM dynamics, which is a characteristic of cancer (36).

Interestingly, our results also showed that smoking-affected patients carried fewer filtered potentially damaging germline variants than those without a smoking history. To some extent, this is consistent with a previous finding that familial mutation carriers reported fewer pack-years than other patients with lung cancer (21). Therefore, without a smoking history, many more innate genetic predisposition factors are needed for the development of pre-invasive and invasive LUAD manifesting as GGNs. We speculated that this observation may also explain the different mutations (46) and growth patterns (47) in smokers and non-smokers with malignant GGN.

Table 1 Gene-based burden testing of the genes with recurrent mutations

Gene	Existing_variation	Family ID	Case ID	Gene-based burden testing (P)[†]	OR	95% CI	GGN-mut[‡]	GGN-wt[§]	Control-mut[¶]	Control-wt[††]	LC without FHLC-mut[‡‡]	LC without FHLC-wt[§§]	Gene-based burden testing (P, not adjusted)[¶¶]
ABCA4	rs61749446, rs1413097229, rs201471607	5, 6, 45	G0005, G0106, G0006, G0045	0.3431	–	–	3	31	33	645	1	38	1
ALDH6A1	rs369485559, COSM5927708, rs370897364	8, 10, 37	G0008, G0010, G0137, G0037	0.0186	8.1050	2.221–28.43	3	31	8	670	0	39	1
ARHGEF10	rs146766107, rs187607027, –	21, 38, 41	G0121, G0021, G0038, G0041	0.1327	–	–	3	31	14	664	1	38	0.36217479
CATSPERG	rs200132227	5, 10, 36	G0005, G0010, G0036	0.2028	–	–	3	31	20	658	1	38	0.46187873
COL9A1	rs1422617430, rs375684014, rs767544695	2, 11, 18	G0102, G0002, G0111, G0011,G 0018	0.2653	–	–	3	31	26	652	0	39	1
CRIPAK	rs528457959	15, 21, 46	G0015, G0121, G0021, G0046	0.2653	–	–	3	31	27	651	2	37	0.66036729
CYP2D6	rs202102799, rs532668079, rs185772085	34, 36, 19, 27	G0134, G0034, G0036, G0019, G0027	0.0291	8.0850	2.678–24.92	4	30	11	667	2	37	0.0984435
DIAPH3	rs145827856, rs760815388, rs770994435	8, 38, 39	G0008, G0038, G0039	0.2653	–	–	3	31	28	650	0	39	1
DPYD	rs570122671, COSM50544, –	4, 8, 41	G0104, G0004, G0008, G0041	0.1401	–	–	3	31	15	663	0	39	1
DTHD1	rs529758698, rs577534478	7, 41, 46	G0007, G0041, G0046	0.0291	13.0300	3.294–52.20	3	31	5	673	0	39	1
DYSF	rs185596534, rs200195517, rs141536854, rs759505768	3, 5, 13, 41	G0103, G0003, G0005, G0113, G0013, G0041	0.9141	–	–	4	30	67	611	2	37	1
ECE2	rs779580606, rs368866385, rs772740984	5, 36, 37	G0005, G0036, G0137, G0037	0.2653	–	–	3	31	26	652	1	38	0.57150265
EP400	rs183260874, rs575639601, rs760508158	3, 6, 41	G0103, G0003, G0106, G0006, G0041	0.5841	–	–	3	31	40	638	2	37	0.33890905
EPPK1	rs782582986, –, –	6, 15, 31	G0106, G0006, G0015, G0031	1.0000	–	–	3	31	79	599	1	38	0.71897502
KIAA1217	rs761928869, rs41279868, rs780371689	7, 20, 30	G0007, G0120, G0020, G0030	0.2653	–	–	3	31	28	650	0	39	1
KRT73	rs116282210	3, 8, 36	G0103, G0003, G0008, G0036	0.1750	–	–	3	31	18	660	0	39	1
MMP9	rs752547204, rs573936612, rs771359021	2, 40, 46	G0102, G0002, G0040, G0046	0.0186	8.1050	2.221–28.43	3	31	8	670	0	39	1
MSH5	rs561487480, rs746903566	18, 32, 38	G0132, G0032, G0018, G0038	0.0491	9.2760	2.492–35.87	3	31	7	671	0	39	1
MYO1H	rs759230534, rs544074593, rs758198897	15, 27, 36	G0015, G0027, G0036	0.2653	–	–	3	31	28	650	0	39	1
MYO7A	rs117966637, rs375182858	4, 5, 6	G0104, G0004, G0005, G0106, G0006	1.0000	–	–	3	31	64	614	1	38	1
MYOM2	rs140558918, rs755061905, –	8, 10, 40	G0008, G0010, G0040	0.9141	–	–	3	31	53	625	3	36	0.21962473
NIPAL1	rs572130928, rs190045000, rs777029979	7, 42, 43	G0007, G0042, G0043	0.1327	–	–	3	31	14	664	0	39	1
OBSCN	rs371324697, rs776567153, rs370234174, rs781156170, rs772564832, rs1378528618, –	1, 3, 5, 8, 31, 36	G0101, G0201, G0001, G0103, G0003, G0005, G0008, G0031, G0031, G0036	1.0000	–	–	6	28	213	465	4	35	0.63787766
PKD1	rs146096401, rs578031762, rs111244530	17, 38, 40	G0117, G0017, G0038, G0040	1.0000	–	–	3	31	123	555	2	37	0.56663464
PLEC	rs543632870, rs782618187, rs549098011, –, –	15, 21, 34, 38, 40	G0015, G0121, G0021, G0134, G0034, G0038, G0040	1.0000	–	–	5	29	166	512	1	38	0.11075302
SAMM50	rs78038328	4, 15, 18	G0004, G0104, G0015, G0018	0.1475	–	–	3	31	16	662	1	38	0.43046147
SHANK3	rs1461926484, rs376862893, –, –, –	5, 27, 43	G0005, G0027, G0027, G0027, G0043	0.0186	21.7700	4.857–94.84	3	31	3	675	1	38	0.20084246
TBL3	rs745933962, rs749858979, rs1232676880	19, 26, 43, 46	G0019, G0026, G0043, G0046	0.0540	–	–	4	30	16	662	0	39	1
TDRD12	rs1350637914	30, 31, 45	G0030, G0031, G0045	0.1182	–	–	3	31	12	666	1	38	0.32510298
TRIO	rs146453151, rs200262568	10, 36, 38	G0010, G0036, G0038	0.2304	–	–	3	31	22	656	2	37	0.15411722
WDR81	rs1485459176, rs774204130, rs146081272	8, 12, 36	G0008, G0112, G0012, G0036	0.5841	–	–	3	31	38	640	0	39	0.61673108

[†], P values of gene-based burden testing in GGNs cohort compared to the healthy cohort; [‡], number of mutant type variants in GGNs cohort; [§], number of wild type variants in GGN cohort; [¶], number of mutant type variants in healthy cohort; [††], number of wild type variants in healthy cohort; [‡‡], number of mutant type variants in sporadic LUAD cohort; [§§], number of wild type variants in sporadic LUAD cohort; [¶¶], P values of gene-based burden testing in sporadic LUAD cohort compared to the healthy cohort. OR, odds ratio; GGN, ground-glass nodule; LC, lung cancer; FHLC, first-degree family history of lung cancer.

Moreover, germline variants in DNA repair genes have been reported in a wide range of cancers. In a real-world study, the pathogenic germline variants of patients with lung cancer were most commonly found in DNA repair genes (48), which are associated with lung cancer through several repair pathways, including chromatin structure, homologous recombination, DNA polymerases, ubiquitination, and changing sensitivity to DNA-damaging agents (49). The Cancer Genome Atlas has reported that 2.5% to 4.5% patients with LUAD carry potential damaging germline variants of 8 genes, which fall most frequently in DNA repair pathways (14). In this study, heritable rare variants in *MSH5* were significantly enriched in this East-Asian population, and a total of 19 DNA repair genes were identified in 30 patients, including *MSH5*, *MSH4*, and *BRCA2* (Table S5). Several candidate variants, including *MSH5* (A685T), *ANKRD* (P429L), *KRT73* (R212C), and *NUPL2* (Y174C), which are found in high-risk loci regions and detected by GWAS, were also identified through the stepwise filtering; this suggesting the rationality of our filter strategy and confirms the existence of heritable potentially pathogenic germline variants in East Asian patients with early-stage LUAD and FHLC.

CYP2D6 is a member of the CYP450 superfamily of enzymes that participates in the metabolism of many common carcinogenetic agents of lung cancer, such as tobacco, nitrosamine, nicotine-derived nitrosamine ketone, nicotine, and cotinine (50,51). Moreover, the A allele and AA genotype of *CYP2D6* rs1065852 are associated with an increased risk of lung cancer development (52). The *CYP2D6* locus has is also associated with a higher risk of lung cancer or carcinogenesis in the Chinese population (52). One of the explanations for this is that the genotypes of this gene are associated with higher carcinogen-DNA adducts and 7-methyl-dGMP levels, which bind to DNA and may induce more gene mutations and increase the lifetime risk of lung cancer, mostly in non-smokers (53). Rare variants of *CYP2D6* were significantly enriched in this early-stage LUAD cohort. SNPs affecting the metabolism of carcinogenetic agents in populations influence the response to carcinogenetic agents of lung cancer. This can partially explain why some patients who were non-smokers still developed LUAD while some heavy smokers were free of lung cancer.

One of the main advantages of this study is that the recruited patients with GGNs were pathologically diagnosed with pre-invasive and invasive LUAD, and they all had first-degree relatives with lung cancer. Chen *et al.* reported that *YAP1*-mutant carriers had a higher predisposition for GGNs (10); however, as the nodules were not pathologically diagnosed, their conclusions should be interpreted with caution. By contrast, we analyzed the genetic susceptibility of GGNs in patients with pathologically confirmed FHLC.

This study has some limitations. First, the number of GGN patients who had first-degree relatives with lung cancer was not large enough, and we could not exclude potential selection bias and statistical power was limited. Second, lack of validation of the identified mutations in a separate large-scale cohort limits the relevance of our findings, but the results of this study can be used as the preliminary basis for further research. Third, it was difficult to provide direct evidence that specific SNP could increase the risk of lung cancer due to due to generally mild effects of a single SNP/gene in the complex pathogenesis of lung cancer. Last, lack of relatives limited the analysis of transmission in the family. However, due to the age-dependent penetrance of cancer, it was difficult to use the "non-cancer" relatives as a true negative control to filter out variants. Therefore, we used the family segregation rule as an alternative of transmission analysis.

In summary, using WES, a stepwise filter strategy, and gene-based burden testing, we presented a global view of germline variants in patients with pre-invasive or invasive LUAD presenting as GGNs. Our results indicated that rare, recurrent, heritable, and potentially highly disruptive risk-conferring variants of *MSH5*, *MMP9*, and *CYP2D6* may have contributed to the formation of LUAD. Non-smoking patients probably have a higher genetic predisposition than smoking-affected patients. In the future, it will be necessary to perform validation studies in a larger cohort and conduct functional verification of potentially high-risk candidate mutations to explore the high-risk genes in this unique lung cancer subtype, find the populations at risk, and guide screening for early-stage LUAD.

Acknowledgments

The authors are grateful to all the patients who participated in this study. The authors also thank Editage (www.editage.cn) for its linguistic revision of the manuscript.

Funding: This research was supported by funding from the National Natural Science Foundation of China (Nos. 81673031, 81872510), Guangdong Provincial People's Hospital Young Talent Project (No. GDPPHYTP201902), High-level Hospital Construction Project (No.

DFJH201801), Research Fund from Guangzhou Science and Technology Bureau (No. 201704020161), GDPH Scientific Research Funds for Leading Medical Talents and Distinguished Young Scholars in Guangdong Province (No. KJ012019449), Guangdong Basic and Applied Basic Research Foundation (No. 2019B1515130002), and Guangdong Provincial Key Laboratory of Lung Cancer Translational Medicine (No. 2017B030314120).

Footnote

Reporting Checklist: The authors have completed the STROBE reporting checklist. Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-21-789/rc

Data Sharing Statement: Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-21-789/dss

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-21-789/coif). WZZ served as an unpaid Associate Editor-in-Chief of *Translational Lung Cancer Research* from February 2021 to January 2023. YLW is a consultant of AstraZeneca, Roche Holdings AG, and Boehringer Ingelheim; he has received honoraria from AstraZeneca, Eli Lilly, Roche Holdings AG, Pfizer, Sanofi, Boehringer Ingelheim, Merck Sharp & Dohme, and Bristol-Myers Squibb; he has received research funding from Boehringer Ingelheim (Inst), Roche Holdings AG (Inst). WZZ has received honoraria from AstraZeneca, Eli Lilly, Pfizer, Roche Holdings AG, and Sanofi. RRC, ZXT, and HXL are full-time employees of GenePlus-Beijing Institute, Beijing, China. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Institutional Review Board of Guangdong Provincial People's Hospital (No. GDREC2019523H), and written informed consent was obtained from all individual participants.

References

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
2. Zang R, Shi JF, Lerut TE, et al. Ten-Year Trends of Clinicopathologic Features and Surgical Treatment of Lung Cancer in China. Ann Thorac Surg 2020;109:389-95.
3. Jemal A, Miller KD, Ma J, et al. Higher Lung Cancer Incidence in Young Women Than Young Men in the United States. N Engl J Med 2018;378:1999-2009.
4. Malhotra J, Malvezzi M, Negri E, et al. Risk factors for lung cancer worldwide. Eur Respir J 2016;48:889-902.
5. Yankelevitz DF, Yip R, Smith JP, et al. CT Screening for Lung Cancer: Nonsolid Nodules in Baseline and Annual Repeat Rounds. Radiology 2015;277:555-64.
6. Travis WD, Brambilla E, Noguchi M, et al. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol 2011;6:244-85.
7. Travis WD, Brambilla E, Burke A, et al. WHO classification of tumours of the lung, pleura, thymus and heart. International Agency for Research on Cancer; 2015.
8. Bell DW, Gore I, Okimoto RA, et al. Inherited susceptibility to lung cancer may be associated with the T790M drug resistance mutation in EGFR. Nat Genet 2005;37:1315-6.
9. Yamamoto H, Higasa K, Sakaguchi M, et al. Novel germline mutation in the transmembrane domain of HER2 in familial lung adenocarcinomas. J Natl Cancer Inst 2014;106:djt338.
10. Chen HY, Yu SL, Ho BC, et al. R331W Missense Mutation of Oncogene YAP1 Is a Germline Risk Allele for Lung Adenocarcinoma With Medical Actionability. J Clin Oncol 2015;33:2303-10.
11. Cheng YI, Gan YC, Liu D, et al. Potential genetic modifiers for somatic EGFR mutation in lung cancer: a meta-analysis and literature review. BMC Cancer

2019;19:1068.
12. Ikeda K, Nomori H, Mori T, et al. Novel germline mutation: EGFR V843I in patient with multiple lung adenocarcinomas and family members with lung cancer. Ann Thorac Surg 2008;85:1430-2.
13. Lu S, Yu Y, Li Z, et al. EGFR and ERBB2 Germline Mutations in Chinese Lung Cancer Patients and Their Roles in Genetic Susceptibility to Cancer. J Thorac Oncol 2019;14:732-6.
14. Parry EM, Gable DL, Stanley SE, et al. Germline Mutations in DNA Repair Genes in Lung Adenocarcinoma. J Thorac Oncol 2017;12:1673-8.
15. Shen H, Zhu M, Wang C. Precision oncology of lung cancer: genetic and genomic differences in Chinese population. NPJ Precis Oncol 2019;3:14.
16. Agarwala V, Flannick J, Sunyaev S, et al. Evaluating empirical bounds on complex disease genetic architecture. Nat Genet 2013;45:1418-27.
17. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biol 2017;18:77.
18. Kang G, Lin D, Hakonarson H, et al. Two-stage extreme phenotype sequencing design for discovering and testing common and rare genetic variants: efficiency and power. Hum Hered 2012;73:139-47.
19. Lamina C. Digging into the extremes: a useful approach for the analysis of rare variants with continuous traits? BMC Proc 2011;5 Suppl 9:S105.
20. Li D, Lewinger JP, Gauderman WJ, et al. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. Genet Epidemiol 2011;35:790-9.
21. Liu Y, Kheradmand F, Davis CF, et al. Focused Analysis of Exome Sequencing Data for Rare Germline Mutations in Familial and Sporadic Lung Cancer. J Thorac Oncol 2016;11:52-61.
22. Jin G, Zhu M, Yin R, et al. Low-frequency coding variants at 6p21.33 and 20q11.21 are associated with lung cancer risk in Chinese populations. Am J Hum Genet 2015;96:832-40.
23. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.
24. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-303.
25. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17:122.
26. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020;581:434-43.
27. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, et al. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061-73.
28. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248-9.
29. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812-4.
30. Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet 2016;48:1581-6.
31. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA.
32. Yu G, Wang LG, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284-7.
33. Yu G, Wang LG, Yan GR, et al. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics 2015;31:608-9.
34. Yu G, He QY. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst 2016;12:477-9.
35. Qiu ZB, Zhang C, Chu XP, et al. Quantifying invasiveness of clinical stage IA lung adenocarcinoma with computed tomography texture features. J Thorac Cardiovasc Surg 2022;163:805-815.e3.
36. Lu P, Weaver VM, Werb Z. The extracellular matrix: a dynamic niche in cancer progression. J Cell Biol 2012;196:395-406.
37. Chen C, Li WJ, Weng JJ, et al. Cancer-associated fibroblasts, matrix metalloproteinase-9 and lymphatic vessel density are associated with progression from adenocarcinoma in situ to invasive adenocarcinoma of the lung. Oncol Lett 2020;20:130.
38. Kazma R, Babron MC, Gaborieau V, et al. Lung cancer and DNA repair genes: multilevel association analysis from the International Lung Cancer Consortium. Carcinogenesis 2012;33:1059-64.
39. Wang Y, Broderick P, Webb E, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. Nat Genet 2008;40:1407-9.
40. Wood RD, Mitchell M, Lindahl T. Human DNA repair

genes, 2005. Mutat Res 2005;577:275-83.
41. Voskarides K, Dweep H, Chrysostomou C. Evidence that DNA repair genes, a family of tumor suppressor genes, are associated with evolution rate and size of genomes. Hum Genomics 2019;13:26.
42. Quail DF, Joyce JA. Microenvironmental regulation of tumor progression and metastasis. Nat Med 2013;19:1423-37.
43. Götte M, Kovalszky I. Extracellular matrix functions in lung cancer. Matrix Biol 2018;73:105-21.
44. Yu H, Mouw JK, Weaver VM. Forcing form and function: biomechanical regulation of tumor evolution. Trends Cell Biol 2011;21:47-56.
45. Liu YN, Guan Y, Shen J, et al. Shp2 positively regulates cigarette smoke-induced epithelial mesenchymal transition by mediating MMP-9 production. Respir Res 2020;21:161.
46. Mao W, Wen Y, Lei H, et al. Isolation and Retrieval of Extracellular Vesicles for Liquid Biopsy of Malignant Ground-Glass Opacity. Anal Chem 2019;91:13729-36.
47. Kobayashi Y, Sakao Y, Deshpande GA, et al. The association between baseline clinical-radiological characteristics and growth of pulmonary nodules with ground-glass opacity. Lung Cancer 2014;83:61-6.
48. Yi Y, Chen R, Dai P, et al. OA 18.01 Paired Tumor-Normal Next-Generation Sequencing (NGS) to Identify Pathogenic/Likely Pathogenic Germline Mutations in Lung Cancer Patients. J Thorac Oncol 2017;12:S1795.
49. Sun S, Liu Y, Eisfeld AK, et al. Identification of Germline Mismatch Repair Gene Mutations in Lung Cancer Patients With Paired Tumor-Normal Next Generation Sequencing: A Retrospective Study. Front Oncol 2019;9:550.
50. Bouchardy C, Benhamou S, Dayer P. The effect of tobacco on lung cancer risk depends on CYP2D6 activity. Cancer Res 1996;56:251-3.
51. Crespi CL, Penman BW, Gelboin HV, et al. A tobacco smoke-derived nitrosamine, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone, is activated by multiple human cytochrome P450s including the polymorphic human cytochrome P4502D6. Carcinogenesis 1991;12:1197-201.
52. Li M, Li A, He R, et al. Gene polymorphism of cytochrome P450 significantly affects lung cancer susceptibility. Cancer Med 2019;8:4892-905.
53. Kato S, Bowman ED, Harrington AM, et al. Human lung carcinogen-DNA adduct levels mediated by genetic polymorphisms in vivo. J Natl Cancer Inst 1995;87:902-7.

Cite this article as: Fu R, Zhang JT, Chen RR, Li H, Tai ZX, Lin HX, Su J, Chu XP, Zhang C, Qiu ZB, Chen ZH, Tang WF, Dong S, Yang XN, Zhang GQ, Zhao GP, Wu YL, Zhong WZ. Identification of heritable rare variants associated with early-stage lung adenocarcinoma risk. Transl Lung Cancer Res 2022;11(4):509-522. doi: 10.21037/tlcr-21-789

Supplementary

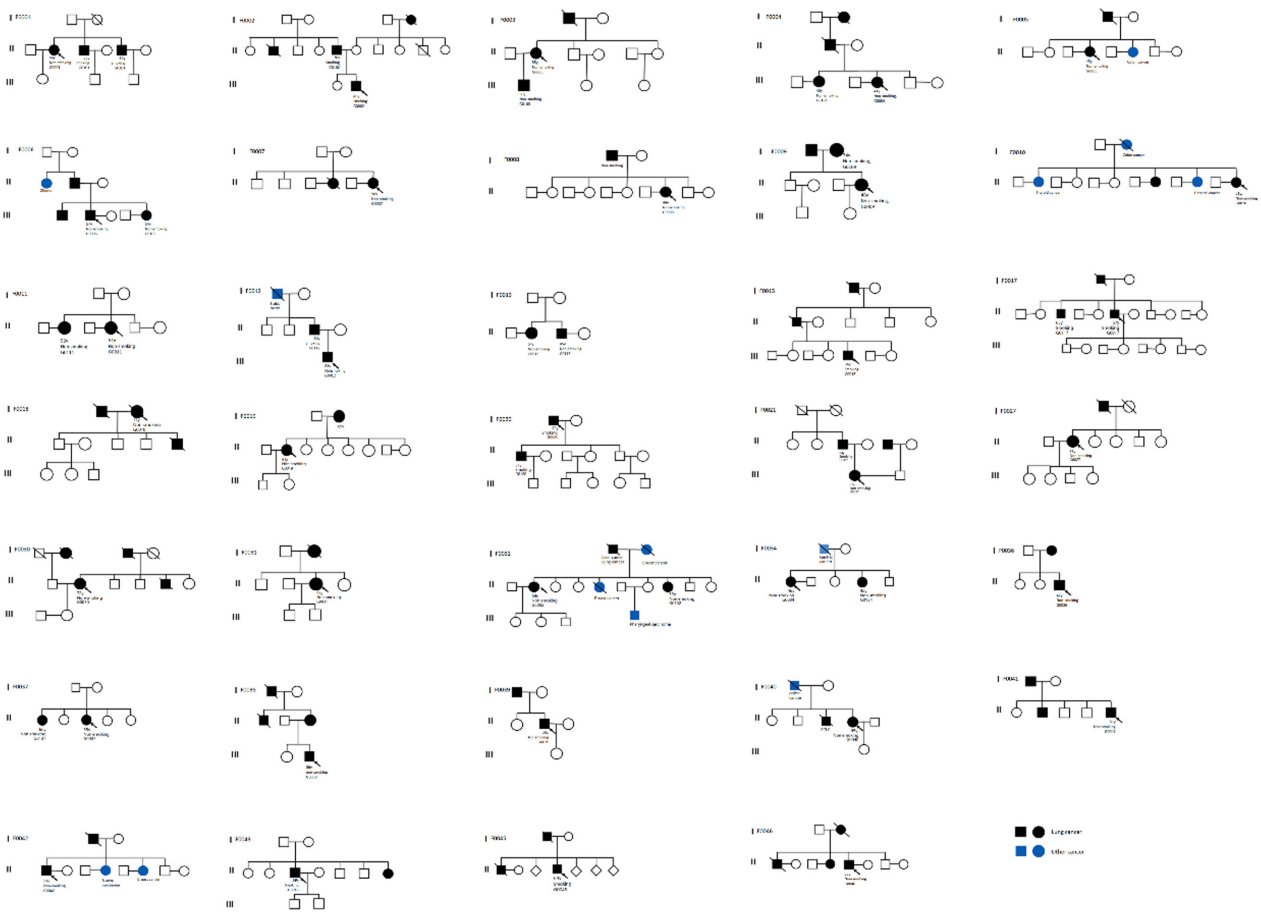

Figure S1 Family pedigree of the 34 GGN families.

https://dx.doi.org/10.21037/tlcr-21-789

Table S1 Clinicopathological characteristics of GGN patients

Characteristic	Patients with GGN (n=50)
Age at diagnosis	
Median	51
Range	30-75
Gender - No. (%)	
Female	27
Male	23
Smoking – No. (%)	
No	37
Yes	13
GGO subtype – No. (%)	
Pure GGO	21
Mixed GGO	23
Others [*1]	6
Numbers of GGO – No. (%)	
1	31
≥2	19
Pathology – No. (%)	
AIS	9
MIA	16
IAC	24
Others [*2]	1
Clinical stage – No. (%)	
0	10
IA1	24
IA2	10
Others [*3]	6

*1 others: solid tumor or solid tumor + pure/mixed GGO; *2 others: mixed mucus adenocarcinoma and non-mucus adenocarcinoma; others *3: IA3 or later stage; GGN, ground-glass nodule; GGO, ground-glass opacity; AIS, adenocarcinoma in situ; MIA, minimally invasive adenocarcinoma; IAC, invasive adenocarcinoma.

https://dx.doi.org/10.21037/tlcr-21-789

Table S2 Clinicopathological characteristics of lung cancer patients without family history of lung cancer

Patient ID	Gender	Cancer family history	Cancer patients in the family	Smoking	Subtypes of lung cancer	Stage	Age at diagnosis	Somatic Actionable mutations
lc001	F	No		No	Adenocarcinoma	I	63	
lc002	F	No		No	Adenocarcinoma	IV	47	EGFR L858R
lc003	F	No		No	Adenocarcinoma	IV	58	EGFR EX19del
lc004	F	No		No	Adenocarcinoma	IV	45	EGFR EX19del
lc005	F	No		No	Adenocarcinoma	IV	54	EGFR L858R
lc006	M	No		Yes	Adenocarcinoma	IV	52	EGFR EX19del
lc007	F	No		No	Adenocarcinoma	IV	49	EGFR L858R
lc008	F	No		No	Adenocarcinoma	III	51	EGFR EX19del
lc009	M	No		Yes	Adenocarcinoma	II	56	
lc010	M	No		No	adenosquamous carcInoma	IV	57	RET fusion
lc011	F	No		No	Adenocarcinoma	II	52	KRAS Q61H
lc012	F	No		No	Adenocarcinoma	II	46	ROS1 fusion
lc013	F	No		No	Adenocarcinoma	II	52	EGFR EX19del
lc014	M	No		No	Adenocarcinoma	I	46	EGFR EX19del
lc015	M	No		No	Adenocarcinoma	II	68	
lc016	F	No		No	Adenocarcinoma	IV	60	
lc017	F	No		No	Adenocarcinoma	IV	22	
lc018	F	No		No	Adenocarcinoma	I	63	EGFR L858R
lc019	M	No		No	Adenocarcinoma	II	70	EGFR L858R
lc020	F	No		No	Adenocarcinoma	I	51	EGFR EX19del
lc021	M	No		Yes	Adenocarcinoma	I	79	EGFR G719S+S768I
lc022	F	No		No	Adenocarcinoma	IV	57	EGFR L858R
lc023	F	No		No	Adenocarcinoma	I	52	EGFR L858R
lc024	M	No		No	Adenocarcinoma	IV	41	EGFR EX19del
lc025	F	No		No	Adenocarcinoma	IV	49	ALK fusion
lc026	F	No		Yes	Adenocarcinoma	IV	62	
lc027	F	No		No	Adenocarcinoma	II	51	EGFR EX19del
lc028	M	Yes	Father, gastric cancer	Yes	Adenocarcinoma	IV	79	EGFR EX19del
lc029	F	No		No	Adenocarcinoma	II	49	BRAF K601E
lc030	M	Yes	Father, CNS cancer	No	Adenocarcinoma	I	64	
lc031	F	No		No	Adenocarcinoma	II	40	ERBB2 EX20Ins
lc032	M	No		Yes	Adenocarcinoma	I	52	
lc033	M	Yes	Father, gastric cancer	No	Adenocarcinoma	I	48	
lc034	F	No		No	Adenocarcinoma	I	50	EGFR L858R
lc035	F	No		No	Adenocarcinoma	I	52	BRAF K601E
lc036	F	No		No	Adenocarcinoma	II	56	EGFR L858R
lc037	F	No		No	Adenocarcinoma	I	50	EGFR L858R
lc038	F	Yes	Mother, breast cancer	No	Adenocarcinoma	IV	45	EGFR EX19del
lc039	F	No		No	Adenocarcinoma	I	46	EGFR EX19del

https://dx.doi.org/10.21037/tlcr-21-789

Table S3 Number of variants during the filtering of the GGN, lung cancer and 678 healthy control cohort

Filters	GGN family cohort	678 healthy people cohort	Patients without lung cancer family history
all loci by WES	435980 SNVs + 119189 Indels	606326 SNVs + 75106 Indels	142387 SNVs+21947 Indels
frameshift + missense + splicing + stop gain	40770 SNVs + 1168 Indels	193556 SNVs + 6118 Indels	39421 SNVs +1044 Indels
MAF<0.01 from ExAC, 1000G,gnomAD and 678 asian WES cohort	13008 SNVs + 440 frameshifts	157129 SNVs + 4643 frameshift	11962 SNVs +396 frameshifts
predicted as damaging and deleterious or framshift	3786 SNVs + 440 frameshifts	46160 SNVs + 4643 frameshift	3456 SNVs +360 frameshifts
presented in all the tested patients of the same family	2325 SNVs + 238 frameshifts	-	-
MAF<0.01 in ALFA database, M-CAP defined as pathogenic or likely pathogenic	1517 SNVs + 238 frameshifts	32329 SNVs + 4643 framefhift	2391 SNVs + 342 frameshifts
Variants in at least 2 families	35 SNVs + 10 frameshifts	-	59 SNVs + 19 frameshifts
Variants in at least 3 families	4 SNVs + 1 frameshift	-	3 SNVS + 8 frameshifts

Table S4 Recurrent rare germline variants identified in the 34 families

SYMBOL	SNV / InDels	Existing_variation	MAF in 678 local healthy people	family	case
ANKRD24	p.Arg1013Gln	rs199779504,COSM3227404, COSM3227405,COSM3227406	0.00737463	8,31	G0008,G0031
ANKRD33	p.Pro429Leu	rs770135401	-	8,37	G0137,G0008,G0037
ARFGAP3	p.Ser317Leu	rs147432132	0.00884956	32,46	G0132,G0023,G0032,G0046
CA5A	p.Pro82Thr	rs377135599	0.00294985	9,17	G0117,G0109,G0017,G0009
CABP1	p.Arg92Trp	rs543428199,COSM5672565	0.00884956	38,45	G0038,G0045
CATSPERG	p.Thr576Asn	rs200132227	0.00147493	5,10,36	G0036,G0010,G0005
CCDC66	p.Thr834AsnfsTer68	rs370165016	-	8,46	G0008,G0046
CKAP2L	p.Pro482Arg	rs1037829641,COSM4084268	0.00147493	36,41	G0036,G0041
CNTROB	p.Arg892Cys	rs151174639	0.00589971	5,15	G0015,G0005
CPPED1	p.Ala192Thr	rs192649616	0.00884956	5,46	G0005,G0046
CRIPAK	p.Cys38SerfsTer369	rs528457959	0.00884956	15,21,46	G0121,G0015,G0021,G0046
CYP2D6	p.Arg441His	rs532668079	-	34,36	G0134,G0034,G0036
DNAH7	p.Val827SerfsTer20	rs752238407	0.00737463	31,45	G0031,G0045
DTHD1	p.Ser754ValfsTer25	rs529758698	0.00737463	7,41	G0007,G0041
FAM178B	p.Gly381Val	rs764132573	0.00147493	2,30	G0102,G0002,G0030
GAS2L2	p.Arg273Cys	rs140842796	0.00294985	15,40	G0015,G0040
GSN	p.His4ArgfsTer86	rs764841269	0.00442478	20,30	G0120,G0030,G0020
KRT73	p.Arg212Cys	rs116282210	0.00147493	3,8,36	G0103,G0008,G0036,G0003
KRTAP5-5	p.Ala53GlyfsTer129	rs762422220	0.00737463	30,41	G0041,G0030
MAPK15	p.Pro140Leu	rs201842849	0.00294985	30,43	G0030,G0043

Table S4 (*continued*)

https://dx.doi.org/10.21037/tlcr-21-789

Table S4 (*continued*)

SYMBOL	SNV / InDels	Existing_variation	MAF in 678 local healthy people	family	case
MCM3AP	p.Leu885Phe	rs201315959	0.00442478	5,36	G0036,G0005
MSH5	p.Ala685Thr	rs561487480	0.00294985	18,32	G0132,G0032,G0018
MTHFD1L	p.Thr619Met	rs143492706	0.00147493	10,42	G0010,G0042
MUC2	p.Thr446Met	rs199865570	0.00737463	15,18	G0015,G0033,G0018
MYO7A	p.Cys1201Ser	rs117966637,CM1212017	0.00442478	4,5	G0104,G0004,G0005
MYOD1	p.His88Arg	rs544592180	0.00442478	4,42	G0104,G0004,G0029,G0042
NUPL2	p.Tyr174Cys	rs199844379	0.00589971	21,40	G0121,G0040,G0023,G0021
ODAM	p.Pro88Ser	rs373877978	0.00147493	36,38	G0036,G0038
PRAMEF1	p.Leu354SerfsTer20	rs531127236	0.00589971	40,41	G0040,G0041
RHBDF2	p.Arg109Cys	rs369829771	0.00147493	11,15	G0111,G0015,G0011
SAMM50	p.Arg267Gln	rs78038328	0.00884956	4,15,18	G0104,G0015,G0004,G0018
SHANK3	p.Ala463GlyfsTer40	-	-	7,27	G0007,G0027
SLC22A12	p.Arg90His	rs121907896,CM042474	0.00737463	31,43	G0031,G0043
SLC2A8	p.Asp119ThrfsTer21	rs749728472	0.00147493	18,34	G0134,G0034,G0018,G0016
SLC6A20	p.Val53Met	rs371916242	-	10,18	G0010,G0018
STON1-GTF2A1L	p.Val617Asp	rs747845774	-	17,27	G0117,G0027,G0017
SYCE1L	p.Arg54Trp	rs368565145	0.00294985	3,43	G0103,G0003,G0043
TAS1R3	p.Ala403Val	rs548456115	-	19,46	G0019,G0046
TCP10	p.Pro269HisfsTer7	rs778595860,COSM216006	0.00442478	8,12	G0112,G0008,G0012
TDRD12	p.Phe897Ser	rs1350637914	-	30,31,45	G0031,G0030,G0045
TINAG	p.Ser324Asn	rs147494351	0.00442478	7,9	G0109,G0007,G0009
TLDC2	p.Arg184Cys	rs148426788	0.00442478	39,41	G0039,G0041
TRIO	p.Pro67Ser	rs146453151	0.00294985	10,36	G0036,G0010
TRPM1	p.Asp842His	rs771110434	0.00147493	10,40	G0040,G0010
VPS33B	p.Ile383Thr	rs149121639,COSM4400756	-	6,38	G0106,G0006,G0038

https://dx.doi.org/10.21037/tlcr-21-789

Figure S2 Chromosomal position, gene structure, protein domain(s), sequencing reads and evolutional conservation analysis of the candidate mutations of *CYP2D6*.

https://dx.doi.org/10.21037/tlcr-21-789

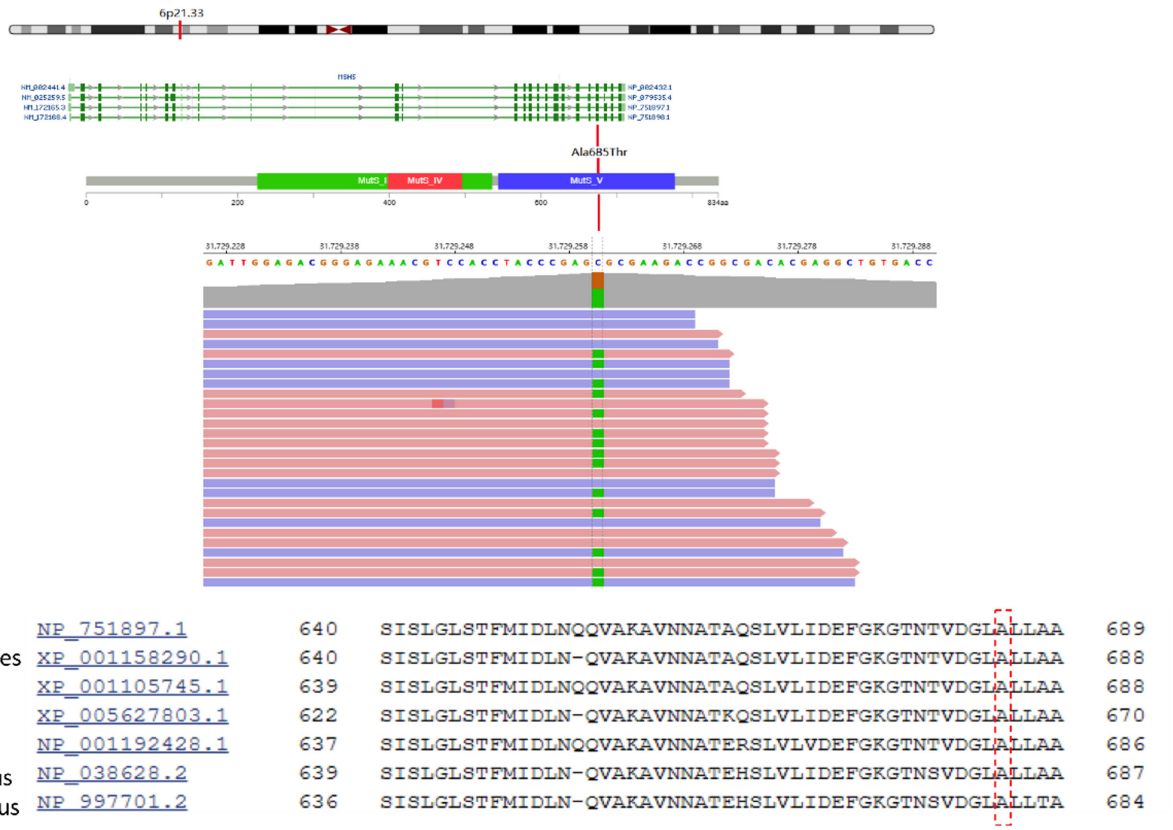

Figure S3 Chromosomal position, gene structure, protein domain(s), sequencing reads and evolutional conservation analysis of the candidate mutations of *MSH5*.

https://dx.doi.org/10.21037/tlcr-21-789

Table S5 Gene-based burden testing of the DNA repair genes with rare mutations

SYMBOL	Existing_variation	Family ID	Case ID	p value	somatic DDR variants (case ID)
APEX1	rs1413851946,rs780293860	27	G0027,G0027	0.1367	
BRCA2	rs200598289	46	G0046	>0.9999	
BRIP1	rs201869624	10	G0010	>0.9999	
FANCD2	rs767860064	46	G0046	0.2176	
FANCI	rs200186938	40	G0040	>0.9999	TP53 p.E28Kfs*14 (G0040)
FANCM	rs202171930,rs148304968	15,31	G0015,G0031	0.2105	TOPBP1 p.G274V (G0031)
GTF2H1	rs191761375	40	G0040	0.3578	TP53 p.E28Kfs*14 (G0040)
MNAT1	rs118051600	11	G0111,G0011	0.3578	
MSH4	rs116141807,rs780475342	21,42	G0121,G0021,G0042	0.1083	
MSH5	rs561487480,rs746903566	18,32,38	G0018,G0132,G0032,G0038	0.0095	REV3L p.T2275S (G0038)
PNKP	rs756933064	17	G0117,G0017	>0.9999	PRKDC p.H1464L (G0117)
POLG	rs796052895	27	G0027	>0.9999	
POLQ	rs759231797	19	G0019	>0.9999	
RAD52	-	15	G0015	0.4736	
RAD54L	rs186059216,-	9,27	G0109,G0009,G0027	0.1407	REV3L p.R2499* (G0109), TP53 p.Y163H (G0109)
SLX4	rs377440877	19	G0019	>0.9999	
XAB2	rs200271935	42	G0042	0.6101	SLX4 p.R237Q (G0042)
XPA	-	10	G0010	0.0933	
XRCC1	rs199748521	45	G0045	0.5236	

https://dx.doi.org/10.21037/tlcr-21-789